4-20-2020

Three Essays on Econometrics

Zhonghui Zhang
University of Connecticut - Storrs, zhangxiao2432@gmail.com

Three Essays on Econometrics

Zhonghui Zhang, PhD

University of Connecticut, **2020**

Abstract: This dissertation contains three chapters. In the first chapter, I investigate why there exist considerable variation in estimates of the coefficient of relative risk aversion (CRRA) and the elasticity of intertemporal substitution (EIS) in the consumption-based asset pricing model with Epstein and Zin (1989) preferences. I find the Epstein and Zin (1989) structure collapses to the time-separable structure when returns are within a reasonable range. I also show the choice of parameters might lead to an "ill-behaved" conditional moment, which might cause either the GMM method to get "stuck," or the estimates do not move much from the starting points. The second and third chapters are about the estimation of latent group heterogeneities in panel data. In chapter two, I propose a Mahalanobis metric based k-means algorithm (KMM) for group membership estimation in linear panel data models with time-varying grouped fixed-effects by Bonhomme and Manresa (2015). The proposed method improves the accuracy of estimates by taking serial correlation and heteroscedasticity into account. I also derive the optimal β for group membership estimation and show that it may be different from the true coefficient parameter. Since the optimal β is not feasible in practice, I propose the data-driven selection method for its implementation. In the third chapter, I develop a novel approach to estimate causal parameters and latent group heterogeneities in two independent steps without an iteration procedure. In the existing literature, the estimation of coefficients depend on the estimation of grouping, and vice versa, which may lead to significant estimation error if any step is misleading. My approach applies to a linear panel data model with either time-varying or interactive group fixed effects.

Three Essays on Econometrics

Zhonghui Zhang

B.S., University of Shanghai for Science and Technology, **2012**

M.A., University at Albany, SUNY, **2015**

M.S., University of Connecticut, **2020**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2020

2020

Approval Page

Doctor of Philosophy Dissertation

Three Essays on Econometrics

Presented by
Zhonghui Zhang, B.S., M.A.,M.S,

Major Advisor _____

Chihwa Kao

Associate Advisor _____

Min Seong Kim

Associate Advisor _____

Jungbin Hwang

University of Connecticut
2020

Acknowledgments

I am lucky that in my five years of Ph.D. life, I met many lovely and inspiring people who had a profound impact on me. It is they who made the tough research life full of happiness and progress. My sincerest thanks go to my advisor Professor Chihwa Kao, for his exceptional sights, profound enlightenment, and constant encouragement. It was a great privilege and honor to work and study under his guidance. I'm extremely grateful for what he has offered me. Without his advice and persistent help, this dissertation would not have been possible.

I would like to thank my committee members, Professor Min Seong Kim and Professor Jungbin Hwang, for providing constructive feedback throughout the whole process of this dissertation. Their excellent lectures helped me build a firm understanding of the econometrics foundations as well as many cutting-edge technologies, which is valuable for my current and future research. I'm also very grateful to all the professors who taught me at the University of Connecticut and my colleagues in the Department of Economics over the past five years.

Finally, I want to thank my mother for her unselfish love throughout my life and giving me the strength to chase my dreams.

Contents

Chapter 1

Is the recursive preference asset pricing model more flexible? A Monte Carlo Study

1.1 Literature review

Research on recursive preferences asset pricing model has been going on for decades. In this literature, the model proposed by Epstein and Zin (1989) and Weil (1989) (EZW hereafter) is widely used because of its higher degree of flexibility with regards to attitudes towards to the CRRA and the EIS. Comparing to the model with time-separable utility function such as

$$U_t = E\left[\sum_{t=0}^{\infty} \beta^t \frac{C_t^{1-\gamma}}{1-\gamma}\middle| \mathcal{F}_t\right], \tag{1.1.1}$$

where γ is both the CRRA and the reciprocal of the EIS. The EZW model has a recursive preferences utility function defined as

$$V_t = \left[(1-\beta)C_t^{1-\rho} + \beta(E\left[V_{t+1}^{1-\theta}|\mathcal{F}_t\right])^{\frac{1-\rho}{1-\theta}}\right]^{\frac{1}{1-\rho}}, \tag{1.1.2}$$

where the CRRA is characterized by θ and agent's EIS is characterized by $\frac{1}{\rho}$. It's easy to show (1.1.1) is a special case of (1.1.2) if we let $\theta = \rho$.

Because of its higher generality, the EZW model is widely used in the empirical study as researchers are interested in estimating the CRRA and the EIS separately. The estimates of those parameters, however, vary widely, and there is still no one widely accepted estimate. Epstein and Zin (1991) attribute this issue to the model's sensitivity to the choice of consumption measure and instrumental variables. Kocherlakota (1990) provides a theoretical proof that, assuming an i.i.d. endowment process, (1.1.1) and (1.1.2) are "observationally equivalent". In other words, using assets returns and aggregate consumption growth data, the empirical investigator cannot disentangle the EIS from CRRA in the EZW setting. Smith (1999) studies the finite sample properties of tests of the EZW model and points out that those tests often have little power to reject the null hypothesis that $\theta = \rho$.

In this paper, we investigate why the estimates of the CRRA and the EIS vary significantly in the literature. We use a GMM type proxy-free method to estimate the parameters, and two sources of data: empirical data and simulated data to verify the results. In both cases, we find that the surface of the GMM objective function is not globally concave, which might lead the most optimization algorithms to fail. More importantly, most local minimum points fall onto the line of $\theta = \rho$. This is a direct evidence to support the point in Kocherlakota (1990) and Smith (1999). Secondly, it may cause infinite or undefined GMM moment and lead to lead to either the estimates do not move from the starting point or no estimates are available if we don't restrict the relationship between the CRRA and the EIS. Lastly, our result shows the choice of instruments is not critical to varying estimates. The rest of the

paper is organized as follows. Section 1.2 introduces the empirical data we use, and the data generation process of simulation. Section 1.3 describes the estimation method and reports the results. Section 1.4 concludes the paper.

1.2 Data

1.2.1 Empirical data

The empirical data we use is from Chen et al. (2013) (CFL hereafter). We provide a brief introduction here, but refer readers to CFL for the detailed description. The aggregate data are quarterly returns and consumption growth from the first quarter of 1952 to the first quarter of 2015. The data of asset returns includes 3-month Treasury bill rate, and six value-weighted portfolios of common stock. For the instruments, we choose the same variables as CFL and keep the notation consistent. The relative bill rate and the excess return on the S&P 500 index are denoted as $RREL$ and $SPEX$, respectively. The consumption-wealth ratio is measured as the cointegrating residual between log consumption, log asset wealth, and log labor income, and is denoted as $\widehat{cay}$.

1.2.2 Simulated data

Time series of the state variables (λ_t, ξ_t): Our data generation process is based on Kocherlakota (1990) and Smith (1999). There are two state variables: annual aggregate consumption growth ($\lambda_t = c_t/c_{t-1}$), and annual aggregate dividend growth ($\xi_t = d_t/d_{t-1}$),

which are generated using the following VAR(2) model

$$\begin{aligned} \ln \lambda_t &= 0.021 + 0.017 \ln \xi_{t-1} - 0.161 \ln \lambda_{t-1} + \varepsilon_t^1, \\ \ln \xi_t &= 0.004 + 0.117 \ln \lambda_{t-1} + 0.414 \ln \xi_{t-1} + \varepsilon_t^2, \\ \Sigma(\varepsilon) &= \begin{bmatrix} 0.01400 & 0.00177 \\ 0.00177 & 0.00120 \end{bmatrix}. \end{aligned} \tag{1.2.1}$$

The error terms $\varepsilon = (\varepsilon^1, \varepsilon^2)$ are assumed to be jointly normally distributed with the covariance matrix $\Sigma(\varepsilon)$. The parameters are calibrated from the data of annual per-capita, real non-durable consumption growth and the annual S&P 500 aggregate, real dividend growth over the period 1888-1978.

Transition matrix from Tauchen's method: We use Tauchen's quadrature method in Tauchen and Hussey (1991) to discretize the continuous stochastic process of $\{\lambda_t, \xi_t : t = 1, \cdots, T\}$ to a finite Markov-chain with state values of $\{\lambda_i, \xi_j : i = 1, \cdots, S, j = 1, \cdots, S\}$ and a transition matrix $\Pi = \{\pi_{ij} = Pr(s_{t+1} = \lambda_j, \xi_j | s_t = \lambda_i, \xi_i) : i = 1, \cdots, S, j = 1, \cdots, S\}$, where S is the number of states.

Three simulated returns: Given the state value of (λ, ξ) and the $S \times S$ transition matrix Π, along with the parameter set (β, θ, ρ), we can then simulate the return on aggregate wealth (R_w), the return on risky asset (R_{sp}), and the risk-free rate (R_f) (which is defined as the return on a claim paying a unit of consumption for sure one period forward) as follows. Let us consider a simple Lucas-style asset pricing model case in which the return is the future price of asset plus the dividend income. Denote $p_{w,t}$, and $d_{w,t}$ to be the price and dividends of the aggregate wealth portfolio in period t, respectively. We assume the representative agent receives only dividends as income and consumes all dividends when received, i.e., $d_{w,t}/d_{w,t-1} = \lambda_t$, which is the same as in Smith (1999). Then the finite-states Markov

representation of the Euler equation takes the form of

$$\begin{aligned} p_{w,i} &= \sum_{j=1}^{S} \pi_{ij}[(p_{w,j} + d_{w,j})m_{ij}(c_i, c_j)], \\ m_{ij} &= \left(\beta \left(\frac{c_j}{c_i}\right)^{-\rho}\right)^{\frac{1-\theta}{1-\rho}} (\mathcal{R}_{w,ij})^{\frac{1-\theta}{1-\rho}-1}, \end{aligned} \tag{1.2.2}$$

where m_{ij} is the stochastic discount factor. Let $v_{w,t} = p_{w,t}/d_{w,t}$ be the price-dividend ratio of the aggregate wealth portfolio. With previous assumption that dividend growth is equal to the consumption growth, equation (1.2.2) can be written as

$$v_{w,i}^{(1-\theta)/(1-\rho)} = \sum_{j=1}^{S} \pi_{ij}[\beta^{(1-\theta)/(1-\rho)} \lambda_j^{1-\theta} (1 + v_{w,j})^{(1-\theta)/(1-\rho)}], \quad i = 1, 2, ..., S. \tag{1.2.3}$$

Equation (1.2.3) comprises of a system of S linear equations in $v_{w,i}$ which can be solve directly. Let v_w^* be the vector of solutions to (1.2.3) given (β, θ, ρ). The state values of returns can be represented as the function of v_w^*. The return on aggregate wealth over state i and j is given by

$$R_{w,ij} = \frac{1 + v_{w,j}^*}{v_{w,i}^*} \lambda_j. \tag{1.2.4}$$

The return on risk-free asset, which is defined as the price in state i of a bond paying sure unity next period is

$$\frac{1}{R_{fi}} = \beta^{(1-\theta)(1-\rho)} \sum_{j=1}^{S} \pi_{ij} \lambda_j^{-\theta} \left(\frac{1 + v_{w,j}^*}{v_{w,i}^*}\right)^{(\rho-\theta)(1-\rho)}. \tag{1.2.5}$$

For the return on risky asset, everything is the same as return on aggregate wealth except that consumption growth is replaced by the simulated S&P 500 dividend stream. The state

returns are given by

$$R_{sp,ij} = \frac{1 + v^*_{w,j}}{v^*_{w,i}} \xi_j. \tag{1.2.6}$$

Once we obtain the vector of state value of three returns, we then use the following algorithm to simulate the time-series of each return:

(1) Draw a random variable u_0 from a uniform distribution $U(0,1)$. Let the initial state, n_0, be the smallest number such that $p(1) + p(2) + \cdots + p(n_0) \geq u_0$, where $p(i)$ is the stationary probability of being in state i, $1 \leq n_0 \leq S$.

