Heterogeneity, Uncertainty and Learning: Semiparametric Identification and Estimation*

Jackson Bunting [†] Paul Diegert[‡] Arnaud Maurel[§]

November 5, 2023

Abstract

We provide semiparametric identification results for a broad class of learning models in which outcomes of interest depend on i) predictable heterogeneity, ii) initially unpredictable heterogeneity that may be revealed over time, and iii) transitory uncertainty. We consider a common environment where the researcher only has access to a short panel on choices and realized outcomes. We establish point-identification of the outcome equation parameters and the distribution of the three types of unobservables, under the standard assumption that unpredictable heterogeneity and uncertainty are normally distributed. We also show that, in the absence of predictable heterogeneity, the model is identified without making any distributional assumption. We then derive the asymptotic properties of a sieve MLE estimator for the model parameters, and devise a tractable profile likelihood based estimation procedure. Monte Carlo simulation results indicate that our estimator exhibits good finite-sample properties.

*Preliminary version. We thank seminar participants at CREST-PSE, LMU Munich, TSE, UC Davis, conference participants at the 2023 IAAE, SETA and SOLE meetings, as well as Victor Aguirregabiria, Xavier D'Haultfoeuille, Yuichi Kitamura, Mauricio Olivares, Yuya Sasaki, Chris Taber and Daniel Wilhelm for useful comments.

[†]Texas A&M University, jbunting@tamu.edu.

[‡]Toulouse School of Economics, paul.diegert@tse-fr.eu.

[§]Duke University, NBER and IZA, arnaud.maurel@duke.edu.

1 Introduction

Learning models, in which agents have imperfect information about their environment and update their beliefs over time, are frequently used in economics. These models have received particular interest in various subfields in empirical microeconomics, including industrial organization and health (see, e.g., Ackerberg, 2003; Coscelli and Shum, 2004; Crawford and Shum, 2005; Abbring and Campbell, 2005; Chan and Hamilton, 2006; Yang, 2020; Aguirregabiria and Jeon, 2020, for a survey in the context of oligopoly competition), as well as in labor economics (see, e.g. Miller, 1984; Antonovics and Golan, 2012; Pastorino, 2015; Hincapié, 2020; Pastorino, 2022) and economics of education (see, e.g. Arcidiacono, 2004; Zafar, 2011; Stinebrickner and Stinebrickner, 2012; Stange, 2012; Thomas, 2019; Kinsler and Pavan, 2021; Arcidiacono et al., 2023). Since the seminal work of Erdem and Keane (1996), learning models have also been popular in the marketing literature (see Ching et al., 2013, for a survey). However, while learning models are often structurally estimated, much remains to be known about the identification of this important class of models.

In this paper we provide new semiparametric identification results for a general class of learning models. Importantly, we consider an environment where the researcher has access to a short panel on choices and realized outcomes only. As such, our results are widely applicable, including in frequent situations where one does not have access to elicited beliefs data, or to a vector of selection-free measurements of latent individual heterogeneity. Specifically, we consider throughout our analysis a potential outcome model of the following form:

$$Y_{it}(d) = \alpha_t(d) + Z_{it}^\intercal \beta_t(d) + \lambda_i^\intercal F_t(d) + \epsilon_{it}(d), \tag{1}$$

where Z_{it} is a vector of explanatory variables associated with individual i in period t, $\theta := (\alpha_t(d), \beta_t(d), F_t(d))$ are unknown parameters, λ_i denotes a vector of latent individual effects, and $\epsilon_{it}(d)$ is an idiosyncratic shock. While interactive fixed effects models of this kind have been the object of much interest in econometrics, a key distinctive feature of the setup considered in this paper is the existence of two different types of individual effects. Namely, we assume the individual effect λ_i consists of two components: $\lambda_{k,i}$ which are supposed to be initially known by the agent, and $\lambda_{u,i}$ which are initially unknown but may be learned over time. We complement this outcome model with a flexible choice model, in which agent i's assignment in period t can depend arbitrarily on contemporaneous and lagged explanatory variables, assignments and realized outcomes. This framework encompasses most of the decision models that have been considered in the learning literature.

