Essays on Dynamic Unobservable Heterogeneity

Silvia Sarpietro

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Economics

University College London

September 23, 2021

I, Silvia Sarpietro, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Date: September 23, 2021

Abstract

A large body of the recent literature has highlighted the importance of unobservable heterogeneity and its dynamics for many questions in Economics. In this thesis, I study the interplay between cross-sectional heterogeneity and dynamics with micro panels, i.e., panel data with many units (N) observed over a relatively smaller number of periods (T). I focus on how to estimate dynamic unobservable heterogeneity and exploit this for the problem of forecasting individual outcomes.

The second chapter, titled "*Dynamic Unobservable Heterogeneity: Income Inequality and Job Polarization*", studies how to use state-space methods to estimate unobserved heterogeneity and its dynamics when using micro panels. I illustrate the methodology with an empirical application to earnings dynamics and job polarization using a novel dataset for the UK.

The third chapter, titled "*Individual Forecast Selection*", continues with an analysis of unobserved heterogeneity for forecasting with panel data. It proposes a new methodology for forecasting that relies on individual forecast selection. For each individual, the approach selects the best forecast out of a class of competing methods, based on the out-of-sample accuracy of the method in one past time period. It is shown that this approach can be minimax-regret optimal relative to choosing the same forecasting model for everyone.

Finally, the last chapter, titled "*Regularized CUE: a Quasi-Likelihood Approach*", discusses some GMM-type estimators used in panel data and a proposed modification to the Continuous Updating Estimator (CUE). Analytical results and Monte Carlo simulations show that this modification has nice finite sample properties: It reduces the finite sample variance of the CUE, restoring its finite sample moments.

Impact Statement

In this thesis, I study some methodologies to estimate and forecast unobserved heterogeneity and investigate how this has changed over time using micro panels.

In recent years, administrative datasets have become increasingly available. While this wealth of data can be instrumental in answering several key questions in Economics, it also introduces modeling challenges. Most of the time, administrative data, which are micro panels, provide rich information on individuals and firms over time. However, micro panels usually have few dimensions of observable heterogeneity: for instance, administrative data on earnings typically lack information on education, marital status, and health conditions. This drawback makes it crucial to model unobservable heterogeneity (e.g., ability, skills) both over time and across individuals and the medium time-series dimension requires careful modeling of the dynamics. Even when several variables are observed, it is still important to assess the role played by unobserved heterogeneity. Unobservable heterogeneity is not only interesting per se, but it also affects several other outcomes of interest. For instance, heterogeneity in earnings dynamics influences predicted mobility out of low earnings; heterogeneity in income profiles conditional on parents' background is crucial to the study of intergenerational mobility.

The second chapter analyzes how state space methods can be used to identify and estimate this unobserved heterogeneity. Moreover, a modeling framework that features pervasive unobservable heterogeneity and dynamics would be useful in addressing new empirical questions using administrative data. An example would be how unobserved heterogeneity and dynamics differ by occupation and how this is related to the observed phenomenon of job polarization, which is the empirical ap-

plication I propose in the second chapter of this thesis to illustrate the methodology.

In the third chapter, we focus on forecasting individual outcomes with micro panels. Forecasting individual outcomes (microforecasting) is a key component of economic, policy, and business decisions and is becoming increasingly prominent in empirical economics. For example, the literature on long-term treatment effects relies on forecasting the effects of treatments such as early childhood interventions or job-training programs. The literature on teacher value-added can be viewed as predicting teacher performance by estimating the unobservable teacher quality. Other relevant examples of applications are forecasting individual incomes for consumption/savings decisions and revenues of banks after regulatory changes. Panel data models seem like the most natural candidates and the fact that we have richer time series dimensions makes modeling the dynamics of individual outcomes increasingly possible.

We investigate how forecasting with micro panel data changes the modeling and econometric of the existing forecasting literature. We propose a methodology that can optimally trade-off time series and cross-sectional information, where optimality is defined according to a minimax-regret criterion.

Finally, GMM-type estimators are widely used in panel data. This motivates the importance of investigating their finite sample properties. In the last chapter, we propose a modification to the Continuous Updating Estimator (CUE) that is shown to have nice finite sample properties. Theoretical results and Monte Carlo simulations show that the proposed estimator provides an attractive alternative to 2-step GMM and CUE in empirical work.

Acknowledgements

I am grateful to several people who greatly inspired and supported me during my PhD studies at UCL.

First of all, I would like to express my gratitude to Raffaella Giacomini, Toru Kitagawa, and Dennis Kristensen, who have been great mentors and supervisors. Raffaella has provided me with invaluable support and assistance. I am incredibly grateful for her guidance throughout these years. She introduced me to the world of research, and our interactions inspired me to work on the topics of this thesis. I feel very fortunate to have benefited from her precious advice during several key moments of my PhD journey. I am also very grateful to Toru and Dennis, for their great advice and research discussion. They have always been supportive and constructive, adding to my research with many insightful comments. It has been great to be mentored by Raffaella, Toru, and Dennis. Their advice and suggestions made me grow as a researcher and greatly contributed to shaping my research projects.

This thesis benefited greatly from the interactions with many other people. I would like to express my gratitude to Irene Botosaru, Simon Lee, Daniel Wilhelm, Martin Weidner, Michela Tincani, who have provided great comments and support at various stages.

I truly enjoyed working in the research environment at UCL and, for this reason, I would like to thank all other professors and PhD students at the Department of Economics at UCL. I would like to mention Mimosa Distefano for her great support and invaluable friendship. I would also like to thank Rubén Poblete-Cazenave, Alessandro Toppeta, Davide Melcangi, Javier Turén Roman, Richard Audoly, Carlo Galli, Michele Giannola, Gavin Kader, and all other PhD students who shared with

me room G06 and have been a constant point of reference for me. I greatly appreciated their ideas, support, and humour.

I would also like to express my gratitude to Prof. Arellano and to Tatiana Rosá García, Diego Astorga, Martín Almuzara, Julio Gálvez, and all other professors and PhD students I met during my visiting at CEMFI, in Madrid, which was a great research and life experience.

I gratefully acknowledge financial support from the ESRC and the Bank of England, and data access from UK Data Service/ONS.

To conclude, special thanks go to my family. I owe hugely to my parents, Salvo and Gina, to my brother Giuseppe and his wife Mariagrazia, and to the newly entrant to the family, little Gloria. I would like to dedicate this thesis and achievement to you, for always being by my side, for the invaluable advice, continuous support and unconditional love. I would also like to thank Deneuve Presicce and Assunta Casalaspro for the amazing friendship. Finally, I would like to thank a special person in my life, Arthur Taburet, for his love and support throughout the entire process.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A large body of the recent literature has highlighted the importance of unobservable heterogeneity and its dynamics for many questions in Economics. In this thesis, I study how to estimate dynamic unobservable heterogeneity and exploit this for the problem of forecasting individual outcomes with micro panels, which are panel data where many units (N) are observed over a relatively smaller number of time periods (T).

In the second chapter, I propose the use of state-space methods as a unified econometric framework for studying heterogeneity and dynamics in *micropanels*, which are typical of administrative data. I formally study identification and inference in models with pervasive unobservable heterogeneity. I show how to consistently estimate the cross-sectional distributions of unobservables in the system and uncover how such heterogeneity has changed over time. A mild parametric assumption on the standardized error term offers key advantages for identification and estimation, and delivers a flexible and general approach. Armed with this framework, I study the relationship between job polarization and earnings inequality, using a novel dataset on UK earnings, the New Earnings Survey Panel Data (NESPD). I analyze how the distributions of unobservables in the earnings process differ across occupations and over time, and separate the role played on inequality by workers' skills, labor market instability, and other types of earnings shocks.

The third chapter is based on a joint project with Raffaella Giacomini and Simon Lee. We propose a new methodology for microforecasting, i.e. forecasting

with micro panels, based on selection of the best forecasting model for each individual. Our approach to forecasting individual outcomes with micro panel data relies on individual forecast selection. For each individual, the approach selects the best forecast out of a class of competing methods, based on the out-of-sample accuracy of the method in one past time period. Our proposed data-driven method uses information about both the individual's past behavior and the behavior of other individuals to deliver a model-based clustering, which can improve the accuracy of decisions based on prediction of individual behavior. We show that this approach can be minimax-regret optimal relative to choosing the same forecasting model for everyone - guarding against large losses when competing forecasts have different accuracies and weakly improving accuracy even when choosing among equally accurate forecasts. In the presence of unobserved heterogeneity, our approach can be viewed as a way to harness the strength - but avoid the tyranny - of the majority by deciding who to pool (or shrink towards the mean). We show that this delivers accuracy gains over state-of-the-art approaches such as Empirical Bayes methods.

Finally, in the fourth chapter, I analyze some GMM-type estimators typically used in panel data and a proposed modification to the Continuous Updating Estimator (CUE). This chapter is based on a joint project with Dennis Kristensen. We propose a regularized version of the Continuously Updated Estimator (CUE), which we call the quasi-likelihood GMM (QL-GMM) estimator, as a solution to the no-moment problem of the CUE. The estimator is obtained by adding the log-determinant of the optimal weighting matrix to the CUE objective function. The motivation for this term is asymptotic: The QL-GMM objective function is the large-sample log-likelihood of the sample moments. The additional term works as a finite-sample penalization. Analytical results, for the linear setting, and extensive Monte Carlo simulations show that QL-GMM restores the finite sample moments of CUE at the cost of slightly bigger biases compared to the CUE in some settings.

Chapter 2

Dynamic Unobservable Heterogeneity: Income Inequality and Job Polarization

2.1 Introduction

In recent years, administrative datasets have become increasingly available. While this wealth of data can be instrumental in answering several key questions in Economics, it also introduces modeling challenges. Most of the time, administrative data are *micropanels*, which are panel data where many units (N) are observed for a medium number of time periods (T), and thus provide rich information on individuals and firms over time. However, micropanels usually have few dimensions of observable heterogeneity: for instance, administrative data on earnings typically lack information on education, marital status, and health conditions, with demographical variables for each worker limited to age and gender. This drawback makes it crucial to model unobservable heterogeneity (e.g., ability, skills) both over time and across individuals, and the medium time-series dimension requires careful modeling of the dynamics. Even when several variables are observed, it is still important to assess the role played by unobserved heterogeneity. Unobservable heterogeneity is not only interesting per se, but it also affects several other outcomes of interest.[1]

[1]For instance, heterogeneity in earnings dynamics influences predicted mobility out of low earnings (Browning et al., 2010); heterogeneity in income profiles conditional on parents' background is

Indeed, many important questions in the earnings literature, covering topics such as wage inequality or insurance against earnings shocks, require an understanding of the interplay between dynamics and heterogeneity.[2] In addition to this, a modeling framework that features pervasive unobservable heterogeneity and dynamics would be useful in addressing new empirical questions using administrative data.

In this chapter, I propose the use of state-space methods as a unified econometric framework for the study of heterogeneity and dynamics in micropanels. I estimate unobservable heterogeneity and uncover how such heterogeneity has changed over time. As a key contribution, I formally study identification and inference in models with pervasive unobservable heterogeneity. Armed with this framework, I analyze how earnings dynamics of UK workers differ across occupations and over time, making use of a novel dataset on UK earnings, the New Earnings Survey Panel Data (NESPD). My approach and findings reconcile empirical evidence of an increase in the 50/10 wage gap (the ratio of median and low wages) and the documented phenomenon of job polarization (increase in employment in low- and high-skill occupations alongside a simultaneous decrease in middle-skill occupations).

Several econometric methods, often applied to the study of earnings dynamics, treat unobservable heterogeneity as nuisance parameters. Following Almuzara (2020) and Botosaru (2020), I depart from the existing approach in the literature and explicitly treat unobservable heterogeneity as the main object of interest. Almuzara (2020) and Botosaru (2020) adopt a non-parametric approach for estimation of unobservable heterogeneity in earnings models. I consider comparatively richer heterogeneity and dynamics, while imposing a mild parametric assumption on the

crucial to the study of intergenerational mobility (see Mello, Nybom, and Stuhler, 2020).

[2]The distinction between transitory and persistent shocks and the trade-off between heterogeneity and persistence are useful in explaining how individual earnings evolve over time and in decomposing residual earnings inequality into different variance components; the persistence of earnings affects the permanent or transitory nature of inequality (MaCurdy (1982), Lillard and Weiss (1979), Meghir and Pistaferri (2004)). The components of the stochastic earnings process drive much of the variation in consumption, savings, and labor supply decisions, (see Guvenen (2007), Guvenen (2009), Heathcote et al. (2010), Arellano et al. (2017)). Moreover, they play a crucial role for the determination of wealth inequality, and for the design of optimal taxation and optimal social insurance. Finally, separating permanent from transitory income shocks is relevant for income mobility studies and to test models of human capital accumulation.

standardized error term. I assume that innovations are Gaussian, but this assumption can be relaxed, and several more flexible distributions, e.g. mixtures of normals, can be considered.[3] Moreover, the approach proposed in this chapter lends itself to several generalizations, such as unbalanced panel data and measurement errors, and can be adapted to accommodate a treatment of heterogeneity as either fixed or random effects. Finally, my proposal for estimating the distribution of interest, which builds upon results in the state-space literature, improves upon existing approaches in that it allows to analyze heterogenous dynamics of income process in a flexible way, separating the dynamic component, modeled as a time varying parameter, from the heterogenous time-invariant part, modeled as a state variable.

Knowledge of the distribution of unobserved heterogeneity allows us to answer interesting economic questions or make policy decisions. For instance, understanding the shape of the skill distribution, and separating this from skill prices and from the heterogenous dynamics of income shocks, is important to investigate the sources of the uneven distribution of labour market outcomes across workers. An example is provided in the empirical application, where I recover the distribution of skills for different categories of workers and disentangle this from the time-varying price of the skill, and from the potentially heterogenous dynamics of the income process (autocovariances and autocorrelations).[4] The findings of large heterogeneity in the dynamics of income process are informative in their own right.

The chapter's contribution is twofold: methodological and empirical.

My *first* contribution is to derive theoretical results on how to adapt state-space methods to the analysis of panel data featuring heterogeneous dynamic structures. The choice of using state-space methods with filtering and smoothing techniques is motivated by their usefulness for estimation and inference about unobservables in dynamic systems. As emphasized by Durbin and Koopman (2012) and Hamilton (1994), state-space methods include efficient computing algorithms that provide (smoothed) estimates of unobservables, while providing flexible and general model-

[3] Note that, when errors are not normally distributed, results from Gaussian state-space analysis are still valid in terms of minimum variance linear unbiased estimation.

[4] Several authors in the income dynamics suggest the importance of heterogeneity in income dynamics, see (Browning et al., 2010)

ing that can incorporate individual explanatory variables, macro shocks, trends, seasonality, and nonlinearities. Another main advantage is that these methods can be used in the presence of data irregularities, e.g. unbalanced panel data and measurement error. The models typically considered in the earnings literature, e.g. ARIMA, are a special case of state-space models but state-space methods include techniques for initialization, filtering, and smoothing. If the goal is to uncover the evolution of the state variables, state-space models are the most natural choice. Multivariate extensions with common parameters and time-varying parameters are much more easily handled in state-space modeling with respect to a pure ARIMA modeling context.

State-space methods have been mainly used in the context of time series models or with macropanels (panel data with few units observed over many time periods), but the unique structure of micropanels requires the development of new econometric tools for analysis. There is a lack of theoretical results on how to extend their use to micropanels for the analysis of heterogeneous dynamic structures.[5] Therefore, I adapt state-space methods to the analysis of unobserved heterogeneity in micropanels and formally study identification and inference in the context of these heterogeneous models. I show how to consistently estimate the cross-sectional distributions of unobservables in the system and uncover how such heterogeneity has changed over time.

A mild parametric assumption on the standardized error term offers substantial advantages for identification and estimation, and delivers a flexible and general approach. Following the literature on state-space methods, I propose an argument for identification based on a large-T approach. In Appendix A.3, I also consider a fixed-T identification approach to establish a comparison with the existing nonparametric literature. I discuss the corresponding estimation procedures and further analyze the asymptotic properties of these distribution estimators. In the existing

[5]Some notable exceptions are the Seemingly Unrelated Times Series Equations (SUTSE) by Commandeur and Koopman (2007), and Dynamic Hierarchical Linear models by Gamerman and Migon (1993) and by Petris and An (2010) but the focus of the analysis is rather different. I build on these models, discuss the differences, and provide theoretical results on how to recover the cross-sectional distribution of heterogeneous components.

literature, properties of the distribution estimators for the individual parameter estimates obtained from state-space models are unknown. Moreover, it is computationally challenging to extend state-space analysis and filtering to heterogeneous micropanels, which feature large N.

As a first step of the analysis, I consider a simple state-space model and treat the history of each individual i as a separate time series. Identification of the parameters relies on a large T argument, while asymptotic properties of the distribution estimators are established under some ratio between N and T. Building on the work of Okui and Yanagi (2020) and Jochmans and Weidner (2018), I derive this ratio and propose a bias correction for small T.

In a second step, I introduce time-varying parameters in the state-space model and further consider extensions where these parameters are assumed to be common across groups of similar individuals. I discuss how the identification results change in this setting. To devise a tractable estimation strategy, I use stratification as a device to reduce the computational burden of a large cross-sectional dimension on filtering and smoothing algorithms. Once I estimate the parameters and state variables of interest, a larger cross-section is used to consistently estimate the distribution of heterogeneous unobservables.

Finally, in Appendix A.3, I consider a fixed-T approach to explore the relationship to the current non-parametric approach, (see Almuzara, 2020, and Botosaru, 2020), which relies on a fixed-T argument for identification of the cross-sectional distribution of unobservables in the model. The main limitation of fixed-T approaches is that the condition for identification may be difficult or even impossible to verify and existing estimation techniques can be computationally expensive. I show how the parametric assumption on the error term can permit achieving identification with a short number of time periods, making the analysis feasible when richer heterogeneity is allowed in the model. I also discuss what the implications of a parametric assumption on error terms are for regular identification of the distribution of unobservables, following the work of Escanciano (2020).[6]

[6]Regular identification of functionals of nonparametric unobserved heterogeneity means identification of these functionals with a finite efficiency bound.

My *second* contribution is to provide new empirical evidence on the phenomenon of job polarization using a novel UK micropanel, the NESPD, and to study it within a dynamic framework. Analysis of job polarization in the literature is typically grounded on a static approach. The literature on job polarization, pioneered by Autor et al. (2006), defines job polarization as a significant increase in employment shares in low-skill occupations and high-skill occupations, associated with a simultaneous decrease in employment shares in middle-skill occupations, which is a pattern that has been been observed and documented in the US and UK over the last 40 years.[7] I use this novel dataset to test several hypotheses on the relation between job polarization and income inequality. The NESPD is a survey directed to the employer, running from 1975 to 2016, with large cross-sectional and time-series dimensions, which allow the earnings process to feature type dependence in a flexible way. Stratification by observables is possible and replaces the first-stage regression of earnings on covariates, which restricts the dependence of earnings on them. I analyze how the distributions of unobservables in earnings processes have evolved over time and across occupations, and separate the role that workers' skills, labor market instability, and other types of earnings shocks have played on inequality. I use the proposed modeling framework to test whether the distribution of individuals' skills among different occupations has evolved over time and by different age groups. Moreover, I investigate how the corresponding skill prices have changed, and how the distributions of permanent and transitory shocks have changed over time and by occupation.

This chapter uses the answers to the above questions to reconcile the empirical evidence that an increase in the 50/10 wage gap (inequality between the low and median wages) has occurred despite the documented phenomenon of job polarization, which would predict the opposite if relative demand is rising in the low-skill jobs relative to middle-skill jobs. The findings can provide key insights to inform policy decisions based on the dynamics of earnings and of their distributions over time, and are relevant to think about the evolution of labor markets and inequality,

[7] Following the literature, occupations are classified into the categories of low-, middle-, and high-skill jobs based on 1976 wage density percentiles.

also during and after the COVID-19 pandemic. Another interesting empirical question is to uncover heterogeneity in firms' productivity and document how this has changed over time.

To conclude, I develop a state-space framework as a new tool for modelers, with several advantages for identification and estimation, which can be used to address questions on dynamic unobservable heterogeneity in many settings.

The outline of the chapter is as follows: Section 2.2 presents an overview of the related literature. Section 2.3 provides the model setup and a sketch of the methodology. In Section 2.4, I establish the argument for identification, while the corresponding estimation procedure and the properties of the distribution estimators are discussed in Section 2.5. Section 2.6 provides a discussion of the Gaussian assumption and further extensions. In Section 2.7, I describe the dataset used for the empirical analysis. In Section 2.8, I present the empirical application and report empirical findings. Finally, Section 3.5 concludes and discusses directions for future research.

2.2 Related literature

There is an extensive literature on state space methods for time series or macropanels, which are panel data with small N and large T (Durbin and Koopman (2012), Hamilton (1994)). However, the unique nature of micropanels requires the development of new econometric tools to make use of state space methods. I contribute to this econometric literature on state-space by adapting existing methods to suit the characteristics of administrative data, i.e. micropanel data, which feature large N. In particular, I derive theoretical results on how to consistently estimate the cross-sectional distribution of unobservables estimated with state space models.

In order to establish the asymptotic properties of (and make inference on) the estimators of the cross-sectional distribution of unobservables, I rely on the literature on heterogeneous dynamic panel data (Okui and Yanagi (2020), Jochmans and Weidner (2018), Mavroeidis et al. (2015)). Okui and Yanagi (2020) propose a model-free approach, whereas Jochmans and Weidner (2018) consider a Gaussian

assumption on error term but obtain similar results. Finally, Mavroeidis et al. (2015) consider heterogeneous AR(1) models with a fixed-T setting. I extend these existing approaches to investigate the asymptotic properties of the estimator of the cross sectional distribution of unobservables, which are estimated in a first-stage using a state-space model.

Panel data factor models, e.g. Bai (2009), are related to the analysis of panel data with state space methods since dynamic factor models are special cases of state-space models where the econometrician specifies dynamic properties for latent factors in the state equation. However, the state vector is small, and the goal of the analysis is to find commonalities in the covariance structure of a high dimensional dataset.

By developing the corresponding fixed-T approach in Appendix A.3, I explore the relation of my methodology with a recent literature on estimation of the cross-sectional distribution of unobservables with panel data for the analysis of earnings processes. Almuzara (2020) and Botosaru (2020) adapt the identification argument in Hu and Schennach (2008), with the aim of identifying the distribution of heterogeneous variance and permanent components in earning processes. I consider a more general process but impose a (flexible) parametric assumption on the error term: in particular, I focus on large dimensions of heterogeneity, with time-varying parameters, and I impose a mild parametric assumption on the standardized error term. Moreover, this approach lends itself to generalizations such as allowing for unbalanced panel data and measurement errors.

This chapter also relates to the literature on earning dynamics. The literature on the analysis of earnings processes is large and can be distinguished into several strands: one strand focuses on the permanent-transitory decomposition of earnings residuals (Abowd and Card (1989), MaCurdy (1982), Lillard and Weiss (1979)); another strand introduces growth-rate heterogeneity, e.g. Baker (1997), Haider (2001), Guvenen (2009); a third strand considers income variance dynamics allowing for conditional heteroskedasticity in permanent and transitory shocks, e.g. Meghir and Pistaferri (2004), Hospido (2012), Botosaru and Sasaki (2018); finally,

nonlinear models have recently been proposed by De Nardi et al. (2016), Arellano et al. (2017). Guvenen et al. (2015) and Browning et al. (2010) introduce pervasive heterogeneity and are the closest to the present chapter. However, Browning et al. (2010) do not consider a transitory-persistent decomposition of earnings shocks and both these chapters do not propose arguments for identification and estimation of the cross-sectional distribution of unobservables.

Finally, I investigate the relationship between wage inequality and job polarization, which has only been analyzed using static approaches in the literature. The literature on job polarization, pioneered by Autor et al. (2006), defines job polarization as a significant increase in employment shares in low-skill occupations and high-skill occupations, associated with a simultaneous decrease in employment shares in middle-skill occupations, which is a pattern that has been been observed and documented in the US and UK over the last 40 years. The phenomenon of job polarization has been documented by Autor et al. (2006) for the US, and by Goos and Manning (2007) for the UK. The literature that supports the hypothesis of skill-biased technical change cannot explain the increase in employment in low- and high-skill occupations alongside a simultaneous decrease in medium-skill occupations (U-shape in figure 1) because it would only predict change in demand for unskilled vs skilled workers. The hypothesis of automation and routinization, advanced by Autor et al. (2006), can explain this U-shape, but contradicts the fact that wages in low-skill jobs have been falling relative to those in medium-skill jobs. Indeed, one would think that the opposite occurs if relative demand is rising in the low-skill jobs relative to middle-skill jobs. The modeling approach developed in my chapter links the literature on earnings dynamics and wage inequality with the literature on job polarization and investigates this puzzle by testing different hypotheses on the equality of distributions of unobservables over time and across occupations.

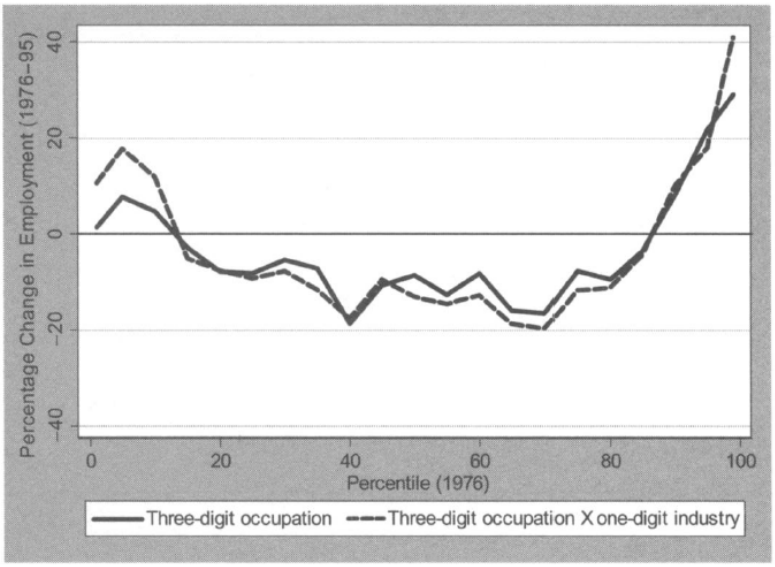


Figure 2.1: Job Polarization in the UK

The graph is taken from Goos and Manning (2007). It shows the impact of job polarization on employment growth by wage percentile. Data are taken from NES using 3-digit SOC90 code. Employment changes are taken between 1976 and 1995. Percentiles are the 1976 wage density percentiles.

2.3 Model Setup

I start by describing a general state-space model and how a model of earning process can be written in terms of a state-space representation. I then discuss the object of interest and sketch the proposed methodology.

2.3.1 State-Space Model

The state-space representation of a dynamic system is used to capture the dynamics of an observable variable, y_{it}, in terms of unobservables, known as the state variables for the system, z_{it}. Consider the following state-space representation to describe the dynamic behavior of y_{it}, for $i = 1,...,N$, and $t = 1,...,T$:

$$\begin{aligned} y_{it} &= A_{it} z_{it} + D_{it} x_{it} + \sigma_i \varepsilon_{it} && \text{(observation equation)} \\ z_{it+1} &= B_{it} z_{it} + R_{it} \eta_{it} && \text{(state equation)} \end{aligned} \tag{2.1}$$

where I name $\tilde{\varepsilon}_{it} \equiv \sigma_i \varepsilon_{it}$ the raw errors and ε_{it} the "standardized" errors; $\varepsilon_{it} \sim \mathcal{N}(0, H_t)$, $\eta_{it} \sim \mathcal{N}(0, S_{it})$; z_{it} denotes the state variables; $\tilde{\varepsilon}_{it}$ and η_{it} are the

errors. I further make an orthogonality assumption on the error terms: ε_{it} and η_{it} are independent of each other and over time. A vector of exogenous observed variables x_{it} can be added to the system. The state equation describes the dynamics of the state vector, while the observation equation relates the observed variables to the state vector. The unobservables of the model are the (potentially time-varying) parameters, the state variables, and the error terms. To complete the system and start the iteration via the Kalman filter I further make the assumption that for each individual i, the initial value of the state vector, z_{i1} is drawn from a normal distribution with mean denoted by $\hat{z}_{i1|0}$ and variance $P_{i1|0}$.[8] Assuming the parameters are known, the Kalman filter recursively calculates the sequences of states $\{\hat{z}_{it+1|t}\}_{t=1}^{T}$ and $\{P_{it+1|t}\}_{t=1}^{T}$ where $\hat{z}_{it+1|t}$ is the optimal forecast of z_{it+1} given the set of all past observations $(y_{it},...,y_{i1},x_{it},...x_{i1})$, and its mean squared forecast error is $P_{it+1|t}$. It does so by first getting the filtered values of the states $\{\hat{z}_{it|t}\}_{t=1}^{T}$ and variances $\{P_{it|t}\}_{t=1}^{T}$. When the interest is in the state vector per se, it is possible to improve inference on it by obtaining the smoothed estimates of the states, i.e. $\{\hat{z}_{it|T}\}_{t=1}^{T}$ and $\{P_{it|T}\}_{t=1}^{T}$, i.e. the expected value of the state when all information through the end of the sample, up to time T, is used, and its corresponding mean square error.[9] When parameters are unknown, maximum likelihood estimation is possible but presupposes the model to be identified.[10]

The model for earnings y_{it}, of an individual i at time t, can be cast in the general model 2.1 with the following state-space representation:

$$y_{it} = [p_t]\alpha_i + z_{it} + \sigma_i\varepsilon_{it} \qquad \text{(observation equation)}$$

$$z_{it+1} = \rho_i z_{it} + \eta_{it} \qquad \text{(state equation)}$$

with $\varepsilon_{it} \sim N(0,H_t)$ and $\eta_{it} \sim N(0,S_{it})$, where A_{it}, z_{it}, B_{it}, and R_{it} in model 2.1 are respectively replaced by $A_{it} = [p_t\, 1]$, $z_{it} = [\alpha_i z_{it}]'$, $B_{it} = [1,0;0,\rho_i]$, $R_{it} = [0,1]$. The

[8] If the vector process z_{i1} is stationary, i.e. if the eigenvalues of B_{it} are all inside the unit circle, then $\hat{z}_{i1|0}$ and $P_{i1|0}$ would be the unconditional mean and variance of this process, respectively. If the system is not stationary or time-varying then they represent the initial guess for z_{i1} and the associated uncertainty.

[9] The general formulas used by the Kalman filter and smoother are provided in Appendix A.1.

[10] Details on the likelihood are provided in Appendix A.1.

term $D_{it}x_{it}$ of the general model 2.1 is here omitted. In this specification, the individual specific component α_i enters the state vector, and the coefficient p_t enters the matrix of parameters A_{it} in the general model described in 2.1. The factor p_t might be included as a measure of skills price. Note that transitory shocks are assumed to be i.i.d. in these models. However, more general moving average representations, which are common in the earnings literature, can be accommodated by augmenting the state vector accordingly. In addition, note that the measurement error is not separately modeled, hence the error term in the observation equation should be interpreted as a mixture of transitory earnings shocks and measurement error.
An extension of this model to include a term $\beta_i t$ can account for an individual's ith specific income growth rate with cross-sectional variance σ^2_β, see HIP model in Guvenen (2009).[11] A model for earnings could further include job-specific (or firm-specific) effects γ_i, for job (or firm) k, with $j_{ik} = 1(\mathrm{K}_i = k)$.[12]

For the rest of the chapter, I focus on model 2.3.1 but the analysis could be potentially extended to the more general model described in 2.1.

2.3.2 Object of interest

The objects of interest are the cross-sectional distributions of unobservables (state variables and parameters in the model described above) and their dynamics over time. I relax the strong assumption of a fully parametric approach to estimate unobserved heterogeneity: I do not impose any restrictions on the cross-sectional distributions of unobservables. I impose a parametric assumption on the standardized error term of the model, in particular I assume that this term is Gaussian, and further discuss the validity and implications of this assumption in Section 2.4 and 2.6.