(2) Let n' be the current state and n'' be the next state to be drawn. Draw u'' from $U(0,1)$ and let the n'' be the smallest number such that $\sum_i^{n'} \pi(n', i) \geq u''$, where $\pi(i,j)$ is the transition probability from state i to state j.

(3) Set $n' = n''$ and then return to step 2 until $t = T$. The time series of three returns are obtained by choosing $R_w(n', n'')$, $R_f(n', n'')$, and $R_{sp}(n', n'')$.

In the literature, the constant discount factor β is not very controversial. In order to reduce the computation burden, we fix it at 0.99, which is a standard value in the literature. The estimates of CRRA (θ) and EIS ($\frac{1}{\rho}$) vary widely. Nestor and Ruben (2015) show that the most widely accepted measures of θ lie between 1 and 3, and ρ is widely agreed to be less than 1. Following Smith (1999), we set the true values of the parameters in the simulation to be $(\theta, \rho) \in \{(0.8, 0.8), (0.8, 1.3), (1.3, 1.3), (1.3, 5.2)\}$. The mean and standard deviation of the simulated state variables and returns are reported in Table (1.2.1). Across all combinations of parameters, the return on the aggregate wealth portfolio (R_w) and the return on risky asset (R_{sp}) have higher return and risk compared to the risk-free asset, but the equity premium ($R_{sp} - R_f$) doesn't vary much. This result provides an evidence for Kocherlakota (1990) arguing that separating the parameters of CRRA and EIS does not solve the equity premium puzzle.

Table 1.2.1: Summary information on state variables and simulated returns from Monte Carlo simulations using the parameter values from Epstein and Zin (1991).

θ, ρ		λ_t	ξ_t	R_w	R_f	R_{sp}	Equity premium
0.80, 0.80	Estimator	1.026	1.007	1.023	1.021	1.031	0.010
	Std.dev.	0.108	0.107	0.128	0.015	0.138	
0.80, 5.20	Estimator	1.030	1.042	1.065	1.028	1.029	0.000
	Std.dev.	0.111	0.142	0.133	0.027	0.133	
1.35, 1.35	Estimator	1.000	1.024	1.042	1.020	1.039	0.019
	Std.dev.	0.094	0.129	0.128	0.025	0.134	
1.35, 5.20	Estimator	1.036	0.988	1.111	1.086	1.116	0.030
	Std.dev.	0.125	0.121	0.188	0.101	0.150	

1.3 Estimation and results

The estimation method: We use a two-steps semi-parametric estimation approach developed by CFL to estimate (β, θ, ρ). This approach directly estimates the unobservable value function without requiring the proxy for $\mathcal{R}_w$. We provide a brief introduction to CFL's method here. For the large sample properties and proof, see Ai and Chen (2003). Recall that the EZW utility function is defined recursively by

$$\begin{aligned} V_t &= \left[(1-\beta)C_t^{1-\rho} + \beta \mathcal{R}_t (V_{t+1})^{1-\rho}\right]^{\frac{1}{1-\rho}}, \\ \mathcal{R}_t\left(V_{t+1}\right) &\equiv \left(E\left[(V_{t+1})^{1-\theta} | \mathcal{F}_t\right]\right)^{\frac{1}{1-\theta}}, \end{aligned} \tag{1.3.1}$$

where $\mathcal{R}_t(V_{t+1})$ is the risk adjustment to the date $t+1$ continuation value function. Dividing both sides by C_t, we have

$$\frac{V_t}{C_t} = \left[(1-\beta) + \beta \mathcal{R}_t \left(\frac{V_{t+1}}{C_{t+1}} \frac{C_{t+1}}{C_t}\right)^{1-\rho}\right]^{\frac{1}{1-\rho}}. \tag{1.3.2}$$

The stochastic discount factor (SDF) is

$$M_{t+1} = \beta \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \left(\frac{\frac{V_{t+1}}{C_{t+1}} \frac{C_{t+1}}{C_t}}{\mathcal{R}_t \left(\frac{V_{t+1}}{C_{t+1}} \frac{C_{t+1}}{C_t} \right)} \right)^{\rho-\theta}. \tag{1.3.3}$$

Rearranging (1.3.2) and plugging it into (1.3.3), the SDF becomes

$$M_{t+1} = \beta \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \left(\frac{\frac{V_{t+1}}{C_{t+1}} \frac{C_{t+1}}{C_t}}{\left(\frac{1}{\beta} \left[\left(\frac{V_t}{C_t} \right)^{1-\rho} - (1-\beta) \right] \right)^{\frac{1}{1-\rho}}} \right)^{\rho-\theta}. \tag{1.3.4}$$

The only latent variable in (1.3.4) is the continuation value function-to-consumption ratio, $\frac{V_t}{C_t}$. CFL estimate $\frac{V_t}{C_t}$ by a sequence of flexible parametric functions, with the number of parameters expanding as the sample size grows. To make it simple, the $\frac{V_t}{C_t}$ is approximated by a linear combination of basis functions, such as

$$\frac{V_t}{C_t} \approx F_t(\cdot, \delta) = a_0(\delta) + \sum_{j=1}^{K_T} a_j(\delta) B_j \left(\frac{V_{t-1}}{C_{t-1}}, \frac{C_t}{C_{t-1}} \right), \tag{1.3.5}$$

where the coefficients $\{a_0, a_1, ..., a_{K_T}\}$ depend on $\delta = (\beta, \theta, \rho)$. The basis functions $\{B_j(\cdot, \cdot) : j = 1, ...K_T\}$ with the dimensionality of K_T have known functional forms, and are independent of δ. We use the following B-splines with degree $m = 3$ as our basis function:

$$B_m(y) = \frac{1}{(m-1)!} \sum_{k=0}^{m} (-1)^k \begin{pmatrix} m \\ k \end{pmatrix} [\max(0, y-k)]^{m-1}. \tag{1.3.6}$$

From now on, the latent value function is completely characterized by the parameters $\{\frac{V_0}{C_0}, a_0, a_1, ..., a_{K_T}\}$.

The estimation is based on the following two-steps GMM method. Given the initial value

of $\delta = (\beta, \theta, \rho)$, we first estimate $\hat{F}_t(\cdot, \delta)$ by minimizing

$$\hat{F}_t(\cdot, \delta) = \underset{F_{K_T} \in \mathcal{V}_{\mathcal{T}}}{\operatorname{argmin}} \left[g_t(\delta, F_t(\delta, \cdot)) \right]' W_t \left[g_t(\delta, F_t(\delta, \cdot)) \right], \tag{1.3.7}$$

where g_t is a $N \times 1$ vector of $g_{i,t}$ defined by

$$g_{i,t}(\delta, F_t) = \frac{1}{T} \sum_{t=1}^{T} \left[\beta \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \left(\frac{F_{t+1} \cdot \frac{C_{t+1}}{C_t}}{\left(\frac{1}{\beta} \left[F_t^{1-\rho} - (1-\beta) \right] \right)^{\frac{1}{1-\rho}}} \right)^{\rho-\theta} R_{i,t+1} - 1 \right] \otimes x_t, \tag{1.3.8}$$

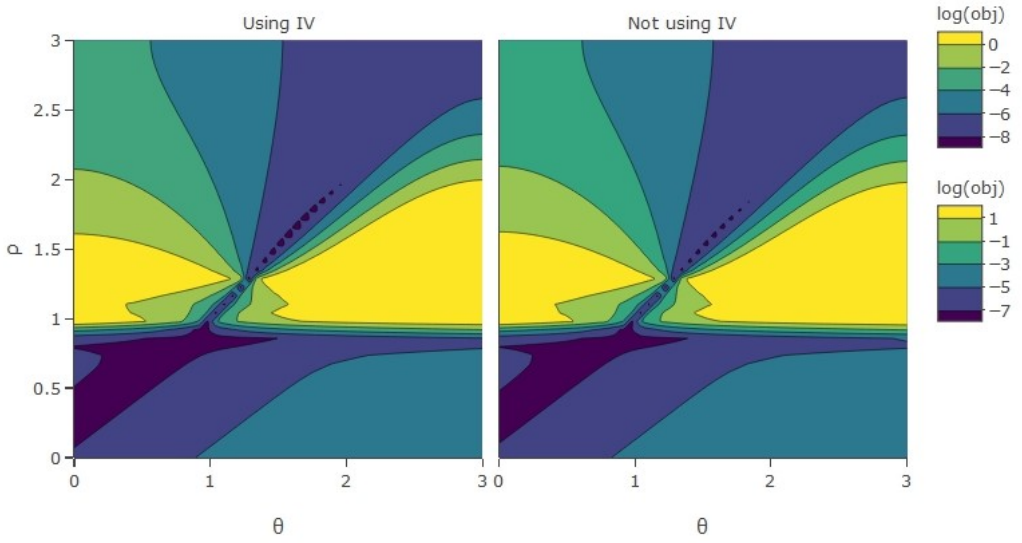
where $\otimes$ is the kronecker product, and x_t is the state variable that captures the information in period t. For the estimation using simulated data, we choose the lag of consumption growth and dividend growth as x_t as they are the "true value" in our simulation. For the empirical data in CFL, we provide the result using the same instruments as in CFL as well as without any instrument. The details will be discussed later in this section. With the optimal $\hat{F}_t$ in hand, the estimators of δ are defined as

$$\hat{\delta} = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}} \left[g_t \left(\delta, \hat{F}_t(\delta) \right) \right]' W_t \left[g_t \left(\delta, \hat{F}_t(\delta) \right) \right], \tag{1.3.9}$$

where $\mathcal{D}$ is the parameter set. The two steps are repeated until the convergence condition is satisfied.

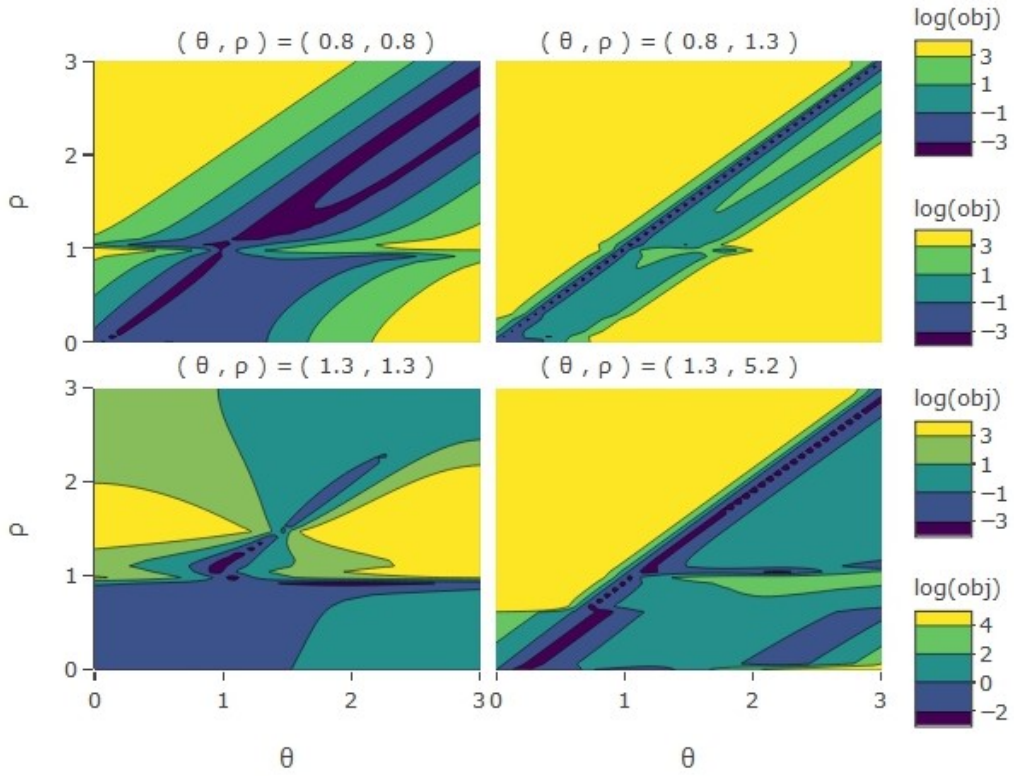
Estimation using the empirical data. Instead of providing a point estimator of (θ, ρ) in the EZW model, we believe showing the surface of the GMM objective function can give readers a better idea why the estimation of these parameters is controversial in the literature. In Figure 1.3.1, the left panel is the surface of GMM object using instruments. We choose $x_t = \left(c\hat{a}y_t, RREL_t, SPEX_t, \frac{C_t}{C_{t-1}} \right)$, which is consistent with CFL.

