TYPE FIXED EFFECTS AND RATIONAL ADDICTION: A GMM FRAMEWORK FOR LATENT TYPE HETEROGENEITY

JORGE A. RIVERO

Abstract. This paper reexamines Rational Addiction (RA) by introducing the type fixed effects (TFE) panel model. The TFE model incorporates heterogeneous coefficients and time-varying patterns of heterogeneity, which reflect differences in preferences and the addiction process. The model assumes the existence of a latent, time-invariant continuous variable referred to as a "type", which drives the heterogeneity in the parameters. Smoothness of the parameters as functions of the type is key to identification, allowing individuals of similar types to have similar parameter values. Correlation between the parameters, covariates, and instruments stem from type heterogeneity. I propose the type fixed effects generalized method of moments (TFE-GMM) estimator and establish consistency. I provide fast computation procedures based on the stochastic gradient descent algorithm. Simulations demonstrate good performance of this estimator. Using yearly household cigarette purchase data to estimate the model shows that most households follow cyclical consumption patterns and insensitivity to prices changes, giving support to educational interventions to curb smoking.

Keywords: rational addiction, cigarette demand, heterogeneous effects, time-varying heterogeneity, panel data, fixed effects, varying coefficients, GMM

JEL codes: C14, C23, C26, H25, I12, I18

Last modified: November 1, 2023, click here (jarivero.com) for an up-to-date draft of this JMP. *Preliminary draft, subject to change, all errors are my own. I am grateful for the support and encouragement from my committee chair Yanqin Fan. Special thanks to Stéphane Bonhomme and Hide Ichimura for their valuable comments.* Correspondence to jrivero@uw.edu. Department of Economics, University of Washington, Box 353330, Seattle, WA 98195.

1. Introduction

The consumption of harmful addictive goods is a paradox involving rational and self-destructive behavior, yielding brief benefits to the consumer while causing prolonged damage to their health. The Rational Addiction (RA) framework initiated by Becker and Murphy [1988] reconciles this dichotomy by incorporating awareness of these beneficial and harmful effects. It carries important policy implications, such as endorsing taxes as a public health measure since it predicts consumers are sensitive to price changes in the long run. However, while the majority of health economists view its real world implications favorably, its empirical support is recognized as weak (Melberg [2009]). Acknowledging the complicated diversity of addiction across individuals, I introduce unobserved type heterogeneity in a RA model that induces type-heterogeneous effects and type-specific time series for each consumer type. The type heterogeneity framework is flexible yet parsimonious by assuming individuals with similar types have similar type parameters, rather than considering complete heterogeneity of parameters. Estimation provides evidence that cigarette consumers binge over years and are less price sensitive than previously thought, implying that educational policies may effectively lengthen periods of temperance, similar to successes in addressing binge-eating disorders (McElroy et al. [2015]).

The basic RA model is of a rational finite-lived forward-looking representative consumer of an addictive good that is aware of the potential harm from consumption. Past consumption accumulates a "consumption capital" or addiction "stock" that incurs disutility and depreciates over time. The agent follows a reinforcement mechanism known as "adjacent complementarity" where the marginal utility of current consumption increases with past consumption and also displays tolerance, which amounts to diminishing marginal returns of past consumption. This is the most popular iteration of the RA model as it incorporates important features of addiction while remaining simple. The model with additional assumptions yields a linear consumption plan that depends on consumption in the previous and subsequent period and most studies, following the lead of Becker et al. [1994], have taken this to the data as a fixed effects model. Both state-level and household studies offer similar conclusions: consumers display saddle-point dynamics, consuming high/low amounts early in their lifetimes, approach an equilibrium and then return or exceed previous levels of consumption for the rest of their lives. However, I claim that the empirical view of rational addiction has been narrow and expanding the scope of dynamics and heterogeneity offers insights into the consumption of addictive goods.

In theory the RA model can generate cyclical consumption patterns, but this feature has not received empirical attention. Becker and Murphy [1988] and Dockner and Feichtinger [1993] show that if there are two addictions stocks accumulating based on past consumption, then the consumer may display binging behavior. As an example, there may be the traditionally considered habituation stock that displays adjacent complementarity as well as an additional "poor health" stock that displays *adjacent substitutability*, meaning that marginal utility will fall when health begins to deteriorate. We should expect the habituation stock to be discounted at a larger rate than health so that eventually the harm surpasses the benefits from satisfying cravings and consumption falls. When health improves, the habituation stock may once again drive consumption to higher levels, implying binging consumption patterns. In addition, nothing is preventing both stocks from exhibiting adjacent complementarity, therefore simultaneously reinforcing future consumption. Therefore, without controlling for this unobserved time-varying process there is potential for bias since past consumption is summarizing two stock variables that are at odds, where a common time factor or individual fixed effect will not suffice for individual-specific time-varying effects.

Household data is preferred over aggregate data since the unit of interest is the individual's behavior and addiction. In addition, aggregate data are usually short panels, which may lead to inconsistent estimators for dynamic panel models (Baltagi and Griffin [2001]). Auld and Grootendorst [2004] demonstrate how aggregate studies may be misleading by estimating an RA model to show milk consumers appear more addicted than cigarette users. Nevertheless, despite using household data, the implied saddle-point dynamics do not explain why many cigarette consumers quit and then return to habit, or exhibit binging throughout their lifetime. The inclusion of a second addiction stock may rectify this issue, but unobserved differences in preferences might also be a driving factor. From a pharmacology perspective, these dynamics are associated to nicotine dependence (Benowitz [2010]) where after periods of abstinence the sensitivity to rewards from nicotine are refreshed, facilitating relapse. The effects of nicotine dependence manifest very differently across the population, for instance, Saunders et al. [2022] survey 3.4 million people and found almost 4,000 genetic variants associated to the use of tobacco, which in addition to cultural norms pose a high degree of unobserved heterogeneity in the data. They may also impact forward looking decision making, sensitivity to prices, and age-related factors that impact consumption in the life-cycle; see Grant et al. [2010] and Carroll [2021].

This discussion alludes to the potential for there being many heterogeneous consumer types with corresponding addiction processes and responses impacting their consumption of addictive goods. To address this, I augment the RA model with "type heterogeneity" written as:

$$C_{it} = \alpha_t(\xi_i) + \theta_l(\xi_i)\, C_{it-1} + \theta_f(\xi_i)\, C_{it+1} + \beta(\xi_i)' X_{it} + U_{it} \qquad (1.1)$$
$$i = 1, \ldots, N; \quad t = 1, \ldots, T,$$

where C_{it} denotes consumption of the addictive good, X_{it} are observable characteristics such as price, income, race, age, etc and U_{it} is an error term. Note that C_{it-1} and C_{it+1} are both endogenous and I assume for now that we have access to valid instruments Z_{it}. The variable ξ_i is a continuous, time-invariant and unobserved variable I call a type that belongs to a compact subset Ξ of the reals and randomly sampled along with observables from some joint density. All covariates and instruments may be arbitrarily correlated with this type and thus with the type-specific parameters $\alpha_t(\xi_i)$ $\theta_l(\xi_i)$, $\theta_f(\xi_i)$ and $\beta(\xi_i)$. I call a panel model with unobserved types, type-specific parameters and arbitrary correlation structure with observables a type fixed effects (TFE) model.

The TFE RA model is a flexible way to incorporate complex forms of consumer heterogeneity towards addictive goods without assuming complete heterogeneity of parameters over individuals and time. The type variable ξ_i governs the partial effects $\theta(\cdot)$ of covariates on consumption, reflecting heterogeneous preferences. This determines a dynamic profile of effects from the testable implications of the RA model for each individual allowing for some segments of the population to follow saddle-point dynamics while others are in cycles. The time-varying term $\alpha_t(\xi_i)$ is known as the type fixed effect and captures the variation from an alternative addiction stock driving unobserved time patterns of consumption. To my knowledge, this is the first paper to propose estimating a RA model with many heterogeneous parameters, controlling for a second addiction stock, and to consider heterogeneous consumption dynamics.

A key identification condition for this model is that those of similar types must have comparable parameter values. The parameters can be regarded as curves (functions) parametrized by the type and in this sense the curves must be sufficiently smooth to allow such comparisons. This is similar to semiparameteric models with varying coefficients (Hastie and Tibshirani [1993]), however the coefficients vary according to an unobservable. Along these lines, $\alpha_t(\cdot)$ for $t = 1, \ldots, T$ can be thought of as a nonparametric specification of the functional form of the heterogeneity, where important examples are two-way fixed effects and interactive fixed effects (see Wooldridge [2010] and Bai [2009]). This observation along with the extension of varying coefficients to an unobserved argument and parameter with dimension growing with the sample are two contributions to the semiparametric literature.

In Section 2, I introduce the TFE-GMM framework for models with type heterogeneity, discuss identification of types, define an estimator, and provide a computational algorithm based on stochastic gradient descent. The model (1.1) is a special example of this framework and additional identification conditions are provided to ensure compatibility with TFE-GMM. In Section 3, I give conditions for consistency of the TFE-GMM estimator for parameters of a TFE linear panel model with endogenous covariates and discuss how consistency for types may be achieved. In Section 4, I present simulation results indicating good performance of the estimator under various data generating processes. In Section 5, I give descriptive statistics regarding the Nielsen household panel data showing heterogeneity in consumption and provide the results from estimating the TFE RA model. Most individuals in the sample are binging and are less price sensitive than previous studies suggest.

Related Literature. This section reviews some of the important related work to the TFE RA model. It is organized by first discussing literature associated to RA followed by econometric techniques.

Cawley and Ruhm [2011] presents models of addiction and habit that includes the RA model. The model has been extensively applied to many goods considered addictive in both aggregate and micro panel studies such as tobacco, alcohol and coffee (Becker et al. [1994], Grossman et al. [1998] and Olekalns and Bardsley [1996], respectively). The RA model is typically tested against the alternative that the coefficients of the lead and lag consumption covariates are positive, so that if the null is rejected then the consumer is taken to follow addiction (lag) and is forward looking (lead). Micro studies are preferred since they capture individual behavior more closely and can be aggregated to describe municipality and state-level dynamics (Chaloupka [1991], Grossman et al. [1998]). They also tend to produce more plausible estimates of the discount factor, but results remain mixed. Laporte et al. [2017] argue with simulation evidence that saddle-point dynamics may cause identification issues when an unstable root is dominant, meaning stability can't be used to pin down values of parameters. Unstable roots are a feature of dynamic micro panel models and so they claim it may be difficult to estimate RA models in general. Considering type heterogeneity allows consumers to display different dynamics and stability properties, which may give more credibility to estimates. The closest application of RA to this paper is found in Fernández-Val and Lee [2013], but they do not include unobserved time-varying fixed effects and heterogeneity of the lead and lag of consumption thereby excluding heterogeneous dynamics.

The TFE model is related to random coefficients models, except it includes both individual and time-varying heterogeneity, see Hsiao [2022] for a text on these and other panel data methods. It is well known that ignoring parameter heterogeneity by imposing fixed coefficients generally results in inconsistent estimation of the mean of random coefficients (Yitzhaki [1996], Heckman and Vytlacil [1998], Angrist et al. [2000], Angrist [2004]) and moreover denies estimation of the other features of the coefficients. However, identification requires additional care such as limiting the degree of heterogeneity of the random coefficients, which may not be justified empirically and in theory (Arellano and Bonhomme [2012], Graham and Powell [2012], Laage [2020]). Nonparametric and semiparametric techniques are prevalent in economics in part for their robustness to misspecification at the cost of requiring a large sample, see Unit 1 of Li and Racine [2007] for kernel smoothing techniques. This is not the first application of a panel model of varying coefficients, however it is the first to include parameters that vary according to an unobservable and also vary across time, see Hoover et al. [1998] and Fan and Zhang [2000] that focus on the case of coefficients as unspecified functions of time. The TFE-GMM criterion function is similar to the criterion of Fernández-Val and Lee [2013], which is based on the two-step procedure of Hansen [1982] but is aggregated on the cross-sectional level using each individual unit's time series GMM criteria. There is finite-sample bias for GMM and this does not exclude the TFE-GMM estimator, however there does exist bias correction measures that may be applied to this case (Newey and Smith [2004]). Incidental parameter bias may affect estimates of dynamic models using short panels and model (1.1) is no exception; see Neyman and Scott [1948], Chamberlain [1980] and Nickell [1981] and Arellano and Hahn [2007] for bias correction approaches for fixed effects models.

There has been a large focus on discrete types, commonly known as groups, with and without time-varying heterogeneity (Sun [2005], Chang-Ching and Serena [2012], Bester and Hansen [2016] among others). Bonhomme and Manresa [2015] introduce the grouped fixed effects (GFE) model and estimation where the heterogeneity forms an unobserved group structure and allow for slope parameters to vary across groups in addition to a group time-varying heterogeneity term similar to TFEs. Identification of groups is similar to identification of types requiring separability in the model, specifically for GFE, that the group fixed effects terms are different in the mean squared sense. Other estimators of this model have been proposed and require this assumption (Chetverikov and Manresa [2022], Mugnier [2022b]) and when there is unobserved group heteroskedasticity there are steeper conditions needed in order to identify groups; see Rivero [2023] that requires group fixed effects separation as a function of group variances. Su et al. [2016] and Mehrabani [2022] consider linear and nonlinear models with unknown group structure where the random coefficients are heterogeneous across groups, but homogenous across individuals within the same group. For linear models with endogeneity they specialize to a penalized GMM (PGMM) estimation framework that contributes to the fused-Lasso literature where some of the individual coefficients share the same value, hence forming groups through penalizing coefficients into clusters. Cheng et al. [2019] estimate a model of time-invariant multi-group heterogeneity and covariates that are endogenous despite the groups. My proposal can be viewed as a continuous extension of the discrete case where, instead of grouped patterns of heterogeneity shared among those in the same group, we have type heterogeneity that is shared within types.

The discreteness assumption may be viewed as too strict and the TFE model may be more flexible in this regard. Additionally, not much is known of the consequences of violating the group structure assumption or how to control for a continuous latent variable of this kind. One example is Bonhomme et al. [2022] who propose a two-step procedure by first discretizing the latent variable by clustering moments of observables that are informative of the types and then estimating parameters and the time-varying heterogeneity terms via maximum likelihood. This approach is particularly useful for nonlinear models where identification can be troublesome. I approach the problem of a continuous variable head-on and avoid the need for additional auxiliary moments by simply relying on the moments that identify parameters, which are smooth curves parametrized by types.

First initiated by Robbins and Monro [1951], the stochastic gradient descent (SGD) algorithm is an incredibly popular technique in machine learning due to the size of data sets and wide applicability, see Bottou [2010]. The version of stochastic gradient descent proposed in this paper is related to the k-means algorithm (Forgy [1965], Lloyd [1982]) where the assignment step in this case is a "soft" assignment that puts a weights on observations depending on relative positions of types of other observations, locally estimating heterogeneous parameters based on proximity in the type space Ξ. SGD can be applied to many extremal estimation problems since all that is required is the form of the gradient or subgradient of an objective function, for example see Lee et al. [2023] which involves quantile regression where there the non differentiability of the criterion is not an obstacle.

2. GMM Framework with Type Heterogeneity

Denote $i \in \{1, \ldots, N\} = \mathcal{N}$ and $t \in \{1, \ldots, T\} = \mathcal{T}$ as the index of individuals and the index of the observations of the individuals, respectively. Suppose we have a balanced panel data set $\{w_{it}\}$ with support $\mathcal{W}$ that is independent and identically distributed (iid) over $i \in \mathcal{N}$ from a density function f, with bounded fourth moments, and stationary and strongly mixing over $t \in \mathcal{T}$ with mixing coefficients that decay exponentially. Let $\theta \in \Theta$ and let $\alpha = (\alpha_1, \ldots, \alpha_T) \in \mathcal{A}^T$ be infinite-dimensional parameters both defined as functions on some common set $\Xi \subset \mathbb{R}$. Both of these parameters are unknown functions of types, which are the realizations of an unobserved random variable ξ_i that has support on the interior of the type space $\mathring{\Xi}$ according to density ν and may be arbitrarily correlated with some or all elements of w_{it}. The set of moment conditions will depend on the true types $\{\xi_i^0\}_{i \in \mathcal{N}}$ that are assumed to be iid draws from a type density ν^0. Let $x \mapsto \|x\|$ denote the standard Euclidean norm for vectors or the L^2 norm for functions, whichever is applicable.