The approach of identifying and estimating the full cross-sectional distribution of unobservables offers several advantages compared to an alternative approach that only targets certain moments. First, once the full distribution is estimated, it is pos-

[11] Following Guvenen (2009), it would be possible to assume that individuals form their beliefs about their heterogeneous intercept and slope and update their beliefs according to the observed income realizations. In the following, I do not consider the worker's optimal learning process.

[12] The New Earnings Survey Panel Data (NESPD) can be merged with the Business Structure Database (BSD) to obtain a matched employer-employee dataset for the UK labour market and include a firm component of pay in the earnings process.

sible to estimate all moments (note that moments beyond the second-order may be of interest), as well as the quantiles and other features of the distribution. Second, this approach offers the possibility of analyzing the dynamics of the distribution, while an approach that only targets moments would require further specifying their evolution over time. An additional advantage is that it would be possible to investigate ex-post which observables predict the estimated heterogeneity without loss of statistical power, which could affect alternative approaches in existing studies, see Lewis et al. (2019).[13]

The identification results are presented in the following section. First, I consider a simpler model of earnings and treat the history of each individual i as a separate time series. I provide the argument for identification of the parameters and states, and of their cross-sectional distribution. The identification of the parameters relies on a large T argument, while the asymptotic properties of the estimator of the cross-sectional distribution of parameters and states are established under some ratio of N and T. I derive this ratio and propose a bias correction method to use when T is small. In the second step of the analysis, I introduce time-varying parameters in the state-space model. I discuss how the identification results change in this setting. As an additional extension, I consider panel data factor models. Finally, in Appendix A.3, I relate to the nonparametric existing approach, which relies on a fixed-T argument for identification of the unobservables in the model and of their cross-sectional distribution. I show how the parametric assumption on the error term can permit to achieve identification with a shorter number of time periods and discuss whether high-level assumptions for identification hold.

Note that the proposed methodology encompasses treatment of unobserved heterogeneity as both fixed effects and random effects, with some differences in the assumptions required to establish the properties of the distribution estimators in the two cases.

[13] It might be an interesting empirical questions to see how much of this unobserved heterogeneity can be explained by observables such as education, marital status, health information. Ideally by using a dataset with lots of observables, one could quantify their contributions to unobserved heterogeneity.

2.4 Identification

In the following Section, I discuss identification of the unobservables in state-space models.

2.4.1 Benchmark Model

First, treat the earnings history of each individual i as a separate time series. In particular, let's consider the time-invariant version of model 2.3.1 and assume that for each individual i, the time series is represented by the state-space model:

$$\begin{aligned} y_{it} &= \alpha_i + z_{it} + \sigma_i \varepsilon_{it} && \text{(observation equation)} \\ z_{i,t+1} &= \rho_i z_{it} + \eta_{it} && \text{(state equation)} \end{aligned} \tag{2.2}$$

with $\varepsilon_{it} \sim N(0,1)$ and $\eta_{it} \sim N(0, \sigma^2_{i,\eta})$. I further make the assumption that the innovations ε_{it} and η_{it} have zero mean, are independent of each other and over time, and independent of α_i. This model decomposes earnings into a deterministic fixed effect, which captures heterogeneity in income profiles due to different unobserved and time invariant characteristics, e.g. ability, and a stochastic term, which has a transitory and a persistent component, which are idiosyncratic and unobserved shocks to income such as health shocks, bonus, promotions. This decomposition and the orthogonality assumption between transitory and persistent components are widely used in the earnings literature.

I first discuss how the model's parameters are identified and how it is possible to identify the cross-sectional distribution of the parameters and state variables.

A state-space model is identified when a change in any of the parameters of the state-space model would imply a different probability distribution for $\{y_{it}\}_{t=1}^{\infty}$. There exist several ways of checking for identification. Burmeister et al. (1986) provide a sufficient condition for identification: a state-space model is minimal if it is completely controllable with respect to the error term (and external variable directly affecting both the observed and the state variables) and completely observable. If the state-space is minimal, then it is identified. The model considered above is observable as the observation matrix has rank equal to the number of state variable

where the observability matrix is defined as: $O = [A'B'A'(B^2)'A'(B^3)'A'...(B^{n-1})'A']$ where n is the number of state variables. Require $\rho \neq 1$ for observability. Under this condition, the above model is also controllable with respect to the error term as the controllability matrix $C = [RBRB^2RB^3R...B^{n-1}R]$ has full rank.

An alternative way of checking identification of a state-space model is to rely on the exact relationship between the reduced form parameters of an ARIMA process and the structural parameters in the state-space model, and use the condition for identification of parameters in ARIMA models. The literature on linear systems has also extensively investigated the question of identification, see Gevers and Wertz (1984) and Wall (1987) for a survey of some of the approaches.

For the above state-space model, it is possible to verify that under stationarity the following holds, $\forall i$:

$$\rho_i = \frac{Cov(y_{it}, y_{it+2})}{Cov(y_{it}, y_{it+1})}$$

$$\sigma_i^2 = Var(y_{it}) - \frac{Cov(y_{it}, y_{it+1})}{\rho_i} = Var(y_{it}) - \frac{Cov(y_{it}, y_{it+1})}{\frac{Cov(y_{it}, y_{it+2})}{Cov(y_{it}, y_{it+1})}}$$

$$\sigma_{i,\eta}^2 = (Var(y_{it}) - \sigma_i^2)(1 - \rho_i^2)$$

$$\alpha_i = E(y_{it})$$

where the mean, variance, and covariances are moments of the distribution of y_{it} taken over time, for each individual i.

Once I establish identification of the model's parameters, which is based on properties of each individual's ith time series, I can exploit the cross-section of the time series to identify the cross sectional distributions of the variables of interest (parameters and states), and analyze the asymptotic properties of these distribution estimators. In line with these results, I derive nonparametric bias correction via split

panel Jackknife methods when T is small.

2.5 Estimation

In this Section, I present the main results on asymptotic properties of the distribution estimators of unobserved estimated from state-space models. I discuss how the parametric assumption on the error terms helps to establish these results. Finally, I provide some details on the estimation procedure, which I adopt in the empirical application. In the Appendix A.3, I consider the alternative fixed-T identification argument and estimation procedure.

2.5.1 Asymptotic Properties of the Distribution Estimators

For the above time-invariant state-space model 2.2, I collect all unknown parameters in a vector $\theta_i = \{\alpha_i, \rho_i, \sigma_i^2, \sigma_{\eta i}^2\}$. Let $\hat{\theta}_i$ be the MLE estimator for the vector of parameters θ_i, obtained as: $\hat{\theta}_i = \arg\max_{\theta_i} Q_T(\theta_i)$, where $Q_T(\theta_i) = T^{-1}\sum_{t=1}^{T} \log f(y_{it};\theta_i) \coloneqq m(w_{it},\theta_i)$ and $f(y_{it};\theta_i)$ is the likelihood from the state-space model as derived in Appendix A.1. Following a similar notation and argument as in Okui and Yanagi (2020), define $\mathbb{P}_N^{\hat{\theta}} \coloneqq N^{-1}\sum_{i=1}^{N} \delta_{\hat{\theta}_i}$, as the empirical measure of $\hat{\theta}_i$, where $\delta_{\hat{\theta}_i}$ is the probability distribution degenerated at $\hat{\theta}_i$. Also, let P_0^{θ} be the probability measure of θ_i. Denote as $\mathbb{F}_N^{\hat{\theta}}$ the empirical distribution function, so $\mathbb{F}_N^{\hat{\theta}}(a) = \mathbb{P}_N^{\hat{\theta}} f$ for $f = 1_{(-\infty,a]}$, where $1_{(-\infty,a]}(x) \coloneqq 1(x \leq a)$ and the class of indicator functions is denoted as $\mathscr{F} \coloneqq \{1_{(-\infty,a]} : a \in \mathbb{R}\}$. Similarly, $\mathbb{F}_0^{\theta}(a) = P_0^{\theta} f$. Finally, denote as $P_T^{\hat{\theta}}$ the probability measure of $\hat{\theta}_i$. In the following, for simplicity of notation, I omit superscripts $\hat{\theta}$ and θ, so $\mathbb{P}_N = \mathbb{P}_N^{\hat{\theta}}$, $\mathbb{F}_N = \mathbb{F}_N^{\hat{\theta}}$, $P_0 = P_0^{\theta}$, $\mathbb{F}_0 = \mathbb{F}_0^{\theta}$, $P_T = P_T^{\hat{\theta}}$, $\mathbb{F}_T = \mathbb{F}_T^{\hat{\theta}}$.

Assumption 1 Assume that $\{\{\varepsilon_{it}\}_{t=1}^{T}, \{\eta_{it}\}_{t=1}^{T}\}_{i=1}^{N}$ is i.i.d. across i and y_{it} is a scalar random variable.

Assumption 2 The true parameters θ_i must be continuously distributed.

Assumption 3 Further, assume that: $|\rho_i| < 1$; θ_i identified, and not on the boundary of parameter space.

Assumptions 2 and 3 state standard and sufficient conditions that are required for the ML estimators of the unknown parameters in the time-invariant Gaussian

state-space model to be consistent and asymptotically normal. In particular, Assumption 3 is required to establish convergence in probability of $\hat{\theta}_i$ to θ_{i0}, as $T \to \infty$. Note that even without normal distributions the quasi maximum likelihood estimates $\hat{\theta}_i$, obtained assuming Gaussian errors, is consistent and asymptotically normal under certain conditions, see White (1982).

Indeed, the above model is a Gaussian time-invariant state space model, which has a stationary underlying state process (ρ_i is assumed to be less than 1 in absolute value), and which has the smallest possible dimension, see Hannan and Deistler (2012). Under these general and sufficient conditions, then the MLE estimator is consistent and asymptotically normal if the true parameters are identified and not at the boundary of the parameter space, see Douc et al. (2014).
These assumptions are not restrictive and are likely to hold within the context of earnings dynamics. The assumption that ρ_i is in absolute value smaller than 1 is reasonable when allowing for lots of unobservable heterogeneity in the earnings process. In the empirical application, I find that estimates of the persistence parameter are smaller than 1. This empirical evidence is consistent with existing findings in the earnings literature: Browning et al. (2010) reject the hypothesis that any worker has a unit root when allowing for pervasive heterogeneity.

Assumption 4 The CDFs of θ_i is thrice boundedly differentiable. The CDFs of $\hat{\theta}_i$ is thrice boundedly differentiable uniformly over T.

Under these assumptions, it is possible to establish uniform consistency and asymptotic normality of the distribution estimator. In the following theorem, I show that the estimator for the distribution of the true individual parameters and states uniformly converges to their true population distribution and it converges in distribution at the rate $N^{3+\varepsilon}/T^4$, where $\varepsilon \in (0, 1/3)$, if the above assumptions hold.

Theorem 1 Under Assumption 1-4, when $N, T \to \infty$: (i) $\sup |\mathbb{P}_N f - P_0 f| \xrightarrow{as} 0$, where $\xrightarrow{as}$ signifies almost sure convergence. Moreover, (ii) when $N, T \to \infty$, with $N^{3+\varepsilon}/T^4 \to 0$ and $\varepsilon \in (0, 1/3)$: $\sqrt{N}(\mathbb{P}_N - P_0) \rightsquigarrow G_{P_0}$ in $l^\infty(\mathscr{F})$, where $\rightsquigarrow$ means weak convergence and $\mathbb{G}_{P_0}$ is a Gaussian process with zero mean and covariance

function $F_0(a_i \wedge a_j) - F_0(a_i)F_0(a_j)$ with $f_i = 1(-\infty, a_i]$ and $f_j = 1(-\infty, a_j]$ for $a_i, a_j \in R$ and $a_i \wedge a_j$ is the minimum of a_i and a_j and where $l^\infty(\mathscr{F})$ is the collection of all bounded real functions on $\mathscr{F}$.

The key idea behind Theorem 1 is that the asymptotic properties of the ML estimator $\hat{\theta}_i$ for each individual's i parameters guarantee that it is possible to bound the norm of the difference between the cross-sectional distribution of the ML estimators and the true distribution of the true parameters, i.e. the term $\sup |P_T f - P_0 f|$. See Appendix A.2 for the proof. This result combines the existing results in the state-space literature with the results of the model free approach in Okui and Yanagi (2020): more specifically, I rely on the results in the existing state-space literature to establish under which assumptions the ML estimator $\hat{\theta}_i$ is consistent and asymptotically normal; I then build on the proof in Okui and Yanagi (2020) to bound the difference between the distribution estimator and the true distribution of the parameters.

Following Okui and Yanagi (2020) and Jochmans and Weidner (2018), when T is small I propose a nonparametric bias correction method via split-panel jackknife (HPJ). I divide the panel along the time series dimensions into two parts and obtain $\hat{F}^{HPJ} = 2\hat{F} - \bar{F}$, where $\hat{F}$ is the estimator obtained using the whole sample, while $\bar{F} = (\hat{F}^1 + \hat{F}^2)/2$ with $\hat{F}^j$ for $j = 1, 2$ being the estimators obtained when using each half of the panel.

2.5.2 Time-varying Model

When adding time-varying parameters in the state-space model for each i, the derivation of the Kalman filter and smoother is essentially the same as for the case of time-invariant model. Note that if the matrices $A_{it}, D_{it}, B_{it}, R_{it}, H_{it}, S_{it}$, in equation 2.1, are generic functions of the stochastic variable x_t, then, even if the error terms are normal, the unconditional distribution of the state variable and of the observation y_{it} is no longer normal, while normality can be established conditionally on the past observations and x_t.

Assumption 3 in Theorem 1 can be modified by using existing results that provide conditions on asymptotic properties of the ML estimator for time-varying state-space models. Indeed, assumption 3 can be relaxed along several dimensions: it is possible to rely on results in Chapter 7 of Jazwinski (1969) for a departure of the time-invariance assumption, and it is further possible to weaken the assumption that $\rho < 1$ for stability of the filter, as in Harvey (1990).

For time-varying parameters that are common across (groups of) individuals, I consider a multivariate version of the state-space model above. I consider stratification by observables and, within each group, I impose common time-varying parameters (e.g. price of skills) and individual-specific parameters. A general model for earnings y_{it} is:

$$\begin{aligned} y_{it} &= p_t(x_i)\alpha_i + z_{it} + \sigma_i \varepsilon_{it} \\ z_{it+1} &= \rho(x_i) z_{it} + \eta_{it} \end{aligned} \tag{2.3}$$

with time-varying variances for error terms: $\varepsilon_{it} \sim N(0, H_t(x_i))$ and $\eta_{it} \sim N(0, S_t(x_i))$ and where x_i are observable covariates (e.g. gender). This is the same model as 2.3.1. The main challenge is that the Kalman filter and smoother can be computationally intense or even infeasible when the cross-sectional dimension N is large. I give proposals on how to deal with these issues in the estimation section.

I leave for future research to derive the identification results and the changes to Assumption 3 in Theorem 1 to establish the asymptotic properties of the ML estimator for the time-varying state-space model in 2.3.

2.5.3 Implementation

State-space estimation and filtering with heterogeneous dynamic panel data pose econometric challenges. Estimation of the distribution of unobservables is performed in 2 stages: a first step of estimation is performed via state-space methods; then, in a second step, I obtain the empirical cross-sectional distribution of unobservables estimated in the first step.

In the first step, estimation of model' s parameters is based on maximum likeli-

hood.[14] I employ the kalman filtering and smoothing algorithm to get smoothed estimates of state variables and error terms.

The econometric challenge in this first step of estimation is on how to deal with state-space models for a dataset featuring a large cross-section N: given recursive nature of filter, at each period inversion of $F_t = Var(v_t|y_{t-1})$, where $v_t = y_t - A_t E[z_t|Y_{t-1}]$ is the innovation, can be problematic, see Durbin and Koopman (2012) (F_t has size N x N, computationally costly with large N). In the models I consider, H_t is diagonal, hence, it is possible to adopt matrix identity for inverse of F_t. Moreover, I perform stratification as a way to avoid intractability while also addressing the issue of not restricting the dependence of earnings on covariates. When introducing time-varying parameters, I impose that within each group some parameters are common and time-varying parameters (e.g. price of skills), while others are individual-specific (e.g. the standard deviation of the shocks as reported in the matrix R_{it} in model 2.1).

For starting the recursions, I implement diffuse initialization as in De Jong et al. (1991), i.e. the uncertainty around initial states is represented in the model with an arbitrarily large covariance matrix for the initial state distribution.[15]

Once (smoothed) estimates of unobservables are obtained, I obtain the empirical cross-sectional distribution of the unobserved components estimated from the state-space models in the second step of the estimation strategy. Note that dimensionality of vector y_t can vary over time. Thus, the methodology can be easily extended to deal with unbalanced panel data.

2.6 Discussion on Gaussian Error and Extension

One might be worried that the parametric assumption about the innovations ε_{it} and η_{it} in models 2.2 and 2.3 is quite restrictive. Horowitz and Markatou (1996) provide empirical evidence that the normal distribution can approximate well the distribution of the permanent component of the income process. However, there might be

[14] Details on the likelihood are provided in Appendix A.1.

[15] Durbin and Koopman (2012) show that initialization of the Kalman filter is not affected the choice of representing the initial state as a random variable with infinite variance as opposed to assuming that it is fixed, unknown and estimated from observations at t=1.

concerns that the parametric assumption is restrictive for the transitory component of earnings shocks. Indeed, there is empirical evidence that the cross-sectional distribution of transitory shocks features negative skewness and high kurtosis. These stylized facts have been documented, among others, by Arellano et al. (2017) as relevant features of the earnings process.

First and importantly, note that when errors are not normally distributed, results from Gaussian state-space analysis are still valid in terms of Minimum Variance Linear Unbiased Estimation (MVLUE): Kalman Filter estimates are not necessarily optimal, but they will have the smallest mean squared errors with respect to all other estimates based on a linear function of the observed variables $(y_{it}, y_{it-1}, ..., y_{i1}, x_{it}, x_{it-1}, ..., x_{i1})$, see Anderson and Moore (1989).

Second, the homogeneity assumptions may explain some of these stylized facts: once allowing for rich heterogeneity, it is unclear whether the residuals will still display the same features. One interesting empirical question is to test to what extent these features are still present when allowing for rich heterogeneity and time-varying parameters. Assuming individual Gaussian shocks with heterogeneous variances allows obtaining flexible cross-sectional distributions and, depending on the cross-sectional distributions of the heterogeneous variances, the resulting cross-sectional distribution might display the above key features.

Finally, extensions to different distributions are feasible within a state-space framework. Alternative assumptions on error terms can be considered by non-Gaussian state-space models; for instance, the error term can be assumed to follow a Mixture of Normals distribution. It would be interesting to see how much the goodness of fit improves when the assumption on Gaussian shocks is relaxed.

2.7 Data

The dataset used for the empirical application is a novel confidential dataset for the UK, the New Earnings Survey Panel Data (NESPD). It is an annual panel, running from 1975 to 2016. All individuals whose National Insurance Number ends in a given pair of digits are included in the survey, making it representative of the UK

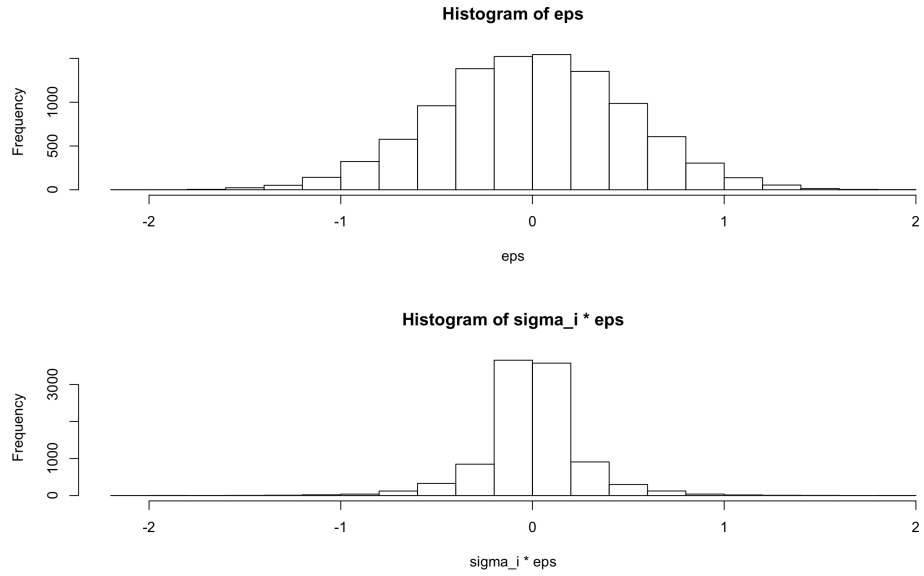


Figure 2.2: Simulated Distribution of Raw Errors

Monte Carlo simulation showing that, despite the assumption of Gaussian errors at individual level, the cross sectional distribution of raw errors can display very high kurtosis (and potentially also skewness) depending on the cross-sectional distribution of heterogeneous variances, σ_i^2.

workforce.[16] It surveys around 1% of the UK workforce.

The questionnaire is directed to the employer, who completes it based on payroll records for the employee; the survey contains information on earnings, hours of work, occupation, industry, gender, age, working area, firms' number of employers, and unionization. This information relates to a specified week in April of each year: the data sample is taken on the 1st of April of each calendar year and concerns complete employee records only. As a result of being directed to the employer, NESPD has a low non-classical measurement error and attrition rate.

For both Standard Industry Classification (SIC) and Standard Occupation Classification (SOC) codes, different classifications have been used over time. I report SIC and SOC codes to the same classification using conversion documents provided by ONS on their website: I map all SIC codes into the SIC07 Division (2-digit); for

[16]It might under-sample part-time workers if their weekly earnings falls below the threshold for paying National Insurance and those that moved jobs recently. Thus, the following categories are likely to be under-sampled: self-employed, some groups of seasonal workers, and those working only few hours irregularly. To address these concerns, one could perform robustness checks using the Labour Force Survey data, which, however, has a much smaller sample size.

those divisions where there are multiple correspondences, I use the information on whether the individual has stayed in the same job in the last twelve months to identify the mapping; I proceed in an analogous way for SOC codes.

I rank occupations by percentiles of the median wage distribution in the starting year and separate them accordingly into three groups: low-skill, medium-skill, and high-skill occupations.[17]

Given large dimensions, this dataset is particularly suited to obtain a flexible treatment of covariates by stratification. I stratify by observables instead of running a first-stage regression on covariates which restricts the dependence of earnings on them. Stratification allows considering a specification for the earning process that features type dependence in a flexible way. In particular, I perform stratification by occupations: high-skill occupations, medium-skill occupations, and low-skill occupations; by age groups; and by gender.

Two time-series dimensions are taken into account. One is the time-series of earnings history for each worker; I keep individuals in the panel only if they have a continuous earning history of at least twelve time periods.[18] The second time-series dimension I think of is the time-series of cross-sections, which can be much longer and is exploited to analyze how unobserved heterogeneity has evolved over time.[19] Therefore, I only consider continuous earning histories for each worker, but the panel is unbalanced because of changes in the composition of the workforce. Note that attrition is not a concern for this dataset as it is the employer who reports information on the employee based on her payroll records.[20]

[17] Another classification might be based on routine task intensity of occupations since one of the main hypotheses put forward to explain job polarization is the bias of recent technological change towards replacing labor in routine tasks (this is called routine-biased technological change, RBTC, by Goos et al. (2014)).

[18] I use the techniques for small bias correction methods described in Section 2.4.

[19] This dimension allows answering several interesting empirical questions, for instance to address questions on the evolution of life cycle inequality over time.

[20] By considering continuous earnings history only, I do not account for transitions into and out of employment. In order to account for the choice of workers, one would probably need a structural model.

2.8 Empirical Application

Over the last 40 years, in the US and UK, there has been a significant increase in employment shares in low-skill occupations and high-skill occupations, and a simultaneous decrease in employment shares in middle-skill occupations. Goos and Manning (2007) document that this phenomenon, known as job polarization, has occurred in the UK since 1975. A likely explanation for it is the automation of some types of jobs only, the middle-skill jobs, which require precision and are easy to be replaced by machines.

In the following figures, the phenomenon of job polarization results in the characteristic U-shape with much a negative change in employment share for middle-skill occupations. Note that this pattern is observed over the whole period, and is not driven by a change in the gender composition of the workforce.[21]

As a result of job polarization, one would expect an increase in wages for both low-skill and high-skill occupations, while a decrease in wages for medium-skill occupations. Indeed, job polarization would predict a rising relative demand in the low-skill relative to middle-skill jobs. However, this has not been the case: on the contrary, earnings inequality also between low and median wages has increased over time. Part of the increase in wage inequality might be justified by the fact that wage growth is monotonically positively related to the quality of jobs. If one includes more controls, the within job inequality significantly reduces. Once one controls for job-specific effects, there should only be between job inequality, not within. However, as suggested by Goos and Manning (2007) the findings that wages in low-skill jobs are falling relative to those in middle-skill jobs presents something of a problem for the routinization hypothesis, as one might expect the opposite if relative demand is rising in the low-skill jobs relative to middle-skill jobs.

The methodology proposed in the chapter is used to shed light on the relation between job polarization and earnings inequality, which is relevant to think about the evolution of labor markets and inequality, also during and after the COVID-19 pandemic. The goal of the empirical analysis is to relate the components and

[21] Results are robust to the chosen level of disaggregation by occupation.

dynamics of the earnings process to the phenomenon of job polarization, which is usually investigated only with a static approach. To this aim, I am going to test different hypotheses on the degree of heterogeneity of the distributions of unobservables, by observables and over time, to shed light on this puzzling empirical evidence.

More specifically, first I am going to consider the time-invariant model used as benchmark model in the analysis. In a second step of the analysis, I am going to introduce time varying parameters, in the form of a time-varying price of skills (p_t) in the model above, and by allowing the variances of the shocks to be time-varying. For both models, for each group obtained by stratification by observables, I use state-space analysis to obtain (smoothed) estimates of unobservables. Finally, I estimate the cross-sectional distribution of the unobservables, potentially for aggregated strata in order to recover a larger cross-sectional dimension needed for inference on distributions. I compare these distributions via tests of the null hypothesis of equal distributions by Kolmogorov-Smirnov test to test for different degrees of heterogeneity.[22]

2.8.1 Toy Model

The following toy model is used to motivate things and illustrate some of the underlying mechanisms that I would like to test.

Consider a model with two types of individuals, $i \in \{LG, HG\}$, where LG stands for low growth type and HG for high growth type. Further, assume that there are 3 types of occupations, $k \in \{LS, MS, HS\}$, i.e. low-skill, medium-sill, and high-skill occupations. The price of the skills in occupation k, at time t, is $\pi_{k,t}$, and $\pi_{LS,t} \leq \pi_{MS,t} \leq \pi_{HS,t}$. The individual i's earnings at time t from occupation k is:

$$y_{i,k,t} = \pi_{k,t}(\alpha_{i,k})$$

where α_{ik} is the heterogeneous level, which is time-invariant. The individual's prob-

[22] There might be a problem of independence if aggregate time effects are taken into account.

lem at time s is:

$$max_k \sum_{t=s}^{T} E(y_{i,k,t})\beta_d^t$$

where β_d is the discount factor. In this scenario one moves from MS to LS occupation if either displaced with probability δ_i or if $\pi_{MS,t}(\alpha_{i,MS}) < \pi_{LS,t}(\alpha_{i,LS})$. Analogously from HS to MS.

I model routinization as a negative demand shock in MS occupation, i.e. $\pi_{MS,t}$ decreases. After this shock, all HG type move from MS occupations to HS occupations, or stay in MS occupations. Vice versa all LG type move from MS occupations to LS occupations, or stay in MS occupations. Assume that, after the shock, for $i = HG$, $\pi_{MS,t}(\alpha_{HG,MS}) \leq \pi_{HS,t}(\alpha_{HG,HS})$ and $\pi_{LS,t}(\alpha_{LG,LS}) \geq \pi_{MS,t}(\alpha_{LG,MS})$. Assuming that there is a nonzero outflow of people from MS occupation, the overall effect would be an increase in inequality.

Now, let's consider a more realistic earnings process by adding the stochastic persistent and transitory components $z_{i,t} + \varepsilon_{i,t}$:

$$y_{i,k,t} = \pi_{k,t}(\alpha_{i,k}) + z_{i,t} + \varepsilon_{i,t}$$

An increase in inequality might occur also if the variances of the stochastic components significantly changed over time and by different type of occupation. This might happen as a result of changes in institutions that have lead to a decline in wages at the bottom of the distribution. In UK there has been a marked decline of both unionization and minimum wage over time.

Several hypotheses can be tested to investigate this phenomenon:

H1: Change in prices of skills by occupation and over time.

H2: Distribution of skills changed by occupation and over time.

H3: Distribution of variance of transitory shocks more concentrated depending on the evolution of unionization and minimum wage by occupation and over time. I investigate this channel given that a possible explanation for increase in inequality might be that change in institutions have been in such a way to lead to a fall in

wages at the bottom of the distribution.

In practice, I consider the time-varying model 2.3:

$$\begin{aligned} y_{it} &= p_t(x_i)\alpha_i + z_{it} + \sigma_i \varepsilon_{it} \\ z_{it+1} &= \rho(x_i) z_{it} + \eta_{it} \end{aligned} \tag{2.4}$$

where x_i is individual's i category of occupation: LS, MS, or HS occupation. I test H1 by comparing the evolution of the prices p_t for workers in LS, MS, and HS occupations. I test H2 and H3 by comparing the distributions of respectively skills, i.e. of the α_i, and variance of shocks ε_{it} and η_{it}, in the different categories of occupation by Kolmogorov Smirnov test of equality of distributions and further compare these distributions over time.

2.8.2 Empirical Findings

The empirical findings provide evidence that earnings dynamics feature considerable unobservable heterogeneity. This is an interesting result in its own right. First, I uncover the amount of unobservable heterogeneity using the simple time-invariant model considered as benchmark model in the theoretical section. I document that workers in middle-skill occupations display significantly different earnings dynamics with respect to workers in other occupations. In particular, as shown in table 2.1, persistence to earnings shocks for workers in middle-skill jobs is on average smaller, over the entire time period. The distribution of persistence has the largest dispersion for workers in low-skill occupations. Moreover, empirical evidence suggests a relatively higher correlation between the skills of workers in middle-skill occupations and the dispersion of earnings shocks they face.

To test the hypotheses presented in the above section, I introduce time-varying parameters in the state-space model. Figure 2.3 displays a pattern of increase in the prices of skills for workers in low- and high-skill occupations, while it shows that the change over time of the skill prices for workers in middle-skill occupations has been unstable.

These preliminary findings can be interpreted as suggestive of a pattern of

		α_i	ρ_i	σ_i^2	α_i	ρ_i	σ_i^2
		1975-1999			2000-2005		
LS	Mean	-0.1909	0.5424	0.0442	-0.3085	0.5028	0.0671
	St. Dev.	0.3112	0.5391	0.0639	0.3534	0.5607	0.1278
	IQR	0.3993	0.7996	0.0422	0.4107	0.8280	0.0724
MS	Mean	-0.1140	0.4620	0.0354	-0.1054	0.4731	0.0368
	St. Dev.	0.2909	0.5526	0.0408	0.3342	0.5440	0.0626
	IQR	0.3719	0.7686	0.0354	0.4760	0.8267	0.0329
HS	Mean	0.1527	0.5095	0.0278	0.2507	0.5926	0.0340
	St. Dev.	0.2873	0.5260	0.0416	0.3879	0.5366	0.0750
	IQR	0.3718	0.7501	0.0259	0.4536	0.7926	0.0293

Table 2.1: Empirical Results for time-invariant model

The table reports the means, standard deviation, and interquartile range (IQR) of the cross-sectional distributions of α_i, ρ_i, and σ_i^2, for workers in LS occupations, MS-occupations, and HS-occupations, for two time windows: 1975-1999, 2000-2005. Split-panel jackknife (HPJ) is used for bias correction.

negative demand shocks in MS occupations over the considered time period. Moreover, there has been a positive shift in the distribution of skills for individual in MS-occupations due to a compositional change in the UK workforce as shown in table 2.2.