Figure 1.3.1: Surface of GMM objective function using data of CFL



In the right panel, we let x_t be identity matrix. As shown, the state variables x_t does not affect the surface too much. In both panels, the surface is not globally concave, and this may cause most of the optimization algorithm to fail. A common solution to this issue is to start with many different initial values. However, this may cause the estimators to heavily depend on the choice of the starting points, especially in the nonlinear case. We conjecture that it is the main reason why the estimation of CRRA and EIS vary widely in the literature. Another important finding is that most local minimum of GMM objective function fall onto the line of $\theta = \rho$, in which case the recursive preference model degenerates to the time-separable model. It is a direct numerical evidence to support the point in Kocherlakota (1990) that though the EZW model is a more general setting than the time-separable model, it has no more explanatory power than the latter one. Last but not least, some combinations of the parameters may lead to the ill-behaved objective function or generate "infinite moment" as it is pointed out in Tauchen (1986). This partially explains why the GMM method is frequently "stuck" during the estimation procedure.

Figure 1.3.2: Surface of GMM objective function using simulated data



Estimation using the simulated data. To verify our results, we run a Monte Carlo simulation with true values of $(\theta, \rho) \in \{(0.8, 0.8), (0.8, 1.3), (1.3, 1.3), (1.3, 5.2)\}$. Using the same GMM estimation method and letting the state variable be $x_t = (\lambda_{t-1}, \xi_{t-1})$, we report the surface of GMM objective function in Figure 1.3.2. Firstly, similar results as in the case using CFL's data show up across all combinations of parameters. In each panel of figure 1.3.2, most of the local minimum points fall onto the line of $\rho = \theta$ even if the true value of ρ, θ are different. This confirms that the empirical investigators with data on asset prices and aggregate consumption cannot disentangle CRRA form EIS. Secondly, for example in the case of $(\theta, \rho) = (0.8, 1.3)$, the frequency of "infinite moment" is very high, which may cause either the GMM method get "stuck" at most initial points, or the estimator does not move from where it starts. Both are commonly encountered problems in the literature.

1.4 Conclusion

This paper is an attempt to numerically investigate "identifiability" of the CRRA and EIS in the EZW model. We find that most of the local minima fall onto the line of $\theta = \rho$ implying that optimization routines cannot disentangle θ from ρ. This result implies that asset pricing model with recursive preference does not have more explanatory power than the model with time-separable utility. Secondly, we find that the GMM moment condition in the case of the EZW model is likely to be "infinite" or undefined unless we impose a substantial restriction on the relation between the CRRA and the EIS. These numerical issues might cause either GMM method to fail or the estimates to remain unchanged from the initial values. Lastly, our first finding is robust to the choice of instrumental variables.

Chapter 2

Mahalanobis Metric Based Clustering for Fixed Effects Model

2.1 Introduction

Panel data is widely used in econometrics. A crucial advantage of using panel data in regression analysis is that researchers can address the endogeneity problem caused by unobserved heterogeneity by using the fact that individuals are observed over time. In this regard, researchers usually introduce individual specific fixed effects and time effects in regression models. This approach is arguably restrictive in that it is valid under the assumption that individual heterogeneity is time invariant and the effect of time effects is the same across individuals, but this may not hold in empirical applications.

We consider a regression model that allows for group specific unobserved heterogeneity that is time varying. Let

$$y_{it} = x_{it}'\beta + \alpha_{g_i t} + v_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T, \tag{2.1.1}$$

where β is a $(k \times 1)$ vector of coefficients, $g_i \in \{1, \ldots, G\}$ is the group membership for

individual i, and α_{gt} is the group specific time effects for group g at time period t. In contrast to standard fixed effects models, we consider the case that both g_i and α_{gt} are unknown to econometricians and need to be estimated. The covariates x_{it} are assumed to be contemporaneously uncorrelated with v_{it}, but may be arbitrarily correlated with α_{g_it}. The number of groups G is small, and we assume it is known throughout this paper. This model is first considered by Bonhomme and Manresa (2015) (BM hereafter) and they propose the grouped fixed effects estimator based on an optimal grouping of individuals using the k-means clustering method. Introducing k-means to panel data models provides researchers an alternative to prior information (e.g., country and county) for studying the potential group pattern. Adding clustering to panel models also has the benefit of addressing incidental parameters problem, see, e.g., Neyman and Scott (1948) and Bester and Hansen (2016).

This paper extends BM's approach in the following ways. First, we use the Mahalanobis metric (Mahalanobis (1936)) as a dissimilarity measure, which could improve clustering if the panel data embed serial correlation and heteroscedasticity when clusters have a similar shape, size, and orientation. The Euclidean metric that the k-means is based on, by contrast, does not consider the correlation between variables which leads it very sensitive to the noise. Figure 1 illustrates this issue. We generate serially correlated data based on the DGP in Section 2.4. The sample size is $(N, T) = (90, 5)$ and each individual is drawn from 3 groups. The left panels of this figure show the estimated grouping using the Euclidean distance, and the right panel uses our KMM algorithm. As shown, the Euclidean metric fails to recover the correct grouping for six individuals even though the clusters are well separated. In contrast, the Mahalanobis metric based clustering estimates all the group memberships correctly in this case. The latter tends to outperform the former since it takes into account the serial correlation of data and is thus unit-free and scale-invariant. Graphically, it first converts the shape of data from ellipse to round and then applies the k-means algorithm to make it easier to assign each individual to a right cluster.

Second, we derive the optimal β to estimate group membership and show that it may be different from the true coefficient parameter. Clustering performs better when groups are well separated. We first define a signal to noise ratio that captures the degree of separation, and find that this ratio may not be maximized at the true β. In practice, this optimal β is not feasible, because it contains unknown parameter values that are very hard to estimate. To address this issue, we propose a data driven algorithm to have a feasible β that amplifies the group signal and hence improves group membership estimation.

The rest of the paper is organized as follows. In Section 2.2, we introduce the proposed estimator and discuss its properties. Section 2.3 derives the optimal β for group membership estimation and proposes its data driven selection procedure. Section 2.4 provides the simulation results, and Section 2.5 concludes the paper and discuss the potential extension.

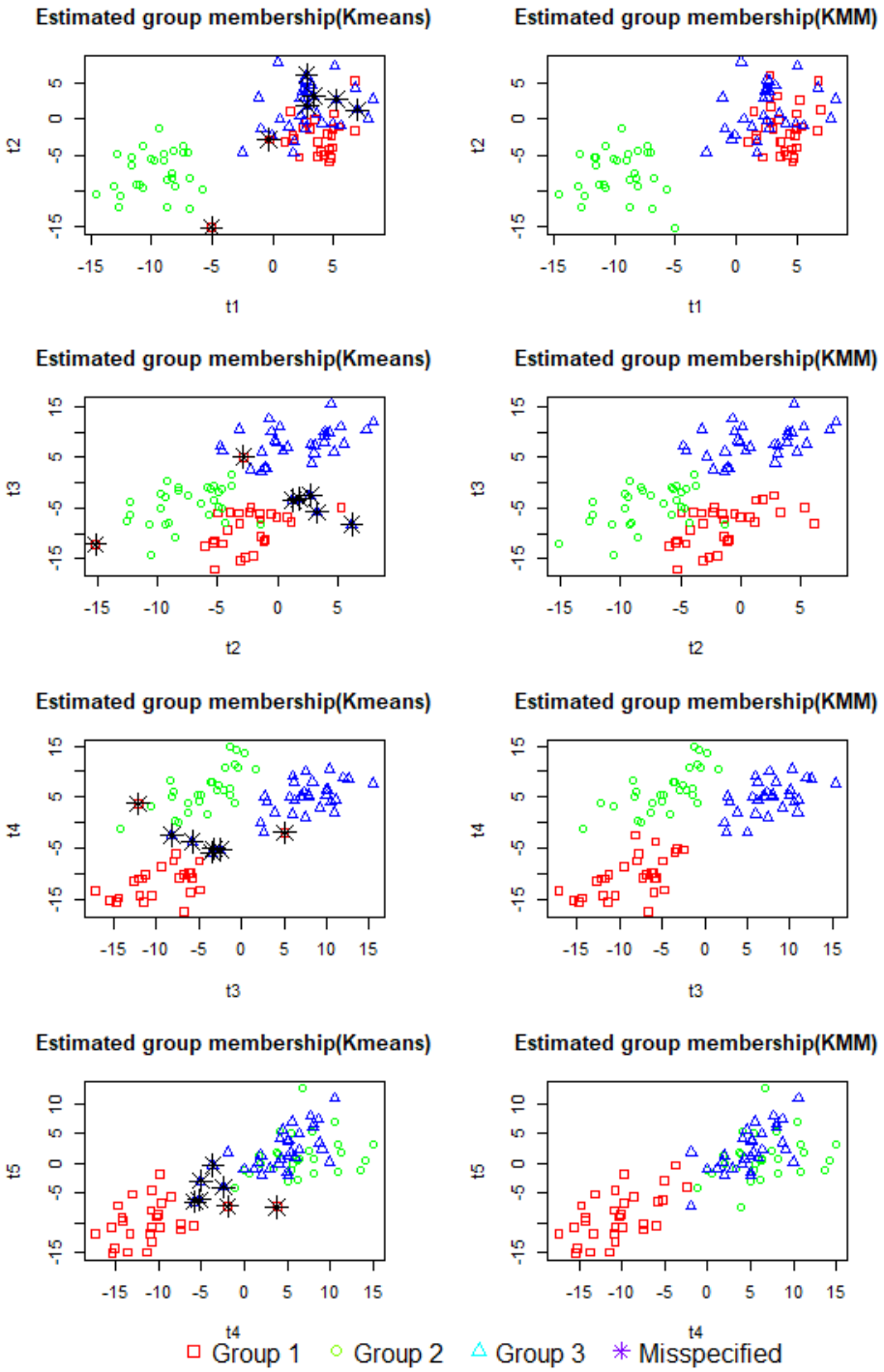
2.2 Mahalanobis metric based grouped fixed effects estimator

We consider a Mahalonobis metric based criterion function to estimate unknown parameters in (2.1.1). Let $Y_i = (y_{i1}, \ldots, y_{iT})'$ and $X_i = (x_{i1}, \ldots, x_{iT})'$. We also denote the parameter spaces for β and α_{gt} as $\Theta_\beta \subset \mathbb{R}^k$ and $A \subset \mathbb{R}$ respectively. Our estimator is defined as

$$(\hat{\beta}, \hat{\gamma}, \hat{\alpha}) = \underset{(\beta,\alpha,\gamma)\in\Theta_\beta\times A^{GT}\times\Gamma_G}{\operatorname{argmin}} \frac{1}{NT}\sum_{g=1}^{G}\sum_{i\in\{g_i=g\}}(Y_i - X_i\beta - \alpha_{g_i})'\widehat{W}^{-1}(Y_i - X_i\beta - \alpha_{g_i}), \quad (2.2.1)$$

where the minimization of the criterion function is taken over the common parameter β, time varying grouped fixed effects $\alpha = \{\alpha_{gt};\ g = 1, ..., G \text{ and } t = 1, ..., T\}$, and group membership $\gamma = \{g_1, \ldots, g_N\} \in \Gamma_G$. $\widehat{W}$ is the $(T \times T)$ sample covariance matrix of the latent component $\{Y_i - X_i\beta;\ i = 1, ..., N\}$ of the regression. Since the criterion function is rescaled by the covariance matrix, the serial correlation and heteroscedasticity of the dataset are

Figure 2.1.1: Clustering using Euclidean (left) and Mahalanobis (right) distance



taken into account. It is obvious that when $\widehat{W}$ is an identity matrix our estimator in (2.2.1) reduces to BM's grouped fixed effects estimator which employs the Euclidean metric based criterion function.