For exposition, we distinguish two kinds of parameters: the type-specific parameters that enter the model directly (θ and α) and the unobserved types ξ_i. The key strategy for identification is to treat the two types of parameters separately by assuming that the types are known and then identifying the type parameters and then working the converse. This defines a system with the all parameters as unique solutions. A similar argument can be found in Bai [2009] for interactive fixed effects models. The endogenous linear model with type fixed effects is an important example for studying rational addiction and will be discussed in the context of the following GMM framework including the necessary identifying assumptions.

2.1. Identification of type-specific parameters.

Assume first that the true types ξ_i^0 are known and suppose that the true parameters $\theta^0(\xi_i^0)$ are identified by the conditional moment conditions:

$$\mathbb{E}\left[g(w_{it}; \theta^0(\xi_i^0), \alpha_t^0(\xi_i^0)) | \xi_i^0\right] = 0 \tag{2.1}$$

for all $(i, t) \in \mathcal{N} \times \mathcal{T}$ where $g : \mathcal{W} \times \Theta \times \mathcal{A} \to \mathbb{R}^\ell$ are known functions with $\ell \geq p$. Suppose that the infinite-dimensional time parameters α_t are identified by:

$$\mathbb{E}\left[\rho_t(w_{it}; \theta^0(\xi_i^0), \alpha_t^0(\xi_i^0)) | \xi_i^0\right] = 0 \tag{2.2}$$

for all $t \in \mathcal{T}$ and, ν^0-almost surely, for all $\xi \in \Xi$ and where $\rho_t : \mathcal{W} \times \Theta \times \mathcal{A} \to \mathbb{R}$ is known. The conditions (2.2) are also included in (2.1). I assume that for all $t \in \mathcal{T}$ the function ρ_t

is strictly monotonic with respect to α_t so that there exists a function φ_t defined by

$$\alpha_t^0(\xi_i^0) = \varphi_t(\xi_i^0; \theta^0) \tag{2.3}$$

where φ_t depends on the sample information at time t. We can solve for the infinite-dimensional parameters by using (2.2) to get (2.3) and essentially substitute it in the moment conditions (2.1) for θ^0. Note that for fixed $i \in \mathcal{N}$ the type ξ_i^0 is a constant random variable so that (2.1) can be written as

$$\mathbb{E}\left[g(w_{it}; \theta^0(\xi_i^0), \alpha_t^0(\xi_i^0))\right] = 0, \quad \text{for any } i \in \mathcal{N}, \tag{2.4}$$

which is an expectation over individual i's time series. This property allows inspection of an individual i's time series information to extract their true type ξ_i^0, which will be important in Section 2.2 for identification of types.

Example 1 (Endogenous linear panel model with type fixed effects)**.** Suppose that $\{w_{it}\} = \{(y_{it}, x_{it}, z_{it})\}$ where $(i,t) \in \mathcal{N} \times \mathcal{T}$, is iid over $i \in \mathcal{N}$ and strictly stationary, strongly mixing processes with exponentially decaying mixing coefficients for all $t \in \mathcal{T}$ and that iid types $\{\xi_i^0\}_{i \in \mathcal{N}}$ are observed. Consider the structural linear model with type fixed effects and type-specific coefficients

$$y_{it} = x_{it}'\theta(\xi_i^0) + \alpha_t(\xi_i^0) + u_{it} \tag{2.5}$$

where for all $i \in \mathcal{N}$ the covariates x_{it} are contemporaneously correlated with ξ_i^0, $\alpha_t(\xi_i^0)$, u_{it} and $\theta(\xi_i^0)$. We assume that we have weakly exogenous instruments z_{it} and that types ξ_i^0 are also exogenous. Specifically, $\mathbb{E}\left[u_{it}|\xi_i^0, \alpha_t(\xi_i^0)\right] = 0$, $\mathbb{E}\left[z_{it}u_{it}|\xi_i^0\right] = 0$ and $\mathrm{Cov}(z_{it}, x_{it}|\xi_i^0) \neq 0$ for all $(i,t) \in \mathcal{N} \times \mathcal{T}$[1]. It is useful to write the model (2.5) as

$$\begin{aligned} y_{it} &= x_{it}'\theta(\xi_i^0) + \sum_{s=1}^{T} \alpha_s(\xi_i^0)\mathbb{1}\{t = s\} + u_{it} \\ &= x_{it}'\theta(\xi_i^0) + \alpha(\xi_i^0)\delta_t + u_{it} \end{aligned}$$

where $\delta_t : \mathcal{T} \to \{0,1\}^T$ is the Kronecker delta function defined by $\delta_t = (\mathbb{1}\{t = s\})_{s=1}^T$, a vector of zeros in each entry except for a 1 in the t entry. Since α_t is an exogenous variable, we use δ_t for $t \in \mathcal{T}$ as an instrument and allow the non constant elements of z_{it} be correlated to $\alpha_t(\xi_i^0)$. Therefore, $\ell = K + T$ where $K \geq p$. Note that the type ξ_i^0 drives the correlation between covariates and the parameters of the model[2].

[1] Weak dependence assumptions can also be made to guarantee consistency, see Section 3.

[2] A coonnection can be made from these type coefficients to stationary random coefficients: $\beta(\xi_i^0) = \beta + \xi_i^0$ where $\xi_i^0 \sim (0, \sigma^2)$ iid so that the mean and variance of the coefficients are the constants $\beta \in \mathbb{R}^p$ and $\sigma^2 \geq 0$, respectively.

The approach of Robinson [1988] for identification of partially linear models is used to identify $\theta(\xi_i^0)$ for any $i \in \mathcal{N}$ and highlight the connection to traditional fixed effects. Taking conditional expectations:

$$\mathbb{E}\left[y_{it}|\xi_i^0\right] = \mathbb{E}\left[x_{it}|\xi_i^0\right]' \theta(\xi_i^0) + \alpha_t(\xi_i^0) \tag{2.6}$$

where we used the property $\mathbb{E}\left[\theta(\xi_i^0)|\xi_i^0\right] = \theta(\xi_i^0)$ and the fact that $\mathbb{E}\left[u_{it}|\xi_i^0\right] = 0$.

Differencing out the type conditional means:

$$y_{it} - \mathbb{E}\left[y_{it}|\xi_i^0\right] = (x_{it} - \mathbb{E}\left[x_{it}|\xi_i^0\right])'\theta(\xi_i^0) + u_{it} \tag{2.7}$$

eliminates the type fixed effects and can be regarded as a “within-type” transformation.

Let $\widetilde{y}_{it} = y_{it} - \mathbb{E}\left[y_{it}|\xi_i^0\right]$ and $\widetilde{x}_{it} = x_{it} - \mathbb{E}\left[x_{it}|\xi_i^0\right]$. Provided that the usual rank conditions hold i.e. $\operatorname*{plim}_{T\to\infty} \sum_{t=1}^{T} z_{it}z_{it}'$ and $\operatorname*{plim}_{T\to\infty} \sum_{t=1}^{T} z_{it}\widetilde{x}_{it}'$ are of full rank for any $i \in \mathcal{N}$, then $\theta^0(\xi_i^0)$ is (over) identified.

The curve $\theta^0(\cdot)$ is also over identified by

$$\mathbb{E}\left[z_{it}\left((y_{it} - \mathbb{E}\left[y_{it}|\xi\right])\right)\middle|\xi_i^0 = \xi\right] = \mathbb{E}\left[z_{it}\left(x_{it} - \mathbb{E}\left[x_{it}|\xi_i^0\right]\right)\middle|\xi_i^0 = \xi\right]' \theta^0(\xi) \tag{2.8}$$

for any $t \in \mathcal{T}$ and ν^0-almost surely $\xi \in \Xi$ provided full rank conditions conditional on types hold. In other words there must be sufficient variation within-types across time. Since in practice the types will be estimated simultaneously with parameters, stronger identification conditions will be imposed in Section 3.

Relating back to the previous section gives us

$$g(w_{it}; \theta(\xi_i^0), \alpha_t(\xi_i^0)) = z_{it}(y_{it} - x_{it}'\theta(\xi_i^0) - \alpha_t(\xi_i^0)) \tag{2.9}$$

$$\rho_t(w_{it}; \theta(\xi_i^0), \alpha_t(\xi_i^0)) = y_{it} - x_{it}'\theta(\xi_i^0) - \alpha_t(\xi_i^0), \tag{2.10}$$

for all $(i,t) \in \mathcal{N} \times \mathcal{T}$ where ρ_t is found from $\mathbb{E}\left[u_{it}|\xi_i^0\right] = 0$ and φ_t for all $t \in \mathcal{T}$ follows:

$$\begin{aligned} \alpha_t(\xi_i^0) &= \mathbb{E}\left[y_{it} - x_{it}'\theta(\xi_i^0)|\xi_i^0\right] && (2.11) \\ &= \mathbb{E}\left[y_{it}|\xi_i^0\right] - \mathbb{E}\left[x_{it}|\xi_i^0\right]' \theta(\xi_i^0) && (2.12) \\ &= \varphi_t(\xi_i^0, \theta(\xi_i^0)) && (2.13) \end{aligned}$$

where the expectation is over the cross-sectional dimension so this is understood in sample terms as an average over individuals with type ξ_i^0 at time t. □

2.2. **Identification of types.** The identification of types amounts to distinguishing patterns between individuals based on their time series. Towards this, I introduce notation to accommodate the individual's time series GMM criteria:

$$g_i(\theta^0(\xi), \alpha^0(\xi)) = \frac{1}{T}\sum_{t=1}^{T} g(w_{it}, \theta^0(\xi), \alpha_t^0(\xi)). \tag{2.14}$$

Fix $i \in \mathcal{N}$ and assume we know the true parameters and the type space Ξ. Then, given individual i's time series moment condition (2.4) and exponentially decaying mixing coefficients, we have $\operatorname*{plim}_{T\to\infty} g_i(\theta^0(\xi), \alpha^0(\xi)) = 0$. With this we can identify their type ξ_i^0 with a simple rule under some invertibility conditions on the moment functions.

Assumption 1. *For any $i \in \mathcal{N}$, $\operatorname*{plim}_{T\to\infty} \left\| g_i(\theta^0(\xi), \alpha^0(\xi)) - g_i(\theta^0(\widetilde{\xi}), \alpha^0(\widetilde{\xi}))\right\| = 0$ if and only if $\xi = \widetilde{\xi}$.*

This assumption allows the individual's GMM time series moments to be informative of types, all else equal. The following uses this and (2.4) to define a rule to assign types to each individual.

Lemma 1. *For all $i \in \mathcal{N}$ and provided* (2.1) *and Assumption 1 hold, the true realized type is $\xi_i^0 = \xi$ if and only if, for any $\widetilde{\xi} \in \Xi$ such that $\widetilde{\xi} \neq \xi$,*

$$0 = \operatorname*{plim}_{T\to\infty} g_i(\theta^0(\xi), \alpha^0(\xi))' W g_i(\theta^0(\xi), \alpha^0(\xi)) < \operatorname*{plim}_{T\to\infty} g_i(\theta^0(\widetilde{\xi}), \alpha^0(\widetilde{\xi}))' W g_i(\theta^0(\widetilde{\xi}), \alpha^0(\widetilde{\xi})) \tag{2.15}$$

where W is any symmetric positive definite matrix that does not depend on Ξ[3].

This lemma implies the conditional distribution of the true types given the time series $w_i = \{w_{it}\}_{t\in\mathcal{T}}$ is a degenerate distribution around the function:

$$F(w_i; \theta^0, \alpha^0) = \operatorname*{argmin}_{\xi\in\Xi} \operatorname*{plim}_{T\to\infty} \left\| g_i(\theta^0(\xi), \alpha^0(\xi))\right\|^2 \tag{2.16}$$

where the weight matrix W is dropped since the value of the minimum is zero due to the moment condition being satisfied. In view of this, let $\nu^0(\xi|w_i) = \delta(\xi - F(w_i; \theta^0, \alpha^0))$, where δ

[3] The weight matrix may depend on the type if errors are heteroskedastic with respect to the type. In this case, identification restrictions may need to be stronger, see Rivero (2023) for an example in the discrete case.

is the the Dirac delta function defined by the property $\int h(x)\delta(x)\,dx = h(0)$ for continuous h with compact support or rapidly shrinking tails[4].

The density ν^0 can be derived from this generalized function under some additional smoothness conditions on the moment functions g and on the parameters (θ^0, α^0). Such conditions guarantee a differentiable F via the implicit function theorem and allow for a change-of-variables to apply, which defines a push forward mapping from the observables to the unobserved types.

Assumption 2 (Smoothness)**.** *The following hold:*

a. The set $\Xi \subset \mathbb{R}$ is connected and compact and its interior $\mathring{\Xi}$ is the support of ν^0.

b. The infinite dimensional parameters are smooth functions on the type space: there exists a constant $M > 0$ such that for all $\xi \in \mathring{\Xi}$ we have $\left\|\frac{\partial^2\theta^0(\xi)}{\partial\xi^2}\right\| < M$ and $\operatorname*{plim}_{T\to\infty} T^{-1}\left\|\frac{\partial^2\alpha^0(\xi)}{\partial^2\xi}\right\|^2 < M$, $\left\|\frac{\partial\theta^0(\xi)}{\partial\xi}\right\| > 0$, and $\operatorname*{plim}_{T\to\infty} T^{-1}\left\|\frac{\partial\alpha^0(\xi)}{\partial\xi}\right\|^2 > 0$.

c. The function $g(w, \theta^0(\xi), \alpha^0(\xi))$ has bounded second derivatives with respect to θ^0, α^0, and w.

d. For any $i \in \mathcal{N}$, if $\xi = F(w, \theta^0, \alpha^0)$ then

$$\operatorname*{plim}_{T\to\infty}\left\|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))}{\partial\theta^0}\cdot\frac{\partial\theta^0(\xi)}{\partial\xi}+\frac{\partial g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))}{\partial\alpha_t^0}\cdot\frac{\partial\alpha_t^0(\xi)}{\partial\xi}\right\|^2 > 0$$

e. There exists constants $a > 0$ and $b > 0$ such that $\mathbb{P}(\|w_{it}\| > m) \le e^{1-(\frac{m}{b})^a}$ for all $(i,t) \in \mathcal{N} \times \mathcal{T}$ and $m > 0$.

Assumption 2(a) requires that the range of F is the interior of Ξ, excluding boundary cases for minima. Assumption 2(b) is a collection of smoothness conditions on the parameters. Bounded second derivatives lean on the interpretation that individuals with similar types will display similar parameter values. First derivatives bounded from zero imply that the parameters are smooth (regular) curves parametrized by the types $\xi \in \Xi$, so geometrically they will not halt or retrace themselves locally. Assumption 2(c) ensures that F

[4] An example is the Gaussian density, which is itself an example of a Schwartz function that has rapidly decreasing derivatives.

is sufficiently smooth and Assumption 2(d) is a convexity condition enabling F to be of a unique minimizer, i.e., injective. Assumption 2(e) imposes a faster-than-any-polynomial tail decay property on observables, which satisfies the property associated to the Dirac delta function so the marginal density ν^0 is well-defined by the change-of-variables. These properties are sufficient for F to be differentiable and a change-of-variables can be applied.

Theorem 1. *Suppose that Assumption 1 and 2 holds. Then, the type density is given as*

$$\nu^0(\xi) = \int_{\mathcal{W}} f(w_i)\delta(\xi - F(w_i;\theta^0,\alpha^0))\,dw_i. \tag{2.17}$$

Theorem 1 identifies the density of types using the individual's type rule (2.15) and smoothness of parameters. It is important to verify on a case-by-case basis that Assumptions 1 and 2 hold. For the endogenous linear model (2.5) the moment functions are known to be sufficiently smooth, but conditions for injectivity (Assumption 1) must be found.