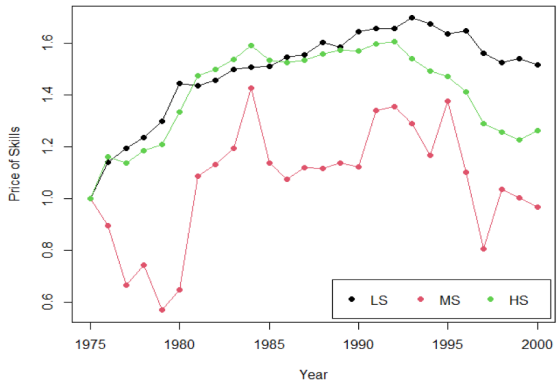


Figure 2.3: Changes in Skill Prices by Occupation

Finally, the dispersion of the variance of transitory shocks has increased over time, comparatively more for workers in LS-occupations, while the variances have not significantly changed over time for workers in MS-occupations. One likely explanation might be that institutions have changed in such a way to lead to a fall

		α_i		
		1975	1985	1995
MS	Mean	-0.0600	-0.0578	-0.0252
	St. Dev.	0.1815	0.1693	0.1271
	IQR	0.1550	0.1438	0.1038

Table 2.2: Empirical Results on the Distribution of Skills for Middle-Skill Occupations

in wages at the bottom of the distribution.

2.9 Conclusions

In this chapter, I propose a formal econometric framework for studying identification and estimation of unobservable heterogeneity and its dynamics. I adapt state-space methods to the analysis of heterogeneous dynamic structures with micro panels. The framework proposed in this chapter allows for rich heterogeneity and dynamics in models, while a mild parametric yet flexible assumption on the distribution of the shocks provides several advantages for identification and estimation.

The framework in this chapter will enable empirical researchers to answer a variety of new empirical questions using administrative data. Moreover, it naturally lends itself to important and useful generalizations such as allowing for common unobservable macro shocks, trends, seasonality, and nonlinearities.

In the empirical application, I use a novel dataset on UK workers, the NESPD, to uncover unobserved heterogeneity in earnings processes and investigate how this is related to the phenomenon of job polarization.

A natural next step in the analysis is to combine the information on UK workers provided by the NESPD with information about the supply side as reported in another novel UK dataset, the Business Structure Database (BSD), which can be merged with NESPD to get a matched employee-employer dataset for UK.

Chapter 3

Individual Forecast Selection

3.1 Introduction

Forecasting individual outcomes (microforecasting) is a key component of economic, policy and business decisions and is becoming increasingly prominent in empirical economics. For example, the literature on long-term treatment effects relies on forecasting the effects of treatments such as early-childhood interventions (García et al., 2020) or job-training programs (Athey et al., 2019). Chamberlain and Hirano (1999) forecast individual incomes for consumption/savings decisions and Liu et al. (2020) forecast revenues of banks after a regulatory change. The literature on teacher value-added (e.g., Kane and Staiger, 2008; Chetty et al., 2014a,b) can be viewed as predicting teacher performance by estimating the unobservable teacher quality.

Micro panel data make it possible to forecast individual outcomes, but present econometric challenges due to the short time series dimension and the few observable characteristics that are typical of these datasets. As a result, existing methods rely on simple models and estimation methods that are based on either the time-series dimension, the cross-sectional dimension ("pooling") or on intermediate approaches such as empirical Bayes ("shrinking towards the mean"). The tradeoffs between these estimators are intuitive. The individual time series is informative about time-invariant unobserved characteristics but provides noisy estimates when it is short. Pooling and empirical Bayes reduce the noise by "borrowing strength

from the majority", but can turn into the "tyranny of the majority" by hiding the unobserved heterogeneity. While the existing literature appears to favour pooling or empirical Bayes methods, this paper shows that using the same forecasting method for all individuals could lead to bad decisions, from loss of accuracy to unfairly penalizing high-performers or rewarding low-performers.

This paper proposes an alternative approach to microforecasting that, instead of using the same forecasting method for all individuals, performs individual forecast selection (henceforth IFS) out of a class of competing forecasting methods. The competing methods could be based on different estimators within the same model, such as a time series estimator versus a pooled (empirical Bayes) estimator, in which case IFS can be viewed as deciding who to pool (who to shrink towards the mean). If the competing forecasts are based on different models, IFS can be viewed as delivering a form of model-based clustering.

The selection of the best forecasting method is based on out-of-sample accuracy over one past time period. For example, suppose we want to forecast at time T the individual outcomes at time $T+1$ and the competing forecasts are the cross-sectional mean or past behaviour. For each individual, we can first establish which of the two forecasts made at time $T-1$ would have been more accurate for the outcome at time T, and then use the same method to forecast the outcome at time $T+1$. Intuitively, the approach suggests using past behaviour for "outliers" - individuals whose unobserved heterogeneity is far from the mean - and for "creatures of habit" - individuals whose behaviour is consistent over time - while it pools (or shrinks towards the mean) everybody else.

We illustrate the theoretical motivation of IFS when outcomes are the sum of time-invariant unobserved heterogeneity and an idiosyncratic shock, which is similar to the setting considered in the teacher valued-added literature mentioned above. Since we use only one time period to choose the best forecast, the forecast selection cannot be consistent. Instead, we investigate the advantages of IFS from a minimax regret perspective. We analyze two cases: IFS for selecting between a time-series (TS) and a cross-sectional (CS) forecast and IFS for selecting between

TS and empirical Bayes. In both cases, we show that IFS can be minimax-regret optimal relative to using the same forecast method (that is, either TS or CS) for all individuals. For example, the relative accuracy of CS, TS and IFS depends on the state-space spanned by the ratio of variances of the individual heterogeneity and of the idiosyncratic shock. No forecast uniformly dominates the others when the state-space is sufficiently large to reflect the uncertainty about the relative magnitudes of these two variances. However, IFS can guard against making large errors over regions of the state-space where the accuracies of CS and TS are very different, while IFS does well even if it makes a mistake when selecting between almost equally accurate TS and CS. Perhaps surprisingly, we show that the presence of outliers in the distribution of unobserved heterogeneity means that there can be an advantage to forecast selection even when TS and CS are indistinguishable in terms of accuracy.

There is a relatively recent literature on microforecasting with panel data. See, e.g., Chamberlain and Hirano (1999) for an earlier reference and Baltagi (2008) for a brief survey. Gu and Koenker (2015) and Liu et al. (2020) show the optimality of empirical Bayes methods for microforecasting in context different from ours.[1] Our findings show that it is possible to further improve on empirical Bayes by selecting which individuals to shrink towards the mean, as long as empirical Bayes does not uniformly dominate the competing method over the state-space. Our work is broadly related to Giannone et al. (2021), who emphasize the role of predictive model uncertainty and show that a single sparse model is limited in economic applications.

This paper can be related to the literature on statistical decision theory for decision making. For example, Manski (2019) emphasizes evaluation of decision rules by its performance across the state-space and advocates the minimax regret criterion. There are only a couple of papers that apply the minimax regret criterion to panel data: handling missing data in sample design (Dominitz and Manski,

[1] Specifically, Gu and Koenker (2015) and Liu et al. (2020) consider a linear dynamic panel data model, whereas we consider a simple static model of the sum of time-invariant unobserved heterogeneity and an idiosyncratic shock with both terms having individual-specific variances. Strictly speaking, the two models are non-nested.

2021) and forecasting discrete outcomes under partial identification or other concerns (Christensen et al., 2020). Their focuses are distinct from ours.

IFS delivers a clustering of individuals according to which forecasting method is more accurate out-of-sample. In contrast, the literature on model-based clustering (e.g., Fröhwirth-Schnatter and Kaufmann, 2008) postulates the existence of a finite number of clusters for the parameters of a given model, and then assigns individuals to different clusters based on a measure of in-sample fit.

In the empirical application, we extend IFS to a richer class of models and estimators for predicting earnings in the Panel Study of Income Dynamics (PSID). Different models of earnings have been proposed in the literature, including models with persistent and transitory income shocks, with possibly time-varying volatility. Understanding which model performs best has potentially useful implications: In macroeconomics (Guvenen (2007), Guvenen (2009), Heathcote et al. (2010), Arellano et al. (2015)), the process for earnings is a key element of models with incomplete markets and, hence, the chosen specification affects the patterns of consumption and labour supply over the life cycle. The earnings process also plays a crucial role for the determination of wealth inequality and for the design of optimal taxation and optimal social insurance. In labour economics (MaCurdy (1982), Lillard and Weiss (1979), Meghir and Pistaferri (2004)), the distinction between persistent and transitory shocks and the trade-off between heterogeneity and persistence can help explain how individual earnings evolve over time and is relevant for studying income mobility and for testing models of human capital accumulation. The literature has used in-sample methods to evaluate some of the proposed models but no general agreement has yet been reached. Our empirical results could be viewed as providing a comparative evaluation of the alternative models considered in the literature, through the different lens of their forecasting performance.

The chapter is organized as follows. Section 3.2 describes the proposed approach. In Section 3.3 we derive the properties of IFS: we analytically show that IFS can be optimal according to a minimax regret criterion under general assumptions. Section 3.4 describes the data used in the empirical applications and reports

the empirical findings. Finally, Section 3.5 concludes.

3.2 Individual Forecast Selection (IFS)

Our goal is microforecasting, that is, for individual i, we aim to forecast the outcome $Y_{i,T+1}$ at time T using panel data $\mathscr{Y}_{N,T} := \{Y_{i,t} : i = 1,\ldots,N, t = 1,\ldots,T\}$.[2] We consider a short panel setup so that we have a large N and short T.

At time T, we have a class of K possible forecasting methods to choose from for each i:

$$\mathscr{F} = \{f_k(\mathscr{Y}_{N,T}), \quad k = 1,...,K\}.$$

A forecast method is generically defined as a function of the panel data available at time T. This allows for forecasts based on models with individual-specific parameters, which are estimated using the individual's time series as well as models with parameters that are common across individuals, which are estimated using pooling techniques. It also allows for the use of Bayesian methods. Forecasts that are based on the same model but rely on different estimators are also viewed as different forecasting methods.

For each individual i, IFS chooses the method that would have given the most accurate forecast at time $T-1$ for the outcome at time T, for example based on a quadratic loss:

$$\widehat{k}_i = \arg\min_{k=1,\ldots,K} (Y_{i,T} - f_k(\mathscr{Y}_{N,T-1}))^2.$$

The one-step-ahead forecast for individual i at time T is then based on the $\widehat{k}_i$-th method:

$$\widehat{Y}_{i,T}^{IFS} = f_{\widehat{k}_i}(\mathscr{Y}_{N,T}).$$

The procedure induces a clustering of individuals based on out-of-sample forecast

[2]For simplicity of notation we focus on a balanced panel, but the method extends to unbalanced panels in a straightforward manner.

accuracy, with the number of clusters determined endogenously. Note that we base selection on the out-of-sample accuracy in one time period. This can be viewed as a worst-case scenario that allows for panels with a very short time dimension (e.g., $T = 2,3$ observations). For longer time-series dimensions, one could instead base the selection on average accuracy computed over a larger out-of-sample window, for example comprising observations at times $T-\ell,...,T$, for some integer $\ell > 0$.

3.3 Minimax regret optimality of IFS

We consider two cases where the competing forecasts from which IFS chooses are in one case a time-series (TS) and a cross-sectional (CS) forecast and in the other case TS and an empirical Bayes (EB) forecast. For each case, we show that IFS can be minimax regret optimal among the three models (TS, CS, IFS in the one case or TS, EB, IFS in the other case) in the context of a simple data-generating process.

3.3.1 Setup

Throughout the paper, we let roman capital letters denote random variables and greek lowercase letters denote parameters or other non-random quantities, respectively. The usual indicator function is denoted by $\mathbb{I}\{A\}$ for an event A. That is, $\mathbb{I}\{A\} = 1$ if A is true and $\mathbb{I}\{A\} = 0$ otherwise.

Assume that for each individual i,

$$Y_{i,t} = A_i + U_{i,t}, i = 1,...,N; t = 1,...,T, \tag{3.1}$$

where $A_i \sim (0, \lambda_i^2)$ and $U_{i,t} \sim (0, \sigma_i^2)$. Suppose that $A_i, U_{i,1}, \ldots, U_{i,T}$ are mutually independent. Furthermore, $Y_{i,t}$ are independent across individuals. However, we do not assume that $Y_{i,t}$ are identically distributed over i. Instead, the variances λ_i^2 and σ_i^2 are heterogenous across individuals. Here, $A_i, U_{i,1}, \ldots, U_{i,T}$ are random variables, whereas λ_i^2 and σ_i^2 are individual-specific parameters. In other words, we take the frequentist approach.

3.3.2 IFS: time-series vs. cross-section

In this section we consider IFS based on the following forecasting methods:

- Time series forecast (TS) $\widehat{Y}_{i,T}^{TS} := Y_{i,T}$,
- Cross sectional forecast (CS) $\widehat{Y}_{i,T}^{CS} := \frac{1}{N}\sum_{j=1}^{N} Y_{j,T}$.

In words, TS predicts $Y_{i,T+1}$ using the most recent individual time-series observation, whereas CS uses the most recent cross-sectional average.

In this section we base IFS on the out-of-sample performance at time $T-1$, instead of time T as described in Section 2. This simplifies the analytical results as it introduces independence between the forecast and the forecast selection rule. Assuming $T \geq 3$, we thus define IFS for predicting $Y_{i,T+1}$ as

$$\begin{aligned}\widehat{Y}_{i,T}^{IFS} := \widehat{Y}_{i,T}^{TS}\mathbb{I}\left\{(Y_{i,T-1}-\widehat{Y}_{i,T-2}^{TS})^2 \leq (Y_{i,T-1}-\widehat{Y}_{i,T-2}^{CS})^2\right\} \\ + \widehat{Y}_{i,T}^{CS}\mathbb{I}\left\{(Y_{i,T-1}-\widehat{Y}_{i,T-2}^{TS})^2 > (Y_{i,T-1}-\widehat{Y}_{i,T-2}^{CS})^2\right\},\end{aligned} \tag{3.2}$$

In words, we make use of the observations at time $T-1$ and $T-2$ to select between TS and CS and employ the observations at time T to forecast $Y_{i,T+1}$, depending on the forecast selection outcome.

3.3.2.1 Minimax Regret

Consider the mean squared forecast error of forecast m under the data-generating process θ_i,

$$\text{MSFE}(m,\theta_i) = \mathbb{E}\left[\left(Y_{i,T+1}-\widehat{Y}_{i,T}^{m}\right)^2\right].$$

Here $m \in \mathscr{M}$, where $\mathscr{M}$ includes TS, CS, and IFS. We focus on $\theta_i = (\lambda_i, \sigma_i)$ regarding the unknown status θ_i of the data-generating process. That is, we are uncertain about individual heterogeneity in terms of the individual-specific variance λ_i^2 of A_i as well as the individual-specific variance σ_i^2 of $U_{i,t}$. Define regret as

$$R(m,\theta_i) := \text{MSFE}(m,\theta_i) - \min_{h\in\mathscr{M}} \text{MSFE}(h,\theta_i).$$

The minimax-regret (MMR) criterion selects the forecast m that minimizes the maximum regret $\max_{\theta_i\in\Theta} R(m,\theta_i)$, where Θ is the set of possible states.

Note that

$$\text{MSFE}(\text{TS}, \theta_i) = \mathbb{E}\left[(Y_{i,T+1} - Y_{iT})^2\right] = 2\sigma_i^2$$

and

$$\text{MSFE}(\text{CS}, \theta_i) = \mathbb{E}\left[\left(Y_{i,T+1} - \frac{1}{N}\sum_{j=1}^{N} Y_{j,T}\right)^2\right] =: \lambda_i^2 + \sigma_i^2 + R_N,$$

where R_N is the remainder term.

For the IFS rule, note that

$$\begin{aligned} Y_{i,T+1} - \widehat{Y}_{i,T}^{IFS} = \left(Y_{i,T+1} - \widehat{Y}_{i,T}^{TS}\right) \mathbb{I}\left\{(Y_{i,T-1} - \widehat{Y}_{i,T-2}^{TS})^2 \le (Y_{i,T-1} - \widehat{Y}_{i,T-2}^{CS})^2\right\} \\ + \left(Y_{i,T+1} - \widehat{Y}_{i,T}^{CS}\right) \mathbb{I}\left\{(Y_{i,T-1} - \widehat{Y}_{i,T-2}^{TS})^2 > (Y_{i,T-1} - \widehat{Y}_{i,T-2}^{CS})^2\right\}. \end{aligned}$$

Thus, the mean squared forecast error for IFS is given by

$$\text{MSFE}(\text{IFS}, \theta_i) = \mathbb{E}\left[\left(Y_{i,T+1} - \widehat{Y}_{i,T}^{IFS}\right)^2\right],$$

where

$$\begin{aligned} \left(Y_{i,T+1} - \widehat{Y}_{i,T}^{IFS}\right)^2 = \left(Y_{i,T+1} - \widehat{Y}_{i,T}^{TS}\right)^2 \mathbb{I}\left\{(Y_{i,T-1} - \widehat{Y}_{i,T-2}^{TS})^2 \le (Y_{i,T-1} - \widehat{Y}_{i,T-2}^{CS})^2\right\} \\ + \left(Y_{i,T+1} - \widehat{Y}_{i,T}^{CS}\right)^2 \mathbb{I}\left\{(Y_{i,T-1} - \widehat{Y}_{i,T-2}^{TS})^2 > (Y_{i,T-1} - \widehat{Y}_{i,T-2}^{CS})^2\right\}. \end{aligned}$$

It seems tedious to analyze this general case directly. We thus consider an important leading case where $T = 3$ and N is large. The TS forecast is thus $Y_{i,3}$. Since N is large, we assume that the CS forecast is 0 as a first-order approximation to simplify the analysis. For IFS, we employ the first two period observations to choose between TS and CS. Namely, we choose TS if $(Y_{i,2} - Y_{i,1})^2 \le Y_{i,2}^2$ and CS if $(Y_{i,2} - Y_{i,1})^2 > Y_{i,2}^2$. Thus, the IFS forecast is

$$\widehat{Y}_{i,3}^{IFS} := Y_{i,3} \mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 \le Y_{i,2}^2\right\}, \tag{3.3}$$

using $\widehat{Y}_{i,3}^{TS} = Y_{i,3}$ and $\widehat{Y}_{i,3}^{CS} = 0$.

Lemma 1 *Assume that $T = 3$, the TS forecast is $Y_{i,3}$, the CS forecast is 0, and the IFS forecast is given by* (3.3)*. Then, the mean squared forecast errors are given by*

$$\begin{aligned}
\mathrm{MSFE}(\mathrm{TS}, \theta_i) &= 2\sigma_i^2, \\
\mathrm{MSFE}(\mathrm{CS}, \theta_i) &= \lambda_i^2 + \sigma_i^2, \\
\mathrm{MSFE}(\mathrm{IFS}, \theta_i) &= \left(\lambda_i^2 + \sigma_i^2\right) + \sigma_i^2 \mathrm{Pr}\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] \\
&\quad - \mathbb{E}\left[A_i^2 \mathrm{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \middle| A_i\right]\right].
\end{aligned}$$

3.3.2.2 Analytical results

To derive analytical results for IFS, we impose the following condition.

Assumption 1 *For each $t = 1, 2$, the distribution of $U_{i,t}$ is absolutely continuous with respect to the Lebesgue measure. In addition, A_i, $U_{i,1}$ and $U_{i,2}$ are mutually independent and satisfy*

$$\mathrm{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \middle| A_i\right] \geq \mathrm{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \ a.s. \tag{3.4}$$

Condition (3.4) in Assumption 1 is the key condition in the paper. It can be rewritten as

$$\mathrm{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \middle| A_i\right] \geq \mathrm{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 | A_i = 0\right] \ a.s.$$

Thus, condition (3.4) seems plausible because it should be easier to satisfy

$$(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2,$$

when A_i deviates from zero. Figure 3.1 demonstrates that this condition is satisfied when $U_{i,1}$ and $U_{i,1}$ are generated independently from $N(0,1)$. In the figure, the blue curve corresponds to $\mathrm{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \middle| A_i = \alpha\right]$, which is the probability of selecting TS in IFS, and the red dotted horizontal line is the value

of $\Pr\left[(U_{i,2}-U_{i,1})^2 \leq (A_i+U_{i,2})^2 | A_i = 0\right]$, which is the probability of selecting TS when A_i equals zero.

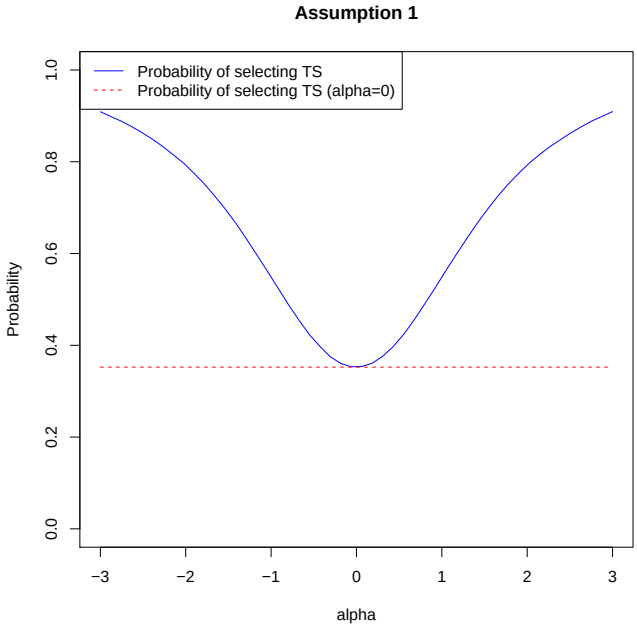


Figure 3.1: Graphical Demonstration of Assumption 1

Note that

$$\begin{aligned}
&\Pr\left[(U_{i,2}-U_{i,1})^2 \leq (A_i+U_{i,2})^2 \middle| A_i\right] \\
&= \mathbb{E}\left[\Pr\left[(U_{i,2}-U_{i,1})^2 \leq (A_i+U_{i,2})^2 \middle| A_i, U_{i,2}\right] \middle| A_i\right] \\
&= \mathbb{E}\left[F\left(|A_i+U_{i,2}| \middle| U_{i,2}\right) - F\left(-|A_i+U_{i,2}| \middle| U_{i,2}\right) \middle| A_i\right],
\end{aligned}$$

where $F(\cdot|U_{i,2})$ is the CDF of $U_{i,2}-U_{i,1}$ conditional on $U_{i,2}$. Thus, a sufficient condition for (3.7) can be obtained if we assume some shape restrictions on $F(\cdot|U_{i,2})$. Namely, for each $a \in \mathbb{R}$,

$$F\left(|a+U_{i,2}| \middle| U_{i,2}\right) - F\left(-|a+U_{i,2}| \middle| U_{i,2}\right) \geq F\left(|U_{i,2}| \middle| U_{i,2}\right) - F\left(-|U_{i,2}| \middle| U_{i,2}\right) \ a.s. \tag{3.5}$$

We now move to the next set of regularity conditions. Intuitively, we expect

that

$$\mathrm{MSFE}(\mathrm{TS},\theta_i) = 2\sigma_i^2 \geq \mathrm{MSFE}(\mathrm{IFS},\theta_i) \text{ if } \lambda_i^2 \text{ is sufficiently small,}$$
$$\mathrm{MSFE}(\mathrm{CS},\theta_i) = \lambda_i^2 + \sigma_i^2 \geq \mathrm{MSFE}(\mathrm{IFS},\theta_i) \text{ if } \lambda_i^2 \text{ is sufficiently large.}$$

We formalize this intuition in the following two assumptions.

Assumption 2 *The individual-specific variance λ_i^2 of A_i is small with respect to the individual-specific variance σ_i^2 of $U_{i,t}$ in the sense that*

$$\frac{\lambda_i^2}{\sigma_i^2} \leq \frac{1-\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1-\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}. \tag{3.6}$$

Assumption 3 *The individual-specific variance λ_i^2 of A_i is large with respect to the individual-specific variance σ_i^2 of $U_{i,t}$ in the sense that*

$$\frac{\lambda_i^2}{\sigma_i^2} \geq \frac{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}. \tag{3.7}$$

A sufficient condition for (3.7) is simply

$$\lambda_i^2 \geq \frac{\sigma_i^2}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}. \tag{3.8}$$

The right-hand side of the inequality above is solely a property of the distribution of idiosyncratic terms $U_{i,t}$'s, independent of the fixed effect A_i. Thus, (3.8) is satisfied if λ_i^2 is sufficiently large.

We now turn to (3.6). Let us further assume that there exists a constant $\eta < 1$ such that $\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] \leq \eta$. This is reasonable since it mainly excludes the case that one can detect the regime for TS perfectly. Then, a sufficient condition for (3.6) is simply

$$\lambda_i^2 \leq \frac{\sigma_i^2(1-\eta)}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}, \tag{3.9}$$

which holds if λ_i^2 is sufficiently small.

The following theorem establishes that IFS performs better than TS if λ_i^2/σ_i^2 is sufficiently small and it performs better than CS if λ_i^2/σ_i^2 is sufficiently large.

Theorem 1 *Let Assumption 1 hold.*

(i) If Assumption 2 holds, $\mathrm{MSFE}(\mathrm{IFS},\theta_i) \leq \mathrm{MSFE}(\mathrm{TS},\theta_i)$.

(ii) If Assumption 3 holds, $\mathrm{MSFE}(\mathrm{IFS},\theta_i) \leq \mathrm{MSFE}(\mathrm{CS},\theta_i)$.

We now consider the minimax-regret analysis. Suppose that $\mathscr{M}$ includes TS, CS and IFS. Then

$$\min_{h\in\mathscr{M}} \mathrm{MSFE}(h,\theta_i) \leq \min_{h\in\{\mathrm{TS},\mathrm{CS}\}} \mathrm{MSFE}(h,\theta_i) = \sigma_i^2 + \min\{\sigma_i^2,\lambda_i^2\}.$$

Furthermore, the regrets for TS and CS are

$$\begin{aligned} R(\mathrm{TS},\theta_i) &\geq \sigma_i^2 - \min\{\sigma_i^2,\lambda_i^2\}, \\ R(\mathrm{CS},\theta_i) &\geq \lambda_i^2 - \min\{\sigma_i^2,\lambda_i^2\}. \end{aligned}$$

Since the ratio between λ_i^2 and σ_i^2 matters, we restrict our attention to the following state space:

$$\Theta = \Theta(\mu) := \{(\sigma_i^2,\lambda_i^2) \in \mathbb{R}_+^2 : 1-\mu \leq \lambda_i^2/\sigma_i^2 \leq 1+\mu \text{ and } \sigma_i^2 = \sigma^2\} \quad (3.10)$$

for some constant $0 < \mu < 1$. In other words, the only relevant quantity is λ_i^2/σ_i^2 and so, without loss of generality, we set σ_i^2 to be common across i to simplify representation of the state space. Here, the ratio λ_i^2/σ_i^2 ranges from $1-\mu$ to $1+\mu$ to avoid the degenerate cases where TS uniformly dominates CS or the other way around.

Note that

$$\begin{aligned} \max_{\theta_i\in\Theta} R(\mathrm{TS},\theta_i) &\geq \max_{\theta_i\in\Theta} \left[(\sigma_i^2-\lambda_i^2)\mathbb{I}\{\sigma_i^2 > \lambda_i^2\}\right] = \sigma^2\mu, \\ \max_{\theta_i\in\Theta} R(\mathrm{CS},\theta_i) &\geq \max_{\theta_i\in\Theta} \left[(\lambda_i^2-\sigma_i^2)\mathbb{I}\{\sigma_i^2 < \lambda_i^2\}\right] = \sigma^2\mu. \end{aligned} \quad (3.11)$$

It follows from Theorem 1 that IFS will do better than CS when λ_i^2/σ_i^2 is sufficiently large and will perform better than TS when λ_i^2/σ_i^2 is sufficiently small. When λ_i^2/σ_i^2 is around 1, the choice between TS and CS is unimportant; as a result, it is expected that IFS will do well even if it makes a mistake in selecting the better model between almost equivalent TS and CS. To formalize this intuition, we impose the following regularity condition.

Assumption 4 *Each pair of* $(\sigma_i^2, \lambda_i^2) \in \Theta$ *satisfies that*

$$\max\left\{\Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right], 1-\Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right]\right\} \times \left(\frac{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right]}{\Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right]} - \frac{1-\Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right]}{1-\Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right]}\right) \le \mu.$$

Under Assumption 1, we have that

$$\Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right] \ge \Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right],$$

which implies that the left-hand side of the inequality in Assumption 4 is always nonnegative. The term $\Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right]$ is the probability of selecting TS over CS, while $\Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right]$ is the probability of selecting TS in the absence of the fixed effect A_i. Assumption 4 requires that $\Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right]$ cannot be too large compared to $\Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right]$.

Figure 3.2 shows under what conditions Assumption 4 is satisfied when (i) $U_{i,1}$ and $U_{i,2}$ are generated independently from $N(0,1)$ and (ii) A_i is randomly drawn from $N(0,\lambda^2)$. In the figure, the x-axis shows possible values of $\lambda_i^2/\sigma_i^2 = \lambda^2$. The blue curve corresponds to the left-hand side of the inequality in Assumption 4. Suppose that the state space $\Theta(\mu)$ is given by $\Theta(\mu) = \{1\} \times [1-\mu, 1+\mu]$ with $\mu = 0.999$. It can be seen that the maximum value of the blue curve on Θ is less than μ. Thus, Assumption 4 is satisfied. However, if we shrink Θ to be too small (e.g., $\Theta(\mu) = \{1\} \times [0.9, 1.1]$), it will not be satisfied. One way to interpret Assumption 4 is that it requires that there be a sufficient degree of uncertainty about the value of λ_i^2/σ_i^2.

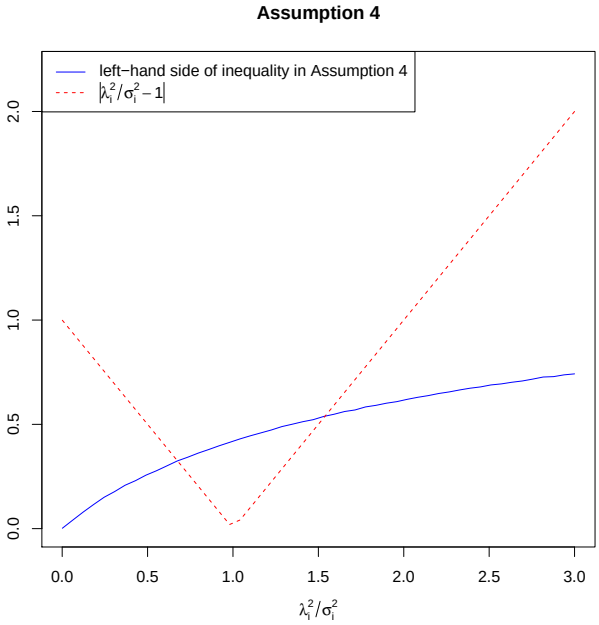


Figure 3.2: Graphical Demonstration of Assumption 4

Before presenting one of the main results in the paper, we strengthen (3.4) in Assumption 1.

Assumption 5 *There exist a set* $\mathscr{A}_i \subset \mathbb{R} \setminus \{0\}$ *and a constant* $0 < c_{\mathscr{A}_i} < \infty$ *such that* $\Pr(A_i \in \mathscr{A}_i) > 0$ *and for* $A_i \in \mathscr{A}_i$,

$$\Pr\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \middle| A_i\right] - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \geq c_{\mathscr{A}_i} \ \ a.s. \quad (3.12)$$

In words, (3.12) requires that there is a subset of the support of A_i such that the probability of selecting TS in IFS with nonzero A_i is, almost surely, higher (by constant $c_{\mathscr{A}_i}$) than the probability of selecting TS when A_i equals zero. We may term the requirement in (3.4) the *weak separability* condition and (3.4) and (3.12) jointly together the *strong separability* condition. Intuitively, Assumption 5 holds if there are individuals whose A_i's are sufficiently different from zero.