We first show that the model is point-identified under two alternative sets of conditions. Our first and main identification result applies to a setup where, consistent with most of the Bayesian learning models that have been considered in the literature, we assume that the idiosyncratic shocks from the outcome equations ($\epsilon_{it}(d)$), as well as the unknown heterogeneity component ($\lambda_{u,i}$), are normally distributed. In contrast, the distribution of the known heterogeneity component ($\lambda_{k,i}$) is left unspecified. From the key observation that the distribution of current realized outcomes conditional on past choices and outcomes is a mixture of normal distributions, we leverage results from Bruni and Koch (1985) to establish identification of the joint distribution of realized outcomes, choices and known heterogeneity component $\lambda_{k,i}$.

We then also show that a pure learning model with only one type of permanent unobserved heterogeneity ($\lambda_{u,i}$) actually remains point-identified without making any distributional assumption. A crucial distinction from the general case is that this model is one of selection on observables only, as individual choices depend on beliefs about $\lambda_{u,i}$ only through prior outcomes, choices and covariates. This feature allows us to build on insights from the interactive fixed effects literature, in particular Freyberger (2018), to establish identification in the pure learning case.

We propose to estimate the model parameters θ via sieve maximum likelihood estimation. We focus on a particular class of functionals of θ, which includes as special cases economically relevant quantities, such as the predictable and unpredictable outcome variances. These variances can in turn be used to evaluate the relative importance of, e.g., uncertainty vs. heterogeneity in the overall lifecycle earnings variability - a question that has been the object of much interest in labor economics (see, e.g., Cunha et al., 2005; Huggett et al., 2011; Cunha and Heckman, 2016). We show that, under mild regularity conditions, the resulting estimators are consistent and asymptotically normal. Monte Carlo simulation results indicate that our estimator exhibits good finite-sample properties. Importantly for practical purposes, our proposed estimator only involves a modest computational cost.

Related literatures

Our paper contributes to several strands of the literature. First and foremost, we add to a set of papers that study the identification of learning models, generally in the context of specific applications (Abbring and Campbell, 2005; Arcidiacono et al., 2023; Gong, 2019; Pastorino, 2022). A key difference with most of the papers in this literature is that we only impose mild restrictions on the choice process. In particular, we remain agnostic about how choices depend on individual beliefs about $\lambda_{u,i}$, while allowing these beliefs to depend arbitrarily on past choices and realized outcomes. Particularly relevant for us is recent work by Pastorino (2022), which establishes formal identification results for a econometric learning model. However, beyond the fact that Pastorino (2022) restricts to the context of workers' and firms' learning, there are two main differences relative to our paper. First, unobserved heterogeneity in that paper is restricted to be discrete, while we allow for both continuous and multivariate unobserved heterogeneity. Second, and importantly, the outcomes which form the basis of learning are assumed to depend on the learned portion of unobserved heterogeneity only. In contrast, in our setup, outcomes may depend on both known and unknown components. Our framework also differs from Gong (2019) in important ways. Notably, while we remain agnostic about how choices depend on agents' beliefs about the distribution of $\lambda_{u,i}$, Gong (2019) assumes that assignment depends on the prior mean only. Gong further imposes various restrictions on the updating rule, while we remain agnostic about how agents form and update their beliefs about $\lambda_{u,i}$.

Our paper also fits into a literature that focuses on the identification and estimation of dynamic discrete choice models in the presence of permanent unobserved heterogeneity (see, among others, Heckman and Navarro, 2007; Hu and Schennach, 2008; Kasahara and Shimotsu, 2009; Arcidiacono and Miller, 2011; Hu and Shum, 2012; Sasaki, 2015; Sasaki and Hu, 2018; Aguirregabiria et al., 2021; Bunting, 2022). Unlike these papers, we focus on a learning framework in which a portion of the permanent individual unobserved heterogeneity is initially unknown to the agents, so that decisions may depend on the unknown component only through the sequence of past outcomes. This asymmetry property plays an important role in our ability to address the deconvolution problem associated with the coexistence of both types of unobserved heterogeneity. A second important difference is that, unlike most papers in this literature (e.g., Hu and Schennach, 2008; Hu and Shum, 2012), we allow for sample selection in our setting.[1] A noteworthy exception is Sasaki (2015) which considers identifications of a dynamic panel model with selection and latent permanent unobserved heterogeneity. Relative to this paper, we allow for multivariate unobserved heterogeneity and for selection to depend on the entire information set (namely all realized outcomes and choices).