We can obtain the proposed estimator in (2.2.1) based on the following two step procedure. First, given some values of (β, α), we estimate group membership, $\hat{\gamma} = (\hat{g}_1, \ldots, \hat{g}_N)$, by applying k-means to the rescaled residuals. That is,

$$(\hat{\gamma}|\beta, \alpha) = \underset{\gamma \in \Gamma_G}{\operatorname{argmin}} \frac{1}{NT} \sum_{g=1}^{G} \sum_{i \in \{g_i = g\}} (Y_i - X_i\beta - \alpha_{g_i})' \widehat{W}^{-1} (Y_i - X_i\beta - \alpha_{g_i}), \tag{2.2.2}$$

where

$$\widehat{W} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - X_i\beta)(Y_i - X_i\beta)'.$$

In practice, we can reduce the computational cost to implement (2.2.2) by using Cholesky decomposition. Given β, we let

$$\widehat{W}^{-1} = \widehat{L}(\beta)\widehat{L}(\beta)',$$

and define $\tilde{Y} = \widehat{L}(\beta)Y$ and $\tilde{X} = \widehat{L}(\beta)X$ given β. We can then obtain $\hat{\gamma}$ from

$$(\hat{\gamma}|\beta, \alpha) = \underset{\gamma \in \Gamma_G}{\operatorname{argmin}} \frac{1}{NT} \sum_{g=1}^{G} \sum_{i \in \{g_i = g\}} (\tilde{Y}_i - \tilde{X}_i\beta - \tilde{\alpha}_{g_i})'(\tilde{Y}_i - \tilde{X}_i\beta - \tilde{\alpha}_{g_i}). \tag{2.2.3}$$

Second, with the estimated group membership, the estimator of (β, α) is given by

$$\left(\hat{\beta}, \hat{\alpha} \middle| \hat{\gamma}\right) = \underset{(\beta,\alpha) \in \Theta_\beta \times A^{GT}}{\operatorname{argmin}} \frac{1}{NT} \sum_{g=1}^{G} \sum_{i \in \{g_i = g\}} (\tilde{Y}_i - \tilde{X}_i\beta - \tilde{\alpha}_{\hat{g}_i})'(\tilde{Y}_i - \tilde{X}_i\beta - \tilde{\alpha}_{\hat{g}_i}).$$

Let's sketch the asymptotics of the proposed estimator without going into the specific

details. We do this by considering the infeasible version $(\overline{\beta}, \overline{\alpha}, \overline{\gamma})$ which is based on the true covariance matrix. Let

$$\left(\overline{\beta}, \overline{\alpha}, \overline{\gamma}\right) = \underset{\beta,\alpha,\gamma}{\operatorname{argmin}} \frac{1}{NT} \sum_{g=1}^{G} \sum_{i \in \{g_i = g\}} \left(Y_i - X_i\beta - \alpha_{gi}\right)' W^{-1} \left(Y_i - X_i\beta - \alpha_{g_i}\right), \tag{2.2.4}$$

where

$$W = E\left[(Y_i - X_i\beta^0)(Y_i - X_i\beta^0)'\right]$$

and β^0 is the true coefficient parameter. The asymptotics of $(\overline{\beta}, \overline{\alpha}, \overline{\gamma})$ can be shown easily by following BM if we impose conditions on W or simply make its "high level" assumptions. That is, if we modify Assumptions 1-3 in BM using $\ddot{Y}_i = LY_i$ and $\ddot{X}_i = LX_i$, where L is by Cholesky decomposition of $W(\beta^0)^{-1}$, i.e., $W(\beta^0)^{-1} = LL'$, then every step in the proof for the asymptotics of BM's estimator will go through for $(\overline{\beta}, \overline{\alpha}, \overline{\gamma})$. The asymptotics of $(\hat{\beta}, \hat{\alpha}, \hat{\gamma})$ is more complicated since it depends on the asymptotic behavior of $\widehat{W}(\hat{\beta})$. We may need consistency of $\widehat{W}(\hat{\beta})$ for W, which can be defined by

$$\left\| \widehat{W}(\hat{\beta})^{-1} W - I_T \right\|_{\max} \xrightarrow{p} 0$$

where $\|B\|_{\max} = \max_{i,j} |b_{ij}|$ for a matrix B whose (i,j)-th element is b_{ij} or other matrix norms $\|\cdot\|$. This is not straightforward because $\widehat{W}(\hat{\beta})$ is a $T \times T$ matrix where $T \to \infty$. In the literature, Fan et al. (2013) study the estimation of a high dimensional covariance matrix in the approximate factor model context, and we may consider an extension of their approach to our model. However, It is beyond the scope of our paper, and we leave it for our future research.

2.3 Implementation

2.3.1 Optimal β for group membership estimation

In this section, we provide the optimal β for group membership estimation by amplifying the group signal in (2.2.3). Let's consider the following simple regression model

$$\begin{aligned} y_{it} &= \beta^0 x_{it} + \alpha^0_{g_i^0 t} + v_{it}, \\ x_{it} &= \delta^0 \alpha^0_{g_i^0 t} + e_{it}, \end{aligned} \tag{2.3.1}$$

where x_{it} is a scalar regressor, $(\beta^0, \alpha^0_{g_i^0 t})$ are the parameter values, and δ^0 is a nonzero coefficient that represents the degree of association between x_{it} and $\alpha^0_{g_i^0 t}$. We assume $\alpha^0_{g_i^0 t}$, v_{it}, and e_{it} are independent of each other and their variances are σ^2_α, σ^2_v, and σ^2_e respectively. The group membership is estimated by applying k-means to the unobserved part $R_{it} = y_{it} - \beta x_{it}$ with some given value of β. After simple calculations, we have

$$R_{it} = (1 + (\beta^0 - \beta)\delta^0)\alpha^0_{g_i^0 t} + (\beta^0 - \beta)^2 e_{it} + v_{it}. \tag{2.3.2}$$

It is straightforward from (2.3.2) that the group information is from the variation of $\alpha^0_{g_i^0 t}$. However, this information might be negligible if we have $1 + (\beta^0 - \beta)\delta^0$ close to 0, i.e., if β is close to $\beta^0 - 1/\delta^0$. In this case, the weak group information causes the estimated grouping $\hat{g}_i$ to be inaccurate and hence leads to erroneous $\hat{\beta}$. In fact, we find that the optimal choice of β for grouping may be different from the true value β^0. Let's refer to the variance of $(1 + (\beta^0 - \beta)\gamma^0)\alpha_g$ as the group signal, and the variance of $(\beta^0 - \beta)e_{it} + v_{it}$ which is the individual level information as the noise for grouping. We then define SNR which is the ratio of signal (group level information) to noise (individual level information) as

$$SNR = \frac{(1 + (\beta^0 - \beta)\delta^0)^2 \sigma^2_\alpha}{(\beta^0 - \beta)^2 \sigma^2_e + \sigma^2_v}. \tag{2.3.3}$$

Taking the FOC with respective to β, we can show that the SNR is maximized at

$$\beta^{opt} = \beta^0 - \frac{\delta^0 \sigma_v^2}{\sigma_e^2}. \tag{2.3.4}$$

(2.3.4) indicates that the true value β^0 is not optimal for estimating group membership in this setting. We report the relationship between SNR and β in panel (D) of Figure 2.3.1. As shown, β^{opt} defined in (2.3.4) maximizes the value of SNR while β^0 provides only about a half of the SNR comparing to β^{opt}. Graphically, we can see that the residuals with β^{opt} in panel (A) are more separated than those with β^0 in panel (C). This implies that the former leads to more accurate group membership estimation when we use the k-means algorithm.

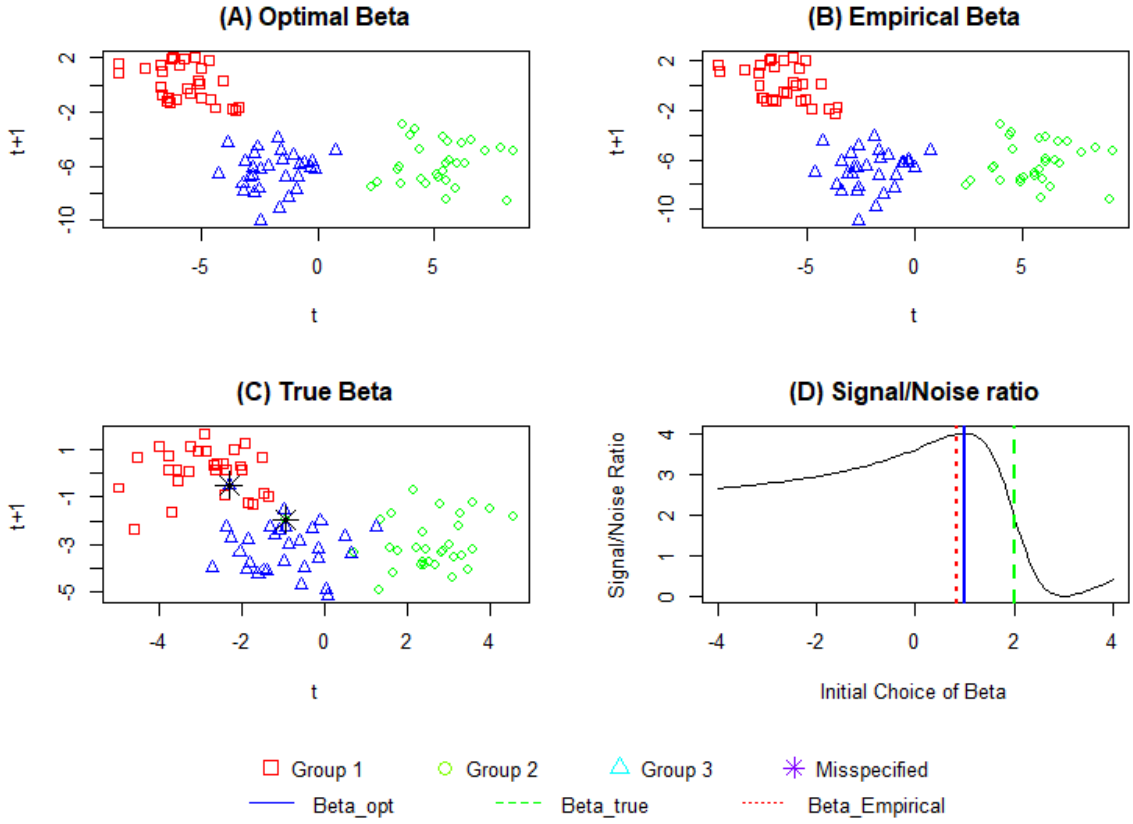
Our finding provides insight on the selection of β for group estimation. For this problem, BM suggest researchers draw many random starting value $(\beta^{(0)}, \alpha^{(0)})$ and select the one that yields the lowest value of the criterion function. This method can be understood as iteratively alternating the "assignment" step and the "update" step until convergence. In the "assignment" step, each individual i is assigned to the group g_i whose group fixed effects $\{\alpha_{gt}, t = 1, ..., T\}$ are closest to the residuals $\{y_{it} - x_{it}'\beta, t = 1, ..., T\}$. In the "update" step, (β, α) are estimated using the LS with previously determined grouping. However, our finding shows that the optimal β in the assignment step is different from the true β, which implies that we may not need BM's iteration procedure.