We will show that IFS minimizes maximum regret under Assumptions 1, 4 and 5. We first establish the following result.

Theorem 2 *Let* $\mathscr{M} = \{\mathrm{TS}, \mathrm{CS}, \mathrm{IFS}\}$. *Let Assumptions 1 and 4 hold. Then,*

$$R(\mathrm{IFS}, \theta_i) \leq \sigma^2 \mu$$

for each $\theta_i \in \Theta$, *which is defined in* (3.10). *Furthermore, the inequality above is strict if Assumption 5 holds additionally.*

Theorem 2 and the inequalities (3.11) together imply the following corollary.

Corollary 1 *Let Assumptions 1 and 4 hold. Then,*

$$\max_{\theta_i \in \Theta} R(\mathrm{IFS}, \theta_i) \leq \min \left\{ \max_{\theta_i \in \Theta} R(\mathrm{TS}, \theta_i), \max_{\theta_i \in \Theta} R(\mathrm{CS}, \theta_i) \right\},$$

where Θ *is defined in* (3.10). *Furthermore, the inequality above is strict if Assumption 5 holds additionally.*

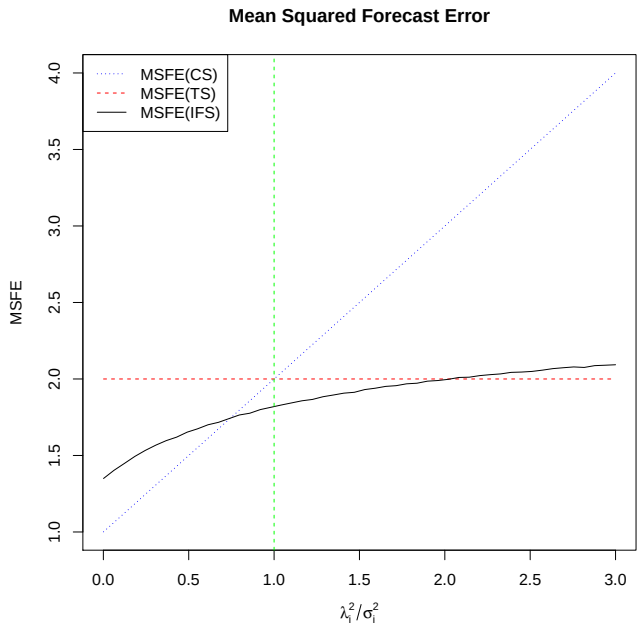


Figure 3.3: Mean Squared Forecast Errors

Corollary 1 implies that IFS minimizes maximum regret. Figures 3.3 and 3.4 show the mean squared forecast error (MSFE) and regret for each of the prediction

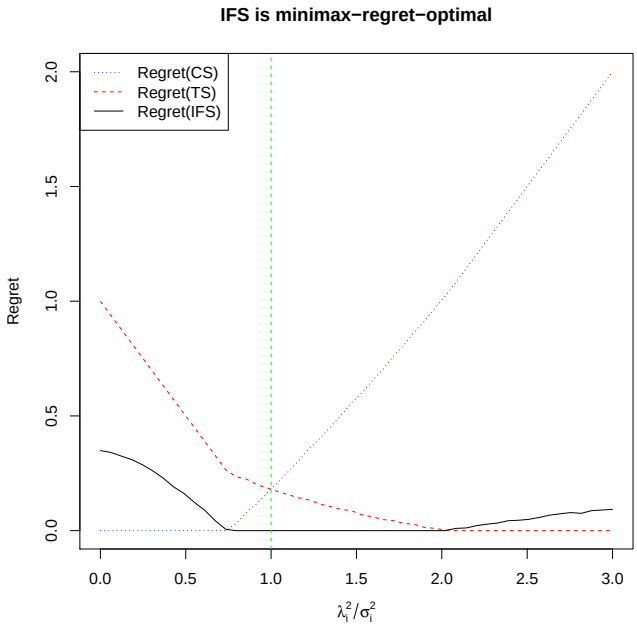


Figure 3.4: IFS is Minimax-Regret Optimal

models when (i) $U_{i,1}$ and $U_{i,2}$ are generated independently from $N(0,1)$ and (ii) A_i is randomly drawn from $N(0,\lambda^2)$, as in Figure 3.2. Note that the MSFE for CS is the smallest when λ^2 is sufficiently small; the MSFE for TS is the smallest when λ^2 is sufficiently large. No prediction model is uniformly superior in terms of MSFE; however, it can be seen in Figure 3.4 that IFS is minimax-regret optimal when the state space is e.g., $\Theta(\mu) = \{1\} \times [0.001, 1.999]$.

3.3.2.3 IFS with equally accurate forecasts

One might ask whether there could be any value of implementing IFS when both TS and CS perform equally well in terms of MSFE. We answer this question in this section by limiting our attention to discrete heterogeneity for A_i.

By Lemma 1, we have that if $\lambda_i^2 = \sigma_i^2 = 1$, the mean squared forecast errors are $\text{MSFE}(\text{TS}, \theta_i) = \text{MSFE}(\text{CS}, \theta_i) = 2$ and

$$\text{MSFE}(\text{IFS}, \theta_i) = 2 + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \mathbb{E}\left[A_i^2 \Pr\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \middle| A_i\right]\right].$$

To analyze MSFE(IFS, θ_i) further, we assume discrete heterogeneity for A_i:

$$A_i = \begin{cases} -(2\delta_i)^{-1/2} & \text{with prob. } \delta_i \\ 0 & \text{with prob. } 1-2\delta_i \\ (2\delta_i)^{-1/2} & \text{with prob. } \delta_i \end{cases} \tag{3.13}$$

and assume that $\delta_i \in (0, 0.5]$. The discrete distribution of A_i is symmetric around zero with three probability mass points and has a constant variance of one, regardless of the value of δ_i. When $\delta_i = 0.5$, A_i is a Rademacher random variable that takes values ± 1 with equal probability. As $\delta_i \to 0$, A_i is mostly zero but can have a very large positive or negative value with very small probability. We interpret individuals with small δ_i's as those whose realized values of individual fixed effects can be outliers.

Theorem 3 *Let Assumption 1 hold and assume that* $\lambda_i^2 = \sigma_i^2 = 1$ *and* A_i *follows the discrete distribution given in* (3.13)*. Then,*

$$\text{MSFE}(\text{IFS}, \theta_i) \leq \text{MSFE}(\text{TS}, \theta_i) = \text{MSFE}(\text{CS}, \theta_i) = 2$$

for any value in $\delta_i \in (0, 0.5]$ *in* (3.13)*. Furthermore, the inequality above holds strictly if* $0 < \delta_i < 0.5$ *and* $\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] < \Pr\left[(U_{i,2} - U_{i,1})^2 \leq \{(2\delta_i)^{-1/2} + U_{i,2}\}^2\right]$.

The theorem shows that (i) IFS is weakly more accurate than TS and CS, even when the two forecasts are equally accurate for any value of $\delta_i \in (0, 0.5]$; (ii) IFS is strictly more accurate than TS and CS if outliers in the distribution of unobserved heterogeneity delivers strict improvement in selecting TS over CS in IFS.

3.3.3 IFS: time series vs. empirical Bayes

We now consider the case where IFS selects between TS and an empirical Bayes forecast. In this section we assume $T = 4$ and that N is large.

For the TS forecast, we assume that the forecast is the mean of the previous

two time-periods: $\widehat{Y}_{i,4}^{TS} := (Y_{i,4} + Y_{i,3})/2$. Then, the MSFE for TS is

$$\text{MSFE}(\text{TS}, \theta_i) = \mathbb{E}\left[\left\{Y_{i,5} - (Y_{i,4} + Y_{i,3})/2\right\}^2\right] = 1.5\sigma_i^2.$$

For the forecast based on cross-sectional information, we go beyond the simple cross-sectional average and consider the following infeasible version of empirical Bayes (EB):

$$\widehat{Y}_{i,4}^{EB} := \frac{1}{N}\sum_{j=1}^{N} Y_{j,4} + \frac{\lambda_i^2}{\lambda_i^2 + \sigma_i^2}\left(Y_{i,4} - \frac{1}{N}\sum_{j=1}^{N} Y_{j,4}\right). \tag{3.14}$$

Since N is large, we again assume that the cross-sectional average is zero in order to simplify the analysis. The EB forecast is thus

$$\widehat{Y}_{i,4}^{EB} = \omega_i Y_{i,4},$$

with $\omega_i \equiv \lambda_i^2/(\lambda_i^2 + \sigma_i^2)$. The MSFE for EB is

$$\begin{aligned}\text{MSFE}(\text{EB}, \theta_i) &= \mathbb{E}\left[\left(Y_{i,5} - \omega_i Y_{i,4}\right)^2\right] \\ &= (1 + \omega_i)\sigma_i^2.\end{aligned}$$

There are three cases: (i) $\lambda_i^2 > \sigma_i^2$, (ii) $\lambda_i^2 < \sigma_i^2$, and (iii) $\lambda_i^2 = \sigma_i^2$. In case (iii), we have that $\omega_i = 0.5$ and $\text{MSFE}(\text{EB}, \theta_i) = \text{MSFE}(\text{TS}, \theta_i) = 1.5\sigma_i^2$. If $\lambda_i^2 > \sigma_i^2$, then $\omega_i > 0.5$ and

$$\text{MSFE}(\text{EB}, \theta_i) > 1.5\sigma_i^2.$$

Therefore, in case (i), TS dominates EB. On the other hand, in case (ii) we have $\omega_i < 0.5$ and

$$\text{MSFE}(\text{EB}, \theta_i) < 1.5\sigma_i^2,$$

which implies that EB dominates TS.

We now define a modified version of the IFS rule based on the out-of-sample performance at time $T-2$, which induces independence between the forecast and the selection rule in order to simplify the analytical results:

$$\widehat{Y}_{i,4}^{IFS} := 0.5(Y_{i,4}+Y_{i,3})\mathbb{I}\left\{(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right\} + \omega_i Y_{j,4}\mathbb{I}\left\{(Y_{i,2}-Y_{i,1})^2 > Y_{i,2}^2\right\}. \tag{3.15}$$

Note that the indicator functions (that is, the forecast selection rule) are the same as those in (3.3). This is because in the current setup, it is still crucial whether $\lambda_i^2 \geq \sigma_i^2$ or not in order to decide which one to choose between TS and EB.

Lemma 2 *Assume that $T=4$, the TS forecast is $(Y_{i,4}+Y_{i,3})/2$, the EB forecast is $\omega_i Y_{i,4}$ with $\omega_i = \lambda_i^2/(\lambda_i^2+\sigma_i^2)$, and the IFS forecast is given by* (3.15)*. Then, the mean squared forecast errors are given by*

$$\begin{aligned}
\text{MSFE}(\text{TS},\theta_i) &= 1.5\sigma_i^2,\\
\text{MSFE}(\text{EB},\theta_i) &= (1+\omega_i)\sigma_i^2,\\
\text{MSFE}(\text{IFS},\theta_i) &= (1+\omega_i)\sigma_i^2 + (0.5-\omega_i^2)\sigma_i^2\text{Pr}\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]\\
&\quad - \mathbb{E}\left[(1-\omega_i)^2 A_i^2 \text{Pr}\left[(U_{i,2}-U_{i,1})^2 \leq (A_i+U_{i,2})^2 \middle| A_i\right]\right].
\end{aligned}$$

We now show minimax regret optimality for IFS when selecting between TS and EB.

3.3.3.1 Analytical results

We make the following assumptions.

Assumption 6 *Suppose that $0 \leq \omega_i \leq 0.5$. Then, the individual-specific variance λ_i^2 of A_i is small with respect to the individual-specific variance σ_i^2 of $U_{i,t}$ in the sense that*

$$2\omega_i \leq \frac{1-\text{Pr}\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1-\text{Pr}\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}. \tag{3.16}$$

Define

$$\nu(\omega) := \frac{\omega(1-\omega)}{0.5-\omega^2}.$$

Assumption 7 *Suppose that* $0.5 < \omega_i < \sqrt{0.5}$. *Then, the individual-specific variance* λ_i^2 *of* A_i *is large with respect to the individual-specific variance* σ_i^2 *of* $U_{i,t}$ *in the sense that*

$$\nu(\omega_i) \geq \frac{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}. \tag{3.17}$$

The following theorem establishes that IFS performs better than TS if ω_i is sufficiently small and performs better than EB if ω_i is sufficiently large.

Theorem 4 *Let Assumption 1 hold.*

(i) If Assumption 6 holds, $\mathrm{MSFE}(\mathrm{IFS}, \theta_i) \leq \mathrm{MSFE}(\mathrm{TS}, \theta_i)$.

(ii) If Assumption 7 holds or $\omega_i \geq \sqrt{0.5}$, $\mathrm{MSFE}(\mathrm{IFS}, \theta_i) \leq \mathrm{MSFE}(\mathrm{EB}, \theta_i)$.

For the minimax regret analysis, let $\mathscr{M}$ include TS, EB and IFS. Then

$$\min_{h\in\mathscr{M}} \mathrm{MSFE}(h, \theta_i) \leq \min_{h\in\{\mathrm{TS},\mathrm{EB}\}} \mathrm{MSFE}(h, \theta_i) = \sigma_i^2 + \min\{0.5, \omega_i\}\sigma_i^2.$$

Furthermore, the regrets for TS and EB are

$$R(\mathrm{TS}, \theta_i) \geq 0.5\sigma_i^2 - \min\{0.5, \omega_i\}\sigma_i^2,$$
$$R(\mathrm{EB}, \theta_i) \geq \omega_i\sigma_i^2 - \min\{0.5, \omega_i\}\sigma_i^2.$$

Since it is crucial whether $\omega_i \geq 0.5$ or not, in this section, we consider the following state space:

$$\Omega = \Omega(\kappa) := \left\{(\sigma_i^2, \lambda_i^2) \in \mathbb{R}_+^2 : \frac{1-\kappa}{2} \leq \omega_i \leq \frac{1+\kappa}{2} \text{ and } \sigma_i^2 = \sigma^2\right\} \tag{3.18}$$

for some constant $0 < \kappa < 1$. In view of the state space given in (3.18),

$$\begin{aligned} \max_{\theta_i \in \Omega} R(\mathrm{TS}, \theta_i) &\geq \sigma^2 \frac{\kappa}{2}, \\ \max_{\theta_i \in \Omega} R(\mathrm{EB}, \theta_i) &\geq \sigma^2 \frac{\kappa}{2}. \end{aligned} \tag{3.19}$$

To establish that IFS minimizes maximum regret, we consider the partition of $\Omega = \Omega_a \cup \Omega_b \cup \Omega_c \cup \Omega_d \cup \Omega_e$:

$$\Omega_a := \left\{ (\sigma_i^2, \lambda_i^2) \in \Omega : \sqrt{0.5} < \omega_i \leq 1 \right\},$$

$$\Omega_b := \left\{ (\sigma_i^2, \lambda_i^2) \in \Omega : 0.5 < \omega_i < \sqrt{0.5} \quad \text{and} \quad v(\omega_i) \geq \frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \right\},$$

$$\Omega_c := \left\{ (\sigma_i^2, \lambda_i^2) \in \Omega : 0.5 < \omega_i < \sqrt{0.5} \quad \text{and} \quad v(\omega_i) < \frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \right\},$$

$$\Omega_d := \left\{ (\sigma_i^2, \lambda_i^2) \in \Omega : 0 \leq \omega_i < 0.5 \quad \text{and} \quad 2\omega_i \leq \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \right\},$$

$$\Omega_e := \left\{ (\sigma_i^2, \lambda_i^2) \in \Omega : 0 \leq \omega_i < 0.5 \quad \text{and} \quad 2\omega_i > \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \right\}.$$

In view of Theorem 4, IFS will perform well if $(\sigma_i^2, \lambda_i^2) \in \Omega_a \cup \Omega_b \cup \Omega_d$. For other cases, we now make an assumption comparable to Assumption 4.

Assumption 8 *(i) If $(\sigma_i^2, \lambda_i^2) \in \Omega_c$, we have that*

$$2\omega_i \left(1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) - \left(1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]\right) \leq \kappa.$$

(ii) If $(\sigma_i^2, \lambda_i^2) \in \Omega_e$, we have that

$$\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - 2\omega_i \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \leq \kappa.$$

Figure 3.5 shows under what conditions Assumption 8 is satisfied again when (i) $U_{i,1}$ and $U_{i,2}$ are generated independently from $N(0,1)$ and (ii) A_i is ran-

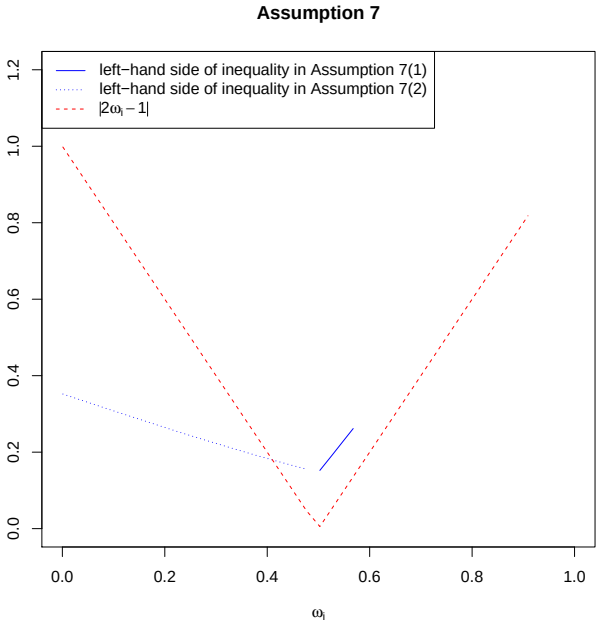


Figure 3.5: Graphical Demonstration of Assumption 8

domly drawn from $N(0,\lambda^2)$. In the figure, the x-axis shows possible values of $\omega = \lambda^2/(\lambda^2+1)$. The blue solid line segment corresponds to the left-hand side of the inequality in Assumption 8 (1) when $(\sigma_i^2,\lambda_i^2) \in \Omega_c$. On Ω_c, it is required that $0.5 < \omega_i < \sqrt{0.5}$ and

$$v(\omega_i) < \frac{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}.$$

It turns out that with normally generated $U_{i,1}$, $U_{i,2}$, and A_i do not satisfy the latter condition once ω_i is greater than 0.6. This is why the blue solid line does now show for values of ω_i larger than 0.6. The blue dotted line segment corresponds to the left-hand side of the inequality in Assumption 8 (2) when $(\sigma_i^2,\lambda_i^2) \in \Omega_e$. It can be seen from Figure 3.5 that the maximum of the left-hand side of the inequalities in Assumption 8 is less than 0.4, which implies that Assumption 8 is satisfied, provided that $\kappa \geq 0.4$ for $\Omega(\kappa) = \left\{(\sigma_i^2,\lambda_i^2) \in \mathbb{R}_+^2 : 1-\kappa \leq 2\omega_i \leq 1+\kappa \text{ and } \sigma_i^2 = \sigma^2\right\}$. Thus, as in Assumption 4, Assumption 8 requires that there be a sufficient degree of uncertainty about the value of λ_i^2/σ_i^2.

We first establish the following result.

Theorem 5 *Let* $\mathscr{M} = \{\text{TS}, \text{EB}, \text{IFS}\}$. *Let Assumptions 1 and 8 hold. Then,*

$$R(\text{IFS}, \theta_i) \leq \sigma^2 \frac{\kappa}{2}$$

for each $\theta_i \in \Omega(\kappa)$, *which is defined in* (3.18). *Furthermore, the inequality above is strict if Assumption 5 holds additionally.*

Theorem 5 and the inequalities in (3.19) together imply that IFS minimizes maximum regret under Assumptions 1 and 8.

Corollary 2 *Let Assumptions 1 and 8 hold. Then,*

$$\max_{\theta_i \in \Omega} R(\text{IFS}, \theta_i) \leq \min \left\{ \max_{\theta_i \in \Omega} R(\text{TS}, \theta_i), \max_{\theta_i \in \Omega} R(\text{EB}, \theta_i) \right\},$$

where Ω *is defined in* (3.18). *Furthermore, the inequality above is strict if Assumption 5 holds additionally.*

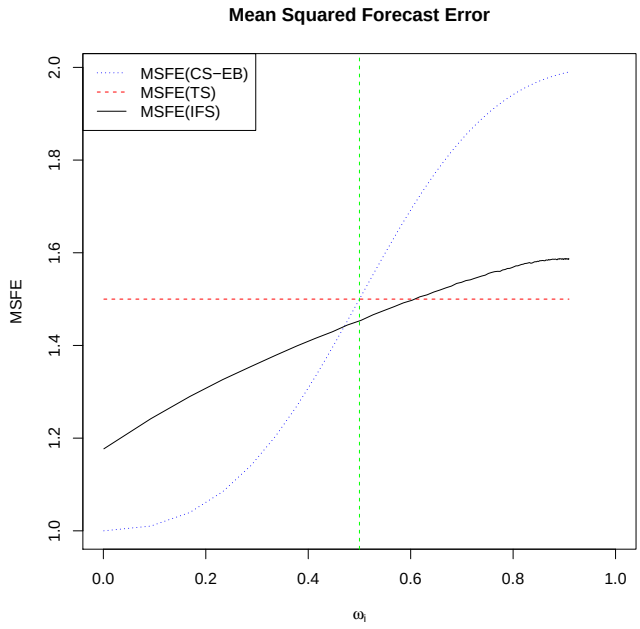


Figure 3.6: Mean Squared Forecast Errors

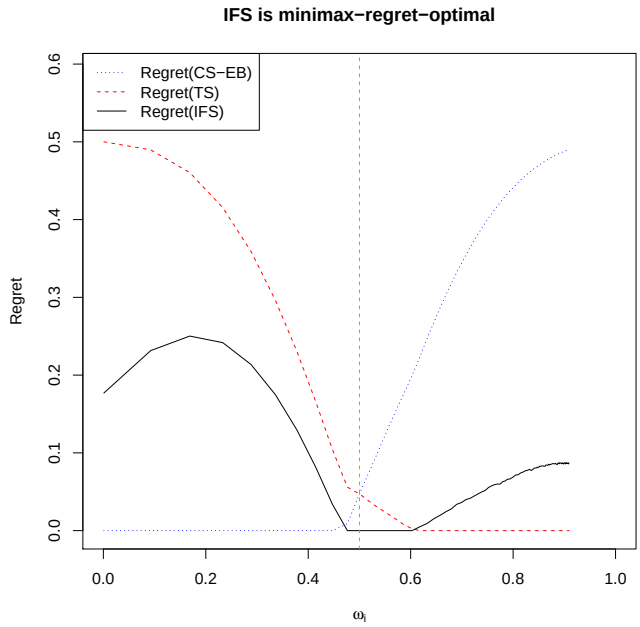


Figure 3.7: IFS is Minimax-Regret Optimal

Corollary 2 implies that IFS again minimizes maximum regret in the current setting. Figures 3.6 and 3.7 show the mean squared forecast error (MSFE) and the regret for each of the prediction models when (i) $U_{i,1}$ and $U_{i,2}$ are generated independently from $N(0,1)$ and (ii) A_i is randomly drawn from $N(0,\lambda^2)$, as in the previous figures. Note that the MSFE for EB is the smallest when $\omega = \lambda^2/(\lambda^2+1)$ is sufficiently small; the MSFE for TS is the smallest when ω is sufficiently large. As in the previous section, no forecast is uniformly superior in terms of MSFE; however, it can be seen in Figure 3.7 that the forecast based on IFS is minimax-regret optimal when the state space is $\Theta(\kappa)$ with a sufficiently large κ, for instance $\kappa = 0.4$.

3.3.4 Extending to a model with nonzero individual-specific means

Recall that our model (3.1) is written as

$$Y_{i,t} = A_i + U_{i,t},$$

where $A_i \sim (0, \lambda_i^2)$ and $U_{i,t} \sim (0, \sigma_i^2)$ and $A_i, U_{i,1}, \ldots, U_{i,T}$ are mutually independent.

Alternatively, we may assume that $A_i \sim (\alpha_i, \lambda_i^2 - \alpha_i^2)$ and $U_{i,t} \sim (0, \sigma_i^2)$, where α_i is individual-specific mean of A_i and $\lambda_i^2 > \alpha_i^2$. Suppose that $N^{-1}\sum_{j=1}^{N} \alpha_j = 0$. Then, all the results presented in the previous subsections remain intact because $\mathbb{E}(A_i^2) = \lambda_i^2$. Hence, we can focus on (3.1) without loss of generality. In other words, λ_i^2 should be interpreted as the *uncentered* second moment of A_i. λ_i^2 can be different across different individuals because they have different individual means or different individual variances (or both).

3.4 Empirical application

3.4.1 Data

We consider microforecasting of earnings using data from the Panel Study of Income Dynamics (PSID) for 1968-1993.[3]

We follow the literature on income dynamics (e.g., Meghir and Pistaferri (2004) and Hospido (2012)) and select a sample of male workers, heads of household, aged between 24 and 55 (inclusive). We drop individuals identifying as Latino, with a spell of self-employment, with zero or top-coded wages and with missing records on race and education. We also require that the change in log earnings is not greater than $+5$ or less than -3.

Following the literature, we work with earnings residuals obtained from a first stage regression of log labor income of an individual i at time t, $Y_{i,t}$, on a set of demographic variables: education, a quadratic polynomial in age, race and year dummies. We denote by $y_{i,t}$ the residuals from this regression. Forecasting earnings residuals is of interest in its own right since earnings residuals measure individual income risk. For instance, accurate forecasting of individual earnings residuals might be of key importance for prospective lenders when reviewing loan applications to decide among a pool of potential loan applicants. Note that the methodology we propose could be used to forecast earnings as well. We do not do it here to avoid

[3] We are using data only up to 1993 because, from 1994, a major revision of the survey disrupted the continuity of PSID files, see Hospido (2015) and Kim et al. (2000). Moreover, after 1997 the PSID switched from an annual to a biannual data collection.

specifying how to model time trend and macro-shocks.

In the following, the goal is to obtain individual one-year-ahead forecasts of the individual outcomes $y_{i,t}$.

3.4.2 Out-of-sample performance of IFS

In this subsection we compare the out-of-sample accuracy of IFS versus using the same forecast method for all individuals. We report results for the balanced samples of $N = 164$ individuals with continuous earnings in all consecutive years for 1968-1993. We first consider forecasts based on a simple static model as in section 3.3.2:

$$y_{i,t} = \alpha_i + \varepsilon_{i,t}. \tag{3.20}$$

For each $T = 1972, ..., 1992$, we produce individual one-step-ahead forecasts by the following methods: Time Series (TS), which forecasts the outcome at time $T+1$ by the average of individual outcomes from time 1 up to time T; Cross Section (CS), which uses the cross-sectional average at time T; a feasible version of the empirical Bayes (EB) forecast in equation (3.14):

$$\widehat{Y}_{i,t}^{EB} := \frac{1}{N}\sum_{j=1}^{N} Y_{j,t} + \frac{\hat{\lambda}_i^2}{\hat{\lambda}_i^2 + \hat{\sigma}_i^2}\left(Y_{i,t} - \frac{1}{N}\sum_{j=1}^{N} Y_{j,t}\right)., \tag{3.21}$$

where $\hat{\lambda}_i^2 + \hat{\sigma}_i^2 = \frac{1}{T-1}\sum_{t=1}^{T}(Y_{i,t} - (\frac{1}{T}\sum_{s=1}^{T} Y_{i,s}))^2$, $\hat{\sigma}_i^2 = \frac{1}{2T}\sum_{t=1}^{T-1}(Y_{i,t} - Y_{i,t+1})^2$, and $\hat{\lambda}_i^2$ is calculated as the difference between these two estimators. Finally, we consider Individual Forecast Selection (IFS) between TS or CS or between CS and EB depending on which of the two methods had the smaller squared error in forecasting the T-outcome at time $T-1$.

We then compare the individual out-of-sample forecasts from each method k, $\{\hat{y}_{i,T}^k\}$ to the actual realizations $\{y_{i,T+1}\}$, for $T = 1972, ..., 1992, i = 1, ..., N$.

We evaluate the out-of-sample accuracy by the average Mean Squared Forecast Error. For each forecasting method k and each individual i, the mean squared

forecast error over the out-of-sample period is

$$MSFE(k,i) = \frac{1}{23}\sum_{T=1970}^{1992}(y_{i,T+1} - \hat{y}_{i,T}^{k})^2.$$

Table 3.1 reports the average of $MSFE(k,i)$ over i for each forecasting method k.

Table 3.1: Out-of-sample Accuracy - Static Model

Method	TS	CS	EB	IFS	
				TS_CS	TS_EB
Avg. MSFE	0.102	0.210	0.175	0.094	0.098

Table 1 shows that, while TS outperforms CS and EB in terms of average MSFE, IFS further improves accuracy by deciding which individuals to pool or shrink towards the mean.

To gain some insight into which individuals are pooled by IFS, in Figure 3.8 we divide the individuals into ten quantiles according to their lagged earnings (the vertical axis) for each year (the horizontal axis). Within each quantile we compute the most frequently selected forecast by IFS: a triangle indicates that for the majority of the individuals in that year and in that quantile IFS selected TS, while a dot indicates that IFS selected CS. Figure 3.8 shows that it is only individuals in the center of the earnings distribution who benefit from pooling.

One possible interpretation of our findings is that in the PSID there is enough unobserved heterogeneity and a long enough time-series dimension to make the time series forecast perform better than pooling everybody or shrinking everybody towards the mean. However, an additional improvement in accuracy can be obtained by IFS, which tends to pool (or shrink towards the mean) individuals with earnings residuals near the center of the distribution.

Finally, we consider a dynamic panel data model for earnings residuals:

$$y_{i,t} = \alpha_i + \rho_i y_{i,t-1} + \varepsilon_{i,t}. \tag{3.22}$$

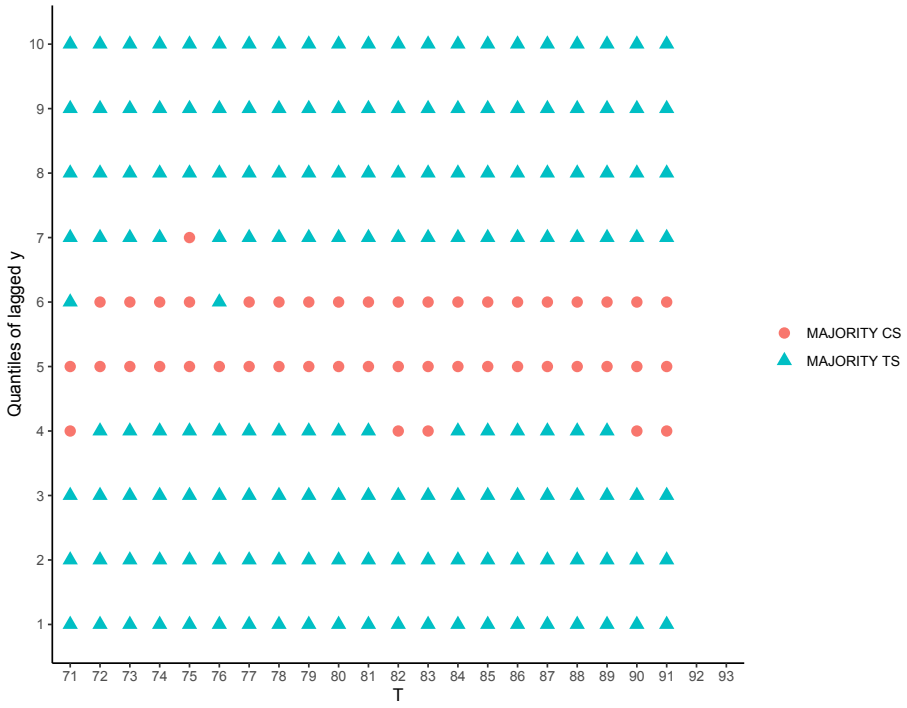


Figure 3.8: Most frequently selected forecast by year and earnings quantiles

We focus again on one-step-ahead forecasts of $y_{i,T+1}$ and consider four forecasting methods based on estimators of the parameters that make different use of the time series and cross-sectional dimensions of the panel:

1. Time Series (TS): $\widehat{\alpha}_i + \widehat{\rho}_i y_{iT}$. This method assumes individual-specific α_i and ρ_i and estimates them using the time series dimension.