Example 1 (Continued)**.** With knowledge of the true parameters θ^0 and α^0, the goal is to write conditions that ensure Assumption 1 holds. Using (2.9) and the model definition (2.5) substituted in for y_{it}, let $\xi, \widetilde{\xi} \in \Xi$ and $\widetilde{\xi} \neq \xi$, and assume without loss of generality that $\xi_i^0 = \xi$. Consider the following:

$$\begin{aligned}
&g_i(\theta^0(\widetilde{\xi}),\alpha^0(\widetilde{\xi}))'Wg_i(\theta^0(\widetilde{\xi}),\alpha^0(\widetilde{\xi})) - g_i(\theta^0(\xi),\alpha^0(\xi))'Wg_i(\theta^0(\xi),\alpha^0(\xi))\\
&= \left\| W^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(u_{it} + x_{it}'\left(\theta^0(\widetilde{\xi}) - \theta^0(\xi)\right) + \left(\alpha_t^0(\widetilde{\xi}) - \alpha_t^0(\xi)\right)\right)\right\|^2 - \left\| W^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}u_{it}\right\|^2\\
&= \left\| W^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\widetilde{\xi}) - \theta^0(\xi)\right) + \left(\alpha_t^0(\widetilde{\xi}) - \alpha_t^0(\xi)\right)\right)\right\|^2 + o_p(1)
\end{aligned}$$

where the $o_p(1)$ term arises from the fact that the instruments are weakly exogenous so cross-terms will vanish asymptotically as T tends to infinity and W is positive definite so there exists a matrix $W^{1/2}$ such that $W = W^{1/2}W^{1/2}$.

We see that this vector must be asymptotically bounded away from the origin whenever $\xi \neq \widetilde{\xi}$. Indeed, letting $c_W > 0$ denote the minimum eigenvalue of W, we can find the lower bound

$$c_W \operatorname*{plim}_{T\to\infty} \left\| \frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\widetilde{\xi}) - \theta^0(\xi)\right) + \left(\alpha_t^0(\widetilde{\xi}) - \alpha_t^0(\xi)\right)\right)\right\|^2 + o_p(1) \tag{2.18}$$

This is precisely the difference in moment functions of Assumption 1 after applying exogeneity of instruments and reveals the need for a "separability" condition between types constructed from the norm (2.18) and moment functions g.

Assumption 3. *There exists a function $C : \Xi \times \Xi \to [0, \infty)$ such that for any $i \in \mathcal{N}$:*

$$\underset{T\to\infty}{\text{plim}} \left\| \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(x_{it}' \left(\theta^0(\widetilde{\xi}) - \theta^0(\xi) \right) + \left(\alpha_t^0(\widetilde{\xi}) - \alpha_t^0(\xi) \right) \right) \right\|^2 \geq C(\xi, \widetilde{\xi}) \tag{2.19}$$

and $C(\xi, \widetilde{\xi}) = 0$ if and only if $\xi = \widetilde{\xi}$.

Assumption 3 satisfies Assumption 1 and is similar to group separability in the group heterogeneous coefficients case of the Appendix of Bonhomme and Manresa [2015][5]. This separability condition is a departure from the others in that it involves a continuous variable and is determined by the time series dependence between observables and type fixed effects. Otherwise, it also rules out perfect collinearity between covariates, instruments and the type fixed effects.

I have shown for arbitrary $\widetilde{\xi} \in \Xi$ for which $\xi \neq \widetilde{\xi}$ that the difference between GMM criterions is bounded away from zero and zero only when $\widetilde{\xi} = \xi = \xi_i^0$, for fixed $i \in \mathcal{N}$. Hence each individuals type is identifiable based on their individual time series and the type density is identified via Assumption 2. Finally, if we know the types we can extract the parameters and, on the other hand, if we know the parameters we can find each individuals type and type density. This defines a system that uniquely determines parameters and types.

The identification conditions do not immediately rule out lagged/forwarded outcomes or time-invariant covariates. These full rank conditions require that covariates and instruments must display sufficient variation *within types* across individuals since the expectation is taken with respect to the cross-sectional dimension unlike in the within-individual transformation of ordinary fixed effects, which transforms variables into functions of the entire time series. This is similar to group fixed effects models that require sufficient within-group variation (Assumption 1.g.) and within-factor variation (Assumption A in Bai [2009]), although they are specialized to their unknown factors so stronger conditions for within-type variation will be provided in Section 3. □

[5] Identification of groups in the panel data with discrete latent variable literature have required this assumption in some way or form, see Bonhomme and Manresa [2015], Cheng et al. [2019], Chetverikov and Manresa [2022], Mugnier [2022b,a].

Recalling that $\alpha^0 = \varphi$ is just-determined by (2.3), the proposed population GMM criterion function is the following:

$$Q = \operatorname*{plim}_{T\to\infty} \mathbb{E}\left[g_i(\theta(\xi_i), \varphi(\xi_i;\theta))' W g_i(\theta(\xi_i), \varphi(\xi_i;\theta))\right] \geq 0 \tag{2.20}$$

with equality if and only if $\theta = \theta^0$, $\alpha = \alpha^0 = \varphi$ and ξ_i is equal in distribution to ξ_i^0. Since ξ_i^0 has a density, we can constrain the set of possible densities to those that are absolutely continuous with respect to the Lebesgue measure and supported on the interior of Ξ. Using the law of iterated expectations we can rewrite this population objective function in terms of the conditional density $\nu(\xi|w_i)$ of candidate type random variables with respect to observables:

$$\begin{aligned} Q &= \operatorname*{plim}_{T\to\infty} \int_\Xi \mathbb{E}\left[g_i(\theta(\xi), \varphi(\xi;\theta)))' W g_i(\theta(\xi), \varphi(\xi;\theta)) | \xi_i = \xi\right] \nu(\xi)\, d\xi & (2.21) \\ &= \operatorname*{plim}_{T\to\infty} \mathbb{E}\left[\int_\Xi g_i(\theta(\xi), \varphi(\xi;\theta))' W g_i(\theta(\xi), \varphi(\xi;\theta)) \nu(\xi|w_i)\, d\xi\right]. & (2.22) \end{aligned}$$

which is zero at the true values of the parameters and $\nu(\xi|w_i) = \nu^0(\xi|w_i)$ as defined in (2.16) since Ξ is compact making this integral well-defined.

2.3. **The Estimator.** To define a sample criterion function from (2.22) I smooth the Dirac mass $\nu^0(\xi|w_i) = \delta(\xi - F(w_i;\theta^0,\alpha^0))$. Let $h > 0$ be a bandwidth that may depend on N and T and let K be a symmetric continuously differentiable density function. Let $\{\widehat{W}_i\}_{i\in\mathcal{N}} \subset \mathbb{R}^{\ell\times\ell}$ be a collection of positive definite weight matrices that may depend on the observables of the sample. I define the one-step type fixed effects GMM (TFE-GMM) estimator as the solution to

$$\left(\widehat{\theta}, \widehat{\mu}\right) = \operatorname*{argmin}_{(\theta,\mu)\in\Theta\times\Xi^N} \frac{1}{N}\sum_{i=1}^N \int_\Xi g_i(\xi;\theta,\mu)' \widehat{W}_i g_i(\xi;\theta,\mu) K_h(\xi - \mu_i)\, d\xi \tag{2.23}$$

where $g_i(\xi;\theta,\mu) = T^{-1}\sum_{t=1}^T g(w_{it};\theta(\xi),\varphi_t(\xi;\theta,\mu))$, $K_h(\xi-\mu) = h^{-1}K((\xi-\mu)/h)$, and for each $t \in T$ the function h_t solves

$$\sum_{i=1}^N \rho_t(w_{it};\theta(\xi),\varphi_t(\xi;\theta,\mu)) \frac{K_h(\xi-\mu_i)}{\sum_{j=1}^N K_h(\xi-\mu_j)} = 0. \tag{2.24}$$

This is the sample analogue of (2.22) except with the type fixed effects parameters replaced by using their moment condition (2.2). This estimator is asymptotically equivalent to the one without making this substitution, which will be used for the asymptotic theory in Section 3. This objective function is similar to the individual GMM criterion of Fernández-Val and Lee [2013] and Cheng et al. [2019] as a cross-sectional conditional average of individual time series GMM criterions, however it is conditioned on a parameter to be estimated. This form of objective function also explicitly separates each individual's GMM criteria to estimate $\theta^0(\xi_i^0)$ for all $i \in \mathcal{N}$, but places weights on each through the type kernel to locally estimate the type fixed effects and the type-specific coefficients with those individuals who are close in the type space. Because of this, μ can be thought of as a location parameter–they cluster individuals with similar types to use their similarities in estimation of the type-dependent parameters.

Example 1 (Continued)**.** The TFE-GMM estimator for the model parameters of (2.5) is defined as the solution to

$$\min_{(\theta,\mu)\in\Theta\times\Xi^N} \frac{1}{N}\sum_{i=1}^N \int_\Xi \left(\sum_{t=1}^T \left(\widetilde{y}_{it} - \widetilde{x}_{it}'\theta(\xi)\right)' z_{it}'\right) \widehat{W}_i \left(\sum_{t=1}^T z_{it}\left(\widetilde{y}_{it} - \widetilde{x}_{it}'\theta(\xi)\right)\right) K_h(\xi-\mu_i)\,d\xi. \tag{2.25}$$

where the within-type transformed variables are formed by

$$\widetilde{d}_{it} = \widetilde{d}_{it}(\xi,\mu) = d_{it} - \frac{1}{N}\sum_{i=1}^N d_{it}\frac{K_h(\xi-\mu_i)}{\sum_{j=1}^N K_h(\xi-\mu_j)}, \quad d\in\{x,y\} \tag{2.26}$$

where the local constant Nadaraya-Watson estimator appears as a plug-in for the conditional expectation given types (2.12). In principle any local polynomial estimator could be used instead, however I consider the simplest example for discussion and computational ease.

Using any kernel K, the first-order conditions for $\widehat{\theta}$ can be taken by using the fact Ξ is an interval subset of $\mathbb{R}$ yielding Euler equations that give us

$$\widehat{\theta}(\xi;\mu) = \left[\sum_{i=1}^N \left(\sum_{t=1}^T z_{it}\widetilde{x}_{it}'\right)' \widehat{W}_i \left(\sum_{t=1}^T z_{it}\widetilde{x}_{it}'\right) K_h(\xi-\mu_i)\right]^{-1} \tag{2.27}$$
$$\times \sum_{i=1}^N \left(\sum_{t=1}^T z_{it}\widetilde{x}_{it}'\right)' \widehat{W}_i \left(\sum_{t=1}^T z_{it}\widetilde{y}_{it}\right) K_h(\xi-\mu_i)$$

revealing that this GMM estimator is a local estimator based on proximity of observations controlled by μ to the type value ξ. The first-order conditions for μ require interchangeability of differentiation and integration, which is satisfied by the smoothness properties of Assumption 2. With a Gaussian kernel function, the partial derivative of the sample GMM criteria $\widehat{Q}$ for any $j \in \mathcal{N}$ is

$$\frac{\partial \widehat{Q}(\theta,\mu)}{\partial \mu_j} = \frac{1}{Nh^2} \int_{\Xi} g_j(\xi;\theta,\mu)' \widehat{W}_i g_j(\xi;\theta,\mu)\,(\xi-\mu_j)\,K_h(\xi-\mu_j)\,d\xi \tag{2.28}$$

$$-\frac{2}{Nh^2} \sum_{i=1}^{N} \int_{\Xi} \left(\frac{1}{T} \sum_{t=1}^{T} z_{it} \left(\widetilde{y}_{jt} - \widetilde{x}_{jt}' \theta(\xi) \right) \right)' \widehat{W}_i g_i(\xi;\theta,\mu)\,(\xi-\mu_j)\,\frac{K_h(\xi-\mu_j)K_h(\xi-\mu_i)}{\sum_{i=1}^{N} K_h(\xi-\mu_i)}\,d\xi. \tag{2.29}$$

□

2.4. **Computation.** For all $i \in \mathcal{N}$ denote

$$\widehat{Q}_i(\xi,\theta,\mu) = g_i(\xi;\theta,\mu)' \widehat{W}_i g_i(\xi;\theta,\mu) K_h(\xi-\mu_i) \tag{2.30}$$

as the individual's time series GMM criteria. To calculate $\left(\widehat{\theta},\widehat{\mu}\right)$ we can use a first order method such as gradient descent provided the gradient of the objective function has a closed-form. We can write a smoothed version of the marginal density of types as follows:

$$\widehat{\nu}(\xi;\mu) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{\xi-\mu_i}{h}\right). \tag{2.31}$$

Then we can write the gradient in terms of a sample average of expectations with respect to the type:

$$\nabla \widehat{Q}(\theta,\mu) = \frac{1}{N} \sum_{i=1}^{N} \int_{\Xi} \frac{\nabla \widehat{Q}_i(\xi,\theta,\mu)}{\widehat{\nu}(\xi;\mu)}\,d\widehat{\nu}(\xi;\mu). \tag{2.32}$$

We cannot use an ordinary gradient descent algorithm since this gradient depends on an integration and the integral calculation depends on the support and density of an unobserved variable. Therefore I turn to a stochastic or online gradient descent approach where samples of this unobserved variable are taken iteratively.

I propose a *double-online* gradient descent where we sample a single type from $\widehat{\nu}^{(s)}$ and a batch of observations w_i at each iterate. In the case of the type, I am using the most crude approximation of the expectation and rely on large numbers of iterations to approximate the gradients. For identification and the asymptotic theory, the type space Ξ was assumed to be known, but in practice it is likely to be unknown. I introduce an additive penalty

$$\sqrt{N^{-1} \sum_{i=1}^{N} \mu_i^2}$$

in the gradient descent to penalize types from being excessively spread out, placing a constraint on their sample variance. I also constrain the types to have sample mean 0, making this an example of penalized SGD where the constraints are expectation constraints on the first and second moments; see Xiao [2019]. This is purely a computational device and not studied in the asymptotic theory, where it is assumed that the type space is known. This empirical correction can be justified by noticing that the types only enter the model through the parameters and only their relative locations matter in the calculations. Therefore there is no need to precisely define the support to obtain estimates, but one must ensure that there are bounds so that estimated types do not travel arbitrarily far away causing numerical instability. The following is the basic algorithm.

Algorithm 1 (SGD for TFE-GMM)**.** *Devise a learning rate schedule η, penalty parameter $\lambda > 0$, convergence threshold $\kappa > 0$, and bandwidth $h > 0$. Initialize $\mu^{(0)}$ randomly and set $s \leftarrow 0$ and $\overline{\mu}^{(s)} = 0$.*

(1) Sample: A type value $\xi^{(s)} \sim \widehat{\nu}(\cdot; \mu^{(s)})$ and a time series w_i randomly from $\mathcal{N}$.

(2) Update parameters: at the sampled value $\xi^{(s)}$ and with all of the observations:

$$\theta^{(s)} = \theta^{(s)}(\xi^{(s)}) \longleftarrow \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \frac{\widehat{Q}_i(\xi^{(s)}, \theta, \mu^{(s)})}{\widehat{\nu}(\xi^{(s)}; \mu^{(s)})}$$

(3) Update types:

$$\mu^{(s+1)} \longleftarrow \mu^{(s)} + \eta^{(s)} \frac{\nabla_\mu \widehat{Q}(\xi^{(s)}, \theta^{(s)}, \mu^{(s)})}{\widehat{\nu}(\xi^{(s)}; \mu^{(s)})} + \eta^{(s)} \lambda \frac{\mu^{(s)}}{\sqrt{N^{-1} \sum_{i=1}^{N} (\mu_i^{(s)})^2}}$$

and update the Ruppert-Polyak averages:

$$\overline{\mu}^{(s+1)} \longleftarrow \left(\frac{s-1}{s} \right) \overline{\mu}^{(s)} + \frac{1}{s} \mu^{(s+1)}.$$

(4) Check: if $\left\| \mu^{(s)} - \mu^{(s+1)} \right\| < \kappa$, then stop and report $\overline{\mu}^{(s+1)}$. Otherwise, set $s \leftarrow s+1$ and go to step 1.