At a high level, our analysis is also connected to the literature that deals with the identification of mixture models (see, e.g., Compiani and Kitamura, 2016; Kitamura and Laage, 2018, and references therein). In particular, central to our main identification result is the observation that the distribution of current outcomes conditional on the sequence of past choices and outcomes is a mixture of normal distributions.

Finally, since the outcome equation in our model involves interactions between

[1]We note that in the case that beliefs are discrete and first-order Markov, one may be able to use the arguments of Hu and Shum (2012) to identify the latent beliefs from variation in the (discrete) choices (c.f., Pastorino, 2022). This approach is not available in our context, since beliefs may be multivariate and continuous and are excluded from the outcome equation.

unobserved individual- and time-specific effects, our paper also fits into the literature that deals with the identification and estimation of panel data models with interactive fixed effects (see, e.g., Bai, 2009; Gobillon and Magnac, 2016; Freyberger, 2018). Among these papers, our identification strategy is most closely related to Freyberger (2018). A fundamental distinction though comes from the fact that Freyberger (2018) considers a selection-free environment. In contrast, individual choices, along with the associated selection issues affecting the potential outcomes, play a central role in our analysis.

Organization of the paper

The remainder of the paper is organized as follows. Section 2 presents the set-up of the model. Section 3 contains our main identification results, both for the general case and for the case of a pure learning model. We discuss in Section 4 the estimation and inference on the parameters of interest, before turning in Section 5 to the implementation of our estimator. We study in Section 6 its finite-sample performances. Finally, Section 7 concludes. The appendix gathers all the proofs.

Notation: $\mathcal{S}(A)$ indicates the support of random variable A. F_A indicates the distribution function of random variable A. For any sequence $(a_1, a_2, \ldots, a_S)$ and $s \leq S$, we let $a^s = (a_1, a_2, \ldots, a_s)$. Upper case letters represent random variables, lower case represent realized values. $A \perp\!\!\!\perp B \mid C$ indicates that A and B are statistically independent conditional upon C.

2 Set-up

Throughout the paper we consider a setup where potential outcomes have an interactive fixed effect structure of the following form:

$$Y_t(d) = \alpha_t(d) + Z_t^{\intercal}\beta_t(d) + \lambda_k F_{kt}(d) + \lambda_u^{\intercal} F_{ut}(d) + \epsilon_t(d), \tag{2}$$

where d represents individual i's assignment in period t, $Y_t(d)$ is a scalar potential outcome variable associated with assignment d, Z_t a vector of explanatory variables, $\theta = (\alpha_t(d), \beta_t(d), F_{kt}(d), F_{ut}(d))$ is a vector of unknown parameters, $\lambda = (\lambda_k, \lambda_u^\intercal)^\intercal$ is the latent individual effect and $\epsilon_t(d)$ is an unobserved random variable. For example, $Y_t(d)$ may represent potential log-wages in occupation d. $Y_t(d)$ may depend on some observed individual and possibly time-varying characteristics (Z_t) as well as on multiple dimensions of unobserved abilities (λ), which might play different roles in different occupations.

Importantly, we allow for two distinct types of latent individual effects. Namely, λ_k is assumed to be known by the agent, while λ_u is initially unknown but may be gradually revealed over time. For example, worker i's log-wage in occupation d at time t, $Y_t(d)$, may depend on her unobserved (to the econometrician) occupation specific productivity, $\lambda_k F_{kt}(d) + \lambda_u^\intercal F_{ut}(d)$. As the worker accumulates more experience, she may update her belief about λ_u, and thus about the initially unknown portion of productivity in each of the possible occupations.