2. Plug-In QMLE (PI): $\widehat{\alpha}_i(\widehat{\rho}_{QMLE}) + \widehat{\rho}_{QMLE} y_{iT}$. This method assumes individual-specific α_i but common ρ and is based on quasi-maximum likelihood estimation of common ρ, which integrates out α_i under some random effects distribution of α_i given initial conditions Y_{i0}. The α_i are then estimated for each unit i by maximum likelihood estimation conditional on $\widehat{\rho}$. [4]

3. Pooled OLS (Pooled): $\widehat{\alpha}_P + \widehat{\rho}_P y_{iT}$. This method assumes common α and ρ and is based on a joint maximum likelihood estimation of the parameters α

[4]Estimation of ρ via QMLE is the same for both PI and EB. The difference between the two is in estimation of the individual-specific α_i, which are shrunk towards some common values in EB.

and ρ:

$$(\widehat{\alpha}_P, \widehat{\rho}_P) = \text{argmin}_{\alpha,\rho} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \rho Y_{it-1} - \alpha)^2$$

4. Empirical Bayes (EB). This method assumes common ρ and shrinks the individual-specific α_i towards some common values. See Appendix B.2 for details on estimation.

We then consider IFS based on all four methods as well as IFS based on pairwise combinations of each "time series" method (TS or PI) with each "cross-sectional" method (Pooled or EB).

We again evaluate performance based on average MSFE across individuals. The results are reported in Table 3.2.

Table 3.2: Out-of-sample Accuracy - Dynamic Model

Method	TS	PI	Pooled	EB	IFS				
					All	TS_Pooled	TS_EB	PI_Pooled	PI_EB
Avg. MSFE	0.081	0.074	0.080	0.074	0.077	0.078	0.074	0.076	0.073

Table 3.2 shows that PI and EB are equally accurate and outperform the remaining forecasts, however performing IFS between these two methods further improves the performance. This finding confirms the usefulness of IFS even when choosing between methods that have very good and equal forecasting accuracy. Performing IFS among all four methods however results in a slight deterioration of the performance, suggesting caution about including poor-performing methods in the class from which IFS chooses.

3.5 Conclusions

There may be no "one-size-fits all" model for forecasting with micro panel data. Individual forecast selection improves forecast accuracy of individual forecasts and is minimax-regret optimal under general assumptions. Theoretical and empirical results using PSID data show that the proposed approach optimally trade-offs time

series and cross-sectional information. They further show that it can deliver accuracy gains over state-of-the-art approaches such as Empirical Bayes methods. Several extensions are of interest: 1) unbalanced panel data; 2) other models typically considered in the earnings literature, e.g. persistent-transitory decomposition with or without heterogeneous slope of earnings profiles (HIP vs RIP); 3) models with heterogeneous persistence; 4) IFS based on in-sample rather than pseudo-out-of sample accuracy for selection of the best forecasting methods; 5) comparison with forecasts based on spike and slab prior.

Chapter 4

Regularized CUE: a Quasi-Likelihood Approach

4.1 Introduction

Two-step generalized method of moments (GMM) is widely used in economics. However, this estimator can suffer from severe biases in finite samples, see e.g. Hansen et al. (1996), Hausman et al. (2011), Newey and Smith (2004). Hansen et al. (1996) proposed the Continuous Updating Estimator (CUE) as a solution to this bias problem and demonstrated through Monte Carlo simulations that CUE indeed significantly reduces the bias problem; Newey and Smith (2004) provided an analytical argument to support this evidence.

Unfortunately, the bias reduction comes at a price: The CUE in some applications exhibits large finite sample variances compared to the 2-step GMM. One possible explanation for this feature is that the CUE suffers from a no-moment problem. There is not a formal proof of this hypothesis in general but Monte Carlo studies have demonstrated that CUE is likely not to have any moments in finite samples. Among others, Guggenberger et al. (2005), Guggenberger (2008), Hausman et al. (2011) document a better performance of CUE in terms of median bias but fatter tails with respect to 2-step GMM, suggesting that CUE might have no moments. Moreover, in the case of linear IV models with homoskedastic errors, CUE is the Limited Information Maximum Likelihood (LIML) estimator, which is known to

have no moments; see Mariano and Sawa (1972), Fuller (1977), and Kinal (1980)).

The no-moments problem does not appear to be particular to the CUE. It belongs to the class of Generalized Empirical Likelihood (GEL) estimators, as shown by Newey and Smith (2004), which also include the Empirical Likelihood (EL) of Owen (1988) and the Exponential Tilting (ET) estimator of Kitamura and Stutzer (1997). Newey and Smith (2004) prove that all GEL estimators eliminate important sources of bias for GMM but Guggenberger (2008), among others, provide Monte Carlo evidence that all GEL estimators are likely not to have moments.

In this chapter, we propose a regularized version of the CUE which we refer to as the quasi-likelihood GMM (QL-GMM) estimator. The estimator is obtained by adding the log-determinant of the optimal weighting matrix to the usual GMM objective function that defines the CUE. The motivation for this term is asymptotic: Assuming that the sample moments satisfy a CLT, the QL-GMM objective function is simply the large-sample log-likelihood of the sample moments. The additional variance term works as a finite-sample penalization that implicitly imposes further restrictions on the resulting estimator.

We show, through simulations, that the regularization reduces finite-sample variances while only adding moderate finite-sample biases. At the same time, since the penalization term vanishes with $1/n$-rate, the QL-GMM is first-order asymptotically equivalent to the corresponding CUE and efficient 2-step GMM. In the special case of linear IV, we analytically demonstrate that the QL-GMM indeed has finite moments.

We conduct extensive Monte Carlo simulations to provide evidence that the new estimator is an attractive alternative to 2-step GMM and CUE. We find that in general QL-GMM has tighter tails than CUE, restoring its finite sample moments, and that this comes with a small price in terms of slightly bigger biases compared to the CUE in some settings. In addition to this, QL-GMM is computationally easier to implement since the penalization term implicitly reduces the parameter space to be searched over. In particular, in contrast to the CUE, we find that the QL-GMM objective function is more regular with a well-defined unique optimizer.

Thus, standard numerical solvers can be used to compute the estimator while the CUE estimator generally requires fine tuning and choosing multiple initial values in order to compute the estimator.

Our proposal is related to Holcblat (2015) and Holcblat and Sowell (2019), who propose the Empirical Saddle Point (ESP) approximation as an alternative to GMM estimators. The ESP estimator corresponds to an MM estimator (or, equivalently, any Generalize Empirical Likelihood (GEL) estimator) shrunk toward parameter values with lower estimated variance. This estimator is however computationally more demanding to implement.

QL-GMM is also related to the Regularized CUE (RCUE) proposed by Hausman et al. (2011), which is meant as a Fuller analogue of the CUE estimator. The authors take as starting point the FOCs of the CUE and then add to these two penalization terms which are meant to regularize the estimator. Hausman et al. (2011) show that the RCUE reduces the dispersion of the CUE in their Monte Carlo simulations and analytically prove that the proposed estimator have finite sample moments in a linear IV setting. However, implementations of RCUE's require the econometrician to specify the penalization terms that enter the RCUE. The performance of RCUE is very sensitive to the chosen penalizations, but Hausman et al. (2011) provide very little guidance for how to choose these. Thus, it is unclear how to achieve good performance of their estimator in practice. In contrast, our penalization term comes out as a natural implication of a given model and sample.

Finally, there is a literature on ridge- and lasso-type modifications of CUE but the focus is on selection of relevant moments and potential weak identification; see, e.g., Caner (2009), Carrasco and Tchuente (2016), Farbmacher (2016). Thus, the motivation and goal of the proposed penalized GMM-type estimators are different from ours. We conjecture that one could potentially combine our proposal with these to obtain a double-penalized version, but we do not pursue this idea here.

The remaining part of the chapter is organized as follows: Section 4.2 sets the stage and develops the proposed estimator, QL-GMM. In Section 4.3, we analyze the asymptotic and finite sample properties of QL-GMM. This section also

discusses some optimization issues and comments about the implementation of our estimator. In Section 4.4, we investigate the finite sample properties of QL-GMM in many settings via Monte Carlo simulations. Finally, Section 4.5 concludes.

4.2 A Modified CUE Estimator

Let z_i, $i = 1,...,n$, be i.i.d. observations from a model specified in terms of a set of $m \geq 1$ moment conditions, $g(z,\theta) \in \mathbb{R}^m$, that identifies the true parameter value, $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ by $\mathbb{E}[g(z,\theta_0)] = 0$, where $\mathbb{E}[\cdot]$ denotes expectation taken with respect to the distribution of z_i. Let $\hat{g}(\theta) = \sum_{i=1}^n g(z_i,\theta)/n$ be the sample moments and $\hat{\Omega}(\theta) \equiv n^{-1}\sum_{i=1}^n g(z_i,\theta)g(z_i,\theta)'$ be the sample covariance of the moments. Given a first-step estimator $\tilde{\theta}$, the 2-step GMM estimator is defined as

$$\hat{\theta}_{\text{GMM}} = \arg\min_{\theta\in\Theta} \hat{g}(\theta)'\hat{\Omega}^{-1}(\tilde{\theta})\hat{g}(\theta),$$

while the CUE proposed by Hansen et al. (1996) solves

$$\hat{\theta}_{\text{CUE}} = \arg\min_{\theta\in\Theta} \hat{g}(\theta)'\hat{\Omega}^{-1}(\theta)\hat{g}(\theta), \tag{4.1}$$

where we assume that the inverse $\hat{\Omega}^{-1}(\tilde{\theta})$ exists (is positive definite). Under standard regularity conditions, the two estimators are first-order asymptotically equivalent but their finite sample performances can be quite different. In particular, $\hat{\theta}_{\text{CUE}}$ tends to have smaller biases but larger variances compared to $\hat{\theta}_{\text{GMM}}$.

To remove the excess variability of $\hat{\theta}_{\text{CUE}}$, we propose to regularize the objective function defining it by adding a term that penalizes "large values" of $\hat{\Omega}(\theta)$. This is motivated by the following asymptotic argument: by the Central Limit Theorem, the set of sample moments $\hat{g}(\theta)$ satisfies $\sqrt{n}\hat{g}(\theta) \to^d N(0,\Omega(\theta))$ assuming that θ is the true data-generating value. Thus, its unknown likelihood function is well-approximated by

$$f_n(\hat{g}(\theta)|\theta) = \frac{n}{(2\pi)^{k/2}\left|\hat{\Omega}(\theta)\right|^{\frac{1}{2}}} \exp\left\{ -\frac{n}{2}\hat{g}(\theta)'\hat{\Omega}^{-1}(\theta)\hat{g}(\theta) \right\}$$

in large samples, where $\left|\hat{\Omega}(\theta)\right| > 0$ denotes the determinant of $\hat{\Omega}(\theta)$. Our QL-GMM estimator is then defined as the corresponding quasi-maximum-likelihood estimator, $\hat{\theta}_{\text{QL}} = \arg\max_{\theta\in\Theta} \log f_n(\hat{g}(\theta)|\theta)$. Observe here that

$$\log f_n(\hat{g}(\theta)|\theta) \propto -\frac{1}{2}\log\left|\hat{\Omega}(\theta)\right| - \frac{n}{2}\hat{g}(\theta)'\hat{\Omega}^{-1}(\theta)\hat{g}(\theta),$$

and so we can rewrite the estimator as

$$\hat{\theta}_{\text{QL}} = \arg\min_{\theta\in\Theta} \hat{g}(\theta)'\hat{\Omega}^{-1}(\theta)\hat{g}(\theta) + \frac{1}{n}\log\left|\hat{\Omega}(\theta)\right|. \tag{4.2}$$

Compared to the CUE objective function, the one of the QL-GMM estimator comes with an added penalization term in the form of $\frac{1}{n}\log\left|\hat{\Omega}(\theta)\right|$. This term penalizes parameter values that generate "large values" of $\left|\hat{\Omega}(\theta)\right|$. As we shall see, this removes certain undesirable features of the CUE estimator in finite samples. At the same time, the penalization term vanishes with rate $1/n$ and so $\hat{\theta}_{\text{QL}}$ is asymptotically first-order equivalent to $\hat{\theta}_{\text{CUE}}$ and $\hat{\theta}_{\text{GMM}}$.

QL-GMM is related to the Regularized CUE (RCUE) proposed by Hausman et al. (2011). The Regularized CUE (RCUE) is meant as the CUE analogue of the Fuller adjustment of linear IV estimators. The class of RCUE's solves a modified version of the FOCs of the CUE where two penalization terms are added. The two penalization terms are not specified and have to be chosen by the econometrician. This has the advantage of giving more flexibility in terms of implementation but the downside is that RCUE requires a careful selection of the tuning parameters and penalty term to achieve good performance. Very little guidance in this regard is offered by the existing literature. Our estimator has a similar structure to that of RCUE and can also be seen as the CUE analogue of the Fuller adjustment of linear IV estimators, with a particular choice of the penalization terms. The particular choice of $\frac{1}{n}\log\left|\hat{\Omega}(\theta)\right|$ is well-motivated and requires no input from the econometrician; it is fully data-driven.

4.3 Properties of QL-GMM

In this section, we analyze the asymptotic and finite sample properties of the proposed estimator. We provide some intuitions on the behavior of QL-GMM in finite sample and prove existence of moments of its finite sample distribution.

4.3.1 Asymptotic Properties

QL-GMM is first-order asymptotically equivalent to CUE: since the additional term in the objective function of QL-GMM is a finite sample correction, to first order there is no large sample efficiency loss:

Theorem 6 *Let z_i, $i = 1,...,n$, be i.i.d. observations and θ be a p-dimensional parameter vector. Assume that θ_0 is the unique solution to $\mathbb{E}[g(z,\theta_0)] = 0$, and is situated in the interior of the compact parameter space $\Theta \subset \mathbb{R}^p$. Assume that the $m \times 1$ vector of moment functions $g(z,\theta)$, with $m \geq p$, is continuously differentiable in a neighborhood of θ_0 with $\mathbb{E}[\sup_{\theta\in\Theta} \|g(z,\theta)\|^\alpha] < \infty$, for some $\alpha > 2$, $\mathbb{E}\Big[\sup_{\theta\in\Theta}\Big\|\frac{\partial g_i(\theta)}{\partial\theta'}\Big\|\Big] < \infty$ and $\mathbb{E}\Big[\frac{\partial g_i(\theta_0)}{\partial\theta'}\Big] \in \mathbb{R}^{m\times p}$ having rank p. Finally, assume that $\Omega(\theta) \equiv \mathbb{E}[g(z,\theta)g(z,\theta)']$ is nonsingular for all θ. Then*

$$\sqrt{n}(\hat{\theta}_{QL} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}\Big[\frac{\partial g(z,\theta_0)}{\partial\theta'}\Big]^{-1}\Omega(\theta_0)\mathbb{E}\Big[\frac{\partial g(z,\theta_0)'}{\partial\theta}\Big]^{-1})$$

Proof. The stated assumptions are standard regularity conditions on the moment functions and their derivatives, which are required for the validity of asymptotic approximation for the GMM-type estimators. As $n \to \infty$ and under the assumption that, for all $\theta \in \Theta$, $\Omega(\theta) > 0$ the additional term in the objective function of QL-GMM vanishes sufficiently fast asymptotically: $\frac{1}{n}\log|\hat{\Omega}(\theta)| \xrightarrow{p} 0$. As a result, $\hat{\theta}_{\text{QL}} \xrightarrow{p} \arg\min_{\theta\in\Theta}\mathbb{E}[g(z,\theta)]'\Omega(\theta)^{-1}\mathbb{E}[g(z,\theta)]$ and $\sqrt{n}(\hat{\theta} - \theta_0) \to N(0,\Sigma(\theta_0))$. ■

We have shown that the proposed estimator and CUE share the same behavior asymptotically. Differences only arise in finite samples.

4.3.2 Tail Behavior of QL-GMM and Properties of its Objective Function

A first inspection of the objective functions that CUE and QL-GMM minimize can shed light on how QL-GMM achieves a smaller dispersion in the estimates.

One main caveat to using CUE is that a large variance covariance matrix, $\hat{\Omega}(\theta)$, makes the criterion function, equation (4.1), small for any value of the sample moment condition $\hat{g}(\theta)$. In Hansen et al. (1996) the authors argue that very large estimates for the CUE estimator can be justified by the shape of its objective function. Assume that the moment conditions are linear in the parameters to be estimated. Then, for the 2-step GMM estimator the criterion function is quadratic in the parameter, while the criterion function for CUE is not. In particular, the objective function of CUE converges for large values of the parameter estimates, leading to extreme outliers for the minimizing value of the parameter.

The objective function of QL-GMM is instead quasi-convex. The regularization term of QL-GMM, $\frac{1}{n}\log|\hat{\Omega}(\theta)|$, increases as the value of the parameter estimate increases when the objective function of CUE converges. This should make the criterion function for QL-GMM, equation (4.2), large enough to potentially eliminate the cases of extreme values in the estimates.

To see this in the context of the IV setting, consider the following setup:

$$\mathbb{E}[\rho(Y,X;\theta)|Z] = 0 \iff \theta = \theta_0$$

for some generalized residual $\rho(\cdot) \in \mathbb{R}$. Choose as moment conditions

$$g_i(\theta) = \rho(Y_i,X_i;\theta)f(Z_i)$$

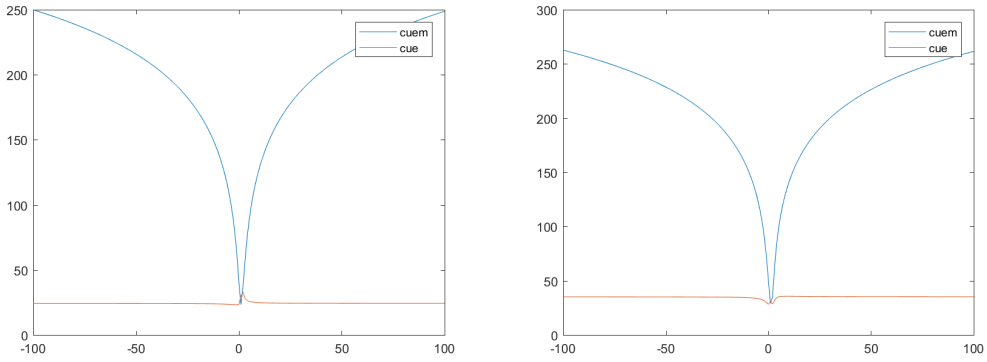
for some function $f(\cdot)$. Let $\hat{\sigma}^2_\varepsilon(\theta) \equiv \hat{\mathbb{E}}[\rho(Y,X;\theta)^2]$, where $\hat{\mathbb{E}}[\rho(Y,X;\theta)^2]$ is the sample analogue estimator of $\mathbb{E}[\rho(Y,X;\theta)^2]$. With a homoskedasticity consistent variance estimator $\hat{\Omega}(\theta) = \hat{\sigma}^2_\varepsilon(\theta)\frac{1}{n}\sum f(Z_i)f(Z_i)'$,

$$\hat{\theta}_{\mathrm{QL}} = \arg\min_{\theta\in\Theta} \frac{1}{n}\log|\hat{\sigma}^2_\varepsilon(\theta)| + \hat{g}(\theta)'\hat{\Omega}^{-1}(\theta)\hat{g}(\theta).$$

Assuming now that, for all values of Y_i, X_i, $\rho(Y_i, X_i; \theta)$ diverges to infinity as $||\theta||$ diverges. It should follow that the sample variance estimator $\hat{\Omega}(\theta)$ will also diverge since $\hat{\sigma}^2_\varepsilon(\theta)$ will diverge. Under this assumption, $\hat{g}(\theta)'(\hat{\Omega}(\theta))^{-1}\hat{g}(\theta)$ converges to some constant for this $\theta \neq \theta_0$. But since $\hat{\sigma}^2_\varepsilon(\theta)$ will diverge, the term $\frac{1}{n}\log|\hat{\sigma}^2_\varepsilon(\theta)|$ will diverge making sure that the diverging $||\theta||$ is not selected as minimizer.

This argument can be extended to any parameter-dependent weighting matrix in equation 4.2: Suppose that there exists a θ_1 (far away from θ_0, the true value) such that $\hat{W}(\theta_1)$ is a singular matrix. We may potentially have that $\hat{g}(\theta_1)'\hat{W}(\theta_1)\hat{g}(\theta_1) = 0$ even if $\hat{g}(\theta_1) \neq 0$. But at the same time, for finite n, $-\frac{1}{n}\log(|\hat{W}(\theta_1)|)$ will converge to $+\infty$ and thereby θ_1 cannot be the minimiser.

The figure below further illustrates this point. It shows the criterion functions for CUE and for QL-GMM in two of the Monte Carlo draws where CUE took extreme values, in the IV setting of Hausman et al. (2011), 4.1a and 4.1b. A description of the Monte Carlo settings is in section 4.4.



(a) **(b)**

Figure 4.1: Criterion Functions of CUE and QL-GMM

These graphs show the criterion functions of CUE and QL-GMM for 2 Monte Carlo draws, in the IV setting of Hausman et al. (2011).

Thus, our estimator shrinks CUE estimates toward parameter values with lower estimated variance.

In particular, the CUE estimator suffers from convergence problems and multimodality. As reported in Imbens et al. (1998), inspection reveals that typically the objective function for the CUE estimator has multiple modes with occasionally the

mode far away from the population value of theta higher than the mode close to the population value. This multimodality justifies the bad performance of CUE in Imbens et al. (1998). Hansen et al. (1996) suggest using the Matlab optimization routine fminu.m that implements a quasi Newton algorithm, which is dependent on an initial setting for the parameters. When this gradient method failed to converge or resulted in unusual estimates, the authors also used the routine fmins.m, which is a simplex search method. As a further check on the obtained numerical results, when extreme parameter estimates were obtained, the authors also examined the continuous updating criterion over a grid of the parameters to obtain additional assurance that the estimated parameters were indeed minimizers of the criterion function. Even with a restricted parameter space, Guggenberger et al. (2005) found that the finite sample criterion function of CUE frequently has global minimum on the boundaries of the parameter space. Therefore, CUE is computationally problematic. Moreover, the CUE optimization procedures are sensitive to the choice of initial conditions. Finally, Peñaranda and Sentana (2015) and Peñaranda and Sentana (2012) propose some alternative intuitive methods that simplify the computation of continuously updated GMM estimators: the optimal CUE can be computed as a minmax criterion based on a certain R^2, optimally computed by means of a sequence of OLS regressions. Also, a careful choice of simple, intuitive consistent parameter estimators that can be used to obtain good initial values can help in estimation of CUE.

Unlike CUE, QL-GMM is not sensitive to the choice of starting values.

Finally, to avoid numerical instabilities in the implementation of our estimator, which may be due to difficulties to calculate the determinant, we consider an equivalent expression for our estimator that replaces $\log(\det(\Omega(\beta)))$ with the principal matrix logarithm $\mathrm{tr}(\mathrm{logm}(\Omega(\beta)))$. Given a matrix A, the principal matrix logarithm of A, denoted as $\mathrm{logm}(A)$, is defined as the unique logarithm for which every eigenvalue has imaginary part lying strictly between $-\pi$ and π.[1]

[1]The principal matrix logarithm of A is defined as the inverse of the matrix exponential of A, $\mathrm{expm}(A)) \equiv V * \mathrm{diag}(\exp(\mathrm{diag}(D)))/V$, if A has a full set of eigenvectors V with corresponding eigenvalues D.

4.3.3 Finite Samples Properties and Existence of Moments

4.3.3.1 Proof of existence of moments in the linear IV setting

Consider a linear model, $y = X\theta + u$, the moment function $g_i(\theta) = z_i'(y_i - x_i\theta)$, and $\hat{\Omega}(\theta) \equiv n^{-1}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)'$, where y is a $n \times 1$ vector, X is a $n \times p$ matrix, θ is a p-dimensional unknown parameter vector, and Z is a $n \times m$ matrix. n is sample size, and m is number of instruments (number of moment conditions).

Consider a homoskedasticity consistent variance estimate:

$$\hat{\Omega}(\hat{\theta}) = (y - X\hat{\theta})'(y - X\hat{\theta})Z'Zn^{-2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\hat{\theta})'(y_i - X_i\hat{\theta})\frac{1}{n}\sum_{i=1}^{n} z_i'z_i$$

and define:

$$\bar{G}(\hat{\theta}_{CUE}) := \left[-X'Z + \frac{X'(y - X\hat{\theta}_{CUE})}{(y - X\hat{\theta}_{CUE})'(y - X\hat{\theta}_{CUE})}\hat{g}(\hat{\theta}_{CUE})' \right] / n^2.$$

We can write the expression for CUE as follows:

$$\hat{\theta}_{CUE} = (\bar{G}(\hat{\theta}_{CUE})'[\hat{\Omega}(\hat{\theta}_{CUE})]^{-1}\bar{G}(\hat{\theta}_{CUE}))^{-1}(\bar{G}(\hat{\theta}_{CUE})'[\hat{\Omega}(\hat{\theta}_{CUE})]^{-1}\hat{g}(0)).$$

With analogous steps, the proposed estimator can be rewritten as:

$$\hat{\theta}_{\mathrm{QL}} = -\left(\bar{G}(\hat{\theta}_{\mathrm{QL}})'[\hat{\Omega}(\hat{\theta}_{\mathrm{QL}})]^{-1}\bar{G}(\hat{\theta}_{\mathrm{QL}}) + \frac{m}{n}\frac{X'X}{\hat{\sigma}_{\varepsilon}^2(\hat{\theta}_{\mathrm{QL}})} \right)^{-1} \left(\bar{G}(\hat{\theta}_{\mathrm{QL}})'[\hat{\Omega}(\hat{\theta}_{\mathrm{QL}})]^{-1}\tilde{g}(0) - \frac{m}{n}\frac{X'y}{\hat{\sigma}_{\varepsilon}^2(\hat{\theta}_{\mathrm{QL}})} \right) \tag{4.3}$$

where $\hat{\sigma}_{\varepsilon}^2(\hat{\theta}_{\mathrm{QL}}) = (y - X\hat{\theta}_{\mathrm{QL}})'(y - X\hat{\theta}_{\mathrm{QL}})$.

The proof to show the existence of moments in the linear IV case is based on the proof that Hausman et al. (2011) use to show that RCUE has finite sample moments. However, their proof is restricted to a specific value for the function they add to the FOC of CUE; more specifically, they derive the proof for the special case that $J(\theta) = \theta$. We need to modify several steps of their proof to adapt it to our case. We can establish the following theorem:

Theorem 7 *Assume linearity of the moment functions in θ and consider the homoskedasticity consistent variance estimate $\hat{\Omega}(\hat{\theta}) = (y - X\hat{\theta})'(y - X\hat{\theta})Z'Zn^{-2}$. Then, the QL-GMM estimator in equation (4.3) is bounded as follows:*

$$||\hat{\theta}_{QL}|| \leq (\tilde{\alpha}_n \tilde{\gamma}_n)^{-1} \left(||g_i(0)||||G_i(\theta)|| \right) - (\tilde{\alpha}_n)^{-1} \left(\frac{m}{n} ||X'y|| \right),$$

where $\tilde{\alpha}_n = \frac{m}{n}\hat{\lambda}_{min}(X'X)$ and $\tilde{\gamma}_n = \hat{\lambda}_{min}(Z'Z)n^{-2}$, where $\hat{\lambda}_{min}(A)$ is the minimum eigenvalue of a generic matrix A.

To use the same argument in Hausman et al. (2011), we analyze equation (4.3). First, consider the denominator. The minimum eigenvalue of X'X, $\hat{\lambda}_{min}(X'X)$, is strictly positive under the assumption that $X'X$ is invertible. Under this assumption, $\frac{m}{n}\frac{\hat{\lambda}_{min}(X'X)}{\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})} > 0$. We impose $\frac{m}{n}\frac{\hat{\lambda}_{min}(X'X)}{\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})}$ as a lower bound to $\left(\bar{G}(\hat{\theta}_{\mathrm{QL}})'[\hat{\Omega}(\hat{\theta}_{\mathrm{QL}})]^{-1}\bar{G}(\hat{\theta}_{\mathrm{QL}}) + \frac{m}{n}\frac{X'X}{\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})} \right)$. In addition to the assumption of invertibility of $X'X$, we need $\bar{G}(\hat{\theta}_{\mathrm{QL}})'[\hat{\Omega}(\hat{\theta}_{\mathrm{QL}})]^{-1}\bar{G}(\hat{\theta}_{\mathrm{QL}})$ to be positive semidefinite with probability 1, as in the proof of Hausman et al. (2011). This requires invertibility of $\hat{\Omega}(\hat{\theta}_{\mathrm{QL}})$, which is also needed to find bounds to the numerator of equation [eq:4] and conclude the proof. We use $\gamma_n = \hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})\hat{\lambda}_{min}(Z'Z)n^{-2}$ as a lower bound to $\hat{\Omega}(\hat{\theta}) = (y - X\hat{\theta})'(y - X\hat{\theta})Z'Zn^{-2}$.

Given the expression for our estimator:

$$\hat{\theta}_{\mathrm{QL}} = -\left(\bar{G}(\hat{\theta}_{\mathrm{QL}})'[\hat{\Omega}(\hat{\theta}_{\mathrm{QL}})]^{-1}\bar{G}(\hat{\theta}_{\mathrm{QL}}) + \frac{m}{n}\frac{X'X}{\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})} \right)^{-1} \left(\bar{G}(\hat{\theta}_{\mathrm{QL}})'[\hat{\Omega}(\hat{\theta}_{\mathrm{QL}})]^{-1}\tilde{g}(0) - \frac{m}{n}\frac{X'y}{\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})} \right)$$

by Cauchy-Schwarz inequality and since $||v + w|| \leq ||v|| + ||w||$ for all v and w:

$$||\hat{\theta}_{\mathrm{QL}}|| \leq (\alpha_n \gamma_n)^{-1} \left(||g_i(0)||||G_i(\theta)|| - \gamma_n \frac{m}{n} \left|\left| \frac{X'y}{\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})} \right|\right| \right)$$

, where $\alpha_n = \frac{m}{n}\frac{\hat{\lambda}_{min}(X'X)}{\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})}$ and $\gamma_n = \hat{\sigma}^2_\varepsilon(\hat{\theta}_{\mathrm{QL}})\hat{\lambda}_{min}(Z'Z)n^{-2}$. The above can be re-

written as:

$$||\hat{\theta}_{\text{QL}}|| \leq (\tilde{\alpha}_n \tilde{\gamma}_n)^{-1} \left(||g_i(0)|| ||G_i(\theta)|| \right) - (\tilde{\alpha}_n)^{-1} \left(\frac{m}{n} ||X'y|| \right)$$

with probability 1, where $\tilde{\alpha}_n = \frac{m}{n}\hat{\lambda}_{min}(X'X)$ and $\tilde{\gamma}_n = \hat{\lambda}_{min}(Z'Z)n^{-2}$. Note that the bounds are conditional on X and Z, as they depend on $\hat{\lambda}_{min}(Z'Z)$ and $\hat{\lambda}_{min}(X'X)$. In order for QL-GMM to have finite moments a necessary and sufficient condition is that the eigenvalues $\hat{\lambda}_{min}(Z'Z)$ and $\hat{\lambda}_{min}(X'X)$ are strictly greater than 0. In practice, given the sample quantities $Z'Z$ and $X'X$, it is possible to check whether this condition holds for any given sample. We have derived conditional bounds (alternatively, one can think about taking the limits of the above expressions and establish asymptotic bounds). Importantly, $\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\text{QL}})$ gets cancelled out, it drops from α_n and γ_n to get $\tilde{\alpha}_n = \frac{m}{n}\hat{\lambda}_{min}(X'X)$ and $\tilde{\gamma}_n = \hat{\lambda}_{min}(Z'Z)n^{-2}$, which depend on observable quantities only and can be examined before proceeding with estimation. The fact that $\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\text{QL}})$ drops in the derivation and does not appear in the bounds is of key relevance. An inspection of the simulation results for the IV setting reveals that the extreme values of CUE are indeed obtained when $\hat{\sigma}^2_\varepsilon(\hat{\theta}_{CUE})$ takes its largest values.