In the case of the endogenous linear model (2.5), Step 2 can be simplified by updating via (2.28). The basic algorithm does not guarantee a global minima and several modifications exist to improve performance. Early stopping and patience techniques can be used to prevent the algorithm from overfitting, for example, the patience parameter controls how long the algorithm will continue to search after finding a local minima; for more on early stopping, see Prechelt [1998]. Ruppert-Polyak averages tend to improve the rate of convergence of the algorithm (Ruppert [1988], Polyak and Juditsky [1992]) and statistical properties (Lee et al. [2023]), although I do not study the asymptotic properties of the estimator resulting from this algorithm; see Lee et al. [2023] for a discussion. Instead of sampling a single w_i, a batch (subsample) of them can be used to reduce the noise in the estimates and improve stability in exchange for a loss in CPU time.

When the objective function is not strictly convex, as in the case of TFE-GMM with respect to μ, there are no guarantees that there will be convergence to the global minima (Kiwiel [2001]). Therefore initialization and the selection of the learning rate schedule are crucial for good performance. I follow a multi-start technique by running the algorithm many times with different initial values and then choosing the result with the smallest minima (Hu et al. [2009], Martí et al. [2016], Ahuja et al. [2020]). I find that initializing the location parameters near estimates of the individual fixed effects from running a regression provides a spread of types that seems to lead to good performance of the algorithm. To set the learning rate schedule, I use the Adaptive Moment Estimation (Adam) class of learning rates (Kingma and Ba [2014]) that combines elements of momentum-based rules that remembers previous updates to keep updates tending in the same direction (Rumelhart et al. [1986]) and Adaptive gradient (AdaGrad) (Duchi et al. [2011]) or root mean square propagation (RMSProp) where the learning rate schedule is adapted to each parameter and decreasing over iterates to decrease the influence of older updates. See Spall [2005] for more classical refinements to the standard algorithm.

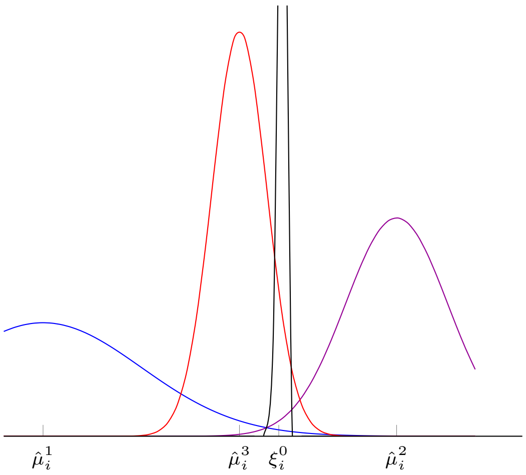


Figure 1. The type estimates $\widehat{\mu}_i^1, \widehat{\mu}_i^2, \widehat{\mu}_i^3$ for ξ_i^0 are representative of a growing sample NT and decreasing bandwidth h. As the sample size grows, the Gaussian kernels concentrate more on their means, i.e., the type estimates while also approaching the true type ξ_i^0 since the integrals in the objective (3.1) will also concentrate on the population moment conditions.

3. Asymptotic Theory: Consistency

Conditions for consistency will be given for the TFE-GMM estimator of the type-specific parameters of the endogenous linear model (2.5) in the case where Ξ is known[6]. I consider the cross-sectional and time dimensions N and T approaching infinity. The bandwidth not only controls the approximation of the type conditional expectations, but also controls the concentration within the integrand directly leveraging the weak convergence property of the kernel function to the Dirac mass. Therefore there must be care in choosing the rate at which h tends to zero.

3.1. **Sketch for consistency.** Recall the definition of the estimator

$$\min_{(\theta,\alpha,\mu)\in\Theta\times\mathcal{A}^T\times\Xi^N}\sum_{i=1}^{N}\int_{\Xi}\left(\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta(\xi)-\alpha_t(\xi)\right)'z_{it}'\right)\widehat{W}_i\left(\sum_{t=1}^{T}z_{it}\left(y_{it}-x_{it}'\theta(\xi)-\alpha_t(\xi)\right)\right)K_h(\xi-\mu_i)\,d\xi \tag{3.1}$$

and the sufficient statistic for some individual i's type:

$$\xi_i^0=F(w_i;\theta^0,\alpha^0)=\operatorname*{argmin}_{\xi\in\Xi}\ \operatorname*{plim}_{T\to\infty}\left\|\frac{1}{T}\sum_{t=1}^{T}g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))\right\|^2 \tag{3.2}$$

where $g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))=z_{it}\left(y_{it}-x_{it}'\theta^0(\xi)-\alpha_t^0(\xi)\right)$.

In the ideal scenario where we have the appropriate conditions, as the dimensions and bandwidth tend to their limits, the sequence of minimization value functions tend to their population counterpart. In particular, as h gets smaller the kernel density, e.g. Gaussian with variance h^2, in the integrand becomes tall and narrow, concentrating on the type estimates μ_i for $i\in\mathcal{N}$ that approximate the sample moment conditions closest to 0. All the while the number of estimated types μ_i becomes large and covers more of the type space so the local constant estimators approach the conditional expectations forming the type fixed effects (2.12) and similarly for θ^0 (2.8). For sufficiently small h and large N,T, the integrands then concentrate around the sufficient statistic (3.2) in the type space, making $\widehat{\mu}$ defined as the minimizer of the GMM criteria a good approximation of the true types. See Figure 1 for an illustration of consistency of the type estimators.

[6] In the case where Ξ is unknown, consistency with respect to the Fréchet distance might be considered which accounts for all possible reparametrizations of the parameter curves (θ,α). This is essentially accounting for different type spaces as different indexing sets for these curves. Reparametrization is similar to relabeling of discrete groups, except that the reparametrizations must preserve smoothness properties of the curves.

3.2. **Consistency of type-specific parameters.** Define a norm for a vector-valued function $\psi : I \to \mathbb{R}^K$ for $K \geq 1$ and $I \subset \mathbb{R}$ compact as

$$\|\psi\|_2 = \left(\int_I \|\psi(u)\|^2 \, du \right)^{1/2} . \tag{3.3}$$

where the inner norm is the Standard Euclidean norm for vectors.

Assumption 4. *There exists $M > 0$ such that*

a. There exists a collection of non random positive definite matrices $\{W_i\}_{i \in \mathcal{N}} \subset \mathbb{R}^{\ell \times \ell}$ such that $\max_{i \in \mathcal{N}} \left\| \widehat{W}_i - W_i \right\| \to_p 0$ for some suitable matrix norm $\|\cdot\|$ and denoting the minimum eigenvalue among W_i for all $i \in \mathcal{N}$ as $\widehat{\tau}$, then $\widehat{\tau} \to \tau > 0$ as $N, T \to \infty$.

b. The parameter spaces are of compactly supported, bounded functions: $\Theta = \{\theta : \Xi \to \mathbb{R} : \|\theta(\xi)\|_2 < \infty, \text{ for all } \xi \in \Xi\}$ and $\mathcal{A} = \{\alpha : \Xi \to \mathbb{R} : |\alpha(\xi)| < \infty, \text{ for all } \xi \in \Xi\}$.

c. For all $(i,t) \in \mathcal{N} \times \mathcal{T}$, $\mathbb{E}\left[\|z_{it}x'_{it}\|\right] \leq M$, $\mathbb{E}\left[\|z_{it}\|^2\right] \leq M$, $\mathbb{E}[u_{it}] = 0$ and $\mathbb{E}\left[u_{it}^4\right] \leq M$.

d. For all $i \in \mathcal{N}$, $\left| \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}\left[z'_{it} z_{is} u_{it} u_{is}\right] \right| \leq M$.

e. The density ν^0 is twice continuously differentiable on its support and the kernel function K must satisfy: $\int_{\mathbb{R}} [K(u)]^2 \, du \leq M$ and $\int f(u) K_b(u - x) \, du \to f(x)$ as $b \to 0^+$ where f is any function that is continuous on a compact domain or a Schwartz function.

f. Let $Z_{it} \in \mathbb{R}^K$ denote the nonconstant elements of z_{it}. For any vector of type assignments $\mu = (\mu_1, \ldots, \mu_N) \in \Xi^N$, define $\widehat{\rho}(\mu, \xi)$ as the minimum eigenvalue of the following $(p + T) \times (p + T)$ matrix:

$$\sum_{i=1}^{N} \frac{K_h(\xi - \mu_i) K_h(\xi - \xi_i^0)}{\sum_{j=1}^{N} K_h(\xi - \mu_j) K_h(\xi - \xi_j^0)} \begin{bmatrix} \frac{1}{T} \sum_{t=1}^{T} Z_{it} x'_{it} & \frac{1}{\sqrt{T}} Z_{i1} & \frac{1}{\sqrt{T}} Z_{i2} & \cdots & \frac{1}{\sqrt{T}} Z_{iT} \\ \frac{1}{\sqrt{T}} Z_{i1} & 1 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{T}} Z_{i2} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{T}} Z_{iT} & 0 & 0 & \cdots & 1 \end{bmatrix}$$

for any $\xi \in \Xi$. Then, for all $\xi \in \Xi$, $\operatorname*{plim}_{N,T \to \infty} \min_{\mu \in \Xi^N} \widehat{\rho}(\mu, \xi) = \rho(\xi) > 0$.

Assumption $4(a)$ requires that the chosen weight matrices converge in probability to positive definite matrices. In the simulations and applications I use the identity matrix, but if θ is just-identified one could use the standard two-stage least squares weight matrix. Assumption $4(b)$ requires the parameters be bounded as functions of the types. Assumption $4(c)$ rules out non stationary processes and perfect collinearity between instruments and covariates. Assumption $4(d)$ is a weak exogeneity condition, bounding the time series dependence between instruments and errors for every individual in the sample. A simple sufficient condition would be instruments are independent from errors. Assumption $4(e)$ contains standard assumptions from kernel density estimation requiring a sufficiently smooth true density and small tailed kernel function. Additionally, only kernels that satisfy weak convergence to the Dirac delta function are permissible as it is crucial to concentrate the estimated types around the true types to obtain consistency of parameter estimators. The Gaussian kernel would satisfy these requirements along with many other commonly used kernel functions.

Assumption $4(f)$ is a relevance condition similar to Bonhomme and Manresa [2015] and their supplementary appendix for group heterogeneous parameters. It requires that z_{it} display sufficient within-type variation over time and across individuals to serve as relevant instruments for covariates x_{it}. In this sense, Assumptions $4(d, f)$ are analogous to the classic conditions for validity of a set of instruments. Notably types to do not enter the former since they are assumed exogenous of the error terms, but types can induce correlation between instruments and covariates.

The following establishes consistency with respect to the norm (3.3).

Theorem 2. *Suppose that Assumption 2 and 4 hold and $h > 0$ approaches 0 as $N, T \to \infty$ and that $Nh \to \infty$. Then, as $N, T \to \infty$,*

$$\left\|\widehat{\theta} - \theta^0\right\|_2^2 \to_p 0 \quad \textit{and} \quad \frac{1}{T}\sum_{t=1}^{T} \left\|\widehat{\alpha}_t - \alpha_t^0\right\|_2^2 \to_p 0.$$

4. Simulation Evidence

In this section, a demonstration of the finite sample properties of the TFE-GMM estimator are presented across relevant data generating processes (DGP). I assume that data is generated randomly by

$$y_{it} = \theta_1(\xi_i)x_{1it} + \theta_2(\xi_i)x_{2it} + \alpha_t(\xi_i) + u_{it} \tag{4.1}$$

$$x_{kit} = \gamma' z_{it} + k^{-1}\alpha_t(\xi_i) + v_{kit}, \quad k \in \{1,2\} \tag{4.2}$$

$$\begin{bmatrix} z_{1it} \\ z_{2it} \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha_t(\xi_i) \\ \alpha_t(\xi_i) \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right), \quad \gamma = (1,-1) \tag{4.3}$$

$$\begin{bmatrix} u_{it} \\ v_{1it} \\ v_{2it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}\right). \tag{4.4}$$

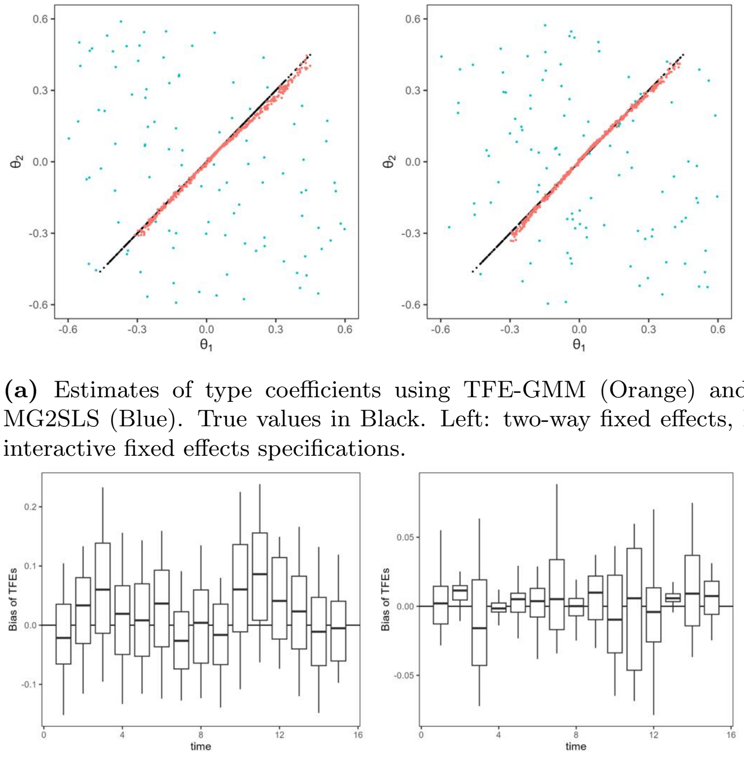
where $N(\cdot,\cdot)$ denotes the normal distribution. I emphasize that the form of this DGP implies covariates x_{it} are correlated with types via $\alpha_t(\xi_i)$ and z_{it} and therefore correlated with coefficients $\theta(\xi_i)$. The additional endogeneity and absence of heteroskedasticity partially justifies the use of the mean group two stage least squares (MG2SLS) estimators of Pesaran and Smith [1995] estimating each individual time series models separately, however it is expected that this estimator performs poorly due to the presence of type heterogeneity. I consider various specifications of type-specific coefficients, such as linear and logarithmic and type fixed effects, including traditional two-way and interactive fixed effects specifications and a dynamic AR(1) form as in the simulation of Mugnier [2022a]. Throughout I sample types from a beta distribution, $\xi_i \sim \text{Beta}(2,2) - 0.5$, reflecting the need for a compact type space. A sample of $(N,T) = (500,15)$ is taken in each simulation and this is repeated 100 times.

Let $\theta_1(\xi) = \theta_2(\xi) = \xi$ and let $\lambda_t \sim N(0, 0.25)$ be a random sample for $t = 1,\ldots,15$. Consider two specifications of TFEs: two-way fixed effects $\alpha_t(\xi_i) = \xi_i + \lambda_t$ and one factor interactive fixed effects $\alpha_t(\xi_i) = \xi_i \times \lambda_t$. Note that estimates result in $N = 500$ pairs of type coefficients and time series of TFEs and so visualization may be challenging. For estimation throughout, I take all of the weight matrices as the identity in the one-step estimation. I also initialize Algorithm 1 at the true values, set penalty to $\lambda = 0.3$ and set the bandwidth to $h = 0.073$, which is obtained from Silverman's rule-of-thumb (Silverman [1986])[7].

Figure 2a shows the geometric properties of the average of the TFE-GMM estimator for the type coefficients along with Figure 2b showing a time-varying box-and-whisker plot displaying bias properties of the average of TFE estimates. At a glance the TFE-GMM bias appears to be mild, while the MG2SLS estimates (in blue) struggle to capture the shape of the line. As for the TFEs, since we are using a local constant plug-in we can

[7] A cross-validation technique may make a more justifiable choice, but this rule-of-thumb seems to work reasonably well likely due to the symmetric and regular type density

interpret the wide range of bias coming from boundary bias, where outlier types have limited information to produce good estimates. This occurs for both coefficients and TFEs, where the coefficients do not adequately reach the boundaries of the line. This suggests that a local linear plug-in for the type conditional expectation functions may be more suitable. Figure 3 displays the statistical properties of the TFE-GMM estimator for specific features of the type-heterogeneous coefficients across various specifications[8]. Figure 5 shows the distribution of the root mean squared error of the average of type heterogeneous coefficients.