(2.3.4) provides the optimal β that minimizes the SNR in BM's context, but we can easily extend this idea to our Mahalanobis metric based group membership estimation. Let $\ddot{x}_{it}, \ddot{y}_{it}$ and $\ddot{\alpha}_{g_it}$ denote the t-th elements of $\ddot{X}_i, \ddot{Y}_i$ and $\ddot{\alpha}_{g_i} = L\alpha_{g_i}$ respectively, and we modify the model in (3.1) based on them. Then, it is straightforward to define the Mahaloanobis metric based optimal β as

$$\beta_M^{opt} = \beta^0 - \frac{\ddot{\delta}^0 \ddot{\sigma}_v^2}{\ddot{\sigma}_e^2}. \tag{2.3.5}$$

where $\ddot{\delta}^0, \ddot{\sigma}_v^2$ and $\ddot{\sigma}_e^2$ are the modified version of the original parameters in (2.3.4).

Figure 2.3.1: Clustering with Different Choice of β



2.3.2 Selection of β based on clustering validation and construction of $\widehat{W}(\hat{\beta})$

While (2.3.3) and (2.3.4) show that the optimal β for group membership estimation, β^{opt}, is different from β^0, β^{opt} is practically not feasible to implement, since it contains unknown parameters which are very hard to estimate. To address this practical issue, we propose a data driven method to find a proxy of β^{opt}, denoted as β^e, that amplifies the group signal and hence improves the performance of k-means. The method employs the internal validation by Brock et al. (2008), which takes only the data set and the clustering as input. We select β^e

by evaluating the internal measures such as connectivity, Dunn index, and silhouette width, which reflect connectedness, compactness, and separation of cluster partitions. We briefly introduce the internal measures we use in the algorithm. For more details, readers can refer to Handl et al. (2005).

The mathematical expression of connectivity is given by

$$Conn(C_1, \ldots, C_G) = \sum_{i=1}^{N} \sum_{j=1}^{K} knn(i, j), \tag{2.3.6}$$

where j represents the j-th nearest neighbor of the observation i. $knn(i, j)$ is zero if i and j are in the same cluster, and $1/j$ otherwise. $C_1, \ldots, C_G$ are G disjoint clusters, and K is the parameter that determines the number of neighbors to compute connectivity. Connectivity attempts to assess how well a given clustering agrees with the connectedness. It is non-negative and a smaller value is preferred in clustering analysis.

The Dunn index (Dunn† (1974)) is another well-known technique that assesses both compactness and separation and computes a final score as the nonlinear combination of these two measures. It is defined as

$$Dunn(C_1, \ldots, C_G) = \frac{\min_{C_k \neq C_l} \min_{i \in C_k, j \in C_j} dist(i, j)}{\max_{C_m \in (C_1, \ldots, C_G)} diam(C_m)}, \tag{2.3.7}$$

which is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The denominator $diam(C_m)$ denotes the maximum distance between observations in cluster C_m. The Dunn index is non-negative with larger values corresponding to highly consistent clustering results.

The last cluster validation criterion is the Silhouettes width by Rousseeuw (1987). It takes both compactness and separation into account, with well-clustered observations having values near 1 and poorly clustered observations having values near -1. For an observation

i, the Silhouettes width is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{2.3.8}$$

where $a(i)$ is the average dissimilarity of i to the rest of observations in the same cluster, and $b(i)$ is the average dissimilarity of i to observations in the nearest neighboring cluster.

From (2.3.2) and (2.3.3), we can consider β as a parameter that affects both intra-cluster homogeneity and inter-cluster separation of R_{it}. Therefore, we may try different values of β and choose the one that optimizes the clustering validation criterion. The procedure to implement β^e is summarized in Algorithm 1. Note that β^e in this algorithm is not for estimation of β^0, but an amplifier to improve the clustering step. Once β^e is obtained, we can use it to estimate group membership and then (β^0, α^0_{gt}).

Algorithm 1

---

procedure CHOOSING THE EMPIRICAL β^e
  Conn = 1e10, Dunn = 0 , Sihouette = 0
  for $\beta_{temp} \in (-\inf, \inf)$ **do**
    Partition $R(\beta_{temp}) = Y - X \cdot \beta_{temp}$
    Compute Conn, Dunn index, and Sihouette given $R(\beta_{temp})$ and clustering
    if Conn *not increase* or Dunn and Sihouette *not decrease* **then** Update $\beta^e = \beta_{temp}$

Given any candidate of β, computing the measures such as connectivity, Dunn index, and Silhouettes requires the partitioning from k-means. It's well known that finding the optimal solution to the k-means clustering problem is NP-hard. The running time of, for example, Lloyd's k-means algorithm is $O(ndki)$, where n is the number of d-dimensional vectors to be clustered; k is the number of clusters; i is the number of iterations needed until convergence. Hence, the entire running time of Algorithm 1 is $O(sndki)$, where s is the number of candidates of β. Thus, for the computational cost to implement β^e, choosing more β's will improve the performance of Algorithm 1 but will also increase the computation time.

We also discuss how to construct the sample covariance matrix $\widehat{W}(\hat{\beta})$ in our procedure. We consider an iteration procedure, because β in the $\widehat{W}$ is latent in the grouping step. Note that our iteration is different from BM's iteration. If we set $\widehat{W}(\hat{\beta})$ to be an identity matrix, then our estimator reduces to the BM's grouped fixed effects estimator. In this case, iteration is not needed, because we can use β^e from Algorithm 1 to estimate grouping, and then estimate $(\hat{\beta}, \hat{\alpha}_{gt})$ given grouping. To implement $\widehat{W}(\hat{\beta})$ for Mahalanobis metric based fixed effects estimation, we provide the following algorithm to estimate $(\hat{\beta}, \hat{\alpha}_{gt})$ and update $\widehat{W}(\hat{\beta})$ until convergence.

Algorithm 2

procedure ESTIMATE β USING KMM
  $\widehat{\beta}^{(0)} = \beta_e^{(0)}$ (from Algorithm 1 given Y, X)
  for $s = 0$ to itermax **do**
    Compute $R^{(s)} = Y - X\widehat{\beta}^{(s)}$
    Compute the sample covariance of $\widehat{W}\left(\widehat{\beta}^{(s)}\right)$
    Cholesky decomposition of $\widehat{W}\left(\widehat{\beta}^{(s)}\right)^{-1} = \widehat{L}\widehat{L}'$
    Rescale data by $\tilde{Y} = \widehat{L}Y$ and $\tilde{X} = \widehat{L}X$
    Update $\beta_e^{(s+1)}$ by Algorithm 1 given $\tilde{Y}, \tilde{X}$
    Partitioning $\tilde{R}^{(s)} = \tilde{Y} - \tilde{X}\beta_e^{(s+1)}$
    Update $\widehat{\beta}^{(s+1)}$ by OLS regression given cluster partition
    if $|\widehat{\beta}^{(s+1)} - \widehat{\beta}^{(s)}| < \epsilon$ **then** Break loop

---

2.4 Monte Carlo simulation

In this section, we examine finite sample properties of our Mahalanobis metric based clustering method and compare with BM's grouped fixed effects method. The following DGP is

employed for simulation.

$$
\begin{aligned}
y_{it} &= x_{it}\beta^0 + \alpha^0_{g_it} + v_{it}, \quad g_i = 1, \ldots, G, i = 1, \ldots, N, t = 1, \ldots, T, \\
\alpha^0_{g_it} &= \rho\alpha^0_{g_it-1} + e^{(1)}_{g_it}, \text{ with } \alpha^0_{g_i1}, e^{(1)}_{g_it} \overset{iid}{\sim} N(0,2), \\
v_{it} &= \theta v_{it-1} + e^{(2)}_{it}, \text{ with } v_{i1}, e^{(2)}_{it} \overset{iid}{\sim} N(0, Uniform(1,2)).
\end{aligned} \tag{2.4.1}
$$

We generate x_{it} by

$$
x_{it} = \delta^0\alpha^0_{g_it} + e^{(3)}_{it}, \text{ with } \delta^0 = 1, e^{(3)}_{it} \overset{iid}{\sim} N(0,1), \tag{2.4.2}
$$

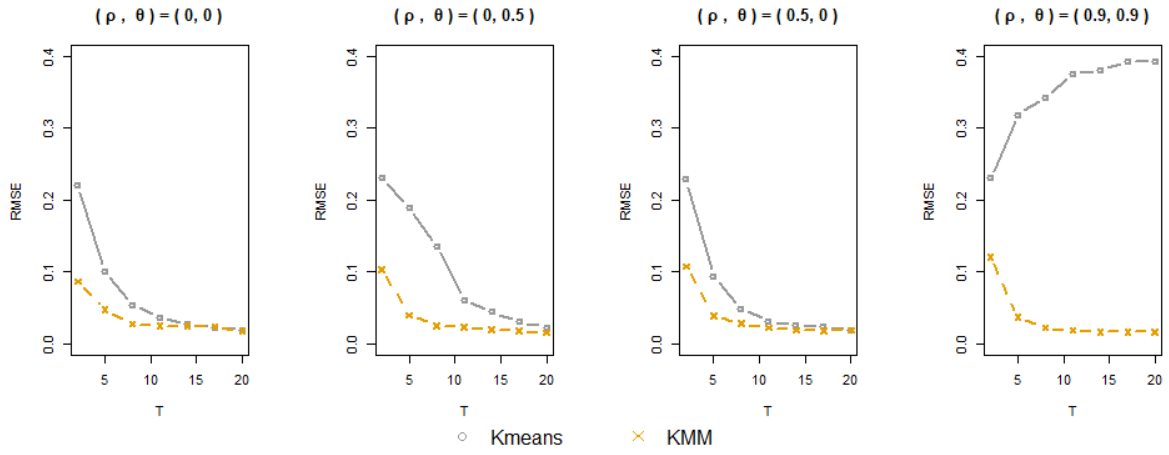
which implies the pooled OLS and standard fixed effects estimators are not valid because x_{it} is correlated with $\alpha^0_{g_it}$. The number of groups G is assumed to be known and each group has the same number of individuals. The number of replications is 1000.

Table 2.4.1: Simulation results for three estimators ($\beta^0 = 1, G = 3$)

			$\hat{\beta}_{Pool}$			$\hat{\beta}_{kmeans}$			$\hat{\beta}$		
N	T	(ρ, θ)	Bias	SD	ERP	Bias	SD	ERP	Bias	SD	ERP
90	5	(0 , 0)	0.7664	0.1046	1.0000	0.0941	0.0894	0.1680	0.0102	0.0756	0.0821
90	5	(0 , 0.5)	0.7681	0.0963	1.0000	0.1692	0.1182	0.4280	0.0078	0.0762	0.0859
90	5	(0.5 , 0)	0.7841	0.1033	1.0000	0.0846	0.0860	0.1350	0.0064	0.0730	0.0848
90	5	(0.9 , 0.9)	0.8395	0.1040	1.0000	0.2922	0.1132	0.5810	0.0123	0.0701	0.0835
90	10	(0 , 0)	0.7869	0.0626	1.0000	0.0230	0.0545	0.0320	0.0062	0.0454	0.0626
90	10	(0 , 0.5)	0.7871	0.0656	1.0000	0.0724	0.0933	0.1830	0.0106	0.0458	0.0959
90	10	(0.5 , 0)	0.8137	0.0642	1.0000	0.0206	0.0554	0.0190	0.0099	0.0474	0.0919
90	10	(0.9 , 0.9)	0.8869	0.0794	1.0000	0.3528	0.1078	0.7450	0.0141	0.0381	0.1089
300	5	(0 , 0)	0.7767	0.0937	1.0000	0.0833	0.0615	0.3390	0.0014	0.0433	0.0534
300	5	(0 , 0.5)	0.7733	0.0922	1.0000	0.1617	0.1003	0.6630	0.0006	0.0413	0.0594
300	5	(0.5 , 0)	0.7899	0.0981	1.0000	0.0759	0.0629	0.2840	0.0012	0.0473	0.0794
300	5	(0.9 , 0.9)	0.8403	0.0913	1.0000	0.2943	0.0973	0.8930	0.0028	0.0413	0.0718
300	10	(0 , 0)	0.7837	0.0604	1.0000	0.0210	0.0339	0.0590	0.0006	0.0255	0.0687
300	10	(0 , 0.5)	0.7850	0.0580	1.0000	0.0578	0.0663	0.2790	0.0034	0.0246	0.0682
300	10	(0.5 , 0)	0.8100	0.0615	1.0000	0.0169	0.0330	0.0390	0.0017	0.0235	0.0538
300	10	(0.9 , 0.9)	0.8891	0.0680	1.0000	0.3545	0.0903	0.9690	0.0039	0.0194	0.0749