Finally, consider how our bounds relate to those obtained by Hausman et al. (2011):

$$||\hat{\theta}_{RCUE}|| \leq (\alpha\gamma)^{-1} ||g_i(0)|| ||G_i(\theta)||$$

with probability 1, where α and γ are two (strictly positive) tuning parameters to be chosen by the econometrician. Unlike these bounds, the bounds on our estimator do not depend on tuning parameters and $\hat{\sigma}^2_\varepsilon(\hat{\theta}_{\text{QL}})$ gets cancelled out. However, unlike the bounds in Hausman et al. (2011), our bounds are conditional on the eigenvalues $\hat{\lambda}_{min}(Z'Z)$ and $\hat{\lambda}_{min}(X'X)$ being away from 0.

4.3.4 Extensions

An interesting setting in empirical applications is when the estimated covariance matrix of the moments is close to singular in some parts of the parameter space. This possibility might lead to the regularisation log-term in equation (4.2) to dominate

the first term. We leave it to future research to investigate whether this could lead to poor performance of the QL-GMM estimator.

4.4 Simulations

We consider several Monte Carlo environments to assess the properties of the new estimator, the QL-GMM, and of 2-stepGMM, iterativeGMM, CUE, and EL estimators, in finite samples. The criteria used to compare estimators are mean and median bias, Root Mean Squared Errors (RMSE), variance of estimators, probability of deviations of the estimator from the parameter value, differences between the 0.95 and the 0.05 quantiles (in absolute value), interquartile range, average computation time, number of failures (or non-convergence fraction). One should be cautious in interpreting the results from the RMSE because, although this measure is potentially highly informative, it might be misleading since the estimators might not have finite moments. The settings for the Monte Carlo simulations are chosen in order to assess the behavior of the estimators in various scenarios, especially when the Gaussian asymptotic theory might provide a poor approximation to the finite sample distribution for GMM.

We assess the finite sample properties of the estimators with both linear and nonlinear models, and check whether these properties are robust to the number of moments, the number of instrument, the dimension of the parameter vector, and when there is weak identification.

4.4.1 Dynamic Panel Data

This simulation design is taken from Blundell and Bond (1998), Bond et al. (2001), Imbens (2002), and Kitamura (2006). Consider the dynamic panel data model:

$$y_{i,t} = \theta_0 y_{i,t-1} + \eta_i + u_{i,t} \qquad i = 1,...,n \;\; t = 2,...,T \tag{4.4}$$

where $\eta_i \sim_{iid} N(0,1)$, $u_{i,t} \sim_{iid} N(0,1)$, the initial value is drawn according to $y_{i,1} = \eta_i/(1-\theta_0) + e_i$ with $e_i \sim_{iid} N(0, 1/(1-\theta_0^2))$.

We use both System (SYS) and Difference (DIF) moment conditions:

$$E[y_{i,t-2}(\Delta y_{i,t} - \theta_0 \Delta y_{i,t-1})] = 0 \quad t = 3, ..., T \tag{4.5}$$

$$E[\Delta y_{i,t-1}(y_{i,t} - \theta_0 y_{i,t-1})] = 0 \quad t = 3, ..., T \tag{4.6}$$

These imply a total of $(T-1)(T-2)/2 + (T-2)$ moments.

Following Kitamura (2006), the panel dimensions are $n = 100$ and $T = 6$, and the number of Monte Carlo replications is 1000 for each design. Note that the derivatives of these moments are stochastic and potentially correlated with the moments.

For the first step of the 2-step GMM estimation, we use the efficient weighting matrix as described in Blundell and Bond (1998) when having DIF moments only; otherwise, we employ the weighting matrix in Bond et al. (2001) for SYS moments.[2] We also report results for the case when an identity matrix is used as weighting matrix in the first step of the 2-step GMM procedure (the estimator should be consistent as the sample size increases, but it might require much larger finite sample sizes to be well behaved). We analyze the following scenarios:

1) Homoskedastic case: First, note that, in the homoskedastic case, with DIF moments only, the 1-step GMM, the 2-step GMM, the CUE, and the new estimator should all be equivalent since the one step efficient weight matrix does not depend on the models' parameters.

Table 4.1 shows the performance of 2-step GMM, CUE, QL-GMM, and EL, when the true parameter value is .9 and both SYS and DIF moments are used. For comparison, it also displays the simulation results for the same setting reported in Kitamura (2006). The proposed estimator, the QL-GMM, has the best performance with respect to all criteria, apart from the mean bias, which, however, is still smaller

[2]Imbens, Spady, Johnson (pag.343) use as initial weight matrix in the first step of the GMM estimation procedure an average of the outer products of the moments evaluated at the true parameter. Although this approach is not feasible in practice, it will lead to overestimate the performance of the GMM estimator.

than that of the 2-step GMM estimator, and the average calculation time, which is slightly larger than that of the CUE.

For the homoskedastic case, we further investigate the performance of the proposed estimator also for the slightly different frameworks considered in Blundell and Bond (1998), and in Imbens (2002). [3] Blundell and Bond (1998) consider 4 and 11 time periods, DIF moments only or both DIF and SYS moments, true parameter value equal to 0, .5, or .9, and 100 or 500 individuals. Results for the 2-step GMM in our simulations are very similar to those in Blundell and Bond (1998). With all moments and small T ($T = 4$), QL-GMM is the best performing for true values equal to 0 or 0.9 when $n = 100, 500$; while for $\theta = 0.5$ the best performing estimators are 2-step GMM when $n = 100$, and 2-step GMM and EL when $n = 500$; QL-GMM is always performing better than CUE. With large T ($T = 11$), QL-GMM always performs better than CUE when $n = 100$, while the two estimators have very similar performances when $n = 500$. [4] Imbens (2002) considers 1434 individuals, 10000 replications, values of the parameter equal to .5, or .9, time periods from 3 to 11. With all moments, for large n (1434) and large T, when the true value is $\theta = .5$ EL has the best performance, CUE and QL-GMM have very similar performances, but when the true value is .9 QL-GMM is the best performing always apart from $T = 11$, when again EL has a better performance. [5]

To summarize, QL-GMM is performing well always for small T no matter what the sample size n, the true value, the initial guess, and the moment conditions are; it has a superior performance when the true value is close to unity apart from the case of large T and (sufficiently) large n where, as predicted by theory, the performance is comparable to that of CUE and EL.

2a) Heteroskedastic case, conditioning on lagged values of y: The heteroskedastic case is taken from Bond et al. (2001) and Kitamura (2006). In the setting above, the $u_{i,t}$ are replaced by a conditionally heteroskedastic process of the

[3] However, neither paper documents the performance of CUE and EL; Blundell and Bond (1998) report the values of the 2-step GMM only, while Imbens (2002) compares the performance of the 2-step GMM, iterated GMM, and ET.

[4] Results for these simulations are available upon requests.

[5] Results for these simulations are available upon requests.

form $u_{i,t}|y_{i,t-1} \sim_{iid} N(0, 0.4 + 0.3y_{i,t-1}^2)$, and the initial condition is generated using fifty pre-sample draws.

2b) Heteroskedastic case, conditioning on lagged values of u: Finally, consider the same setting as above where the error terms $u_{i,t}$ are replaced by a conditionally heteroskedastic process of the form $u_{i,t}|u_{i,t-1} \sim_{iid} N(0, 0.4 + 0.3u_{i,t-1}^2)$.
Table 4.1 shows the performance of 2-step GMM, CUE, QL-GMM, and EL also for these heteroskedastic cases and compares these results to those in Kitamura (2006). Results for the heteroskedastic cases confirm the superior performance of QL-GMM with respect to CUE in terms of MAE and RMSE. The mean bias of QL-GMM is slightly larger than that of CUE, but, as in the homoskedastic case, smaller when compared to that of the 2-step GMM. These simulations results accord well with our theoretical findings.

4.4.2 IV

In the following, we compare the performance of 2-step GMM, CUE, and QL-GMM using the IV setting described in Hausman et al. (2011). The baseline design is the following:

$$y_i = x_i\beta + \varepsilon_i \tag{4.7}$$

$$x_i = z_{i1}\pi + v_i \tag{4.8}$$

$$\varepsilon_i = \rho v_i + \sqrt{1-\rho^2}(\phi\theta_{1i} + \sqrt{1-\phi^2}\theta_{2i}) \tag{4.9}$$

with $\theta_{1i} \sim N(0, z_{1i}^2)$ and $v_i, \theta_{2i} \sim N(0,1)$. [6] We use the same performance criteria as those in Hausman et al. (2011): MSE, median bias, interquartile range, nine decile range, mean bias, variance of estimates. The results comparing the performance of 2-step GMM, CUE, and QL-GMM, according to the above criteria are reported in the tables below (4.2-7). The results show that CUE in this setting obtain extreme

[6] For other details about the setting, see the design in Hausman et al. (2007).

		Simulations			Kitamura	
		Hom.	Heter. condit. on u	Heter. condit. on y	Hom.	Heter. condit. on y
GMM2	Mean Bias	.042	.062	-.200	.014	-.253
	MAE	.074	.082	.259	.071	.261
	RMSE	.091	.097	.317	.096	.364
	$Pr(AE > 0.1)$	.245	.245	.769	.296	.815
	AvgT	.0814	.0877	.1023		
CUE	Mean Bias	.001	-.003	-.080	.001	-.080
	MAE	.087	.101	.175	.084	.148
	RMSE	.111	.128	.260	.113	.264
	$Pr(AE > 0.1)$	.305	.305	.564	.390	.643
	AvgT	.0379	.0420	.0456		
QL-GMM	Mean Bias	.024	.038	-.112		
	MAE	.072	.082	.168		
	RMSE	.087	.128	.224		
	$Pr(AE > 0.1)$	.202	.243	.586		
	AvgT	.0389	.0449	.0465		
EL	Mean Bias	-.005	-.009	-.075	-.005	-.059
	MAE	.086	.101	.144	.080	.119
	RMSE	.110	.127	.189	.113	.189
	$Pr(AE > 0.1)$	.314	.381	.539	.370	.570
	AvgT	.0668	.0757	.0898		

Table 4.1: Simulations - Dynamic Panel Data Model

Comparison with Kitamura, 1000 replications, 100 individuals, 6 time periods, $\theta = .9$ (using Matlab algorithm fminsims). We report the performance of the GMM2, CUE, QL-GMM, and EL estimators, evaluated according to the following criteria: Mean Bias, MAE (Mean Absolute Error), RMSE (Root Mean Square Error), Pr($AE > 0.1$) (Probability that the Absolute Error is greater than .1), AvgT (Average Computing Time), for the homoskedastic and heteroskedastic cases.

values in several simulations. This large dispersion reflects the no-moment problem that affects the estimator. On the contrary, QL-GMM achieves the objective of restoring the finite sample moments of the CUE: QL-GMM is the best performing estimator among all in terms of dispersion, as measured by the variance of estimates, the nine decile range, and the interquartile range. In terms of MSE, QL-GMM has also a superior performance compared to that of CUE and 2-step GMM, for all values of the concentration parameter (CP), and target R^2, when the number of

instruments (M) is 5-10, while the 2-step GMM outperforms QL-GMM in terms of MSE as M increases ($M \geq 30$). Finally, the price of the reduced dispersion of QL-GMM is the larger median bias compared to that of CUE and, in this setting, also to that of the 2-step GMM estimator. We plan to assess the sensitivity of these results to a different choice of the sample sizes.

rsqTarget	CP	M	2-step GMM	CUE	QL-GMM
0	8	5	0.0918	-0.0013	0.1200
0	8	10	0.1600	0.0056	0.1762
0	8	30	0.2323	0.0043	0.2410
0	8	50	0.2577	0.0163	0.2647
0	16	5	0.0529	0.0043	0.0746
0	16	10	0.1019	-0.0005	0.1176
0	16	30	0.1949	0.0118	0.2065
0	16	50	0.2254	0.0126	0.2388
0	32	5	0.0262	0.0000	0.0384
0	32	10	0.0629	0.0016	0.0732
0	32	30	0.1389	-0.0047	0.1516
0	32	50	0.1800	-0.0044	0.1961
0.2	8	5	0.0866	-0.0051	0.1159
0.2	8	10	0.1556	0.0025	0.1701
0.2	8	30	0.2339	-0.0002	0.2404
0.2	8	50	0.2557	-0.0047	0.2616
0.2	16	5	0.0531	0.0019	0.0722
0.2	16	10	0.1003	-0.0065	0.1165
0.2	16	30	0.1929	-0.0035	0.2053
0.2	16	50	0.2279	0.0151	0.2403
0.2	32	5	0.0217	-0.0036	0.0362
0.2	32	10	0.0595	0.0000	0.0732
0.2	32	30	0.1425	0.0021	0.1547
0.2	32	50	0.1819	0.0120	0.1969

Table 4.2: Simulations - IV Setting, Median Bias
Median bias, $\rho = 0.3$, $T = 400$, 6250 replications.

4.4.3 Modified asset pricing model

Finally, we consider the design in Kitamura et al. (2013) and in Ragusa (2011), which is based on the model of Hall and Horowitz (1996): the modified asset pricing

rsqTarget	CP	M	2-step GMM	CUE	QL-GMM
0	8	5	0.1215	2.8E+28	0.0750
0	8	10	0.0716	6.9E+27	0.0552
0	8	30	0.0291	1.2E+28	0.0258
0	8	50	0.0191	7.4E+27	0.0178
0	16	5	0.0625	2.5E+27	0.0476
0	16	10	0.0452	6.5E+26	0.0375
0	16	30	0.0251	2.2E+27	0.0230
0	16	50	0.0169	6.8E+27	0.0161
0	32	5	0.0323	1.4E+25	0.0278
0	32	10	0.0272	6.7E+25	0.0242
0	32	30	0.0182	1.7E+25	0.0171
0	32	50	0.0137	1.7E+26	0.0139
0.2	8	5	0.1209	6.1E+27	0.0714
0.2	8	10	0.0708	1.3E+28	0.0538
0.2	8	30	0.0301	7.5E+27	0.0263
0.2	8	50	0.0199	8.8E+27	0.0180
0.2	16	5	0.0652	1.0E+27	0.0487
0.2	16	10	0.0456	2.2E+27	0.0384
0.2	16	30	0.0237	1.3E+27	0.0223
0.2	16	50	0.0167	3.0E+27	0.0165
0.2	32	5	0.0313	4.1E-02	0.0270
0.2	32	10	0.0262	4.2E+25	0.0236
0.2	32	30	0.0179	7.0E+25	0.0169
0.2	32	50	0.0137	1.6E+26	0.0135

Table 4.3: Simulations - IV Setting, Variance
Variance of estimates, $\rho = 0.3$, $T = 400$, 6250 replications.

model. Let $x = (x_1, x_2)' \sim N(0, 0.4^2 I)$ and consider the moment function:

$$g(x,\theta) = (exp[-0.72 - \theta(x_1 + x_2) + 3x_2] - 1) \begin{bmatrix} 1 \\ x_2 \end{bmatrix} \tag{4.10}$$

The true parameter value is $\theta_0 = .3$ and $x \sim (0, \Sigma_x)$, where $\Sigma_x = diag(.16, .16)$.

We consider the perturbed case as in Kitamura et al. (2013) and the focus in this comparison is rather different: we assess the robustness of the estimators to 64 perturbations to the variance-covariance matrix of the variables as in Kitamura et al. (2013), i.e. perturbations to the DGP (measurement errors, ...). Thus, $x \sim (0, \Sigma_{\delta,\rho})$

rsqTarget	CP	M	2-step GMM	CUE	QL-GMM
0	8	5	1.1157	2.2134	0.8828
0	8	10	0.8684	3.2612	0.7638
0	8	30	0.5594	7.9174	0.5303
0	8	50	0.4569	7.6525	0.4375
0	16	5	0.8124	1.0919	0.7179
0	16	10	0.6881	1.3528	0.6329
0	16	30	0.5181	2.3578	0.4983
0	16	50	0.4230	2.9509	0.4115
0	32	5	0.5835	0.6627	0.5462
0	32	10	0.5353	0.7314	0.5036
0	32	30	0.4477	1.0838	0.4319
0	32	50	0.3871	1.4661	0.3865
0.2	8	5	1.0574	2.1229	0.8658
0.2	8	10	0.8657	3.1881	0.7581
0.2	8	30	0.5726	6.3375	0.5372
0.2	8	50	0.4604	6.9555	0.4374
0.2	16	5	0.8066	1.0786	0.7139
0.2	16	10	0.7038	1.3138	0.6467
0.2	16	30	0.5071	2.3302	0.4861
0.2	16	50	0.4230	3.0055	0.4215
0.2	32	5	0.5849	0.6556	0.5446
0.2	32	10	0.5289	0.7211	0.5014
0.2	32	30	0.4402	1.1341	0.4280
0.2	32	50	0.3835	1.4051	0.3755

Table 4.4: Simulations - IV Setting, Nine Decile Range
Nine decile range 0.05 to 0.95, $\rho = 0.3$, $T = 400$, 6250 replications.

where $\Sigma_{\delta,\rho} = .16 \begin{bmatrix} (1+\delta)^2 & \rho(1+\delta) \\ \rho(1+\delta) & 1 \end{bmatrix}$ where the unperturbed case is obtained for $\delta = \rho = 0$. In the simulation, Kitamura et al. (2013) set $\rho = .1\sqrt{2}cos(2\pi\omega)$ and $\delta = .25sin(2\pi\omega)$ and let ω vary over $\omega_j = j/64$ for $j = 0, ..., 63$. For calculation of the 2-step GMM, the identity matrix is used in the first step of estimation. QL-GMM has the best performance among all estimators in terms of deviation probabilities (Pr) for a specific range of value (a significant number of perturbations). It shows a significant improvement over the performance of the CUE (even when truncated values are considered), as can be seen from the Figure 4.2. Note that in Kitamura et al. (2013) the main finding from simulation results is that GMM-type estimators (2-step GMM and CUE) tend to be highly sensitive to data perturbations. Poor per-

rsqTarget	CP	M	2-step GMM	CUE	QL-GMM
0	8	5	0.0759	-8.4E+12	0.1205
0	8	10	0.1546	-6.8E+12	0.1759
0	8	30	0.2336	-9.7E+12	0.2429
0	8	50	0.2563	-8.3E+12	0.2641
0	16	5	0.0429	-9.1E+11	0.0715
0	16	10	0.0978	-1.0E+12	0.1182
0	16	30	0.1951	-1.4E+12	0.2087
0	16	50	0.2272	-4.9E+12	0.2406
0	32	5	0.0208	-4.7E+10	0.0374
0	32	10	0.0575	-1.0E+11	0.0723
0	32	30	0.1379	-1.1E+11	0.1519
0	32	50	0.1804	-5.0E+11	0.1976
0.2	8	5	0.0686	-3.8E+12	0.1162
0.2	8	10	0.1510	-7.9E+12	0.1722
0.2	8	30	0.2329	-9.0E+12	0.2433
0.2	8	50	0.2568	-8.0E+12	0.2631
0.2	16	5	0.0404	-1.1E+11	0.0703
0.2	16	10	0.0953	-8.5E+11	0.1159
0.2	16	30	0.1920	-2.3E+12	0.2066
0.2	16	50	0.2253	-3.5E+12	0.2388
0.2	32	5	0.0161	-1.5E-02	0.0327
0.2	32	10	0.0585	-1.3E+11	0.0724
0.2	32	30	0.1411	-1.4E+11	0.1550
0.2	32	50	0.1820	-5.2E+11	0.1983

Table 4.5: Simulations - IV Setting, Mean Bias
Mean Bias, $\rho = 0.3$, $T = 400$, 6250 replications.

formance might be partly explained by the shape of the criterion function (quadratic form, high sensitivity to noises). The additional term in the objective function of QL-GMM leads desirable properties in terms of robustness to perturbations of the DGP.

4.5 Conclusions

In this chapter, we propose a new estimator, the QL-GMM estimator, as a solution to the "no moments" problem of CUE. QL-GMM significantly reduces the wide dispersion of the estimates of CUE in finite samples, while only adding moderate finite-sample biases.

Our theoretical findings show that, in the linear IV setting, the proposed modifi-

rsqTarget	CP	M	2-step GMM	CUE	QL-GMM
0	8	5	0.1272	2.8E+28	0.0895
0	8	10	0.0955	7.0E+27	0.0862
0	8	30	0.0836	1.2E+28	0.0848
0	8	50	0.0848	7.5E+27	0.0875
0	16	5	0.0643	2.5E+27	0.0527
0	16	10	0.0548	6.5E+26	0.0515
0	16	30	0.0632	2.2E+27	0.0666
0	16	50	0.0686	6.9E+27	0.0740
0	32	5	0.0327	1.4E+25	0.0292
0	32	10	0.0305	6.7E+25	0.0294
0	32	30	0.0372	1.7E+25	0.0402
0	32	50	0.0462	1.7E+26	0.0530
0.2	8	5	0.1256	6.1E+27	0.0849
0.2	8	10	0.0936	1.3E+28	0.0835
0.2	8	30	0.0843	7.6E+27	0.0855
0.2	8	50	0.0858	8.8E+27	0.0872
0.2	16	5	0.0668	1.0E+27	0.0536
0.2	16	10	0.0547	2.2E+27	0.0518
0.2	16	30	0.0606	1.3E+27	0.0649
0.2	16	50	0.0675	3.0E+27	0.0735
0.2	32	5	0.0316	4.2E-02	0.0281
0.2	32	10	0.0297	4.2E+25	0.0288
0.2	32	30	0.0378	7.0E+25	0.0410
0.2	32	50	0.0469	1.6E+26	0.0528

Table 4.6: Simulations - IV Setting, Mean Square Error
Mean Square error, $\rho = 0.3$, $T = 400$, 6250 replications.

cation restores the finite sample moments of CUE. In future research, we plan to derive these desirable properties also in the nonlinear setting.

With respect to the regularized CUE proposed in the recent literature, QL-GMM has a more regular objective function and does not require fine tuning and choosing multiple initial values in order to compute the estimator.

Monte Carlo simulations in this chapter confirm our theoretical findings and show that the new estimator provides an attractive alternative to 2-step GMM and CUE in empirical work.

rsqTarget	CP	M	2-step GMM	CUE	QL-GMM
0	8	5	0.3930	0.5766	0.3297
0	8	10	0.3406	0.7309	0.3054
0	8	30	0.2235	0.9371	0.2130
0	8	50	0.1834	0.9702	0.1785
0	16	5	0.3094	0.3802	0.2757
0	16	10	0.2679	0.4250	0.2475
0	16	30	0.2074	0.6127	0.1999
0	16	50	0.1767	0.7362	0.1705
0	32	5	0.2325	0.2585	0.2159
0	32	10	0.2119	0.2715	0.2000
0	32	30	0.1771	0.3740	0.1724
0	32	50	0.1557	0.4683	0.1578
0.2	8	5	0.4011	0.5847	0.3358
0.2	8	10	0.3291	0.7038	0.2994
0.2	8	30	0.2261	0.9297	0.2119
0.2	8	50	0.1894	1.0069	0.1781
0.2	16	5	0.3123	0.3837	0.2784
0.2	16	10	0.2682	0.4316	0.2490
0.2	16	30	0.2017	0.6154	0.1974
0.2	16	50	0.1715	0.7215	0.1690
0.2	32	5	0.2270	0.2519	0.2106
0.2	32	10	0.2120	0.2750	0.1993
0.2	32	30	0.1757	0.3787	0.1720
0.2	32	50	0.1571	0.4599	0.1551

Table 4.7: Simulations - IV Setting, Interquartile Range
Interquartile range 0.25 to 0.75, $\rho = 0.3$, $T = 400$, 6250 replications.

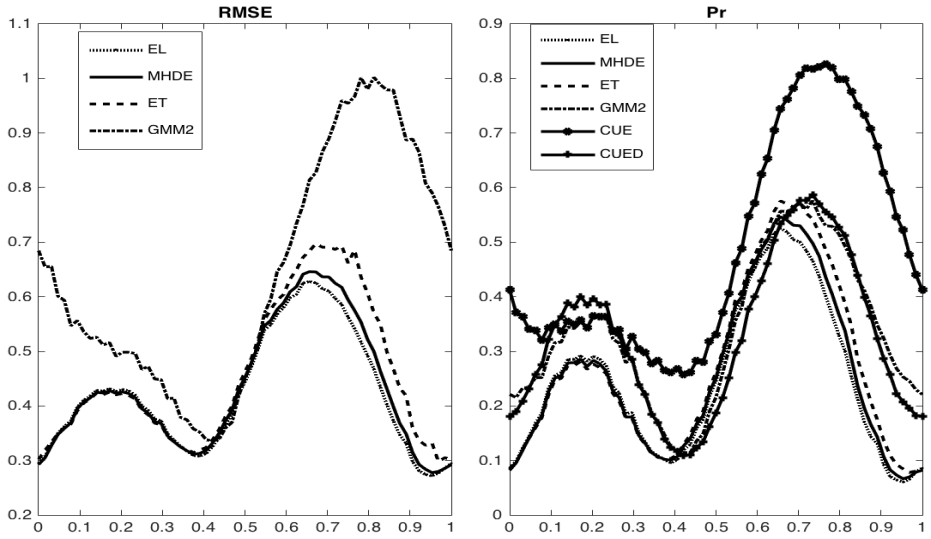


Figure 4.2: Simulations - Modified Asset Pricing Model

The graph shows the performance of QL-GMM in terms of deviation probabilities and RMSE.

Appendix A

Appendix - Chapter 2

A.1 Kalman Filter and Smoother

The Kalman filter is an algorithm that recursively calculates $\{\hat{z}_{it+1|t}\}_{t=1}^{T}$ and $\{P_{it+1|t}\}_{t=1}^{T}$ and given the initial $\hat{z}_{i1|0}$ and $P_{i1|0}$, it is implemented by iterating on the following two equations:

$$\hat{z}_{it+1|t} = B_{it}\hat{z}_{it|t-1} + B_{it}P_{it|t-1}A_t(A_tP_{it|t-1}A_t + \sigma_iH_t)^{-1}(y_{it} - A_t\hat{z}_{it|t-1} - D_tx_{it})$$
$$P_{it+1|t} = B_{it}P_{it|t-1}B'_{it} - B_{it}P_{it|t-1}A_t(A_tP_{it|t-1}A_t + \sigma_iH_t)^{-1}A'_tP_{it|t-1}B'_{it} + S_{it}$$

given that:

$$\hat{z}_{it|t} = \hat{z}_{it|t-1} + P_{it|t-1}A_t(A_tP_{it|t-1}A_t + \sigma_iH_t)^{-1}(y_{it} - A_t\hat{z}_{it|t-1} - D_tx_{it})$$
$$P_{it|t} = P_{it|t-1} - P_{it|t-1}A_t(A_tP_{it|t-1}A_t + \sigma_iH_t)^{-1}A'_tP_{it|t-1}$$

and

$$\hat{z}_{it+1|t} = B_{it}\hat{z}_{it|t}$$
$$P_{it+1|t} = B_{it}P_{it|t}B'_{it} + S_{it}$$

Once I run the Kalman filter and get the sequences $\{\hat{z}_{it+1|t}\}_{t=1}^{T}$ and $\{P_{it+1|t}\}_{t=1}^{T}$, and $\{\hat{z}_{it|t}\}_{t=1}^{T}$ and $\{P_{it|t}\}_{t=1}^{T}$, it is possible to proceed in reverse order in order to calculate the sequence of smoothed estimates $\{\hat{z}_{it|T}\}_{t=1}^{T}$ and their corresponding

mean squared errors $\{P_{it|T}\}_{t=1}^{T}$, as follows:

$$\hat{z}_{it|T} = \hat{z}_{it|t} + P_{it|t}B_{it}'P_{it+1|t}^{-1}(\hat{z}_{it+1|T} - \hat{z}_{it+1|t})$$
$$P_{it|T} = P_{it|t} + P_{it|t}B_{it}'P_{it+1|t}^{-1}(P_{it+1|T} - P_{it+1|t})P_{it+1|t}^{-1}B_{it}P_{it|t}$$

for $t = T-1, T-2, ..., 1$, while $\hat{z}_{iT|T}$ and $P_{iT|T}$ are set equal to the terminal state of the sequence obtained with the Kalman filter and associated variance.

The above recursions are made assuming that the matrices of parameters are known. However, typically parameters are unknown. Denote by θ_i the vector containing all the unknown elements in these matrices for individual i. When one needs to estimate the parameter vector θ_i, one builds the likelihood for the observations y_{it} given its past values and the observables $x_{it}, x_{it-1}, ..., x_{i1}$, for an initial arbitrary guess on θ_i, θ_{i0}. In particular, $y_{it}|x_{it}, ..., x_{i1}, y_{it-1}, ..., y_{i1}; \theta_{i0} \sim \mathcal{N}(\mu_{it}(\theta_{i0}), \Sigma_{it}(\theta_{i0}))$, where $\mu_{it}(\theta_{i0} = D_{it}(\theta_{i0})x_{it} + A_{it}(\theta_{i0})\hat{z}_{it|t-1}(\theta_{i0})$ and $\Sigma_{it}(\theta_{i0}) = A_{it}(\theta_{i0})P_{t|t-1}(\theta_{i0})A_{it}(\theta_{i0})' + \sigma_i(\theta_{i0})H_t(\theta_{i0})$. Based on this, the value of the log-likelihood is:

$$\sum_{t=1}^{T} log f(y_{it}|x_{it}, ..., x_{i1}, y_{it-1}, ..., y_{i1}; \theta_{i0}) =$$
$$k - \frac{1}{2}\sum_{t=1}^{T} log|\Sigma_{it}(\theta_{i0})| - \frac{1}{2}\sum_{t=1}^{T}[y_{it} - \mu_{it}(\theta_{i0})]'\Sigma(\theta_{i0})^{-1}[y_{it} - \mu_{it}(\theta_{i0})]$$

where k is a constant, and the likelihood is evaluated at the initial guess for the unknown parameters. For alternative guesses one proceed to maximize the value of the log-likelihood by numerical method and find the Maximum Likelihood estimates of θ_{i0}. Many alternative optimization techniques exist, one attractive option is the EM algorithm of Watson and Engle (1983).

A.2 Proof of Theorem 1

The proof of Theorem 1 has two parts, one for uniform convergence (i) and the other for convergence in distribution (ii): (i) As in Okui and Yanagi (2020), the proof for uniform convergence starts from the following triangle inequality:
$\sup_{f\in\mathscr{F}} |\mathbb{P}_N f - P_0 f| \leq \sup_{f\in\mathscr{F}} |\mathbb{P}_N f - P_T f| + \sup_{f\in\mathscr{F}} |P_T f - P_0 f|$. The goal is to

show that the term on the left-hand side is bounded by 0. This proof is composed of two steps: in a first step, I bound the second term on the right-hand side by using the convergence in distribution of the MLE estimator. In a second step, I follow Okui and Yanagi (2020) and bound the first term using a modification of the steps in the Glivenko-Cantelli theorem that accounts for the fact that the true distribution of $\hat{\theta}_i$ changes as T increases. In particular, in the first step, I use Assumption 3 to ensure that $\hat{\theta}_i$ converges to θ_i in distribution. Moreover, given that θ_i is continuously distributed by Assumption 2, then Lemma 2.11 in van der Vaart (1998) implies that $\sup_{f\in\mathscr{F}} |P_T f - P_0 f| \to 0$. The second part of the proof is exactly as in Okui and Yanagi (2020) to show that the first term almost surely converges to 0. The assumptions required for this step are assumption 1, condition 1.5 in Hu et al. (1989) and Condition 1.6 in Hu et al. (1989) when I set $X = 2$ in Condition 1.6, which are both satisfied here.