(a) Estimates of type coefficients using TFE-GMM (Orange) and with MG2SLS (Blue). True values in Black. Left: two-way fixed effects, Right: interactive fixed effects specifications.

(b) Per-time period box-and-whisker for bias of TFE estimates using TFE-GMM. Left: two-way fixed effects, Right: interactive fixed effects specifications.

Figure 2. Simulation results specifying linear relationship among type coefficients and traditional fixed effects (two-way and interactive).

For robustness to bandwidth choice I repeated some of the experiments with different bandwidths, penalty values and growing T and reported these additional results in Appendix A.

[8] The MG2SLS estimates are excluded since they are heavily biased numerically and descriptively based on the plots of Figure 2a and 4.

Notably, intuition from over and under smoothing a kernel density estimate seems to carry over to TFE-GMM and when T grows and h is kept fixed, boundary biases may improve but the line created by estimates shifts away from the true values. This indicates that consistency for type estimates relies on choosing a bandwidth that depends on both N, T.

Bias	$\mathbb{E}\left[\theta_k(\xi_i)\right]$		$\mathrm{Var}\left(\theta_k(\xi_i)\right)$	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.068	0.053	-0.01	-0.01
$\theta_1 = \theta_2$ & TWFE	0.067	0.056	-0.007	-0.009
$\theta_1 = \theta_2$ & IFE	0.059	0.056	-0.010	-0.009
$\theta_1 = \theta_2$ & AR(1), $\rho = 0.9$	0.066	0.051	-0.009	-0.007
$\log(\theta_1) = \theta_2$ & AR(1), $\rho = 0.75$	0.062	0.065	-0.006	-0.083

Figure 3. Bias of the mean and variance estimators of the type-heterogeneous coefficients over 100 simulations across different specifications. Bandwidth chosen is $h = 0.073$. Estimator performs just as well with traditional TFE specifications as with AR(1) form. The concavity of the second parameter introduces more downward bias when estimating the variance (bottom row).

For the next experiment, let $\theta_1(\xi) = \xi$ and $\theta_2(\xi) = \log(\theta_1(\xi))$ so that the curve appears as the graph of the natural logarithm. Take TFEs as an AR(1) process with idiosyncratic shocks: $\alpha_t(\xi_i) = 0.75\,\alpha_{t-1}(\xi_i) + U_{\xi_i t}$ where $U_{\xi_i t} \sim \mathrm{Unif}(-0.1, 0.1)$ is specific to $i = 1, \ldots, 500$. Figure 4 shows the geometry of the curve and average of TFE-GMM estimates. The bias is more apparent in segments with more curvature and sparsity of points, which is expected given intuition regarding concavity/curvature and its effect on statistical estimates. Figure 4 also shows moderately small biases even at the boundaries and a single example TFE drawn randomly from the sample shows that it follows the true path reasonably well.

5. Rational Addiction with Type Fixed Effects

I apply the TFE-GMM framework to estimate a RA model with type fixed effects using household cigarette purchase data. I start with a background on the testable implications of the RA model and then a description of the data, providing some initial evidence of cyclical

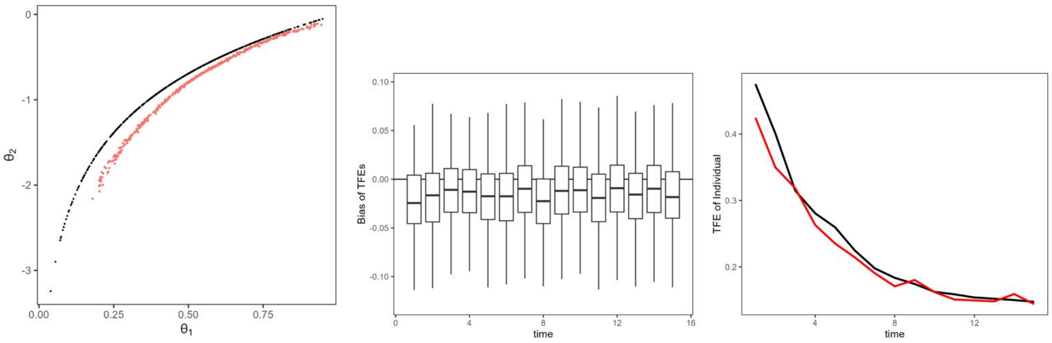


Figure 4. Left: Estimates of type coefficients using TFE-GMM (Orange). True values in Black. Middle: Per-time period box-and-whisker for bias of TFE estimates using TFE-GMM, Right: Single example of one estimated TFE (Red) against true process (Black).

RMSE	Mean		Median		25th Percentile		75th Percentile	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.110	0.107	0.105	0.010	0.086	0.086	0.138	0.134
$\theta_1 = \theta_2$ & TWFE	0.112	0.107	0.107	0.010	0.091	0.084	0.140	0.136
$\theta_1 = \theta_2$ & IFE	0.105	0.105	0.098	0.096	0.083	0.082	0.134	0.131
$\theta_1 = \theta_2$ & AR(1), $\rho = 0.9$	0.111	0.102	0.103	0.099	0.087	0.085	0.139	0.124
$\log(\theta_1) = \theta_2$ & AR(1), $\rho = 0.75$	0.106	0.171	0.099	0.122	0.086	0.101	0.127	0.225

Figure 5. Features of the root mean squared error distribution of the type-heterogeneous coefficients TFE-GMM estimates over 100 simulations across different specifications. Bandwidth chosen is $h = 0.073$. Estimator performs just as well with traditional TFE specifications as with AR(1) form. Concavity of the second parameter makes estimation more difficult, resulting in larger ranges of RMSE (bottom row).

consumption and price insensitivity. I follow with an explanation of how exogenous variation from types, type-specific parameters and dynamic prices as instruments address the endogeneity problem of the dynamic model. Lastly I provide the results of the estimation.

5.1. **Background.** The RA model is of a representative consumer that maximizes their discounted lifetime utility subject to their budget by choosing a consumption plan accompanied by an accumulation of addiction stock that also enters the inter temporal utility function. Consumption of the addictive good provides utility while addiction causes disutility, that is to say that addiction is an undesirable byproduct of consumption. Addiction is viewed as adjacent complementarity between past and current consumption reinforcing habit, and tolerance, which is modeled as diminishing marginal utility from consumption. The addiction stock also depreciates according to some positive factor so that periods of abstinence will lead to addiction vanishing. It is typical to assume a lifetime budget and quadratic utility to simplify the analysis. For complete details regarding the derivations of the RA model, see Ferguson [2000].

With these simplifying assumptions, the consumption plan is linear in past and future consumption C_t along with other explanatory variables X_t such as price, income, age, etc ,and an error term:

$$C_t = \theta_l\, C_{t-1} + \theta_f\, C_{t+1} + \beta' X_t + e_t. \tag{5.1}$$

It is common to test the null that (θ_f, θ_l) are zero against the alternative they are positive. When they are positive, it is taken that there is addiction (θ_l) and that consumers are forward-looking when deciding on consumption in the current period (θ_f). Along with positive values for the lead and lag coefficients (θ_f, θ_l), there are a few other testable implications of RA using these coefficients that are not always analyzed, but are important for our purposes. The conditions as given in Laporte et al. [2017] and Becker and Murphy [1988] are

$$\theta_l\, \theta_f < 0.25, \tag{5.2}$$

which requires that the roots of the second order difference equation associated to the optimal control problem are real-valued, and

$$\theta_l + \theta_f < 1, \tag{5.3}$$

which together with (5.2) guarantees that the consumption path exhibits saddle-point dynamics. In the event that (5.2) is violated, then the roots are complex indicating cyclical dynamics so that the consumer has periods of high consumption followed by periods of low consumption and so on.

The assumption of finite horizon has strict consequences on what kind of consumption paths a rational addict can follow, specifically for the traditional saddle-point dynamics in this application. Figure 6 shows that the finite-lived consumer facing a saddle-point equilibrium may, for example, consume a large amount initially and taper off towards the equilibrium quantity as a consequence of being mindful of the harmful effects from addiction. Towards the end of their life, they find it optimal to ramp up consumption and ignore the side effects. A similar story can be made for the bottom curve in the diagram where in their middle age they consume the most, but then recede back to early life levels of consumption. In summary, the consumer plans out their consumption in view of the harmful addiction they may develop.

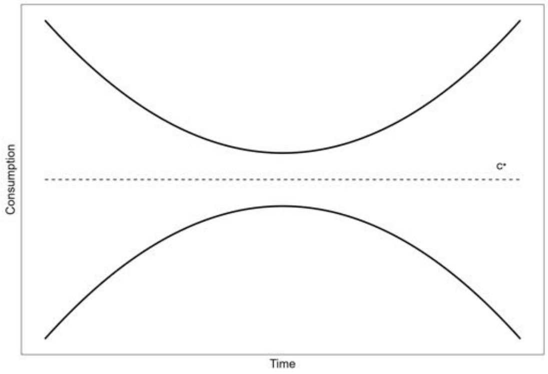


Figure 6. Two examples (top and bottom curves) of optimal consumption plan under saddle-point dynamics. At first, the consumer tends to the equilibrium quantity but over time the unstable branch dominates leading the consumer to veer off.

In micro panels, there are many individuals at different points in the life cycle and on their idiosyncratic consumption paths. Estimation of (5.1) with ordinary fixed effects regression will therefore yield weighted averages of these many different points and dynamics of consumption. To make matters worse, if indeed consumers follow a saddle-point equilibrium, the presence of the unstable root will have, for example, different age cohorts of individuals follow stable and unstable branches at different points in time. In segments where both are occurring, the unstable root prevents identification of the parameters of interest due to combinations with a non stationary process. The type augmented (1.1) is better equipped to reconcile individuals on different time paths with varying dynamics. There is no a priori reason why saddle-point dynamics is the only possibility with rational addiction so inclusion of the type fixed effect may capture a potential second addiction stock that may exhibit adjacent complementarity or substitutability generating other behaviors such as binging, which can be detected on an individual-by-individual basis by checking (5.2) and (5.3).

5.2. **Data.** The data is sourced from the Nielsen Consumer Panel that follows household weekly purchases using a survey-provided scanner. I aggregated to the year-level by summing the quantity of packs purchased and taking the average price paid per pack in the year. I ignore substitution to other goods such as e-cigarettes. I maintain a balanced panel with $N = 3,296$ households and $T = 16$ years between 2004 and 2019. The data also contains demographic information such as the household size, head of household employment, education, marital status, race, sex and state of occupancy. Income is recorded in interval ranges and values reported are from 2 years prior to the panel year.

Figure 7 presents aggregate summary statistics. At the start of the sample period nearly 80% of individuals are over 45 years old and another non inclusive 80% have some college education, either partially or fully completed, or with post graduate study. A majority of the sample (83%) are recorded as White while 8% are Black and 4% Asian. Only 5% of the sample identify as Hispanic. Figure 8 shows the location density of respondents in the USA and displays somewhat representative population density according to state size in the contiguous states.

Variable	Mean	Median	St. Error	Min	Max
Packs Purchased	20.27	14	24.78	0.20	603
Price	5.70	5.28	2.28	1.35	18.69
Income	50−59.9	60−69.9	–	< 5	> 200
Age	50−54	55−64	–	< 25	> 65

Figure 7. Descriptive statistics of the sample. Price and income in 2012 USDs. Income is in thousands and income and age are given in ranges.

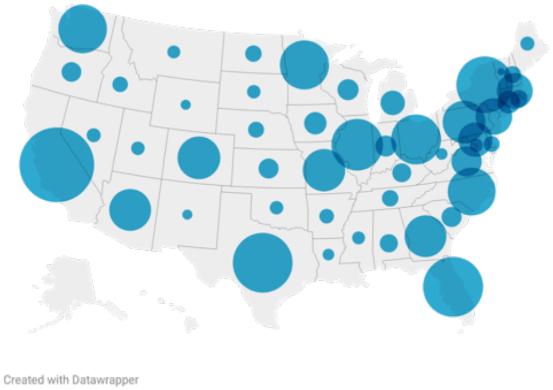


Figure 8. Size of circles indicate density of respondents in each state.

Figure 9 displays the median amount of packs purchased along with the median average unit price paid over the sample period. It appears that the law of demand holds in the aggregate, purchases of cigarettes decline as prices rise. This can also be said for most of the age cohorts as Figure 10 shows the same patterns. However, Figure 11 shows some individual's consumption profiles that appear much different than the aggregate time series. Not all time series strictly follow law of demand and some display cyclical patterns. The fixed effects two-stage least squares estimates are $(0.247, 0.369)$ for the lag and lead consumption covariates, respectively, and are significant and positive meaning that we would conclude cigarette smokers are rationally addicted and follow saddle-point dynamics since they satisfy (5.2) and (5.3). However, judging by the consumption profiles in Figure 11, there is still important unexplained variation. For instance, these estimates may represent those on saddle-point paths, but they do not represent the binging consumers.

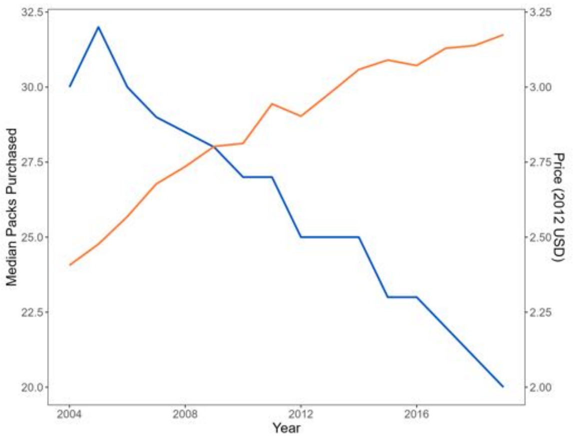


Figure 9. Blue: Median packs purchased for entire sample, Orange: Median of sample of average unit price paid. Left vertical axis follows purchases and right vertical axis follows price in 2012 USDs. Law of demand holds in the aggregate.