Turning to the choice and learning process, the only restriction placed on an individual's assignment in period t, which we denote by D_t, is that it does not directly depend on the unknown component of latent heterogeneity. Specifically, we impose the following restriction:

$$D_t \perp\!\!\!\perp \lambda_u \mid Z^t, Y^{t-1}, D^{t-1}, \lambda_k. \tag{3}$$

The above conditional independence condition highlights the asymmetry between the two types of latent effects: assignments may directly depend on the known component of the latent effect λ_k, but not on the unknown component of the latent effect λ_u. However, we allow the assignment rule to depend arbitrarily on lagged covariates, outcomes and choices. As a result, we do not restrict how agents form their beliefs about λ_u, provided that such beliefs are a measurable function of Z^t, Y^{t-1}, D^{t-1} and λ_k. We also remain agnostic about how assignments depend on agents' beliefs over λ_u.

In particular, this framework is consistent with a setup where agents are rational and Bayesian updaters, so that beliefs coincide with the objective distribution of λ_u conditional on their information set at a given point in time, which may include all realized variables and model parameters. Alternatively, this also accommodates situations where individual decisions may not involve beliefs over the distribution of λ_u, or depend instead on myopic beliefs that are formed based on the prior-period choice and outcome. This further allows for heterogeneous beliefs formation, where, for instance, some agents may have rational expectations about their unobserved characteristic λ_u, while others may have biased beliefs.

Finally, we denote the conditional choice probability (CCP) function as

$$h_t(d^t, z^t, y^{t-1}, v_k) \coloneqq \Pr(D_t = d \mid Z^t = z^t, Y^{t-1} = y^{t-1}, D^{t-1} = d^{t-1}, \lambda_k = v_k).$$

These CCPs play a central role in our identification analysis. In the following section, we provide sufficient conditions under which the CCPs - which are latent objects because of the conditioning on $\lambda_k = v_k$ - are identified. In empirical applications it is very common to impose some structure on the choice process. In particular, it is standard to assume that

$$D_t = \underset{d \in \mathcal{S}(D_t)}{\arg\max} \left\{ v_t(d, Z_t, \lambda_k, X_t) + \eta_t(d) \right\},$$

where v_t is known up to a finite-dimensional vector of parameters, X_t are sufficient statistics for the conditional distribution of λ_u at time t, and η_t follows a known distribution. Having identified the CCPs, one can then apply standard identification arguments from the dynamic discrete choice literature (Magnac and Thesmar, 2002) to identify the primitives of the choice model.

2.1 Uncertainty and Learning

The key feature of the model is the distinction between the three forms of unobserved heterogeneity: (1) permanent heterogeneity that is known to the agent, λ_k, (2) permanent heterogeneity that is initially unknown to the agent, λ_u, and (3) idiosyncratic time-varying shocks, ϵ. This provides a framework for quantifying the importance of uncertainty in outcomes. At $t = 1$, the variance in future outcomes can be decomposed orthogonally into a component that depends on (λ_u, ϵ) and a component that depends on λ_k. This is the approach taken in Cuhna and Heckman (2005, 2016), who decompose the variance in lifetime earnings into a component that is predictable at the time of deciding whether go to college and a component that is not.

In our setup, the importance of uncertainty can change over time as agents learn about λ_u by observing realized outcomes and covariates and use this information to make choices. We provide in Appendix A a class of variance decomposition parameters which includes both the $t = 1$ orthogonal decomposition as well as other decompositions that incorporate these learning and selection effects. These decompositions each provide different ways of quantifying the importance of uncertainty to future wages. Identification of the model implies the identification of all of these parameters. After establishing identification of the model, we special attention to estimation and inference of broad class of functionals that encompasses these kinds of variance decompositions.

3 Identification

We provide in Subsection 3.1 a high-level overview of the proposed identification strategies. We then discuss identification in the leading case with both known and unknown unobserved heterogeneity (Subsection 3.2), before turning to the pure learning case where the only source of permanent unobserved heterogeneity is assumed to be initially unknown to the agent (Subsection 3.3).