(ii) The proof for convergence in distribution follows a similar logic. □

A.3 Relation to Non-Parametric Literature

Finally, I consider a fixed-T approach to establish a comparison with nonparametric estimation (Almuzara, 2020) and analyze how the results differ when I impose a parametric assumption on the error term. Consider the following simple process for log labor income of individual i at time t:

$$y_{it} = z_{it} + \sigma_i \varepsilon_{i,t} \tag{A.1}$$

$$z_{it} = z_{it-1} + \eta_{it} \tag{A.2}$$

where z_{it} and $\varepsilon_{i,t}$ are unobserved components; $E(\sigma_i^2) = 1$, and the initial level of the random walk is $z_{i1} = z_i$. But impose $\varepsilon_{i,t} \sim N(0, \sigma_\varepsilon^2)$; the distribution of the raw errors $\tilde{\varepsilon}_{i,t} = \sigma_i \varepsilon_{i,t}$ is quite flexible, depending on the distribution of heterogeneous variance. It is a special case of the general state-space model above. In the following, I show first identification of the moments of the cross-sectional distribution of (σ_i^2, z_i), and then identification of their joint distribution.

With stationarity only, I need $T \geq 3$ for identification of $Cov(z_i, \sigma_i^2)$ and $T \geq 4$ for $Var(\sigma_i^2)$ (Almuzara, 2020).

$$Cov(y_{it}, y_{it+k}) = \begin{cases} \sigma_z^2 + \sigma_\varepsilon^2 & \text{if } k = 0,\ t = 1 \\ \sigma_z^2 + \sum_{s=2}^{k} \sigma_\eta^2 + \sigma_\varepsilon^2 & \text{if } k = 0,\ t > 1 \\ \sigma_z^2 & \text{if } k > 0,\ t = 1 \\ \sigma_z^2 + \sum_{s=2}^{k} \sigma_\eta^2 & \text{if } k > 0,\ t > 1 \end{cases}$$

$$Cov(z_i, \sigma_i^2) = \frac{Cov(y_{it}, (\Delta y_{i\tau+1})^2)}{2\sigma_\varepsilon^2} \qquad \tau > t+1$$

$$Var(\sigma_i^2) = \frac{Cov((\Delta y_{it})^2, (\Delta y_{i\tau+2})^2)}{4\sigma_\varepsilon^4} \qquad \tau > t+1$$

It is possible to reduce the minimum number of time periods required for identification, T, when assuming Gaussian shocks: under this assumption, I need $T \geq 2$ for identification.

$$Cov(y_{it}, y_{it+k}) = \begin{cases} \sigma_z^2 + \sigma_\varepsilon^2 & \text{if } k = 0,\ t = 1 \\ \sigma_z^2 + \sum_{s=2}^{k} \sigma_\eta^2 + \sigma_\varepsilon^2 & \text{if } k = 0,\ t > 1 \\ \sigma_z^2 & \text{if } k > 0,\ t = 1 \\ \sigma_z^2 + \sum_{s=2}^{k} \sigma_\eta^2 & \text{if } k > 0,\ t > 1 \end{cases}$$

$$Cov(z_i, \sigma_i^2) = \frac{Cov(y_{it}, (\Delta y_{it+1})^2)}{2\sigma_\varepsilon^2}$$

$$Var(\sigma_i^2) = \frac{Var((\Delta y_{it})^2) - 4(1 - \sigma_\varepsilon^4) + \sigma_\eta^2(\sigma_\eta^2 - 8\sigma_\varepsilon^2)}{8 + 6\sigma_\varepsilon^4}$$

For the latter use Gaussian nature of η but can relax this assumption using the moments $E[y_{it+1}^4] - E[y_{it}^4]$.

As for identification of the cross sectional distribution of the unobservables (σ_i^2, z_i) under Gaussian error, the argument in Hu and Schennach (2008) would simplify here as there is no need for instruments. Let's denote by y earnings, by x lagged earnings, and by x^* the unobservables of interest (σ_i^2, z_i).

$$f(y,x) = \int f(y|x^*) f(x|x^*) f(x^*) dx^*$$

Note that $f(y|x^*)$ and $f(x|x^*)$ are known up to parameters. Then, it is possible to identify the unobserved distribution of interest $f(x^*)$ with just (y,x), no need for additional z, by solving the above for $f(x^*)$ in terms of known objects. Identifiability requires the integral operator to be invertible, this is a completeness condition. If I define y to be two-dimensional I do not need x and identification of $f(x^*)$ is obtained as follows:

$$f(y) = \int f(y|x^*) f(x^*) dx^*$$

Without the parametric assumption on the error term, I need to introduce the variable z, which is further lags or leads of y, i.e. more time periods are required (5 time periods for this simple model, see argument in Almuzara (2020)). Note the analogy with the logic of Mavroeidis et al. (2015), which is based on a fixed-T setting and require a parametric assumption on the distribution of error term. Consider again the simple state-space model:

$$y_{it} = z_{it} + \sigma_i \varepsilon_{i,t} \tag{A.3}$$

$$z_{it} = z_{it-1} + \eta_{it} \tag{A.4}$$

Identification relies on the equality:

$$f_{Y_T,...,Y_2|Y_1}(y_T,...,y_2|y_1) = \int\int\int f_{\zeta,\sigma_\varepsilon,\sigma_\eta|Y_1}(z,s_\varepsilon,s_\eta|y_1) \\ f_{Y_T,...,Y_2|\zeta,\sigma_\varepsilon,\sigma_\eta,Y_1}(y_T,...,y_2|z,s_\varepsilon,s_\eta,y_1)dzds_\varepsilon ds_\eta$$

Provided that the solution exists, one can recover the unknown primitive $f_{\zeta,\sigma|Y_1=y_1}$ by solving the linear equation:

$$f_{\zeta,\sigma_\varepsilon,\sigma_\eta|Y_1}(z,s_\varepsilon,s_\eta|Y_1=y_1) = L^{-1} f_{Y_T,...,Y_2|Y_1=y_1}$$

where L is the linear integral operator:

$$L(\xi)(Y_T,...,Y_2) \\ = \int\int\int \xi(z,s_\varepsilon,s_\eta) \\ f_{Y_T,...,Y_2|\zeta,\sigma_\varepsilon,\sigma_\eta}(y_T,...,y_2|z,s)dzds_\varepsilon ds_\eta$$

For identification, need the linear operator $L : \mathscr{L}^2(F_{\zeta,\sigma|Y1=y1}) \to \mathscr{L}^2(F_{Y_T,...,Y_2|Y1=y1})$ to be complete, i.e. $Lf = 0$ in $\mathscr{L}^2(F_{Y_T,...,Y_2|Y1=y1})$ implies $f = 0$ in $\mathscr{L}^2(F_{\zeta,\sigma_\varepsilon,\sigma_\eta|Y1=y1})$.

[On the conditions for identification, the L^2-completeness conditions can be very difficult or impossible to test.[1] The paper of Andrews (2011) proposes a class of distributions satisfying this conditions but it doesn't extend to multivariate case. Characterization of completeness via characteristic function as in D'Haultfoeuille (2011) may extend to multivariate cases. See also paper of Seely on Completeness for a Family of Multivariate Normal Distributions, given that both ε and η are normally distributed.] It is possible to use the argument in Newey and Powell (2003) to this case given the assumption of normality in the univariate case. Extension to the multivariate case can be established using the results in Lemma 7 of Hu and Schennach, which reduce a multivariate completeness problem to a single variate one, under some independence assumptions on the endogenous variables. On the other

[1] Canay et al. (2013) conclude that no nontrivial tests for testing completeness conditions in nonparametric models with endogeneity involving mean independence restrictions exist.

hand, the Gaussian likelihood might introduce an irregular identification problem (Escanciano, 2020), which results in instability in the estimates; one way of dealing with this issue would be to employ sieve methods with incomplete sieve basis. Estimation details will be discussed in the following section.

For the fixed-T approach, the corresponding estimation approach is based on sieve nonparametric maximum likelihood, see Mavroeidis et al. (2015):

$$max_{\theta \in \Theta_{k(N)}} \sum_{i=1}^{N} log \int \int \int f_{\zeta,\sigma_\varepsilon,\sigma_\eta,Y_1:\theta}(z, s_\varepsilon, s_\eta, y_{i1})$$
$$f_{Y_T,...,Y_2|\zeta,\sigma_\varepsilon,\sigma_\eta,Y_1}(y_T, ..., y_2|z, s_\varepsilon, s_\eta, y_1) dz ds_\varepsilon ds_\eta$$

where $\Theta_{k(N)}$ denotes a sieve space whose dimension $k(N)$ increases with the sample size N; and $\Theta \subset \mathscr{L}^1(F_{\zeta,\sigma_\varepsilon,\sigma_\eta,Y1=y1})$.

Appendix B

Appendix - Chapter 3

B.1 Proofs

Proof. [Proof of Lemma 1] The MSFEs for the TS and the CS are immediate. To obtain the MSFE for the IFS, first, write

$$Y_{i,4} - \widehat{Y}_{i,3} = (Y_{i,4} - Y_{i,3})\,\mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right\} + Y_{i,4}\mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 > Y_{i,2}^2\right\}.$$

Then,

$$\left(Y_{i,4} - \widehat{Y}_{i,3}\right)^2 = (Y_{i,4} - Y_{i,3})^2\,\mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right\} + Y_{i,4}^2\mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 > Y_{i,2}^2\right\}.$$

Thus,

$$\begin{aligned}
\text{MSFE}(\text{IFS}, \theta_i) &= \mathbb{E}\left[\left(Y_{i,4} - \widehat{Y}_{i,3}\right)^2\right] \\
&= \mathbb{E}\left[(Y_{i,4} - Y_{i,3})^2\right]\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] + \mathbb{E}\left[Y_{i,4}^2\mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 > Y_{i,2}^2\right\}\right] \\
&= 2\sigma_i^2\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] + \left(\lambda_i^2 + \sigma_i^2\right) - \mathbb{E}\left[Y_{i,4}^2\mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right\}\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[Y_{i,4}^2 \mathbb{I}\left\{(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_{i,4}^2 \middle| A_i\right] \Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2 \middle| A_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(A_i+U_{i,4})^2 \middle| A_i\right] \Pr\left[(U_{i,2}-U_{i,1})^2 \le (A_i+U_{i,2})^2 \middle| A_i\right]\right] \\
&= \mathbb{E}\left[\left(A_i^2+\sigma_i^2\right) \Pr\left[(U_{i,2}-U_{i,1})^2 \le (A_i+U_{i,2})^2 \middle| A_i\right]\right] \\
&= \mathbb{E}\left[A_i^2 \Pr\left[(U_{i,2}-U_{i,1})^2 \le (A_i+U_{i,2})^2 \middle| A_i\right]\right] + \sigma_i^2 \Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right].
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\mathrm{MSFE}(\mathrm{IFS}, \theta_i) \\
&= \sigma_i^2 \Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right] + \left(\lambda_i^2+\sigma_i^2\right) - \mathbb{E}\left[A_i^2 \Pr\left[(U_{i,2}-U_{i,1})^2 \le (A_i+U_{i,2})^2 \middle| A_i\right]\right],
\end{aligned}$$

which proves the lemma. ■

Proof. [Proof of Theorem 1] By Lemma 1, we have that

$$\begin{aligned}
&\mathrm{MSFE}(\mathrm{IFS}, \theta_i) \\
&= \sigma_i^2 \Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right] + \left(\lambda_i^2+\sigma_i^2\right) - \mathbb{E}\left[A_i^2 \Pr\left[(U_{i,2}-U_{i,1})^2 \le (A_i+U_{i,2})^2 \middle| A_i\right]\right].
\end{aligned}$$

Under (3.4),

$$\begin{aligned}
&\mathrm{MSFE}(\mathrm{IFS}, \theta_i) \\
&\le \sigma_i^2 \left\{1 + \Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right]\right\} + \lambda_i^2 \left\{1 - \Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right]\right\}.
\end{aligned}$$

Therefore, $\mathrm{MSFE}(\mathrm{IFS}, \theta_i) \le \mathrm{MSFE}(\mathrm{TS}, \theta_i)$, provided that

$$\sigma_i^2 \left\{1 + \Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right]\right\} + \lambda_i^2 \left\{1 - \Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right]\right\} \le 2\sigma_i^2,$$

which is equivalent to (3.6). This proves the first conclusion of the theorem.

Similarly, $\text{MSFE}(\text{IFS}, \theta_i) \leq \text{MSFE}(\text{CS}, \theta_i)$, provided that

$$\sigma_i^2 \left\{1 + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]\right\} + \lambda_i^2 \left\{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right\} \leq \sigma_i^2 + \lambda_i^2,$$

which is equivalent to (3.7). This proves the second conclusion of the theorem. ■

Proof. [Proof of Theorem 2] Note that

$$R(\text{IFS}, \theta_i) = \max\left\{0, \text{MSFE}(\text{IFS}, \theta_i) - \sigma_i^2 - \min\{\sigma_i^2, \lambda_i^2\}\right\}.$$

If $\text{MSFE}(\text{IFS}, \theta_i) < \sigma_i^2 + \min\{\sigma_i^2, \lambda_i^2\}$, then $R(\text{IFS}, \theta_i) = 0$. In this case, there is nothing left to prove. Hence, it suffices to assume that $\text{MSFE}(\text{IFS}, \theta_i) \geq \sigma_i^2 + \min\{\sigma_i^2, \lambda_i^2\}$.

In the proof of Theorem 1, we have that under (3.4),

$$\begin{aligned} &\text{MSFE}(\text{IFS}, \theta_i) \\ &\leq \sigma_i^2 \left\{1 + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]\right\} + \lambda_i^2 \left\{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right\}. \end{aligned} \tag{B.1}$$

Therefore,

$$\begin{aligned} &\text{MSFE}(\text{IFS}, \theta_i) - \sigma_i^2 - \min\{\sigma_i^2, \lambda_i^2\} \\ &\leq \lambda_i^2 - \min\{\sigma_i^2, \lambda_i^2\} + \sigma_i^2 \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \lambda_i^2 \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \\ &= \sigma_i^2 \left(\frac{\lambda_i^2}{\sigma_i^2} - \min\left\{1, \frac{\lambda_i^2}{\sigma_i^2}\right\} + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2} \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) \\ &= \sigma_i^2 \left(\frac{\lambda_i^2}{\sigma_i^2} - \min\left\{1, \frac{\lambda_i^2}{\sigma_i^2}\right\} + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2} \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) \mathbb{I}(\lambda_i^2 \leq \sigma_i^2) \\ &+ \sigma_i^2 \left(\frac{\lambda_i^2}{\sigma_i^2} - \min\left\{1, \frac{\lambda_i^2}{\sigma_i^2}\right\} + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2} \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) \mathbb{I}(\lambda_i^2 > \sigma_i^2) \\ &= \sigma_i^2 \left(\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2} \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) \mathbb{I}(\lambda_i^2 \leq \sigma_i^2) \\ &+ \sigma_i^2 \left(\frac{\lambda_i^2}{\sigma_i^2} - 1 + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2} \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) \mathbb{I}(\lambda_i^2 > \sigma_i^2). \end{aligned}$$

Consider the following three subsets of Θ.

$$\Theta_1 := \left\{ (\sigma_i^2, \lambda_i^2) \in \Theta : \frac{\lambda_i^2}{\sigma_i^2} > \frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \right\},$$

$$\Theta_2 := \left\{ (\sigma_i^2, \lambda_i^2) \in \Theta : \frac{\lambda_i^2}{\sigma_i^2} < \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \right\},$$

$$\Theta_3 := \left\{ (\sigma_i^2, \lambda_i^2) \in \Theta : \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \leq \frac{\lambda_i^2}{\sigma_i^2} \leq \frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \right\}.$$

Under Assumption 1, we have that

$$\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] \geq \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right],$$

which implies that

$$\frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} \geq \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]}. \tag{B.2}$$

Thus, Θ can be partitioned into $\Theta = \Theta_1 \cup \Theta_2 \cup \Theta_3$.

Note that $(\sigma_i^2, \lambda_i^2) \in \Theta_1$ iff

$$\sigma_i^2 \left\{ -1 + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] \right\} + \lambda_i^2 \left\{ 1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \right\} < \lambda_i^2 - \sigma_i^2.$$

Therefore, on Θ_1, we have that

$$\text{MSFE}(\text{IFS}, \theta_i) - \sigma_i^2 - \min\{\sigma_i^2, \lambda_i^2\} < (\lambda_i^2 - \sigma_i^2)\mathbb{I}(\lambda_i^2 > \sigma_i^2) \leq \sigma^2 \mu.$$

Turing to the second case, note that $(\sigma_i^2, \lambda_i^2) \in \Theta_2$ iff

$$\sigma_i^2 \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \lambda_i^2 \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] < \sigma_i^2 - \lambda_i^2.$$

Therefore, on Θ_2, we have that

$$\text{MSFE}(\text{IFS}, \theta_i) - \sigma_i^2 - \min\{\sigma_i^2, \lambda_i^2\} < (\sigma_i^2 - \lambda_i^2)\mathbb{I}(\lambda_i^2 \leq \sigma_i^2) \leq \sigma^2\mu.$$

We now move to Θ_3. On Θ_3, we have that

$$\begin{aligned}
&\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2}\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \\
&\leq \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]}\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \\
&= \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\left(\frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} - \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]}\right) \\
&\leq \mu,
\end{aligned}$$

where the last inequality follows from Assumption 4. Furthermore, on Θ_3,

$$\begin{aligned}
&\frac{\lambda_i^2}{\sigma_i^2} - 1 + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2}\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \\
&\leq \frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]}\left(1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) - \left(1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]\right) \\
&= \left(1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right)\left(\frac{\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]} - \frac{1 - \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1 - \Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]}\right) \\
&\leq \mu,
\end{aligned}$$

again using Assumption 4. Therefore,

$$\begin{aligned}
&\text{MSFE}(\text{IFS}, \theta_i) - \sigma_i^2 - \min\{\sigma_i^2, \lambda_i^2\} \\
&\leq \sigma_i^2\left(\Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2}\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right)\mathbb{I}(\lambda_i^2 \leq \sigma_i^2) \\
&+ \sigma_i^2\left(\frac{\lambda_i^2}{\sigma_i^2} - 1 + \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \frac{\lambda_i^2}{\sigma_i^2}\Pr\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right)\mathbb{I}(\lambda_i^2 > \sigma_i^2) \\
&\leq \sigma^2\mu.
\end{aligned}$$

In conclusion, we have shown that $R(\text{IFS}, \theta_i) \leq \sigma^2\mu$ for each $\theta_i \in \Theta$. This proves the first conclusion of the theorem. The second conclusion follows from the fact that the inequality in (B.1) will be strict if Assumption 5 holds additionally. ■

Proof. [Proof of Corollary 1] It follows directly from Theorem 2 and the inequalities in (3.11). ■

Proof. [Proof of Lemma 2] The MSFEs for the TS and the CS are given in the main text. For the IFS, note that

$$
\begin{aligned}
&Y_{i,5} - \widehat{Y}_{i,4} \\
&= \big(Y_{i,5} - 0.5(Y_{i,4} + Y_{i,3})\big)\, \mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right\} + \big(Y_{i,5} - \omega_i Y_{i,4}\big)\, \mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 > Y_{i,2}^2\right\}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\text{MSFE}(\text{IFS}, \theta_i) \\
&= 1.5\sigma_i^2 \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] + (1 + \omega_i)\sigma_i^2 - \mathbb{E}\left[\big(Y_{i,5} - \omega_i Y_{i,4}\big)^2\, \mathbb{I}\left\{(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right\}\right].
\end{aligned}
$$

Then, repeating the arguments identical to those in the proof of Lemma 1, we have that

$$
\begin{aligned}
&\text{MSFE}(\text{IFS}, \theta_i) \\
&= (0.5 - \omega_i^2)\sigma_i^2 \Pr\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] + (1 + \omega_i)\sigma_i^2 \\
&- \mathbb{E}\left[(1 - \omega_i)^2 A_i^2 \Pr\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \big| A_i\right]\right],
\end{aligned}
$$

which proves the lemma. ■

Proof. [Proof of Theorem 3] Define

$$
\zeta_i(a) := \Pr\left[(U_{i,2} - U_{i,1})^2 \leq (A_i + U_{i,2})^2 \big| A_i = a\right]
$$

Using this notation, write

$$\begin{aligned}\Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right] &= \mathbb{E}\left[\Pr\left[(U_{i,2}-U_{i,1})^2 \le (A_i+U_{i,2})^2 \middle| A_i\right]\right] \\ &= \left\{\zeta_i[(2\delta_i)^{-1/2}] + \zeta_i[-(2\delta_i)^{-1/2}]\right\}\delta_i + \zeta_i(0)(1-2\delta_i)\end{aligned}$$

and

$$\mathbb{E}\left[A_i^2 \Pr\left[(U_{i,2}-U_{i,1})^2 \le (A_i+U_{i,2})^2 \middle| A_i\right]\right] = 0.5\left\{\zeta_i[(2\delta_i)^{-1/2}] + \zeta_i[-(2\delta_i)^{-1/2}]\right\}.$$

Thus,

$$\begin{aligned}&\text{MSFE}(\text{IFS},\theta_i) - \text{MSFE}(\text{TS},\theta_i) \\ &= \text{MSFE}(\text{IFS},\theta_i) - \text{MSFE}(\text{CS},\theta_i) \\ &= \left\{\zeta_i[(2\delta_i)^{-1/2}] + \zeta_i[-(2\delta_i)^{-1/2}]\right\}(\delta_i - 0.5) + \zeta_i(0)(1-2\delta_i) \\ &= \frac{1-2\delta_i}{2}\left[\left\{\zeta_i(0) - \zeta_i[(2\delta_i)^{-1/2}]\right\} + \left\{\zeta_i(0) - \zeta_i[-(2\delta_i)^{-1/2}]\right\}\right] \\ &\le 0,\end{aligned}$$

where the last inequality follows from Assumption 1, because Assumption 1 implies that $\zeta_i(a) - \zeta_i(0) \ge 0$ almost surely for any a. This proves the first conclusion of the theorem. The second conclusion follows from the strengthened condition, namely, the individual heterogeneity parameter $\delta_i \in (0, 0.5)$ is restricted to satisfy $\zeta_i(0) - \zeta_i[(2\delta_i)^{-1/2}] < 0$. ■

Proof. [Proof of Theorem 4] It follows from (3.4) in Assumption 1 that

$$\begin{aligned}\text{MSFE}(\text{IFS},\theta_i) \le (1+\omega_i)\sigma_i^2 + (0.5-\omega_i^2)\sigma_i^2 \Pr\left[(Y_{i,2}-Y_{i,1})^2 \le Y_{i,2}^2\right] \\ -(1-\omega_i)^2\lambda_i^2 \Pr\left[(U_{i,2}-U_{i,1})^2 \le U_{i,2}^2\right].\end{aligned} \tag{B.3}$$

Therefore, $\text{MSFE}(\text{IFS}, \theta_i) \leq \text{MSFE}(\text{CS}, \theta_i)$, provided that

$$(0.5-\omega_i^2)\sigma_i^2 \Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] - (1-\omega_i)^2\lambda_i^2 \Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right] \leq 0. \tag{B.4}$$

If $\omega_i^2 \geq 0.5$, (B.4) holds trivially. If $\omega_i^2 < 0.5$, (B.4) is equivalent to

$$\frac{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]} \leq \nu(\omega_i)$$

where

$$\nu(\omega) := \frac{\omega(1-\omega)}{0.5-\omega^2}.$$

Note that $\nu(\omega)$ is strictly increasing and $\nu(0.5) = 1$. Hence, $\nu(\omega_i) > 1$ if $\omega_i > 0.5$. In conclusion, we have proved the second conclusion of the theorem.

Analogously, $\text{MSFE}(\text{IFS}, \theta_i) \leq \text{MSFE}(\text{TS}, \theta_i)$, provided that

$$\omega_i\sigma_i^2 + (0.5-\omega_i^2)\sigma_i^2 \Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] - (1-\omega_i)^2\lambda_i^2 \Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right] \leq 0.5\sigma_i^2,$$

which is equivalent to

$$(0.5-\omega_i^2)\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] + \omega_i - \omega_i(1-\omega_i)\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right] \leq 0.5,$$

or

$$\begin{aligned} 2\omega_i\left(1-\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]\right) - \left(1-\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]\right) \\ - 2\omega_i^2\left(\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] - \Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]\right) \leq 0. \end{aligned}$$

Recall that (3.4) implies that

$$\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] \geq \Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right].$$

Thus, it suffices to assume that

$$2\omega_i \leq \frac{\left(1-\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]\right)}{\left(1-\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]\right)}.$$

Here, it is necessary to assume that $\omega_i \leq 0.5$ since the term on the right-hand side of the inequality above is less than or equal to 1. Therefore, we have proved the first conclusion of the theorem. ■

Proof. [Proof of Theorem 5]

As in the proof of Theorem 2, it suffices to consider the case that

$$\text{MSFE}(\text{IFS}, \theta_i) \geq \sigma_i^2 + \min\{0.5, \omega_i\}\sigma_i^2.$$

Recall that Ω is partitioned into $\Omega_a \cup \Omega_b \cup \Omega_c \cup \Omega_d \cup \Omega_e$:

$$\Omega_a := \left\{(\sigma_i^2, \lambda_i^2) \in \Omega : \sqrt{0.5} < \omega_i \leq 1\right\},$$

$$\Omega_b := \left\{(\sigma_i^2, \lambda_i^2) \in \Omega : 0.5 < \omega_i < \sqrt{0.5} \quad \text{and} \quad \nu(\omega_i) \geq \frac{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}\right\},$$

$$\Omega_c := \left\{(\sigma_i^2, \lambda_i^2) \in \Omega : 0.5 < \omega_i < \sqrt{0.5} \quad \text{and} \quad \nu(\omega_i) < \frac{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}\right\},$$

$$\Omega_d := \left\{(\sigma_i^2, \lambda_i^2) \in \Omega : 0 \leq \omega_i < 0.5 \quad \text{and} \quad 2\omega_i \leq \frac{1-\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1-\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}\right\},$$

$$\Omega_e := \left\{(\sigma_i^2, \lambda_i^2) \in \Omega : 0 \leq \omega_i < 0.5 \quad \text{and} \quad 2\omega_i > \frac{1-\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right]}{1-\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]}\right\}.$$

Define

$$\Delta(\theta_i) := \text{MSFE}(\text{IFS}, \theta_i) - \sigma_i^2 - \min\{0.5, \omega_i\}\sigma_i^2.$$

On $\Omega_a \cup \Omega_b$, we have that by (B.3) and the definitions of Ω_a and Ω_b,

$$\begin{aligned}
\Delta(\theta_i) &= \text{MSFE}(\text{IFS}, \theta_i) - 1.5\sigma_i^2 \\
&\leq (\omega_i - 0.5)\sigma_i^2 + (0.5 - \omega_i^2)\sigma_i^2 \text{Pr}\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - (1-\omega_i)^2\lambda_i^2 \text{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \\
&\leq (\omega_i - 0.5)\sigma_i^2 \\
&\leq \sigma^2 \frac{\kappa}{2}.
\end{aligned}$$

On Ω_c, we have that

$$\begin{aligned}
\Delta(\theta_i) &\leq \sigma_i^2 \left\{(\omega_i - 0.5) + (0.5 - \omega_i^2)\text{Pr}\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \omega_i(1-\omega_i)\text{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right\} \\
&\leq \frac{\sigma_i^2}{2}\Big\{2\omega_i\left(1 - \text{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) - \left(1 - \text{Pr}\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]\right) \\
&\qquad - 2\omega_i^2\left(\text{Pr}\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - \text{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right)\Big\} \\
&\leq \frac{\sigma_i^2}{2}\Big\{2\omega_i\left(1 - \text{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right]\right) - \left(1 - \text{Pr}\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right]\right)\Big\} \\
&\leq \sigma^2 \frac{\kappa}{2},
\end{aligned}$$

where the last inequality follows from Assumption 8. On Ω_d, we have that by (B.3) and the definition of Ω_d,

$$\begin{aligned}
\Delta(\theta_i) &= \text{MSFE}(\text{IFS}, \theta_i) - (1 + \omega_i)\sigma_i^2 \\
&\leq (0.5 - \omega_i^2)\sigma_i^2 \text{Pr}\left[(Y_{i,2} - Y_{i,1})^2 \leq Y_{i,2}^2\right] - (1-\omega_i)^2\lambda_i^2 \text{Pr}\left[(U_{i,2} - U_{i,1})^2 \leq U_{i,2}^2\right] \\
&\leq (0.5 - \omega_i)\sigma_i^2 \\
&\leq \sigma^2 \frac{\kappa}{2}.
\end{aligned}$$

On Ω_e, we have that

$$\begin{aligned}
\Delta(\theta_i) &\leq \frac{\sigma_i^2}{2}\Big\{(1-2\omega_i^2)\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] - 2\omega_i(1-\omega_i)\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]\Big\} \\
&\leq \frac{\sigma_i^2}{2}\Big\{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] - 2\omega_i\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right] \\
&\qquad - 2\omega_i^2\left(\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] - \Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]\right)\Big\} \\
&\leq \frac{\sigma_i^2}{2}\Big\{\Pr\left[(Y_{i,2}-Y_{i,1})^2 \leq Y_{i,2}^2\right] - 2\omega_i\Pr\left[(U_{i,2}-U_{i,1})^2 \leq U_{i,2}^2\right]\Big\} \\
&\leq \sigma^2\frac{\kappa}{2},
\end{aligned}$$

where again the last inequality follows from Assumption 8. Therefore, we have proved the weak inequality version of the theorem. A strict inequality version of the theorem can be established as in the proof of Theorem 2 by making the inequality in (B.3) strict under (3.12). ■

Proof. [Proof of Corollary 2] Theorem 5 and the inequalities in (3.19) directly imply the corollary. ■

B.2 Empirical Bayes

Using the same notation as in LMS, consider the dynamic panel data model

$$Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it} \qquad U_{it} \sim N(0,\sigma^2)$$

The oracle forecast is

$$\hat{Y}_{iT+1}^{opt} = \mathbb{E}_{\theta,\pi,\mathscr{Y}_i}[\lambda_i] + \rho Y_{iT}$$

where $\theta = (\rho,\sigma)$ and $\mathscr{Y}_i$ is the observed trajectory.

Use Tweedie's formula for the posterior mean

$$\mathbb{E}_{\theta,\pi,\mathscr{Y}_i} = \hat{\lambda}_i(\rho) + \frac{\sigma^2}{T}\frac{\partial}{\partial\hat{\lambda}_i(\rho)}\ln p(\hat{\lambda}_i(\rho), y_{i0})$$

LMS approximate the oracle forecast using an empirical Bayes approach. They replace the unknown objects θ and $p(\hat{\lambda}_i(\rho), y_{i0})$ by estimates that exploit the cross-

sectional information. Need consistency of the homogenous parameter θ as a key condition. In the code, use a QMLE estimator of θ that integrates out the heterogeneous intercepts λ_i under the misspecified correlated random effects distribution $\lambda_i|Y_{i0} \sim N(\phi_0 + \phi_1 Y_{i0}, \Omega)$. Then, for estimation of the density $p(\hat{\lambda}_i(\rho), y_{i0})$ use either kernel methods, a mixture approximation, or nonparametric maximum likelihood. In the code of LMS, two EB forecasts are proposed: both use QMLE estimators of θ and then estimate the density using either kernel methods or mixture approximation.

I am using the EB forecast based on kernel methods, i.e., the one that employs the QMLE estimator of θ and then kernel methods for estimation of the density $p(\hat{\lambda}_i(\rho), y_{i0})$. Note that the estimator for common θ is the same as that used for the plug-in QMLE estimator.