5.3. **Identification.** The model (1.1) contains lagged and forwarded outcomes and type heterogeneity that does not address all endogeneity concerns. First, I assume that inclusion of the type fixed effect eliminates components of the error term that would be serially correlated. The error term in this case can be interpreted as life-cycle shocks such as loss of job or other personal period-specific shocks. The type fixed effects contains the components of addiction that are not directly caused by habit captured by lag consumption, which may include a stock of health: physical or mental. Regardless of what it may be, the habit stock

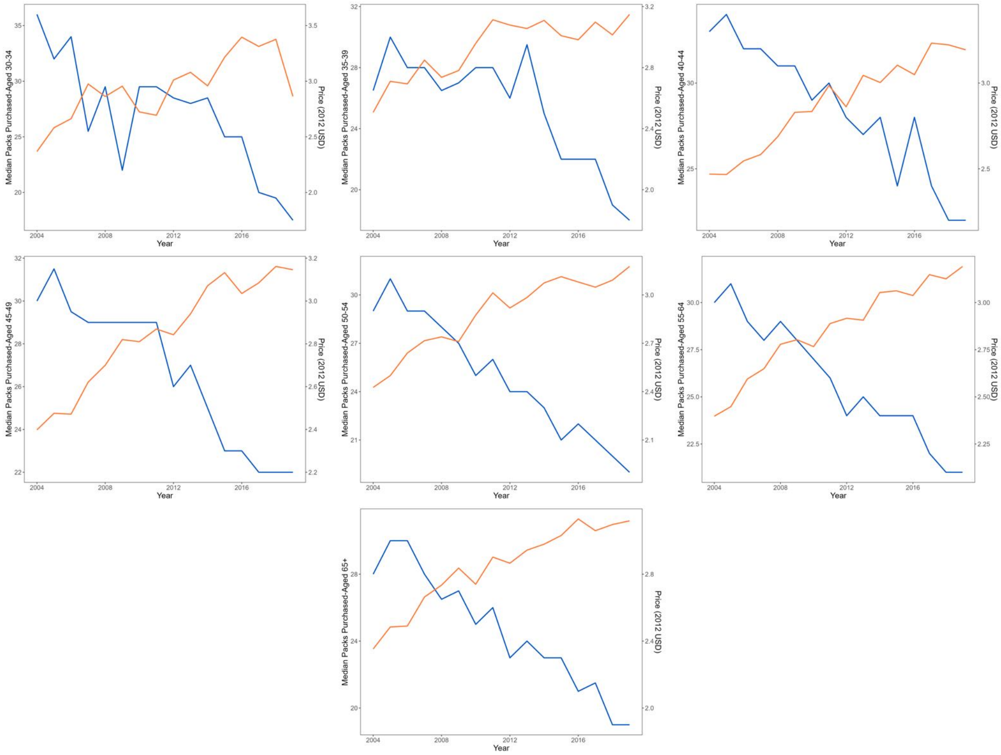


Figure 10. Blue: Median packs purchased for entire sample, Orange: Median of sample of average unit price paid. Left vertical axis follows purchases and right vertical axis follows price in 2012 USDs. Law of demand holds across age cohorts.

may be arbitrarily correlated with this stock, meaning the coefficient on past consumption will be biased if unaccounted for. Type heterogeneity is an important feature of this model to allow for different addiction profiles. For example, both stocks could be positively correlated with consumption in which case the consumer could be classified as fully addicted, borrowing the naming convention from Dockner and Feichtinger [1993]. Another example, while the habit stock is positively associated to consumption, the other could be a health stock that is negatively correlated with consumption indicating a partial addiction. These properties depend upon the realization of the type of the consumer in the sample period, putting them on these different trajectories and if type heterogeneity is ignored then estimates of the coefficient corresponding to lag consumption could be understated in the case of full addiction and overstated in the case of partial addiction. Cigarettes are highly addictive and harmful goods due to nicotine dependence (Benowitz [2010]) so it is expected that omission of type heterogeneity produces estimates that are generally downward biased.

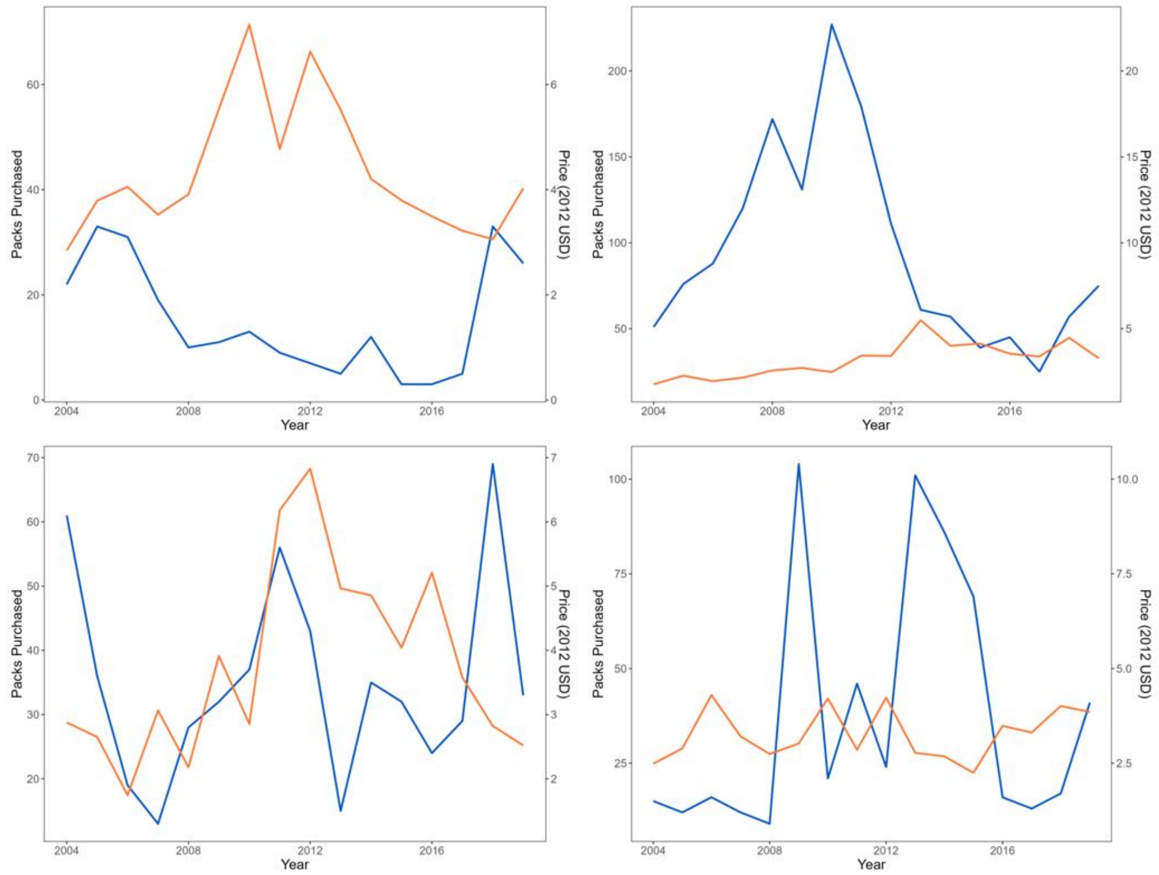


Figure 11. Blue: Median packs purchased for entire sample, Orange: Median of sample of average unit price paid. Left vertical axis follows purchases and right vertical axis follows price in 2012 USDs. Individuals follow complicated consumption paths that do not strictly satisfy the law of demand. Cycles evident in bottom row. Top row appears similar to saddle-point dynamics.

The inclusion of the lead of consumption is important in the RA model and induces endogeneity since the consumer realizes their life-cycle shock for the period and uses that information to infer the amount of cigarettes to consume for next period. The rationally addicted consumer also realizes at the beginning of the period their health stock level, how much they smoked last period, prices and income level. Then, when considering how much to smoke in the current period, they can infer the consequences of smoking today on their health stock in the next period followed by how their addiction would affect future consumption. To correct for this I follow tradition and use the lead and lag of prices and taxes as instruments for lead and lagged consumption. Justification for using future prices as instruments come from the fact that price increases on cigarettes are announced in advance due to government tax hikes and past prices on their own may be a poor instrument (Nelson and Startz [1990]). Figure 12 summarizes this discussion in a diagram.

$$\begin{tikzcd}
\alpha_{t-1}(\xi_i) \arrow[rr, bend left, red] \arrow[d] & P_{it-1} \arrow[r] & C_{it-1} \arrow[d, blue] \arrow[dll, blue] & U_{it-1} \arrow[l] \\
\alpha_t(\xi_i) \arrow[rr, bend left, red] \arrow[d] & P_{it} \arrow[r] & C_{it} \arrow[d, dashed, blue] \arrow[dll, blue] & U_{it} \arrow[l] \\
\alpha_{t+1}(\xi_i) \arrow[rr, bend left, red] & P_{it+1} \arrow[r] & C_{it+1} \arrow[u, shift right] & U_{it+1} \arrow[l]
\end{tikzcd}$$

Figure 12. Diagram illustrating identification. Blue arrows indicate direct effects from habit and red arrows indicate effects from the health stock. Dashed line indicates the additional endogeneity concern despite controlling for type heterogeneity.

5.4. **Results.** Consumption is measured per household member and all relevant values are in 2012 USDs. Figure 13 shows the estimated type-specific parameters (θ_f, θ_l) in blue, the region corresponding to violation of (5.2) in gray, and region corresponding to traditional evidence for the RA model in green. In the sample, 73% of individuals lie in the gray area, which corresponds to cyclical behavior while only 24% are what typically is taken as rational addicts. There are no explosive processes in the sample so all are on stable paths.

Figure 14 shows the absolute price and income effects on consumption of cigarettes and, given discussions in Section 4, may be bias at the boundary. In the figure 95% of the sample is shown since there are significant outliers that would prevent visualization of the concentration of effects at the origin. Most of the sample would show less than a quarter of a pack drop or increase in demand for a dollar increase in price indicating a practically insignificant effect. This implies policy that targets consumption through price interventions such as a sin tax would effectively raise revenue, but may do little at the individual level to curb consumption among entrenched smokers.

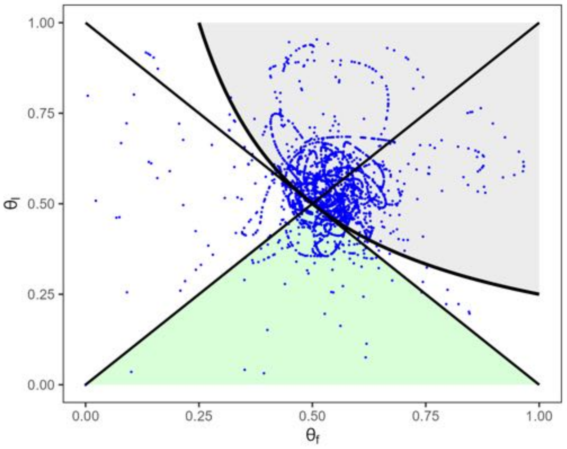


Figure 13. Blue: Estimates of the $N = 3,296$ type heterogeneous lag and lead coefficients using bandwidth $h = 0.128$. The gray region is associated to cyclical dynamics, while the green is traditional saddle-point dynamics of rational addiction.

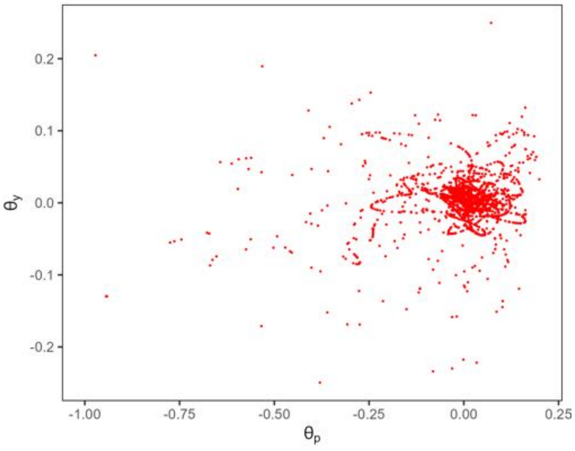


Figure 14. Estimates of the $N = 3,296$ type heterogeneous price and income coefficients using bandwidth $h = 0.128$. Only 95 % of the sample is shown. Values appear to be practically concentrated near the origin indicating weak sensitivity to changes in price and earnings.

The binging behavior of individuals in this sample provide support for educational intervention programs and their insensitivity to price changes could raise revenue for these programs. Examples of cost-effective policies could be further emphasizing that health care providers inform patients that smoking cessation will increase life expectancy or treat an illness (Yu et al. [2004]) or focus on cognitive or interpersonal therapy (Stenberg et al. [2018]).

Figure 15 shows 8 individual's consumption profiles and type fixed effects. A feature of these plots is that periods of intense consumption are preceded by periods of low value of TFE. In some cases, they move in opposite directions just before the dramatic spike. Spikes in TFE usually lead to drops in consumption in the next period. Overall, the movement between the two curves is not always positively associated which indicates that the unobserved components of addiction have a complicated and different effect for each individual. Therefore it is not always possible to interpret the TFE as a single force that strictly exhibits either adjacent complementarity or substitutability. Nonetheless, the TFE RA model is a powerful tool to analyze consumption of addictive goods by controlling for complicated latent structures. Additionally, the GMM framework with type heterogeneity has potential to analyze demand with substitute nicotine products and model quitting, which I leave for future work.

References

K. Ahuja, A. Dhurandhar, and K. R. Varshney. Learning to initialize gradient descent using gradient descent, 2020.

J. D. Angrist. Treatment effect heterogeneity in theory and practice. *The economic journal*, 114(494):C52–C83, 2004.

J. D. Angrist, K. Graddy, and G. W. Imbens. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527, 2000.

M. Arellano and S. Bonhomme. Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, 79(3):987–1020, 2012.

M. Arellano and J. Hahn. Understanding bias in nonlinear panel models: Some recent developments. *Econometric Society Monographs*, 43:381, 2007.

M. Auld and P. Grootendorst. An empirical analysis of milk addiction. *Journal of Health Economics*, 23(6):1117–1133, 2004. ISSN 0167-6296. doi: https://doi.org/10.1016/j.jhealeco.2004.02.003. URL `https://www.sciencedirect.com/science/article/pii/S0167629604000414`.

J. Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009. doi: https://doi.org/10.3982/ECTA6135. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6135`.

B. Baltagi and J. M. Griffin. The econometrics of rational addiction: The case of cigarettes. *Journal of Business Economic Statistics*, 19(4):449–54, 2001. URL `https:`

//EconPapers.repec.org/RePEc:bes:jnlbes:v:19:y:2001:i:4:p:449-54.

G. Becker and K. Murphy. A theory of rational addiction. *Journal of Political Economy*, 96(4):675–700, 1988. URL https://EconPapers.repec.org/RePEc:ucp:jpolec:v:96:y:1988:i:4:p:675-700.

G. Becker, M. Grossman, and K. Murphy. An empirical analysis of cigarette addiction. *American Economic Review*, 84(3):396–418, 1994. URL https://EconPapers.repec.org/RePEc:aea:aecrev:v:84:y:1994:i:3:p:396-418.

N. L. Benowitz. Nicotine addiction. *New England Journal of Medicine*, 362(24):2295–2303, 2010.

C. A. Bester and C. B. Hansen. Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 190(1):197–208, 2016. URL https://EconPapers.repec.org/RePEc:eee:econom:v:190:y:2016:i:1:p:197-208.

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

S. Bonhomme, T. Lamadon, and E. Manresa. Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643, 2022.

J. F. Bonnans and A. Shapiro. Perturbation analysis of optimization problems. In *Springer Series in Operations Research*, 2000. URL https://api.semanticscholar.org/CorpusID:118751821.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

K. M. Carroll. The profound heterogeneity of substance use disorders: Implications for treatment development. *Current Directions in Psychological Science*, 30(4):358–364, 2021. doi: 10.1177/09637214211026984. URL https://doi.org/10.1177/09637214211026984.

J. Cawley and C. J. Ruhm. Chapter three - the economics of risky health behaviors. In M. V. Pauly, T. G. Mcguire, and P. P. Barros, editors, *Handbook of Health Economics*, volume 2 of *Handbook of Health Economics*, pages 95–199. Elsevier, 2011. doi: https://doi.org/10.1016/B978-0-444-53592-4.00003-7. URL https://www.sciencedirect.com/science/article/pii/B9780444535924000037.

F. Chaloupka. Rational addictive behavior and cigarette smoking. *Journal of Political Economy*, 99(4):722–42, 1991. URL https://EconPapers.repec.org/RePEc:ucp:jpolec:v:99:y:1991:i:4:p:722-42.

G. Chamberlain. Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47(1):225–238, 1980. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2297110.

L. Chang-Ching and N. Serena. Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown. *Journal of Econometric Methods*, 1(1):1–14, August 2012. doi: 10.1515/2156-6674.1000. URL https://ideas.repec.org/a/bpj/jecome/v1y2012i1p14n1.html.

X. Cheng, F. Schorfheide, and P. Shao. Clustering for multi-dimensional heterogeneity. 2019.

D. Chetverikov and E. Manresa. Spectral and post-spectral estimators for grouped panel data models. *arXiv preprint arXiv:2212.13324*, 2022.

E. J. Dockner and G. Feichtinger. Cyclical consumption patterns and rational addiction. *The American Economic Review*, 83(1):256–263, 1993.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

J. Fan and J.-T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(2):303–322, 2000.

B. S. Ferguson. Interpreting the rational addiction model. *Health Economics*, 9(7):587–598, 2000. doi: https://doi.org/10.1002/1099-1050(200010)9:7⟨587::AID-HEC538⟩3.0.CO;2-J.