Table 2.4.1 shows the bias, standard deviation (SD) and empirical rejection probability

Figure 2.4.1: RMSEs of $\hat{\beta}$ and $\hat{\beta}_{kmeans}$ with different T (N = 300)



(ERP) at the 5% level given different serial correlation scenarios. $\hat{\beta}_{Pool}$ and $\hat{\beta}_{kmeans}$ denote the pooled OLS estimator and BM's grouped fixed effects estimator respectively. We use the t-statistics and normal critical values to simulate the ERPs. For the t-statistics, we employ the clustered standard errors proposed by Arellano (1987) to allow for serial correlation. As we can see from the table, $\hat{\beta}_{Pool}$, which entirely ignores endogeneity caused by group heterogeneity, suffers from a substantial bias. The associated inference is not valid at all either as the ERPs are all 1. We can relieve this problems by introducing BM's estimator. However, the table indicates that $\hat{\beta}_{kmeans}$ is still severely biased and the inference tends to be misleading when serial correlation is high. We notice that the serial correlation of v_{it} has a much bigger impact on the accuracy of $\hat{\beta}_{kmeans}$ than the serial correlation of α_{gt}. Finally, the proposed estimator, $\hat{\beta}$, which is obtained by our Mahalanobis metric based clustering method, performs very well. It has the smallest bias, SD, and ERP under all the different correlation structures. We also find that our estimator becomes more accurate as N increases. This is well expected because $\hat{\beta}$ depends on the sample covariance $\widehat{W}$ which is estimated from the cross section dimension of panel data.

Figure 2.4.1 compares the root mean square errors (RMSEs) of $\hat{\beta}_{kmeans}$ and $\hat{\beta}$ with differ-

ent T. From the figure, we can see that $\hat{\beta}$ outperforms $\hat{\beta}_{kmeans}$ substantially when T is small, and this gap decreases as T increases when the serial correlation is absent or moderate. However, when $(\rho, \theta) = (0.9, 0.9)$, the RMSE of $\hat{\beta}_{kmeans}$ is shown to increase as T grows. When serial correlation is very high, the noise accumulates as T increases and finally dominants the group signal, so BM's Euclidean metric based group membership estimation works poorly with T large. The proposed procedure relieves this problem because the Mahalanobis metric takes serial correlation into account.

2.5 Conclusion

In this paper, we propose Mahalanobis metric based k-means algorithm to estimate group membership in linear panel data models with time varying group fixed effects. Our estimator improves upon Bonhomme and Manresa's (2015) grouped fixed effects estimator by accounting for serial correlation and heteroscedasticity. We derive the optimal β for group membership estimation by maximizing the signal to noise ratio and show that it may be different from the true β. Since the optimal that β is not feasible in practice, we propose the data driven selection method for its implementation.

Chapter 3

Estimation of latent group heterogeneities using a truncated singular value decomposition (TSVD) method

3.1 Introduction

Clustering in economics is an inevitable trend as there is ample evidence that individuals, firms, countries have group-level heterogeneities that are unobservable to econometricians. Many approaches have been developed in economics literature to estimate such group patterns together with the causal parameter. For example, Bonhomme and Manresa (2015) first introduce the kmeans method to determine the time-varying fixed effects in the panel data model. Su et al. (2016) develop a classifier-Lasso (C-Lasso) that shrinks individual coefficients to the unknown group-specific coefficients, which achieves classification and estimation simultaneously. In both methods mentioned above, the consistent estimator of

coefficient requires the estimator of grouping and vice versa. We propose a novel truncated singular value decomposition (TSVD) method, which can consistently estimate the causal parameter without knowing the group membership. All we need to know is the number of groups. In other words, our approach has an advantage that the estimation of coefficient and the estimation of grouping are two independent steps, which significantly improve the estimation accuracy. Also, our approach does not need an iteration procedure, and therefore remarkably reduce the computation cost. Lastly, our method applies to the linear panel data model with either time-varying group fixed effects in Bonhomme and Manresa (2015) or interactive fixed-effect model in Kim (2018). The estimation performance in both cases is presented.

The rest of the paper is organized as follows. In Section 3.2, we briefly introduce the two models in Bonhomme and Manresa (2015) and Kim (2018), and defined our estimator in each model. We also show that both estimators can be represented as a matrix factorization problem. Section 3.3 describes our TSVD approach. The simulation results are reported in Section 3.4.

3.2 Models

3.2.1 Panel data model with latent group heterogeneities

We consider the following panel data model that allows for group specific unobserved heterogeneity that is time varying:

$$Y_i = X_i\beta^0 + \alpha^0_{g_i} + \epsilon_i, \tag{3.2.1}$$

where β^0 is a $(k \times 1)$ vector of coefficients, $Y_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (x_{i1}, \ldots, x_{iT})'$, and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$. $g_i \in \Gamma^G = \{1, \ldots, G\}$ is the group membership for individual i, and

$\alpha^0_{g_i} = (\alpha^0_{g_{i1}}, \ldots, \alpha^0_{g_{iT}})'$ is the group specific time effects for group g if $g_i = g$. In matrix notation,

$$Y = X\beta^0 + \alpha^0 + \epsilon, \tag{3.2.2}$$

$Y = (Y_1, \ldots, Y_N)$ is $(T \times N)$ and X is a three-dimensional matrix with k sheets $(T \times N \times k)$, the lth sheet of which is associated with the lth element of β^0. The product $X\beta^0$ is $(T \times N)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ is $(T \times N)$. The group effects $\alpha^0 = (\alpha^0_{g_1}, \ldots, \alpha^0_{g_N})$ is also $(T \times N)$ with $\alpha^0_{g_i} = \alpha^0_{g_j}$, if $g_i = g_j, i, j \in (1, \ldots, N)$. we consider the case that both g_i and $\alpha^0_{g_i t}$ are unknown to econometricians and need to be estimated. The covariates X are assumed to be independent with ϵ, but may be arbitrarily correlated with α^0. The number of groups G is small, and we assume it is known throughout this paper.

We define our estimator as

$$(\widehat{\beta}, \widehat{\gamma}, \widehat{\alpha}) = \underset{(\beta,\alpha,\gamma)\in\mathbb{R}^k\times\mathbb{R}^{G\times T}\times\Gamma_G}{\operatorname{argmin}} \frac{1}{NT}\sum_{g=1}^{G}\sum_{i\in\{g_i=g\}} (Y_i - X_i\beta - \alpha_{g_i})'W^{-1}(Y_i - X_i\beta - \alpha_{g_i}), \tag{3.2.3}$$

where

$$W = \mathbb{E}\left[\left(Y_i - X_i\beta^0\right)\left(Y_i - X_i\beta^0\right)'\right]. \tag{3.2.4}$$

The minimization of the criterion function is taken over the common parameter β, time varying grouped fixed effects α and group membership $\gamma = \{g_1, \ldots, g_N\}$. W is the $(T \times T)$ population covariance matrix of the latent component $\{Y_i - X_i\beta^0;\ i = 1, ..., N\}$ of the regression. Since the criterion function is rescaled by the covariance matrix, the serial correlation and heteroscedasticity of the dataset are taken into account.

For convenience, we introduce a $G \times N$ matrix Z of binary indicators such that

$$z_{ig} = \begin{cases} 1, & \textit{individual } i \textit{ is from group } g \\ 0, & \textit{otherwise.} \end{cases}$$

We consider the case that each individual can only belong to exactly one group and assume that the G groups have distinct group fixed effects $\{\alpha_g \equiv \alpha_{g_i=g}; i = 1, \ldots, N, g = 1, \ldots, G\}$, then Z satisfies $\sum_{g=1}^{G} z_{ig} = 1$ and $\sum_{i=1}^{N} z_{ig} = n_g$, where n_g represents the number of individuals within group g.

We can rewrite (3.2.3) as

$$\begin{aligned} (\widehat{\beta}, \widehat{Z}, \widehat{\alpha}) &= \underset{\beta,Z,\alpha}{\operatorname{argmin}} \frac{1}{NT} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} (Y_i - X_i\beta - \alpha_g)' W^{-1} (Y_i - X_i\beta - \alpha_g) \\ &= \underset{\beta,Z,\alpha}{\operatorname{argmin}} \frac{1}{NT} \Bigg\{ \underbrace{\sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} (Y_i - X_i\beta)' W^{-1} (Y_i - X_i\beta)}_{A_1} \\ &\quad \underbrace{-2 \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} (Y_i - X_i\beta)' W^{-1} \alpha_g}_{A_2} + \underbrace{\sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \alpha_g' W^{-1} \alpha_g}_{A_3} \Bigg\}, \end{aligned} \tag{3.2.5}$$

where

$$\begin{aligned} A_1 &= \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} (Y_i - X_i\beta)' W^{-1} (Y_i - X_i\beta) \\ &= \sum_{i=1}^{N} \left|\left| W^{-\frac{1}{2}} (Y_i - X_i\beta) \right|\right|^2 \\ &= \operatorname{Tr}\left[(Y - X\beta)' W^{-1} (Y - X\beta) \right], \end{aligned} \tag{3.2.6}$$

$$
\begin{aligned}
A_2 &= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}(Y_i - X_i\beta)'W^{-1}\alpha_g \\
&= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\sum_{t=1}^{T}(y_{it} - x_{it}\beta)\left(W^{-1}\right)_{tt}\alpha_{gt} \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}\beta)\left(W^{-1}\right)_{tt}\sum_{g=1}^{G} z_{ig}\alpha_{gt} \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}\beta)\left(W^{-1}\right)_{tt}(\alpha Z)_{it} \\
&= \sum_{i=1}^{N}\left((Y - X\beta)'W^{-1}\alpha Z\right)_{ii} \\
&= \mathrm{Tr}\left[(Y - X\beta)'W^{-1}\alpha Z\right],
\end{aligned}
\tag{3.2.7}
$$

and

$$
\begin{aligned}
A_3 &= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\alpha_g' W^{-1}\alpha_g \\
&= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\left|\left|W^{-\frac{1}{2}}\alpha_g\right|\right|^2 \\
&= \sum_{g=1}^{G}\left|\left|W^{-\frac{1}{2}}\alpha_g\right|\right|^2 n_g.
\end{aligned}
\tag{3.2.8}
$$

Let's define another set of estimators by

$$\begin{aligned}(\tilde{\beta},\tilde{Z},\tilde{\alpha}) &= \underset{\beta,Z,\alpha}{\operatorname{argmin}} \frac{1}{NT}\left|\left|W^{-\frac{1}{2}}\left(Y-X\beta-\alpha Z\right)\right|\right|^2 \\ &= \underset{\beta,Z,\alpha}{\operatorname{argmin}} \frac{1}{NT}\operatorname{Tr}\left[\left(W^{-\frac{1}{2}}\left(Y-X\beta-\alpha Z\right)\right)'\left(W^{-\frac{1}{2}}\left(Y-X\beta-\alpha Z\right)\right)\right] \\ &= \underset{\beta,Z,\alpha}{\operatorname{argmin}} \frac{1}{NT}\left\{\underbrace{\operatorname{Tr}\left[(Y-X\beta)'W^{-1}(Y-X\beta)\right]}_{B_1}\right. \\ &\qquad \left. -2\underbrace{\operatorname{Tr}\left[(Y-X\beta)'W^{-1}\alpha Z\right]}_{B_2}+\underbrace{\operatorname{Tr}\left[Z'\alpha'W^{-1}\alpha Z\right]}_{B_3}\right\}.\end{aligned} \tag{3.2.9}$$