Appendix C

Appendix - Chapter 4

C.1 Invariance

Consider 3 types of invariance as in the literature (see e.g. Schennach (2007), Hillier (1990)): 1) Consider general parameter dependent (nonsingular) linear transformation $A(\beta)$ of the vector of moment conditions $g_i(\beta)$. 2) Same as in 1) but assume that the linear transformation does not depend on β, i.e. with $A(\beta) = A$. 3) Finally, consider a general/arbitrary one-to-one differentiable reparameterization $\beta = T(\theta)$ of the moment conditions (invariance in this case means that $\hat{\theta}$ obtained from the reparameterized moment conditions satisfies $\hat{\beta} = T(\hat{\theta})$).

Let's consider 1) first.

Define $h_i(\beta) \equiv A(\beta)g_i(\beta)$, $\hat{h}(\beta) \equiv 1/n\sum_{i=1}^{n} A(\beta)g_i(\beta)$, and $\hat{\Omega}_h(\beta) \equiv 1/n\sum_{i=1}^{n} h_i(\beta)h_i(\beta)' = A(\beta)[1/n\sum_{i=1}^{n} h_i(\beta)h_i(\beta)']^{-1}A(\beta)' = A(\beta)\hat{\Omega}(\beta)A(\beta)'$. Then, define the objective function for the CUE as:

$$\tilde{Q}(\beta) = \hat{h}(\beta)'\hat{\Omega}_h(\beta)^{-1}\hat{h}(\beta)$$

It is possible to see that:

$$\begin{aligned}\tilde{Q}(\beta) &= \hat{h}(\beta)'\hat{\Omega}_h(\beta)^{-1}\hat{h}(\beta) \\ &= \hat{g}(\beta)'A(\beta)'[A(\beta)\hat{\Omega}(\beta)A(\beta)']^{-1}A(\beta)\hat{g}(\beta) \\ &= \hat{g}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{g}(\beta) = Q(\beta)\end{aligned}$$

i.e. the CUE is invariant to parameter dependent nonsingular linear transformation $A(\beta)$ of the vector of moment conditions.

In the case of the proposed estimator:

$$\tilde{Q}(\beta)_{QL} = \frac{1}{n}log|\hat{\Omega}_h(\beta)| + \hat{h}(\beta)'\hat{\Omega}_h(\beta)^{-1}\hat{h}(\beta)$$

The first term is:

$$\frac{1}{n}log|\hat{\Omega}_h(\beta)| = \frac{2}{n}log|A(\beta)| + \frac{1}{n}log|\hat{\Omega}(\beta)|$$

Hence:

$$\tilde{Q}(\beta)_{QL} = \frac{2}{n}log|A(\beta)| + Q(\beta)_{QL}$$

Let's consider 2).

Now for the proposed estimator:

$$\tilde{Q}(\beta)_{QL} = \frac{1}{n}log|\hat{\Omega}_h(\beta)| + \hat{h}(\beta)'\hat{\Omega}_h(\beta)^{-1}\hat{h}(\beta)$$

The first term is:

$$\frac{1}{n}log|\hat{\Omega}_h(\beta)| = \frac{2}{n}log|A| + \frac{1}{n}log|\hat{\Omega}(\beta)|$$

Hence:

$$\tilde{Q}(\beta)_{QL} = k + Q(\beta)_{QL}$$

In this case, QL-GMM is invariant.

We can conclude that QL-GMM is invariant only to linear transformation of the moment function that are not parameter dependent, of the form $h_i(\beta) \equiv Ag_i(\beta)$ for a nonsingular, fixed, *mxm* matrix A, while it is not invariant to linear transformation of this type if the matrix $A(\beta)$ is a function of the unknown parameter β. Note that the standard 2step GMM estimator is not invariant to 1) and 2); the IGMM (iterated GMM) estimator shares the same behavior as QL-GMM, being invariant to 2), but not to 1); CUE, instead, is invariant to both type of transformations.

Now consider 3).

This case is probably the most puzzling because MLE estimators are invariant to this type of transformations. In general, it holds for any estimator where β is the extremum of a differentiable objective function:

$$\frac{\delta}{\delta\theta} log\hat{L}(T(\theta)) = \frac{\delta}{\delta\theta} T(\theta)' \frac{\delta}{\delta\beta} log\hat{L}(T(\beta)) = 0$$

iff

$$\frac{\delta}{\delta\beta} log\hat{L}(T(\beta)) = 0$$

since $\frac{\delta}{\delta\theta} T(\theta)'$ has full rank ($T(\theta)$ being one-to-one).

Take the definition and example in Hayashi textbook (ch.7 on invariance of MLE): an extremum estimator is invariant iff $\tilde{Q}_n(\lambda) = Q_n(\tau^{-1}(\lambda)) \; \forall \lambda \in \Lambda$, where $\tilde{Q}_n(\lambda)$ is the objective function associated with the reparameterized model.

Prove that in the linear case model, $y_t = \theta_0 z_t + \varepsilon_t$ with scalar θ_0 and z_t, the 2-step GMM estimator is not invariant, while CUE is so. $E[x_t(y_t - \theta_0 z_t)] = 0$ with

x_t that can be a vector. For 2-step GMM estimator the objective function is:

$$Q_n(\theta) = \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right)' \hat{W} \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right)$$

Assume that $\Theta = R_{++}$ (i.e. $\theta_0 > 0$) and consider the reparameterization $\lambda = 1/\theta$. The linear equation can be rewritten as $z_t = \lambda_0 y_t - \lambda_0 \varepsilon_t$ and $E[x_t(z_t - \lambda_0 y_t)] = 0$. Now the objective function is:

$$\tilde{Q}_n(\lambda) = \left(\frac{1}{T}\sum_{t=1}^{T} x_t(z_t - \lambda y_t)\right)' \hat{W} \left(\frac{1}{T}\sum_{t=1}^{T} x_t(z_t - \lambda y_t)\right)$$

We can see that $Q_n(\theta) \neq \tilde{Q}_n(1/\theta)$ or, equivalently, $Q_n(1/\lambda) \neq \tilde{Q}_n(\lambda)$. The 2-step GMM estimator is not invariant to this reparameterization. Let's now prove invariance of CUE in the same setting:

$$\begin{aligned} Q_n(\theta) &= \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right)' \left[\sum_{t=1}^{T} x_t(y_t - \theta z_t)(y_t - \theta z_t)' x_t'\right] \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right) \\ &= \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right)' \left[\sum_{t=1}^{T} x_t x_t'(y_t - \theta z_t)^2\right] \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right) \end{aligned}$$

and

$$\tilde{Q}_n(\lambda) = \left(\frac{1}{T}\sum_{t=1}^{T} x_t(z_t - \lambda y_t)\right)' \left[\sum_{t=1}^{T} x_t x_t'(z_t - \lambda y_t)^2\right] \left(\frac{1}{T}\sum_{t=1}^{T} x_t(z_t - \lambda y_t)\right)$$

For invariance need to prove $Q_n(\theta) = \tilde{Q}_n(1/\theta)$.

$$\tilde{Q}_n(1/\theta) = \left(\frac{1}{T}\sum_{t=1}^{T} x_t(z_t - 1/\theta y_t)\right)' \left[\sum_{t=1}^{T} x_t x_t'(z_t - 1/\theta y_t)^2\right] \left(\frac{1}{T}\sum_{t=1}^{T} x_t(z_t - 1/\theta y_t)\right)$$

By multiplying and dividing by $\theta > 0$ and by -1 we get:

$$\tilde{Q}_n(1/\theta) = \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right)' \left[\sum_{t=1}^{T} x_t x_t'(y_t - \theta z_t)^2\right] \left(\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \theta z_t)\right) = Q_n(\theta)$$

Finally, check the behavior of the proposed estimator in the same setting. Add to the objective function $Q_n(\theta)$ the term:

$$\frac{1}{T}log|\hat{\Omega}_h(\theta)| = \frac{1}{T}log\left|\sum_{t=1}^{T} x_t x_t'(y_t - \theta z_t)^2\right|$$

After reparametrization, add to $\tilde{Q}_n(\lambda)$ the term:

$$\frac{1}{T}log\left|\sum_{t=1}^{T} x_t x_t'(z_t - \lambda y_t)^2\right|$$

For invariance need to prove $Q_n^M(\theta) \neq \tilde{Q}_n^M(1/\theta)$. As for the additional term, we will have:

$$\begin{aligned}\frac{1}{T}log\left|\sum_{t=1}^{T} x_t x_t'(y_t - \theta z_t)^2 \frac{1}{\theta^2}\right| &= \frac{1}{T}log\left|\sum_{t=1}^{T} x_t x_t'(y_t - \theta z_t)^2\right|\left|\frac{1}{\theta^2}\right| \\ &= \frac{1}{T}log\left|\sum_{t=1}^{T} x_t x_t'(y_t - \theta z_t)^2\right| + \frac{1}{T}log\frac{1}{\theta^2}\end{aligned}$$

The second term is what makes the proposed estimator not invariant to this reparameterization. Note that for linear transformation, i.e for transformation of the type $\gamma = \alpha\theta$, with α being a constant $\neq 0$, the estimator satisfies invariance. 2-step GMM estimator would not be invariant even to linear transformation.

C.2 Extensions

We consider also a slight modification of the proposed estimator, which adds the Jacobian of the transformation of the parameter θ into $g(\theta)$ when writing the pdf from the distribution of the moment function. This modification is inspired by the the NLFI maximum likelihood estimator, see e.g. Amemiya, 1977, on asymptotic theory of nonlinear estimation and non linear simultaneous equation systems. There the argument is the following. Assume $u_{it} \sim N(0,\Sigma)$, we can write the log likelihood function of the system of nonlinear equations:

$$f_{it}(y_t, x_t, \alpha_i) = u_{it}$$

as:

$$L^* = k - \frac{T}{2} log|\Sigma| + \sum_{t=1}^{T} log\left|\left|\frac{\delta f_t}{\delta y_t'}\right|\right| - \frac{1}{2}\sum_{t=1}^{T} f_t'\Sigma^{-1} f_t$$

where y_t is an N- vector of endogenous variables, x_t is a vector of exogenous variables, and α_i is a K_i-vector of unknown parameters.

In this case, one would get the same expression for the objective function we have plus an additional term containing the Jacobian of the moment with respect to the parameter. The additional term should look like: $-\frac{2}{n} log\left|\left|\frac{\delta}{\delta\theta}\hat{g}(\theta)\right|\right|$. But the problem is on inversion of $\hat{g}(\theta)$, given that we are in the overidentified case.

$$\hat{h} \equiv \hat{g}(\theta) \sim N\left(0, \frac{\hat{\Omega}(\theta)}{n}\right)$$

The pdf of this distribution will be:

$$\frac{1}{2\pi|\hat{\Omega}(\theta)/n|^{1/2}} exp\left\{ -\frac{1}{2}\hat{h}'(\hat{\Omega}(\theta)/n)^{-1}\hat{h} \right\}$$

Use the transformation $\hat{h} \equiv \hat{g}(\theta)$ and the fact that:

$$F_\Theta(\theta) = Pr\{\Theta \leq \theta\} = Pr\{\hat{g}^{-1}(\hat{h}) \leq \theta\} = Pr\{\hat{h} \leq \hat{g}(\theta)\} = F_h(\hat{g}(\theta))$$

$$f_\Theta(\theta) = F_\Theta'(\theta) = \frac{\delta}{\delta\theta} F_h(\hat{g}(\theta)) = f_h(\hat{g}(\theta))\left|\left|\frac{\delta}{\delta\theta}\hat{g}(\theta)\right|\right|$$

Hence:

$$\frac{1}{2\pi|\hat{\Omega}/n|^{1/2}} exp\left\{ -\frac{1}{2}\hat{g}(\theta)'(\hat{\Omega}/n)^{-1}\hat{g}(\theta) \right\}\left|\left|\frac{\delta}{\delta\theta}\hat{g}(\theta)\right|\right|$$

In log terms:

$$k - \frac{1}{2} log|\hat{\Omega}| - \frac{1}{2}\hat{g}(\theta)'(\hat{\Omega}/n)^{-1}\hat{g}(\theta) + log\left|\left|\frac{\delta}{\delta\theta}\hat{g}(\theta)\right|\right|$$

$$-log|\hat{\Omega}| - \hat{g}(\theta)'(\hat{\Omega}/n)^{-1}\hat{g}(\theta) + 2log\left|\left|\frac{\delta}{\delta\theta}\hat{g}(\theta)\right|\right|$$

the estimator would be obtained as:

$$\hat{\beta} = \arg\min_{\beta\in\mathscr{B}} \frac{1}{n} log|\hat{\Omega}| + \hat{g}(\theta)'(\hat{\Omega})^{-1}\hat{g}(\theta) - \frac{2}{n} log\left|\left|\frac{\delta}{\delta\theta}g(\theta)\right|\right|$$

One would need to impose at least local monotonicity, at the true parameter value. An investigation of the properties of this modification is left for future research.

Bibliography

ABOWD, J. M. AND D. CARD (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57, 411–445.

ACEMOGLU, D. AND J. A. ROBINSON (2014): "The rise and fall of general laws of capitalism," *Unpublished Paper, MIT, Department of Economics*.

ANATOLYEV, S. (2005): "GMM, GEL, serial correlation, and asymptotic bias," *Econometrica*, 73, 983–1002.

ANDERSON, B. AND J. MOORE (1989): "Optimal control-linear optimal control," *Prcntice—Hall, Englewood Cliffs*.

ARELLANO, M., R. BLUNDELL, AND S. BONHOMME (2017): "Earnings and consumption dynamics: a nonlinear panel data framework," *Econometrica*, 85, 693–734.

ARELLANO, M., R. W. BLUNDELL, AND S. BONHOMME (2015): "Earnings and consumption dynamics: a nonlinear panel data framework," .

ATHEY, S., R. CHETTY, G. W. IMBENS, AND H. KANG (2019): "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," Working Paper 26463, National Bureau of Economic Research.

AUTOR, D. H., L. F. KATZ, AND M. S. KEARNEY (2006): "The polarization of the US labor market," *The American economic review*, 96, 189–194.

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

BAKER, M. (1997): "Growth-rate heterogeneity and the covariance structure of life-cycle earnings," *Journal of labor Economics*, 338–375.

BALTAGI, B. H. (2008): "Forecasting with panel data," *Journal of Forecasting*, 27, 153–173.

BLUNDELL, R. AND S. BOND (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of econometrics*, 87, 115–143.

BLUNDELL, R., L. PISTAFERRI, AND I. PRESTON (2008): "Consumption inequality and partial insurance," *The American Economic Review*, 98, 1887–1921.

BOND, S., C. BOWSHER, AND F. WINDMEIJER (2001): "Criterion-based inference for GMM in autoregressive panel data models," *Economics Letters*, 73, 379–388.

BONHOMME, S. AND J.-M. ROBIN (2010): "Generalized non-parametric deconvolution with an application to earnings dynamics," *The Review of Economic Studies*, 77, 491–533.

BOTOSARU, I. AND Y. SASAKI (2018): "Nonparametric heteroskedasticity in persistent panel processes: An application to earnings dynamics," *Journal of Econometrics*, 203, 283–296.

BROWNING, M., M. EJRNAES, AND J. ALVAREZ (2010): "Modelling income processes with lots of heterogeneity," *The Review of Economic Studies*, 77, 1353–1381.

BURMEISTER, E., K. D. WALL, AND J. D. HAMILTON (1986): "Estimation of unobserved expected monthly inflation using Kalman filtering," *Journal of Business & Economic Statistics*, 4, 147–160.

CANAY, I. A., A. SANTOS, AND A. M. SHAIKH (2013): "On the testability of identification in some nonparametric models with endogeneity," *Econometrica*, 81, 2535–2559.

CANER, M. (2009): "Lasso-type GMM estimator," *Econometric Theory*, 25, 270–290.

CARRASCO, M. AND G. TCHUENTE (2016): "Efficient estimation with many weak instruments using regularization techniques," *Econometric Reviews*, 35, 1609–1637.

CHAMBERLAIN, G. (1984): "Panel data," *Handbook of econometrics*, 2, 1247–1318.

CHAMBERLAIN, G. AND K. HIRANO (1999): "Predictive distributions based on longitudinal earnings data," *Annales d'Economie et de Statistique*, 211–242.

CHAO, J. C., J. A. HAUSMAN, W. K. NEWEY, N. R. SWANSON, AND T. WOUTERSEN (2012): "An expository note on the existence of moments of Fuller and HFUL estimators," in *Essays in Honor of Jerry Hausman*, Emerald Group Publishing Limited, 87–106.

CHETTY, R., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014a): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632.

——— (2014b): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–79.

CHRISTENSEN, T., H. R. MOON, AND F. SCHORFHEIDE (2020): "Robust Forecasting," arXiv Working Paper arXiv:2011.03153 [econ.EM], `https://arxiv.org/abs/2011.03153`.

CLARK, T. E. AND M. W. MCCRACKEN (2001): "Tests of equal forecast accuracy and encompassing for nested models," *Journal of econometrics*, 105, 85–110.

COMMANDEUR, J. J. AND S. J. KOOPMAN (2007): *An introduction to state space time series analysis*, Oxford University Press.

CREEL, M. AND D. KRISTENSEN (2011): "Indirect likelihood inference," .

——— (2013): "Indirect likelihood inference (revised)," .

DALY, M., D. HRYSHKO, AND I. MANOVSKII (2014): "Reconciling estimates of earnings processes in growth rates and levels," .

DE JONG, P. ET AL. (1991): "The diffuse Kalman filter," *The Annals of Statistics*, 19, 1073–1083.

DE NARDI, M., G. FELLA, AND G. P. PARDO (2016): "The implications of richer earnings dynamics for consumption, wealth, and welfare," Tech. rep., National Bureau of Economic Research.

DEATON, A. (1991): "SAVING AND LIQUIDITY CONSTRAINTS," *Econometrica*, 59, 1221–1248.

DOMINITZ, J. AND C. F. MANSKI (2021): "Minimax–regret sample design in anticipation of missing data, with application to panel data," *Journal of Econometrics*.

DONALD, S. G. AND W. K. NEWEY (2000): "A jackknife interpretation of the continuous updating estimator," *Economics Letters*, 67, 239–243.

DOUC, R., E. MOULINES, AND D. STOFFER (2014): *Nonlinear time series: Theory, methods and applications with R examples*, CRC press.

DURBIN, J. AND S. J. KOOPMAN (2012): *Time series analysis by state space methods*, Oxford university press.

EJRNÆS, M. AND M. BROWNING (2014): "The persistent–transitory representation for earnings processes," *Quantitative Economics*, 5, 555–581.

FARBMACHER, H. (2016): "A normalization of the CUE when some parameters are weakly identified," .

FARCOMENI, A. AND A. PUNZO (2019): "Robust model-based clustering with mild and gross outliers," *Test*, 1–19.

FIELDS, G. S. AND E. A. OK (1996): "The meaning and measurement of income mobility," *Journal of Economic Theory*, 71, 349–377.

——— (1999): "The measurement of income mobility: an introduction to the literature," in *Handbook of income inequality measurement*, Springer, 557–598.

FRÖHWIRTH-SCHNATTER, S. AND S. KAUFMANN (2008): "Model-based clustering of multiple time series," *Journal of Business & Economic Statistics*, 26, 78–89.

FULLER, W. A. (1977): "Some properties of a modification of the limited information estimator," *Econometrica (pre-1986)*, 45, 939.

GABAIX, X., J.-M. LASRY, P.-L. LIONS, AND B. MOLL (2016): "The dynamics of inequality," *Econometrica*, 84, 2071–2111.

GAMERMAN, D. AND H. S. MIGON (1993): "Dynamic hierarchical models," *Journal of the Royal Statistical Society: Series B (Methodological)*, 55, 629–642.

GARCÍA, J. L., J. J. HECKMAN, D. E. LEAF, AND M. J. PRADOS (2020): "Quantifying the Life-Cycle Benefits of an Influential Early-Childhood Program," *Journal of Political Economy*, 128, 2502–2541.

GEVERS, M. AND V. WERTZ (1984): "Uniquely identifiable state-space and ARMA parametrizations for multivariable linear systems," *Automatica*, 20, 333–347.

GIACOMINI, R. AND H. WHITE (2006): "Tests of conditional predictive ability," *Econometrica*, 74, 1545–1578.

GIANNONE, D., M. LENZA, AND G. E. PRIMICERI (2021): "Economic Predictions with Big Data: The Illusion of Sparsity," *Econometrica*, forthcoming.

GOOS, M. AND A. MANNING (2007): "Lousy and lovely jobs: The rising polarization of work in Britain," *The review of economics and statistics*, 89, 118–133.

GOOS, M., A. MANNING, AND A. SALOMONS (2014): "Explaining job polarization: Routine-biased technological change and offshoring," *American economic review*, 104, 2509–26.

GU, J. AND R. KOENKER (2015): "Unobserved heterogeneity in income dynamics: an empirical Bayes perspective," *Journal of Business & Economic Statistics*.

GUGGENBERGER, P. (2008): "Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator," *Econometric Reviews*, 27, 526–541.

GUGGENBERGER, P. ET AL. (2005): "Monte-carlo evidence suggesting a no moment problem of the continuous updating estimator," *Economics Bulletin*, 3, 1–6.

GUVENEN, F. (2007): "Learning your earning: Are labor income shocks really very persistent?" *The American economic review*, 687–712.

——— (2009): "An empirical investigation of labor income processes," *Review of Economic dynamics*, 12, 58–79.

GUVENEN, F., F. KARAHAN, S. OZKAN, AND J. SONG (2015): "What do data on millions of US workers reveal about life-cycle earnings risk?" Tech. rep., National Bureau of Economic Research.

GUVENEN, F. AND A. A. SMITH (2014): "Inferring Labor Income Risk and Partial Insurance From Economic Choices," *Econometrica*, 82, 2085–2129.

HAIDER, S. J. (2001): "Earnings instability and earnings inequality of males in the United States: 1967–1991," *Journal of labor Economics*, 19, 799–836.

HALL, P. AND J. L. HOROWITZ (1996): "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrica: Journal of the Econometric Society*, 891–916.

HAMILTON, J. D. (1985): "Uncovering financial market expectations of inflation," *Journal of Political Economy*, 93, 1224–1241.

——— (1994): "State-space models," *Handbook of econometrics*, 4, 3039–3080.

HANNAN, E. J. AND M. DEISTLER (2012): *The statistical theory of linear systems*, SIAM.

HANSEN, L. P., J. HEATON, AND A. YARON (1996): "Finite-sample properties of some alternative GMM estimators," *Journal of Business & Economic Statistics*, 14, 262–280.

HANSEN, L. P. AND K. J. SINGLETON (1982): "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica: Journal of the Econometric Society*, 1269–1286.

HANSEN, P. R., A. LUNDE, AND J. M. NASON (2011): "The model confidence set," *Econometrica*, 79, 453–497.

HARVEY, A. C. (1990): *Forecasting, structural time series models and the Kalman filter*, Cambridge university press.

HAUSMAN, J., R. LEWIS, K. MENZEL, AND W. NEWEY (2011): "Properties of the CUE estimator and a modification with moments," *Journal of econometrics*, 165, 45–57.

HAUSMAN, J. A., W. K. NEWEY, T. WOUTERSEN, J. C. CHAO, AND N. R. SWANSON (2012): "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 3, 211–255.

HEATHCOTE, J., F. PERRI, AND G. L. VIOLANTE (2010): "Unequal we stand: An empirical analysis of economic inequality in the United States, 1967–2006," *Review of Economic dynamics*, 13, 15–51.

HILLIER, G. H. (1990): "On the normalization of structural equations: Properties of direction estimators," *Econometrica: Journal of the Econometric Society*, 1181–1194.

HOFFMANN, F. (????): "HIP, RIP and the Robustness of Empirical Earnings Processes," .

HOLCBLAT, B. (2015): "On the Empirical Saddlepoint Approximation with Application to Asset Pricing," *Available at SSRN 2082423*.

HOLCBLAT, B. AND F. SOWELL (2019): "The Empirical Saddlepoint Estimator," *arXiv preprint arXiv:1905.06977*.

HOROWITZ, J. L. AND M. MARKATOU (1996): "Semiparametric estimation of regression models for panel data," *The Review of Economic Studies*, 63, 145–168.

HOSPIDO, L. (2012): "Modelling heterogeneity and dynamics in the volatility of individual wages," *Journal of Applied Econometrics*, 27, 386–414.

——— (2015): "Wage dynamics in the presence of unobserved individual and job heterogeneity," *Labour Economics*, 33, 81–93.

HRYSHKO, D. (2012): "Labor income profiles are not heterogeneous: Evidence from income growth rates," *Quantitative Economics*, 3, 177–209.

HU, Y. AND S. M. SCHENNACH (2008): "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 76, 195–216.

IMBENS, G., R. H. SPADY, AND P. JOHNSON (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–358.

IMBENS, G. W. (2002): "Generalized method of moments and empirical likelihood," *Journal of Business & Economic Statistics*, 20, 493–506.

JANTTI, M. AND S. P. JENKINS (2013): "Income mobility," .

JAZWINSKI, A. H. (1969): "Adaptive filtering," *Automatica*, 5, 475–485.

JIANG, W., B. TURNBULL, ET AL. (2004): "The indirect method: inference based on intermediate statistics—a synthesis and examples," *Statistical Science*, 19, 239–263.

JOCHMANS, K. AND M. WEIDNER (2018): "Inference on a distribution from noisy draws," *arXiv preprint arXiv:1803.04991*.

KANE, T. J. AND D. O. STAIGER (2008): "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," Working Paper 14607, National Bureau of Economic Research.

KARAHAN, F. AND S. OZKAN (2013): "On the persistence of income shocks over the life cycle: Evidence, theory, and implications," *Review of Economic Dynamics*, 16, 452–476.

KIM, Y.-S., T. LOUP, J. LUPTON, AND F. P. STAFFORD (2000): "Notes on the 'Income Plus' Files: 1994-1997 Family Income and Components Files," *Documentation, the Panel Study of Income Dynamics (http://www. isr. umich. edu/src/psid/income94-97/y-pls-notes. htm)*.

KINAL, T. W. (1980): "The existence of moments of k-class estimators," *Econometrica: Journal of the Econometric Society*, 241–249.

KITAMURA, Y. (2006): "Empirical likelihood methods in econometrics: Theory and practice," .

KITAMURA, Y., T. OTSU, AND K. EVDOKIMOV (2013): "Robustness, infinitesimal neighborhoods, and moment restrictions," *Econometrica*, 81, 1185–1201.

KITAMURA, Y. AND M. STUTZER (1997): "An information-theoretic alternative to generalized method of moments estimation," *Econometrica: Journal of the Econometric Society*, 861–874.

KLEINBERG, J., J. LUDWIG, S. MULLAINATHAN, AND Z. OBERMEYER (2015): "Prediction policy problems," *The American Economic Review*, 105, 491–495.

KOCHERLAKOTA, N. R. (1990): "On tests of representative consumer asset pricing models," *Journal of Monetary Economics*, 26, 285–304.

KOROBILIS, D. (2016): "Prior selection for panel vector autoregressions," *Computational Statistics & Data Analysis*, 101, 110–120.

LEWIS, D. J., D. MELCANGI, AND L. PILOSSOPH (2019): "Latent heterogeneity in the marginal propensity to consume," *FRB of New York Staff Report*.

LILLARD, L. A. AND Y. WEISS (1979): "Components of variation in panel earnings data: American scientists 1960-70," *Econometrica: Journal of the Econometric Society*, 437–454.

LIU, L. (2017): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," Tech. rep., Working paper, University of Pennsylvania.

LIU, L., H. R. MOON, AND F. SCHORFHEIDE (2016): "Forecasting with Dynamic Panel Data Models," .

——— (2020): "Forecasting With Dynamic Panel Data Models," *Econometrica*, 88, 171–201.

MACURDY, T. E. (1982): "The use of time series processes to model the error structure of earnings in a longitudinal data analysis," *Journal of econometrics*, 18, 83–114.

MANSKI, C. F. (2019): "Econometrics For Decision Making: Building Foundations Sketched By Haavelmo And Wald," Tech. rep., National Bureau of Economic Research.

MARIANO, R. S. (1972): "The existence of moments of the ordinary least squares and two-stage least squares estimators," *Econometrica: Journal of the Econometric Society*, 643–652.

MARIANO, R. S. AND T. SAWA (1972): "The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included

endogenous variables," *Journal of the American Statistical Association*, 67, 159–163.

MAVROEIDIS, S., Y. SASAKI, AND I. WELCH (2015): "Estimation of heterogeneous autoregressive parameters with short panel data," *Journal of Econometrics*, 188, 219–235.

MEGHIR, C. AND L. PISTAFERRI (2004): "Income variance dynamics and heterogeneity," *Econometrica*, 72, 1–32.

MOON, H. R. AND M. WEIDNER (2018): "Nuclear norm regularized estimation of panel regression models," *arXiv preprint arXiv:1810.10987*.

NEWEY, W. K. AND R. J. SMITH (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 219–255.

OKUI, R. AND T. YANAGI (2020): "Kernel estimation for panel data with heterogeneous dynamics," *The Econometrics Journal*, 23, 156–175.

OWEN, A. B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237–249.

PEÑARANDA, F. AND E. SENTANA (2012): "Spanning tests in return and stochastic discount factor mean–variance frontiers: A unifying approach," *Journal of Econometrics*, 170, 303–324.

——— (2015): "A unifying approach to the empirical evaluation of asset pricing models," *Review of Economics and Statistics*, 97, 412–435.

PETRIS, G. AND R. AN (2010): "An R package for dynamic linear models," *Journal of Statistical Software*, 36, 1–16.

PHILLIPS, P. C. (1983): "Exact small sample theory in the simultaneous equations model," *Handbook of econometrics*, 1, 449–516.

POSTEL-VINAY, F. AND H. TURON (2010): "On-the-job search, productivity shocks, and the individual earnings process," *International Economic Review*, 51, 599–629.

PRIMICERI, G. E. AND T. VAN RENS (2009): "Heterogeneous life-cycle profiles, income risk and consumption inequality," *Journal of monetary Economics*, 56, 20–39.

RAGUSA, G. (2011): "Minimum divergence, generalized empirical likelihoods, and higher order expansions," *Econometric Reviews*, 30, 406–456.

SAEZ, E. (2001): "Using elasticities to derive optimal income tax rates," *The review of economic studies*, 68, 205–229.

SAWA, T. (1969): "The exact sampling distribution of ordinary least squares and two-stage least squares estimators," *Journal of the American Statistical association*, 64, 923–937.

SCHENNACH, S. M. (2007): "Point estimation with exponentially tilted empirical likelihood," *The Annals of Statistics*, 634–672.

SENTANA, E. ET AL. (2015): "Finite underidentification," Tech. rep.

SHORROCKS, A. F. (1978): "The measurement of mobility," *Econometrica: Journal of the Econometric Society*, 1013–1024.

TAUCHEN, G. (1986): "Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data," *Journal of Business & Economic Statistics*, 4, 397–416.

TAUCHEN, G. AND R. HUSSEY (1991): "Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models," *Econometrica: Journal of the Econometric Society*, 371–396.

TOPEL, R. H. AND M. P. WARD (1992): "Job Mobility and the Careers of Young Men," *The Quarterly Journal of Economics*, 439–479.

WALL, K. D. (1987): "IDENTIFICATION THEORY FOR VARYING COEFFICIENT REGRESSION MODELS 1," *Journal of Time Series Analysis*, 8, 359–371.

WHITE, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, 1–25.

WRIGHT, J. H. (2003): "Detecting lack of identification in GMM," *Econometric theory*, 19, 322–330.

Statement of Conjoint Work

Note on the joint work in Silvia Sarpietro's thesis "Essays on Dynamic Unobservable Heterogeneity".

Chapter 2, "Dynamic Unobservable Heterogeneity: Income Inequality and Job Polarization", is single-authored by Silvia Sarpietro.

Chapter 3, "Individual Forecast Selection", was undertaken as joint work with Raffaella Giacomini and Simon Lee.

Chapter 4, "Regularized CUE: a Quasi-Likelihood Approach", was undertaken as joint work with Dennis Kristensen.

Contributions from the authors were equal in the case of the shared chapters.