I. Fernández-Val and J. Lee. Panel data models with nonadditive unobserved heterogeneity: Estimation and inference. *Quantitative Economics*, 4(3):453–481, 2013.

E. W. Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

B. S. Graham and J. L. Powell. Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152, 2012.

J. E. Grant, M. N. Potenza, A. Weinstein, and D. A. Gorelick. Introduction to behavioral addictions. *The American Journal of Drug and Alcohol Abuse*, 36(5):233–241, 2010. doi: 10.3109/00952990.2010.491884. URL https://doi.org/10.3109/00952990.2010.491884. PMID: 20560821.

M. Grossman, F. Chaloupka, and I. Sirtalan. An empirical analysis of alcohol addiction: Results from the monitoring the future panels. *Economic Inquiry*, 36(1):39–48, 1998.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993. ISSN 00359246. URL `http://www.jstor.org/stable/2345993`.

J. Heckman and E. Vytlacil. Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, pages 974–987, 1998.

D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998.

C. Hsiao. *Analysis of panel data.* Cambridge university press, 4th edition, 2022.

X. Hu, R. Shonkwiler, and M. C. Spruill. Random restarts in global optimization. 2009.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

K. C. Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical programming*, 90:1–25, 2001.

L. Laage. A correlated random coefficient panel model with time-varying endogeneity. *arXiv preprint arXiv:2003.09367*, 2020.

A. Laporte, A. R. Dass, and B. S. Ferguson. Is the Rational Addiction model inherently impossible to estimate? *Journal of Health Economics*, 54(C):161–175, 2017. doi: 10.1016/j.jhealeco.2016.1. URL `https://ideas.repec.org/a/eee/jhecon/v54y2017icp161-175.html`.

S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast inference for quantile regression with tens of millions of observations, 2023.

Q. Li and J. S. Racine. *Nonparametric econometrics: theory and practice.* Princeton University Press, 2007.

S. Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

R. Martí, J. Lozano, A. Mendiburu, and L. Hernando. Multi-start methods. handbook of heuristics, 1–21. doi: 10.1007. 2016.

S. L. McElroy, A. I. Guerdjikova, N. Mori, M. R. Munoz, and P. E. Keck. Overview of the treatment of binge eating disorder. *CNS Spectrums*, 20(6):546–556, 2015. doi: 10.1017/S1092852915000759.

A. Mehrabani. Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 2022. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2022.12.002. URL `https://www.sciencedirect.com/science/article/pii/S030440762200207X`.

H. O. Melberg. Rational addiction theory: a survey of opinions. HERO Online Working Paper Series 2008:7, University of Oslo, Health Economics Research Programme, June 2009. URL `https://ideas.repec.org/p/hhs/oslohe/2008_007.html`.

M. Mugnier. Unobserved clusters of time-varying heterogeneity in nonlinear panel data models. *Job Market Paper*, 2022a.

M. Mugnier. A simple and computationally trivial estimator for grouped fixed effects models. *Working Paper*, 2022b.

C. Nelson and R. Startz. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business*, 63(1):S125–40, 1990. URL `https://EconPapers.repec.org/RePEc:ucp:jnlbus:v:63:y:1990:i:1:p:s125-40`.

W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

S. Nickell. Biases in dynamic models with fixed effects. *Econometrica: Journal of the econometric society*, pages 1417–1426, 1981.

N. Olekalns and P. Bardsley. Rational addiction to caffeine: An analysis of coffee consumption. *Journal of Political Economy*, 104(5):1100–1104, 1996. ISSN 00223808, 1537534X. URL `http://www.jstor.org/stable/2138954`.

M. Pesaran and R. Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68(1):79–113, 1995. URL `https://EconPapers.repec.org/RePEc:eee:econom:v:68:y:1995:i:1:p:79-113`.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

L. Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, 1998. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(98)00010-0. URL `https://www.sciencedirect.com/science/article/pii/S0893608098000100`.

J. A. Rivero. Unobserved grouped heteroskedasticity and fixed effects, 2023.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

P. M. Robinson. Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

G. R. Saunders, X. Wang, F. Chen, S.-K. Jang, M. Liu, C. Wang, S. Gao, Y. Jiang, C. Khunsriraksakul, J. M. Otto, et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*, 612(7941):720–724, 2022.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons, 2005.

U. Stenberg, A. Vågan, M. Flink, V. Lynggaard, K. Fredriksen, K. F. Westermann, and F. Gallefoss. Health economic evaluations of patient education interventions a scoping review of the literature. *Patient Education and Counseling*, 101(6):1006–1035, 2018. ISSN 0738-3991. doi: https://doi.org/10.1016/j.pec.2018.01.006. URL `https://www.sciencedirect.com/science/article/pii/S0738399118300065`.

L. Su, Z. Shi, and P. C. B. Phillips. Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264, 2016. doi: https://doi.org/10.3982/ECTA12560. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA12560`.

Y. Sun. Estimation and inference in panel structure models. *Econometrics eJournal*, 2005.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

X. Xiao. Penalized stochastic gradient methods for stochastic convex optimization with expectation constraints. *Optimization-online*, 2019.

S. Yitzhaki. On using linear regressions in welfare economics. *Journal of Business & Economic Statistics*, 14(4):478–486, 1996.

C.-M. Yu, C.-P. Lau, J. Chau, S. McGhee, S.-L. Kong, B. M.-Y. Cheung, and L. S.-W. Li. A short course of cardiac rehabilitation program is highly cost effective in improving long-term quality of life in patients with recent myocardial infarction or percutaneous coronary intervention. *Archives of Physical Medicine and Rehabilitation*, 85(12):1915–1922, 2004. ISSN 0003-9993. doi: https://doi.org/10.1016/j.apmr.2004.05.010. URL `https://www.sciencedirect.com/science/article/pii/S0003999304010883`.

Appendix A. Plots & Additional Results

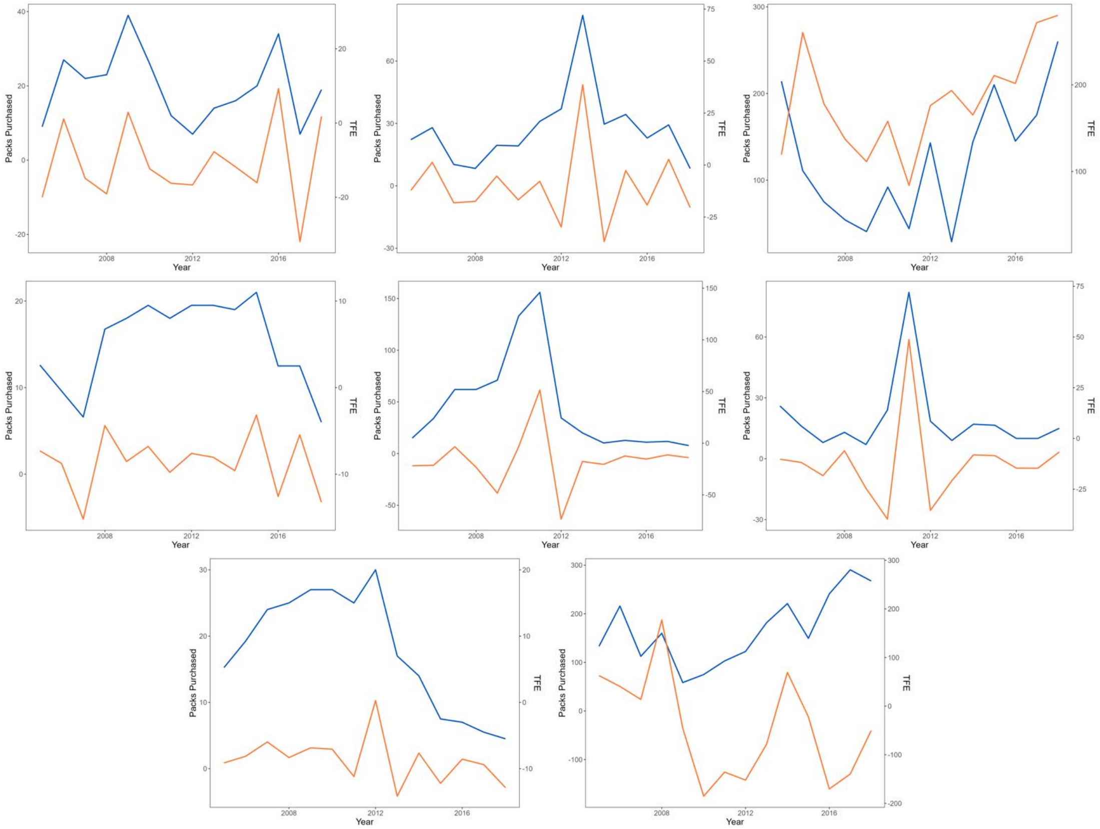


Figure 15. Blue: Packs purchased per household member, Orange: Type fixed effect estimate. Left vertical axis follows purchases and right vertical axis follows TFE. Complicated relationship between consumption and TFE, where at times they follow one another, but for others they seem moving in opposite directions. A feature of these plots is that periods of intense consumption are preceded by periods of low value of TFE. In some cases, they move in opposite directions just before the dramatic spike. Spikes in TFE usually lead to drops in consumption in the next period.

Bias	$\mathbb{E}[\theta_k(\xi_i)]$		$\text{Var}(\theta_k(\xi_i))$	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.073	0.047	-0.024	-0.024
$\theta_1 = \theta_2$ & TWFE	0.073	0.050	-0.023	-0.025
$\theta_1 = \theta_2$ & IFE	0.058	0.056	-0.025	-0.025
$\theta_1 = \theta_2$ & AR(1), $\rho = 0.9$	0.066	0.051	-0.009	-0.007
$\log(\theta_1) = \theta_2$ & AR(1), $\rho = 0.75$	0.062	0.065	-0.006	-0.083

Figure 16. Bias of the mean and variance estimators of the type-heterogeneous coefficients over 100 simulations across different specifications. Bandwidth chosen is $h = 0.15$. Larger bandwidth than the simulation presented in the body of the paper is in line with oversmoothing: estimates are more biased over most of the categories.

Bias	$\mathbb{E}[\theta_k(\xi_i)]$		$\text{Var}(\theta_k(\xi_i))$	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.072	0.049	0.014	0.015
$\theta_1 = \theta_2$ & TWFE	0.074	0.049	0.018	0.015
$\theta_1 = \theta_2$ & IFE	0.059	0.055	0.012	0.011

Figure 17. Bias of the mean and variance estimators of the type-heterogeneous coefficients over 100 simulations across different specifications. Bandwidth chosen is $h = 0.01$. The smaller bandwidth changes the direction of the bias on the variance estimator and in magnitude it is the smallest among the rest of the results.

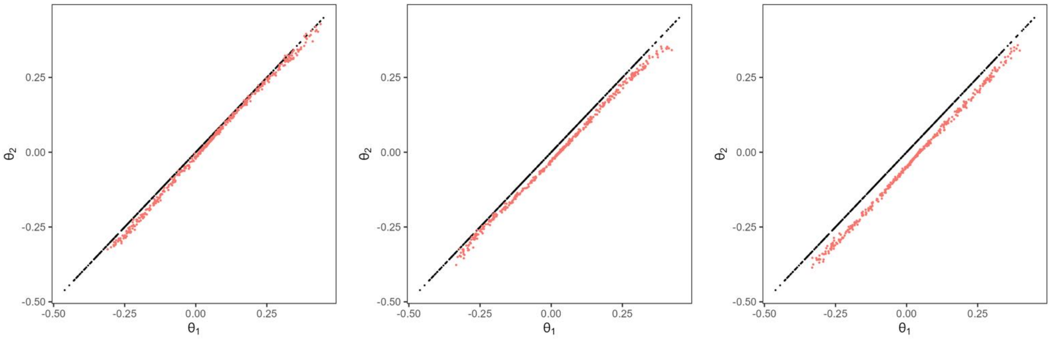


Figure 18. AR(1) TFE with coefficient 0.9. As T increases from 15, 50, 100 with $N = 100$ and $h = 0.07$ fixed. This may indicate that h must also depend on T as T tends to ∞ to ensure bias vanishes asymptotically.

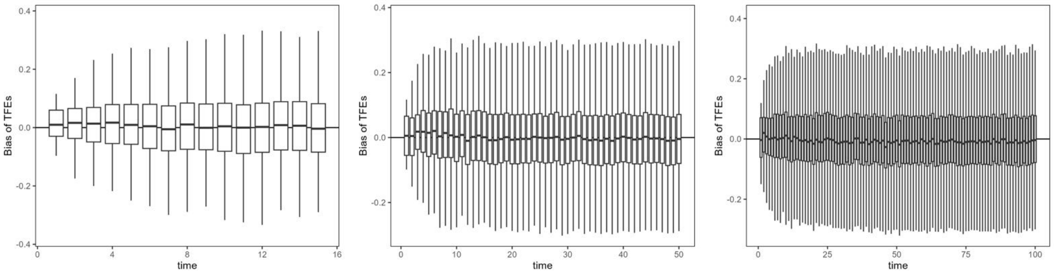


Figure 19. AR(1) TFE with coefficient 0.9. As T increases from 15, 50, 100 with $N = 100$ and $h = 0.07$ fixed. The TFE estimates do not seem to be affected by fixed N and h as T grows. This may be due to the fact that the TFEs are cross-sectional conditional averages for each $t = 1, \ldots, T$.

Appendix B. Proofs

B.1. **Proof of Theorem 1.** Recall the function

$$\xi_i^0 = F(w_i; \theta^0, \alpha^0) = \underset{\xi \in \Xi}{\operatorname{argmin}} \left\| \mathbb{E}\left[g(w_{it}; \theta^0(\xi), \alpha_t^0(\xi))\right] \right\|^2 . \tag{B.1}$$

where the expectation is taken over the time-series dimension throughout.

The conclusion of the theorem requires that a change-of-variables applies to $F(\cdot; \theta^0 \alpha^0)$ in order to move from the space of observables to the unobserved types. Therefore we will need to show that F is differentiable as a function of w; see Bonnans and Shapiro [2000] for general arguments of functions of this form. Since g is a continuous function and w_{it} is well-behaved (Assumptions 2 (c) and (e)), differentiation and integration can be interchanged.

Let

$$G(w, \xi) = \left\| \mathbb{E}\left[g(w_{it}; \theta^0(\xi), \alpha_t^0(\xi))\right] \right\|^2 \tag{B.2}$$

and, by definition of F, a necessary condition for a minimum is

$$\frac{\partial G(w, F(w; \theta^0, \alpha^0))}{\partial \xi} = 0 \tag{B.3}$$

where the derivative exists by Assumption 2(b, c), $F(w; \theta^0, \alpha^0) \in \mathring{\Xi}$ by Assumption 2 (a), and Assumption 2(d) will ensure the first-order condition produces the unique absolute minimum in the interior for any $w \in \mathcal{W}$.

Along the lines of the implicit function theorem, taking the partial derivative with respect to w of this first-order condition will yield a system of equations we can solve for $\dfrac{\partial F(w;\theta^0,\alpha^0)}{\partial w}$ provided differentiability conditions are met.