For a give indicator matrix Z, we can concentrate out α by solving

$$\begin{aligned}\frac{\partial}{\partial\alpha}\left|\left|W^{-\frac{1}{2}}\left(Y-X\beta-\alpha Z\right)\right|\right|^2 &= \frac{\partial}{\partial\alpha}\left(B_1+B_2+B_3\right) \\ &= 2\left(W^{-1}\alpha ZZ'-W^{-1}(Y-X\beta)Z'\right) \\ &= 0,\end{aligned} \tag{3.2.10}$$

which leads to

$$\tilde{\alpha} = (Y-X\beta)\,Z'\left(ZZ'\right)^{-1}. \tag{3.2.11}$$

Thus, the minimization problem in (3.2.9) is equivalent to

$$\begin{aligned}(\tilde{\beta},\tilde{Z},\tilde{\alpha}) &= \underset{\beta,Z,\alpha}{\operatorname{argmin}} \frac{1}{NT}\left|\left|W^{-\frac{1}{2}}\left(Y-X\beta-\alpha Z\right)\right|\right|^2 \\ &= \underset{\beta,Z}{\operatorname{argmin}} \frac{1}{NT}\left|\left|W^{-\frac{1}{2}}\left((Y-X\beta)-(Y-X\beta)Z'\left(ZZ'\right)^{-1}Z\right)\right|\right|^2 \\ &= \underset{\beta,Z}{\operatorname{argmin}} \frac{1}{NT}\left|\left|W^{-\frac{1}{2}}(Y-X\beta)\left(\mathbf{I}_N-Z'\left(ZZ'\right)^{-1}Z\right)\right|\right|^2.\end{aligned} \tag{3.2.12}$$

Next, we are going to show that $(\widehat{\beta},\widehat{Z},\widehat{\alpha})$ and $(\tilde{\beta},\tilde{Z},\tilde{\alpha})$ are the same. This can be done by

showing that $B_1 = A_1, B_2 = A_2$ (obviously), and

$$\begin{aligned} B_3 &= \operatorname{Tr}\left[Z'\alpha' W^{-1}\alpha Z\right] \\ &= \operatorname{Tr}\left[\alpha' W^{-1}\alpha ZZ'\right] \\ &= \sum_{g=1}^{G}\left(\alpha' W^{-1}\alpha ZZ'\right)_{gg} \\ &= \sum_{g=1}^{G}\sum_{l=1}^{G}\left(\alpha' W^{-1}\alpha\right)_{gl}\left(ZZ'\right)_{lg} \\ &= \sum_{g=1}^{G}\left(\alpha' W^{-1}\alpha\right)_{gg}\left(ZZ'\right)_{gg} \\ &= \sum_{g=1}^{G}\left|\left|W^{-\frac{1}{2}}\alpha_g\right|\right|^2 n_g \\ &= A_3. \end{aligned} \tag{3.2.13}$$

In (3.2.13), we use the fact that ZZ' is diagonal, and the diagonal elements represent the number of individuals within each group.

We have shown that the estimators defined in (3.2.3) can be represented as

$$\begin{aligned} (\widehat{\beta}, \widehat{Z}) = \underset{\beta, Z}{\operatorname{argmin}} \frac{1}{NT}\left|\left|W^{-\frac{1}{2}}(Y - X\beta)\left(\mathbf{I}_N - Z'\left(ZZ'\right)^{-1}Z\right)\right|\right|^2 \\ \text{with} \quad \widehat{\alpha} = (Y - X\beta)\, Z'\left(ZZ'\right)^{-1}. \end{aligned} \tag{3.2.14}$$

3.2.2 Factor model with latent grouped patterns

Kim (2018) further imposes a factor structure on the unobservables and consider the following model:

$$Y_i = X_i\beta^0 + F^0\lambda_{g_i} + \epsilon_i. \tag{3.2.15}$$

Or, in matrix form,

$$Y = X\beta^0 + F^0\Lambda' + \epsilon, \tag{3.2.16}$$

where $F^0 = (F_1^0, \ldots, F_T^0)'$ and $\Lambda = (\lambda_{g_1}, \ldots, \lambda_{g_n})'$ are $(T \times r)$ and $(N \times r)$ matrices respectively, and r is the number of principle components. We assume that there are unknown group patterns exist in Λ, and λ_{g_i} is common within each group.

Similarly, we define our estimator as

$$(\widehat{\beta},\widehat{\gamma},\widehat{F},\widehat{\Lambda}) = \underset{(\beta,F,\Lambda,\gamma)\in\Theta_\beta\times\mathbb{R}^{T\times r}\times\mathbb{R}^{N\times r}\times\Gamma_G}{\operatorname{argmin}} \frac{1}{NT}\sum_{g=1}^{G}\sum_{i\in\{g_i=g\}}(Y_i - X_i\beta - F\lambda_{g_i})'W^{-1}(Y_i - X_i\beta - F\lambda_{g_i}), \tag{3.2.17}$$

where W has been defined in (3.2.4). We keep notations the same and denote $\{\lambda_g \equiv \lambda_{g_i=g}; i = 1, \ldots, N, g = 1, \ldots, G\}$, then (3.2.17) can be rewritten as

$$\begin{aligned}(\widehat{\beta}, \widehat{Z}, \widehat{F}, \widehat{\Lambda}) &= \underset{\beta,Z,F,\Lambda}{\operatorname{argmin}} \frac{1}{NT}\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}(Y_i - X_i\beta - F\lambda_g)'W^{-1}(Y_i - X_i\beta - F\lambda_g) \\ &= \underset{\beta,Z,F,\Lambda}{\operatorname{argmin}} \frac{1}{NT}\left|\left|W^{-\frac{1}{2}}(Y - X\beta - F\Lambda'Z)\right|\right|^2.\end{aligned} \tag{3.2.18}$$

The proof is similar to previous section if we replace α with $F\Lambda'$. To make F and Λ separately identifiable, we follow the literature and employ the following normalization:

$$(F'F)/T = \mathbf{I}_r \text{ and } \Lambda'\Lambda = \text{diagonal}. \tag{3.2.19}$$

Then, solving for Λ from the following equation

$$\begin{aligned}&\frac{\partial}{\partial\Lambda}\left|\left|W^{-\frac{1}{2}}\left(Y-X\beta-F\Lambda'Z\right)\right|\right|^2\\=&2\left(ZZ'\Lambda F'W^{-1}F-Z(Y-X\beta)'W^{-1}F\right)\\=&0,\end{aligned}\tag{3.2.20}$$

which leads to

$$\Lambda=\left(ZZ'\right)^{-1}Z\left(Y-X\beta\right)'F/T.\tag{3.2.21}$$

Hence, after concentrating Λ out of (3.2.18), we can see that the estimators defined in (3.2.18) can also be represented as

$$\begin{aligned}(\widehat{\beta},\widehat{Z},\widehat{F})&=\underset{\beta,Z,F}{\operatorname{argmin}}\frac{1}{NT}\left|\left|W^{-\frac{1}{2}}\left(Y-X\beta-F\left(F'(Y-X\beta)Z'(ZZ')^{-1}/T\right)Z\right)\right|\right|^2\\&=\underset{\beta,Z,F}{\operatorname{argmin}}\frac{1}{NT}\left|\left|W^{-\frac{1}{2}}\left(Y-X\beta-\left(FF'/T\right)(Y-X\beta)Z'(ZZ')^{-1}Z\right)\right|\right|^2\\&=\underset{\beta,Z}{\operatorname{argmin}}\frac{1}{NT}\left|\left|W^{-\frac{1}{2}}\left(Y-X\beta\right)\left(\mathbf{I}_N-Z'(ZZ')^{-1}Z\right)\right|\right|^2.\end{aligned}\tag{3.2.22}$$

However, we have to impose additional assumptions because of the factor structure of the unobservables. We will discuss this part in details in the next section.

3.3 TSVD method

From now on, we use $P_A=A(A'A)^{-1}A'$ and $M_A=\mathbf{I}_m-A'(AA')^{-1}A$, where $\mathbf{I}_m$ is the $m\times m$ identity matrix and $(A'A)^{-1}$ denotes some generalized inverse in case A is not of full column rank. We have shown that the estimators defined in (3.2.3) and (3.2.17) can be represented

by the following estimators:

$$(\widehat{\beta}, \widehat{Z}) = \underset{\beta \in \mathbb{R}^k, Z \in \mathbb{R}^{G \times N}}{\operatorname{argmin}} \frac{1}{NT} \left|\left| W^{-\frac{1}{2}} \left(Y - X\beta\right) M_Z \right|\right|^2. \tag{3.3.1}$$

3.3.1 Estimating β:

We first assume W is an identity matrix for simplicity, and will come back to this issue and discuss the general case of weighting matrix. Let $\mathcal{P} = \min(N, T)$, and the SVD of the matrix $(Y - X\beta)$ be

$$Y - X\beta = U\Sigma V' \tag{3.3.2}$$

and partition $U, \Sigma = \operatorname{Diag}(\sigma_1, \ldots, \sigma_{\mathcal{P}})$ and V as follows:

$$U =: [U_G \quad U_\perp], \quad \Sigma =: \begin{bmatrix} \Sigma_G & 0 \\ 0 & \Sigma_\perp \end{bmatrix}, \quad \text{and } V =: [V_G \quad V_\perp],$$

where $U_G \in \mathbb{R}^{\mathbb{T} \times \mathbb{G}}$ and $V_G \in \mathbb{R}^{\mathbb{N} \times \mathbb{G}}$ are the left and right singular vectors matrices corresponding to the largest G singular values, and $\Sigma_G \in \mathbb{R}^{\mathbb{G} \times \mathbb{G}}$ contains the largest G singular values. Therefore,

$$Y - X\beta = U_G \Sigma_G V_G' + U_\perp \Sigma_\perp V_\perp' \tag{3.3.3}$$

By the Eckart–Young–Mirsky theorem, $\widehat{\alpha} = U_G \Sigma_G V_G'$ minimizes $||Y - X\beta - \alpha||^2$ such that $\text{Rank}(\alpha) = G$. Thus,

$$
\begin{aligned}
&\frac{1}{NT} ||(Y - X\beta) M_Z||^2 \\
&= \frac{1}{NT} \left|\left| \left(Y - X\beta - \underbrace{(Y - X\beta) P_Z}_{\alpha} \right) \right|\right|^2 \\
&\geq \min_{\alpha} \frac{1}{NT} ||(Y - X\beta - \alpha)||^2 \\
&= \frac{1}{NT} ||(Y - X\beta - U_G \Sigma_G V_G')||^2 \\
&= \frac{1}{NT} ||(U_\perp \Sigma_\perp V_\perp')||^2 \\
&= \frac{1}{NT} \text{Tr}\left[(U_\perp \Sigma_\perp V_\perp')' U_\perp \Sigma_\perp V_\perp' \right] \\
&= \frac{1}{NT} \text{Tr}\left[\Sigma_\perp' \underbrace{U_\perp' U_\perp}_{I_{T-G}} \Sigma_\perp \underbrace{V_\perp' V_\perp}_{I_{N-G}} \right] \\
&= \frac{1}{NT} \text{Tr}\left[\Sigma_\perp' \Sigma_\perp \right] \\
&= \frac{1}{NT} \sum_{\text{r=G+1}}^{\mathcal{P}} \sigma_r^2 \left(Y - X\beta \right),
\end{aligned}
\tag{3.3.4}
$$

where $\sigma_r(\cdot)$ is the rth largest singular values of the matrix argument. Hence, give any (β, Z), we have shown that the original objective function in (3.3.4) is lower bounded by $\frac{1}{NT} \sum_{\text{r=G+1}}^{\mathcal{P}} \sigma_r^2 \left(Y - X\beta \right)$, which is independent from the grouping Z. Define $\mathcal{L}^0(\beta)$ as our objective function with centering, namely,

$$
\mathcal{L}^0(\beta) = \frac{1}{NT} \sum_{\text{r=G+1}}^{\mathcal{P}} \sigma_r^2 \left(Y - X\beta \right) - \frac{1}{NT} \sum_{\text{r=G+1}}^{\mathcal{P}} \sigma_r^2 \left(Y - X\beta^0 \right). \tag{3.3.5}
$$