3.1 Overview

The identification of the model discussed in Section 2 boils down to identifying the linear interactive fixed effects model, $Y_t(d_t) = \beta_t(d_t)^\intercal Z_t + F_t(d_t)^\intercal \lambda + \epsilon_t$ from the distribution of realized outcomes Y^T. To illustrate the problem, suppose that $\mathcal{S}(D_t) = \{0,1\}$ for all t and Z_t is time invariant. Let $D := \prod_{t=1}^{T} D_t$, and $Y(1) := (Y_1(1), \ldots, Y_T(1))$, and suppose one wants to identify the distribution of $Y(1)$, which is censored for $D = 0$. The relationship between the uncensored and censored distributions can be characterized as follows,

$$f_{Y|Z,D}(y|z,1)\frac{f_{D|Z}(1|z)}{f_{D|Y(1),Z}(1|y,z)} = f_{Y(1)|Z}(y|z)$$

The observed conditional density $f_{Y|Z,D}(y|z,1)$ is weighted by the term $\frac{f_{D|Z}(1|z)}{f_{D|Y,Z}(1|y,z)}$, which reflects selection.

The sequential nature of the choice process provides one strategy for identifying these selection weights. Namely, assuming that individuals can only use information on past outcomes and choices to make their choices yields:

$$f_{D_t|Y(1)D^{t-1},Z}(1|y^t,1,z) = f_{D_t|Y^{t-1}(1),D^{t-1},Z}(1|y^{t-1},1,z) \tag{4}$$

The right hand side is identified from the joint distribution of (D^t, Y^{t-1}, Z) conditional on $D^{t-1} = 1$. It follows that the inverse selection weight, $f_{D|Y,Z}$ is identified as follows

$$f_{D|Y(1),Z}(1|y,z) = f_{D_t|Y^{t-1},D^{t-1},Z}(1|y,1,z)f_{D^{t-1}|Y^{t-2},D^{t-2},Z}(1|y,1,z)\cdots f_{D_1|Z}(1|z)$$

We pursue this identification approach in Section 3.3 in a version of the model we call *pure learning*. The assumption underlying (4) is motivated by assuming that agents learn about a latent variable by observing their outcomes. In that context, choices in period t depend on the agent's beliefs about the latent variable and therefore all the prior outcomes and choices.

The exclusion restriction in (4) can break down, however, when agents base their decisions on variables that are not observed in the data. We propose in Section 3.2 an identification strategy that can be used in such situations. In particular, we show that maintaining a normality assumption commonly made in the learning literature is sufficient to identify the joint distribution (Y^T, D^T, Z) in a first step. One can then identify the model parameters in a second step, along the lines of the reweighting strategy discussed above.

3.2 Known and unknown heterogeneity

This section provides sufficient conditions for identification of the model discussed in Section 2. The first assumption (KL1) imposes that any correlation in the observed outcomes and choices over time and across assignments is due to the latent effect λ. It also imposes that the transition of the control variables, Z_t, does not depend on unobservables.

Assumption KL1. Equation (2) holds. Further, for any $d \in \mathcal{S}(D_t)$

$$F_{\epsilon_t(d) D_t Z_t | Y^{t-1} D^{t-1} Z^{t-1} \lambda} = F_{\epsilon_t(d)} F_{D_t | Y^{t-1} D^{t-1} Z^t \lambda_k} F_{Z_t | Y^{t-1} D^{t-1} Z^{t-1}}.$$

Assumption KL2 imposes that the unknown component of the individual effect, λ_u, is drawn from a multivariate normal distribution, and that the random shock in the outcome equation is normally distributed.

Assumption KL2. $\lambda_u \mid (Z_1 = z_1, \lambda_k = v_k) \sim N\left(0, \Sigma_u(z_1)\right)$ and $\epsilon_t(d) \sim N(0, \sigma_t(d)^2)$.