The derivative of G is given as

$$\begin{aligned}\frac{\partial G(w,\xi)}{\partial \xi} =& 2\mathbb{E}\left[\frac{\partial g(w_{it};\theta^0(\xi),\alpha_t(\xi))}{\partial \theta^0}\cdot\frac{\partial \theta^0(\xi)}{\partial \xi}\right]'\mathbb{E}\left[g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))\right]\\ &+2\mathbb{E}\left[\frac{\partial g(w_{it};\theta^0(\xi),\alpha_t(\xi))}{\partial \alpha_t^0}\cdot\frac{\partial \alpha_t^0(\xi)}{\partial \xi}\right]'\mathbb{E}\left[g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))\right].\end{aligned}$$

and assuming for a moment that we are allowed to take the derivative of (B.3), we get

$$\frac{\partial^2 G(w,F(w;\theta^0,\alpha^0))}{\partial \xi^2}\cdot\frac{\partial F(w;\theta^0,\alpha^0)}{\partial w}+\frac{\partial^2 G(w,F(w;\theta^0,\alpha^0))}{\partial \xi\partial w}=0. \tag{B.4}$$

All that remains to show is that the second partial exists and is non zero, and the cross-partial derivative exists and, by Assumption (b,c), they must exist since g is twice differentiable in its arguments and both θ^0 and α^0 are twice-differentiable and bounded. Evaluated at $\xi=F(w;\theta^0,\alpha^0)$ it is

$$\frac{\partial^2 G(w,F(w;\theta^0,\alpha^0))}{\partial \xi^2}=2\left\|\mathbb{E}\left[\frac{\partial g(w_{it};\theta^0(\xi),\alpha_t(\xi))}{\partial \theta^0}\cdot\frac{\partial \theta^0(\xi)}{\partial \xi}+\frac{\partial g(w_{it};\theta^0(\xi),\alpha_t(\xi))}{\partial \alpha_t^0}\cdot\frac{\partial \alpha_t^0(\xi)}{\partial \xi}\right]\right\|^2\Bigg|_{\xi=F(w;\theta^0,\alpha^0)} \tag{B.5}$$

since the term with second order derivatives will retain the moment condition (2.4) and, by definition, it is zero at $\xi=F(w;\theta^0,\alpha^0)$. By Assumption 2 (d), (B.5) is positive so that the gradient of F with respect to w is well-defined by the implicit function theorem. By Assumption 2 (a) and convexity of G at $\xi=F(w;\theta^0,\alpha^0)$, $w\mapsto F(\cdot;\theta^0,\alpha^0)\in\mathring{\Xi}$ is injective: for any $w\in\mathcal{W}$ a unique ξ is produced by F as an argmin function by convexity. Therefore, the conditions for a change-of-variables is satisfied.

B.2. **Proof of consistency of $(\widehat{\theta},\widehat{\alpha})$ in endogenous linear model** (2.5)**.**

This proof follows the strategy of Bonhomme and Manresa [2015] in their Appendix covering the group heterogeneous coefficients case. Let $\mu=(\mu_1,\mu_2,\ldots,\mu_N)\in\Xi^N$ denote a vector of types assigned to each individual in the sample. Denote $\xi^0=(\xi_1^0,\ldots,\xi_N^0)\in\Xi^N$ as the population types. Let $\{\widehat{W}_i\}_{i\in\mathcal{N}}$ be a collection of positive definite matrices. Recall that for any positive definite matrix W, there exists a unique positive definite matrix $W^{1/2}$ such that $W=W^{1/2}W^{1/2}$ so that there exists a collection $\{\widehat{W}_i^{1/2}\}_{i\in\mathcal{N}}$ such that $\widehat{W}_i=\widehat{W}_i^{1/2}\widehat{W}_i^{1/2}$

for all $i \in \mathcal{N}$. Therefore for any $h > 0$ we can rewrite the TFE-GMM objective function as

$$\widehat{Q}(\theta, \alpha, \mu) = \frac{1}{N} \sum_{i=1}^{N} \int_{\Xi} \widehat{Q}_i(\xi, \theta, \alpha, \mu) K_h(\xi - \mu_i) \, d\xi \tag{B.6}$$

where, for any $i \in \mathcal{N}$, the individual's GMM criterion is

$$\widehat{Q}_i(\xi, \theta, \alpha, \mu) = \left\| \widehat{W}_i^{1/2} \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(y_{it} - x_{it}' \theta(\xi) - \alpha_t(\xi) \right) \right\|^2 . \tag{B.7}$$

Using the true DGP (2.5) for y_{it} we can rewrite this as

$$\widehat{Q}_i(\xi, \theta, \alpha, \mu) = \left\| \widehat{W}_i^{1/2} \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(u_{it} + x_{it}' \left(\theta^0(\xi_i^0) - \theta(\xi) \right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi) \right) \right) \right\|^2 . \tag{B.8}$$

Then, define an auxiliary objective function as

$$\widetilde{Q}(\theta, \alpha, \mu) = \frac{1}{N} \sum_{i=1}^{N} \int_{\Xi} \widetilde{Q}_i(\xi, \theta, \alpha, \mu) K_h(\xi - \mu_i) \, d\xi \tag{B.9}$$

where, for any $i \in \mathcal{N}$,

$$\widetilde{Q}_i(\xi, \theta, \alpha, \mu) = \left\| \widehat{W}_i^{1/2} \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(x_{it}' \left(\theta^0(\xi_i^0) - \theta(\xi) \right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi) \right) \right) \right\|^2 + \left\| \widehat{W}_i^{1/2} \frac{1}{T} \sum_{t=1}^{T} z_{it} u_{it} \right\|^2 . \tag{B.10}$$

I show that $\widehat{Q}$ is uniformly convergent to $\widetilde{Q}$ as N, T tend to infinity and h tends to zero.

Lemma 2. *Let Assumptions 2 and 4 hold and suppose $h \to 0$ as $N, T \to \infty$ and $Nh \to \infty$. Then,*

$$\operatorname*{plim}_{N,T \to \infty} \sup_{(\theta, \alpha, \mu) \in \Theta \times \mathcal{A} \times \Xi^N} |\widehat{Q}(\theta, \alpha, \mu) - \widetilde{Q}(\theta, \alpha, \mu)| = 0 \tag{B.11}$$

Proof. Expanding the $\widehat{Q}_i$ for any $i \in \mathcal{N}$ using bilinearity of the inner product gives

$$\widehat{Q}_i(\xi, \theta, \alpha, \mu) = \left\| \widehat{W}_i^{1/2} \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(x_{it}' \left(\theta^0(\xi_i^0) - \theta(\xi) \right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi) \right) \right) \right\|^2 \tag{B.12}$$

$$+2 \left(\frac{1}{T} \sum_{t=1}^{T} z_{it} u_{it} \right)' \widehat{W}_i \left(\frac{1}{T} \sum_{t=1}^{T} z_{it} \left(x_{it}' \left(\theta^0(\xi_i^0) - \theta(\xi) \right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi) \right) \right) \right) \tag{B.13}$$

$$+ \left\| \widehat{W}_i^{1/2} \frac{1}{T} \sum_{t=1}^{T} z_{it} u_{it} \right\|^2 \tag{B.14}$$

so that the difference for each individual GMM criterions for any parameter values is (B.13). Therefore, the difference in (B.11) is

$$\frac{2}{N}\sum_{i=1}^{N}\int_{\Xi}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}u_{it}\right)'\widehat{W}_i\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\xi)\right)\right)\right)K_h(\xi-\mu_i)\,d\xi. \tag{B.15}$$

Next the integral is reduced using the mean value theorem where I assume without loss of generality that the Lebesgue measure of Ξ is 1. By Assumption $2(a)$ Ξ must be an interval and by Assumption $2(c)$ the integrand must be continuous as a function of ξ so there exists $\overline{\xi}\in\mathring{\Xi}$ such that (B.15) is

$$\frac{2}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}u_{it}\right)'\widehat{W}_i\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\overline{\xi})\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\overline{\xi})\right)\right)\right)K_h(\overline{\xi}-\mu_i). \tag{B.16}$$

Since the sample is iid over the cross-sectional dimension, we can study the limiting behavior for each term in the inner product. For the first, consider

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T}z_{it}u_{it}\right\|^2\right]=\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[z_{it}'z_{is}u_{it}u_{is}\right]\leq\frac{M}{T}\to 0 \tag{B.17}$$

as $T\to\infty$ by Assumption $4(d)$ so by Jensen's inequality the first term vanishes asymptotically. Next, we need to ensure that the other term is bounded. Consider it in two parts:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}z_{it}x_{it}'\left(\theta^0(\xi_i^0)-\theta(\overline{\xi})\right)\right\|\leq\left(\frac{1}{T}\sum_{t=1}^{T}\left\|z_{it}x_{it}'\right\|\right)\left\|\theta^0(\xi_i^0)-\theta(\overline{\xi})\right\| \tag{B.18}$$

where Jensen's and Cauchy-Schwarz were applied successively. Since θ^0 and θ are both functions in Θ, by Assumption $4(b)$, it must be that $\left\|\theta^0(\xi_i^0)-\theta(\overline{\xi})\right\|\leq\eta$ for some scalar $\eta>0$. Therefore, by Assumption $4(c)$,

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|z_{it}x_{it}'\right\|\right]\right)\left\|\theta^0(\xi_i^0)-\theta(\overline{\xi})\right\|\leq\frac{\eta}{T}\sum_{i=1}^{T}M=\eta M. \tag{B.19}$$

Lastly,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}z_{it}\left(\alpha_t^0(\xi_i^0)-\alpha_t(\overline{\xi})\right)\right\|^2\leq\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_t^0(\xi_i^0)-\alpha_t(\overline{\xi})\right)^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}\left\|z_{it}\right\|^2\right)$$

where $\mathcal{A}$ is a space of bounded functions by Assumption 4(b) so

$$\mathbb{E}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(\alpha_t^0(\xi_i^0)-\alpha_t(\bar{\xi})\right)\right\|^2 \leq \eta M. \tag{B.20}$$

Then, by applying Cauchy-Schwarz inequality on (B.16), by Assumption 4(e) on the kernel function and using the above derived bounds finishes the proof. □

The next lemma shows that the auxiliary objective function is uniquely minimized at the true values.

Lemma 3. *Suppose that Assumption 2 and 4 holds. Then, there exists a* $C > 0$ *for any* $(\theta, \alpha, \mu) \in \Theta \times \mathcal{A} \times \Xi^N$,

$$\widetilde{Q}(\theta, \alpha, \mu) - \widetilde{Q}(\theta^0, \alpha^0, \xi^0) \geq C\left[\left\|\theta^0-\theta\right\|_2^2 + \frac{1}{T}\sum_{t=1}^{T}\left\|\alpha^0-\alpha\right\|_2^2\right] + o_p(1), \tag{B.21}$$

where the vector-function norm is defined as (3.3).

Proof. I begin with arguing that the auxiliary objective function vanishes at the true values asymptotically as $N, T \to \infty$ and $h \to 0$. Consider the following:

$$\widetilde{Q}(\theta^0, \alpha^0, \xi^0) = \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta^0(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t^0(\xi)\right)\right)\right\|^2 K_h(\xi-\xi_i^0)\,d\xi$$
$$+\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}u_{it}\right\|^2 \int_{\Xi} K_h(\xi-\xi_i^0)\,d\xi$$

The second term in the sum is $o_p(1)$ by Assumption 4(d) as argued in the proof of Lemma 2 and the fact that the kernel is a bounded function on Ξ. For N and T sufficiently large, h will be small and, since the integrand is a continuous function on compact Ξ, by Assumption 4(e) the kernel function will weakly converge to the Dirac delta function. Therefore, the limit will be the integrand evaluated at $\xi = \xi_i^0$ so $\widetilde{Q}(\theta^0, \alpha^0, \xi^0) \to_p 0$ as $N, T \to \infty$.

Then, denoting $\widehat{c}_i$ as the minimum eigenvalue of $\widehat{W}_i$ for all $i \in \mathcal{N}$ and $\widehat{c} = \min_{i \in \mathcal{N}} \widehat{c}_i$, the difference can be written as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)\,d\xi + o_p(1)\\
\geq &\widehat{c}\frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)\,d\xi + o_p(1)\\
= &\lim_{b\to 0}\frac{1}{N}\sum_{i=1}^{N}\widehat{c}_i\int_{\Xi\times\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\widetilde{\xi})-\theta(\xi)\right)+\left(\alpha_t^0(\widetilde{\xi})-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)K_b(\widetilde{\xi}-\xi_i^0)\,d\xi\,d\widetilde{\xi} + o_p(1)\\
\geq &\lim_{b\to 0}\frac{1}{N}\sum_{i=1}^{N}\widehat{c}_i\int_{\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)\,d\xi + o_p(1)
\end{aligned}$$

using the fact of weak convergence of the kernel to the Dirac delta function and that $\Xi \times \Xi \supset \{(\xi, \widetilde{\xi}) \in \Xi \times \Xi : \xi = \widetilde{\xi}\}$ so one of the variables of integration is eliminated. Now, using Jensen's inequality:

$$\begin{aligned}
&\lim_{b\to 0}\frac{1}{N}\sum_{i=1}^{N}\widehat{c}_i\int_{\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)\,d\xi + o_p(1)\\
\geq &\lim_{b\to 0}\widehat{c}\int_{\Xi}\widehat{P}(\xi,\mu)\left\|\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\frac{K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)}{\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0)}\right\|^2 d\xi + o_p(1)\\
&\geq \lim_{b\to 0}\widehat{c}\,\widehat{p}(\mu)\int_{\Xi}\left\|\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\frac{K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)}{\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0)}\right\|^2 d\xi + o_p(1)
\end{aligned}$$

where

$$\widehat{P}(\xi,\mu) = \frac{1}{N}\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0) \geq \min_{\xi\in\Xi}\widehat{P}(\xi,\mu) = \widehat{p}(\mu) \geq 0.$$

Let $Z_{it} \in \mathbb{R}^k$ denote the non constant elements of z_{it} and let

$$\mathcal{M}(\mu,\xi) = \sum_{i=1}^{N} \frac{K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)}{\sum_{j=1}^{N} K_h(\xi-\mu_j)K_b(\xi-\xi_j^0)} \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} Z_{it}x_{it}' & \frac{1}{\sqrt{T}}Z_{i1} & \frac{1}{\sqrt{T}}Z_{i2} & \cdots & \frac{1}{\sqrt{T}}Z_{iT} \\ \frac{1}{\sqrt{T}}Z_{i1} & 1 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{T}}Z_{i2} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{T}}Z_{iT} & 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{B.22}$$

which is a $(p+T)\times(p+T)$ matrix. Then,

$$\begin{aligned} &\lim_{b\to 0}\widehat{c}\,\widehat{p}(\mu)\int_{\Xi}\left\|\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\frac{K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)}{\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0)}\right\|^2 d\xi + o_p(1) \\ =&\lim_{b\to 0}\widehat{c}\,\widehat{p}(\mu)\int_{\Xi}\begin{bmatrix}\theta^0(\xi)-\theta(\xi) \\ \frac{1}{\sqrt{T}}\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\end{bmatrix}'\mathcal{M}(\mu,\xi)\begin{bmatrix}\theta^0(\xi)-\theta(\xi) \\ \frac{1}{\sqrt{T}}\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\end{bmatrix} d\xi + o_p(1) \\ \geq&\lim_{b\to 0}\widehat{c}\min_{\mu\in\Xi^N}\widehat{p}(\mu)\min_{\mu\in\Xi^N,\xi\in\Xi}\widehat{\rho}(\mu,\xi)\int_{\Xi}\left[\left\|\theta^0(\xi)-\theta(\xi)\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha^0(\xi)-\alpha(\xi)\right\|^2\right] d\xi + o_p(1) \\ =&C\left[\left\|\theta^0-\theta\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha^0-\alpha\right\|^2\right]+o_p(1) \end{aligned}$$

where $C = \lim_{b\to 0}\widehat{c}\min_{\mu\in\Xi^N}\widehat{p}(\mu)\min_{\mu\in\Xi^N,\xi\in\Xi}\widehat{\rho}(\mu,\xi) > 0$ by Assumption 4 (a, f), thus completing the proof. □

To show consistency of the parameters, by Lemma 2 and 3 and the definition of the TFE-GMM estimator (3.1) as the minimizer of $\widehat{Q}$, we have that

$$\widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\mu}) = \widehat{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\mu}) + o_p(1) \leq \widehat{Q}(\theta^0,\alpha^0,\mu^0) + o_p(1) = \widetilde{Q}(\theta^0,\alpha^0,\mu^0) + o_p(1) \tag{B.23}$$

so, because $C > 0$,

$$o_p(1) \leq C\left[\left\|\theta^0-\widehat{\theta}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha^0-\widehat{\alpha}\right\|^2\right]+o_p(1) \leq \widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\mu})-\widetilde{Q}(\theta^0,\alpha^0,\mu^0) \leq o_p(1). \tag{B.24}$$

Therefore it must be that $\left\|\theta^0-\widehat{\theta}\right\|^2 = o_p(1)$ and $\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha^0-\widehat{\alpha}\right\|^2 = o_p(1)$. □