Note that the second term is for the purpose of centering and does not depend on β. We estimate β^0 by

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\, \mathcal{L}^0(\beta). \tag{3.3.6}$$

3.3.2 Estimating grouping and group heterogeneities

Once $\widehat{\beta}$ is obtained, we denote $\widehat{R} = Y - X\widehat{\beta}$ and estimate group membership by

$$\widehat{Z} = \underset{Z \in \mathbb{R}^{G \times N}}{\operatorname{argmin}} \frac{1}{NT} \left|\left| W^{-\frac{1}{2}} \widehat{R} M_Z \right|\right|^2. \tag{3.3.7}$$

This is applying k-means method on $\widehat{R}$ using Mahalanobis distance as a dissimilarity measure. Note that the weighting matrix does not affect the consistency of our estimator, but it does improve the estimation of grouping by reducing the serial-correlation and heteroskedasticity issue in the data. Kao et al. (2020) propose a Mahalanobis metric based clustering method, which is using a particular value of β to improve the performance of k-means. But how to estimate the weighting matrix is not well studied yet. Our approach has an advantage that we don't rely on grouping information to consistently estimate β^0, all we need is the number of groups G. Therefore, we can consistently estimate the covariance matrix by

$$\widehat{W} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - X_i\widehat{\beta})(Y_i - X_i\widehat{\beta})', \tag{3.3.8}$$

and then estimate the group membership by

$$\widehat{Z} = \underset{Z \in \mathbb{R}^{G \times N}}{\operatorname{argmin}} \frac{1}{NT} \left|\left| \widehat{W}^{-\frac{1}{2}} \widehat{R} M_Z \right|\right|^2. \tag{3.3.9}$$

Given the optimal $\widehat{\beta}$ and $\widehat{Z}$, we can further find the group specific time-varying fixed effect by

$$\widehat{\alpha} = \widehat{R}\widehat{Z}' \left(\widehat{Z}\widehat{Z}'\right)^{-1}, \tag{3.3.10}$$

if the model is specified as in (3.2.1).

When the model is specified as in (3.2.15), we have to make the following assumptions on the unobservable component because of the factor structure.

ASSUMPTION SF (Strong Factors):

(1) The true number of factors r is known (r can be relaxed to $\min(G, r)$).

(2) We have $0 < \mathbb{E}(\Lambda'\Lambda) < \infty$.

(3) We have $0 < \mathbb{E}(F^{0\prime}F^0) < \infty$.

Assumption SF(1) is required since, beside the factor structure, our model also impose group patterns on factor loadings, and therefore rises a new issue of comparing G and r (because $\text{Rank}(F^0\Lambda') = \min(G, r, T)$). Assumption SF(2) and SF(3) are regularly made in the literature on large N and T factor models, including Bai and Ng (2002) and Bai (2009), which implies the existence of r factors.

Let's first consider the case when $G \geq r$, in that case, we can not only consistently estimate β^0, but also F^0 and Λ. The estimator for F^0 is the first r eigenvectors (multiplied by $\sqrt{T}$ due to the restriction $F'F/T = I$) associated with the first r largest eigenvalues of the matrix $\widehat{R}\widehat{R}'$, while the estimator of the group-level factor loadings is

$$\widehat{\Lambda} = \left(\widehat{Z}\widehat{Z}'\right)^{-1} \widehat{Z}\widehat{R}'\widehat{F}/T. \tag{3.3.11}$$

Our approach is implicitly the same with the method proposed in Moon and Weidner (2015), if we consider G as a upper bound of the true number of factors. The trivial bound $G = \min(N, T)$ is not allowed. When $G < r$, our estimators for β^0 and Z remain consistent as Assumption A and B are still satisfied. However, Assumption SF(2) will be violated since $\text{Rank}\,(\Lambda'\Lambda) = G$, which implies that $\mathbb{E}\,(\Lambda'\Lambda) \geq 0$. As a result, F^0 and Λ are not well identified separately.

3.4 Monte Carlo simulation

In this section, we examine the performance of our TSVD approach. The following two DGP are employed for simulation.

DGP1(Time-varying fixed effects):

$$\begin{aligned}
y_{it} &= x_{it}\beta_1^0 + z_{it}\beta_2^0 + \alpha_{g_it}^0 + v_{it}, \quad g_i = 1, \ldots, G, i = 1, \ldots, N, t = 1, \ldots, T, \\
\alpha_{g_it}^0 &= \rho\alpha_{g_it-1}^0 + e_{g_it}^{(1)}, \text{ with } \alpha_{g_i1}^0, e_{g_it}^{(1)} \overset{iid}{\sim} N(0,1), \\
v_{it} &= \theta v_{it-1} + e_{it}^{(2)}, \text{ with } v_{i1}, e_{it}^{(2)} \overset{iid}{\sim} N(0, Uniform(1,2)), \\
x_{it} &= \alpha_{g_it}^0 + e_{it}^{(3)}, \text{ with } e_{it}^{(3)} \overset{iid}{\sim} N(0,1), \\
z_{it} &= \frac{1}{2}\alpha_{g_it}^0 + e_{it}^{(4)}, \text{ with } e_{it}^{(4)} \overset{iid}{\sim} N(0,1).
\end{aligned} \tag{3.4.1}$$

DGP2(Interactive fixed effects):

$$\begin{aligned}
y_{it} &= x_{it}\beta_1^0 + z_{it}\beta_2^0 + F_t^{0\prime}\lambda_{g_i} + v_{it}, \quad g_i = 1, \ldots, G, i = 1, \ldots, N, t = 1, \ldots, T, \\
F_t^0 &= \rho F_{t-1}^0 + e_t^{(1)}, \text{ with } F_1^0, e_t^{(1)} \overset{iid}{\sim} N(0,1), \\
\lambda_{g_i} &= e_{g_i}^{(2)}, \text{ with } e_{g_i}^{(2)} \overset{iid}{\sim} N(0,1),
\end{aligned} \tag{3.4.2}$$

where x_{it}, z_{it}, and v_{it} are the same as in DGP1 except for replacing $\alpha_{g_it}^0$ with $F_t^{0\prime}\lambda_{g_i}$ in x_{it} and z_{it}. Note that the pooled OLS and standard fixed effects estimators are not valid in such

cases because x_{it} is correlated with $\alpha_{g_it}^0$ or $F_t^{0\prime}\lambda_{g_i}$. The number of groups G is assumed to be known and each group has the same number of individuals. The number of replications is 1000.

Table 3.4.1: Simulation results for $\widehat{\beta}$ $(DGP1, \beta_1^0 = \beta_2^0 = 1, G = 10)$

			$\widehat{\beta}_1$			$\widehat{\beta}_2$		
N	T	$(\rho,\, \theta)$	Bias	SD	ERP	Bias	SD	ERP
100	30	(0,0)	0.0754	0.0328	0.6	0.0435	0.0332	0.2
100	30	(0.9,0)	0.0552	0.0605	0.04	0.0321	0.0603	0
200	30	(0,0)	0.0329	0.0221	0.3	0.0225	0.0223	0.14
200	30	(0.9,0)	0.0272	0.0398	0	0.0178	0.0397	0

Table 3.4.2: Simulation results for $\widehat{\beta}$ $(DGP2, \beta_1^0 = \beta_2^0 = 1, G = 10)$

				$\widehat{\beta}_1$			$\widehat{\beta}_2$		
N	T	$(\rho,\, \theta)$	r	Bias	SD	ERP	Bias	SD	ERP
100	30	(0,0)	5	0.0192	0.0681	0	0.0182	0.0651	0
100	30	(0.9,0)	5	0.0187	0.1473	0	0.0199	0.1425	0
200	30	(0,0)	5	0.0125	0.0431	0	0.0145	0.0422	0
200	30	(0.9,0)	5	0.0138	0.0972	0	0.0130	0.0930	0
100	30	(0,0)	15	0.0267	0.1124	0	0.0206	0.1101	0
100	30	(0.9,0)	15	0.0227	0.2182	0	0.0193	0.2298	0
200	30	(0,0)	15	0.0153	0.0819	0	0.0121	0.0784	0
200	30	(0.9,0)	15	0.0150	0.1514	0	0.0117	0.1595	0

Tables 3.4.1 and 3.4.2 show the bias, standard deviation (SD) and empirical rejection probability (ERP) at the 5% level given different serial correlation scenarios and DGPs. We use the t-statistics and normal critical values to simulate the ERPs. For the t-statistics, we employ the clustered standard errors proposed by Arellano (1987) to allow for serial correlation. We do not impose any serial correlation on the error term for the moment, namely, $theta = 0$, because it will lead to a bias of the estimator. The bias-correction estimator will be discussed in the future. From 3.4.1 and 3.4.2, our estimator has excellent

finite sample performance in both cases of models with time-varying and interactive group fixed effects. Notably, in the case of the interactive group fixed-effect model, our approach accurately estimates the causal parameters regardless of the number of factors. We also note that the serial correlation in group heterogeneities enhances the group signal and hence improves the estimation of coefficients. Lastly, our estimator becomes more accurate as N increases. This is well expected because of the Law of Large Numbers.

Bibliography

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795–1843.

Arellano, M., 1987. Practitioners'corner: Computing robust standard errors for within-groups estimators. Oxford bulletin of Economics and Statistics 49, 431–434.

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

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221. doi:10.1111/1468-0262.00273.

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

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

Brock, G., Pihur, V., Datta, S., Datta, S., 2008. clvalid: An R package for cluster validation. Journal of Statistical Software, Articles 25, 1–22.

Chen, X., Favilukis, J., Ludvigson, S.C., 2013. An estimation of economic models with recursive preferences. Quantitative Economics 4, 39–83.

Dunn†, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4, 95–104.

Epstein, L.G., Zin, S.E., 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. Econometrica 57, 937–969.

Epstein, L.G., Zin, S.E., 1991. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis. Journal of Political Economy 99, 263–286.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75, 603–680.

Handl, J., Knowles, J., Kell, D.B., 2005. Computational cluster validation in post-genomic data analysis. Bioinformatics 21, 3201–3212.

Kao, C., Kim, M.S., Zhang, Z., 2020. Mahalanobis metric based clustering for fixed effects model. Sankhya B doi:10.1007/s13571-019-00211-z.

Kim, M.S., 2018. Policy analysis using panel and multilevel regression models with group interactive fixed effects. Working Paper .

Kocherlakota, N.R., 1990. Disentangling the coefficient of relative risk aversion from the elasticity of intertemporal substitution: An irrelevance result. The Journal of Finance 45, 175–190.

Mahalanobis, P., 1936. On the generalized distance in statistics. Proceedings of National Institute of Sciences (India) 2, 49–55.

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

Nestor, G., Ruben, H.M., 2015. Risk aversion at the country level. Review 97, 53–66.

Neyman, J., Scott, E.L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53 – 65.

Smith, D.C., 1999. Finite sample properties of tests of the epstein–zin asset pricing model. Journal of Econometrics 93, 113 – 148.

Su, L., Shi, Z., Phillips, P.C.B., 2016. Identifying latent structures in panel data. Econometrica 84, 2215–2264. doi:10.3982/ECTA12560.

Tauchen, G., 1986. Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data. Journal of Business and Economic Statistics 4, 397–416.

Tauchen, G., Hussey, R., 1991. Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models. Econometrica 59, 371–396.

Weil, P., 1989. The equity premium puzzle and the risk-free rate puzzle. Journal of Monetary Economics 24, 401 – 421.