Assumption KL2 leads to a specific functional form for the posterior distribution, namely the Gaussian conjugate distribution. We summarize this result in Lemma 1. To do so, we define (E_t, Σ_t) recursively as follows. First, $(E_1, \Sigma_1) = (0, \Sigma_u(Z_1))$. Second,

$$\begin{aligned}
\Sigma_{t+1} &= \left(\Sigma_t^{-1} + F_{ut}(D_t) F_{ut}(D_t)^\intercal \sigma_t^{-2}(D_t)\right)^{-1} \\
E_{t+1} &= \Sigma_{t+1} \left(\Sigma_t^{-1} E_t + F_{ut}(D_t) \frac{Y_t - \alpha_t(D_t) - Z_t^\intercal \beta_t(D_t) - \lambda_k F_{kt}(D_t)}{\sigma_t^2(D_t)}\right).
\end{aligned}$$

Lemma 1. Let Assumptions KL1 and KL2 hold. Then λ_u conditional upon $(D^{t-1}, Y^{t-1}, Z^t, \lambda_k)$ is distributed $N(E_t, \Sigma_t)$.

Since λ_u conditional on $(Y^{t-1}, D^{t-1}, Z^t, \lambda_k)$ follows a normal distribution with mean E_t and variance-covariance matrix Σ_t, one can use (E_t, Σ_t) as a sufficient statistic for λ_u at time t. Notice that (E_t, Σ_t) is a deterministic function of

$(D^{t-1}, Y^{t-1}, Z^t, \lambda_k)$ and $\theta_1 = \big((\alpha_t, \beta_t, F_{kt}, F_{ut}, \sigma_t)_{t=1}^T, \Sigma_u\big) \in \Theta_1$. Furthermore, we can express (E_t, Σ_t) non-recursively[2] as:

$$\Sigma_{t+1} = \left(\Sigma_u^{-1}(Z_1) + \sum_{s=1}^{t} F_{us}(D_s) F_{us}(D_s)^\intercal \sigma_s^{-2}(D_s)\right)^{-1}$$
$$E_{t+1} = \Sigma_{t+1}\left(\sum_{s=1}^{t} F_{us} \frac{Y_s - \alpha_s(D_s) - Z_s^\intercal \beta_s(D_s) - \lambda_k F_{ks}(D_s)}{\sigma_s^2(D_s)}\right)$$

Suppose $\lambda_u \in \mathbb{R}^p$. Our three remaining assumptions are as follows.

Assumption KL3. (A) For some d_1, $\alpha_1(d_1) = 0$, $F_{k1}(d_1) = 1$. (B) For some $(d_1, d_2, \ldots, d_p)$, $(F_{u1}(d_1) F_{u2}(d_2) \ldots F_{up}(d_p)) = I_{p\times p}$.

Assumption KL3 is a normalization on the finite dimensional parameters. This type of assumption is standard in interactive fixed effect models (Freyberger, 2018), since no scale assumption is placed on the distribution of the unknown latent effects. For example, one could replace Assumption KL3 (A) by a zero mean restriction on the latent individual effect λ, and a unitary variance assumption on the known component of the latent effect λ_k. We also impose Assumption KL3 (B) since the unknown latent effect is inherently scale free.

Assumption KL4. (A) Θ_1 is a compact set. (B) $\mathrm{Supp}(\lambda_k)$ is a compact set. (C) For each t, $F_{ut}^\intercal(d_t)\Sigma_t F_{ut}(d_t) + \sigma_t^2(d_t) \neq 0$, $\sigma_t(d_t) \neq 0$ and $\Sigma_u(z_1)$ is non-singular. (D) $f_{\lambda_k | Y^{t-1}, Z^t, D^t}(v_k; y^{t-1}, z^t, d^t) > 0$ for all for all t and v_k in the support of λ_k. (E) For each t and d_t, the variance-covariance matrix of $(1, Z_t)$ conditional on $D_t = d_t$ is non-singular.

Assumption KL4 places support restrictions on various objects of the model. In particular, Part (B) imposes that the known latent factor λ_k has compact support.

[2]Our identification result would go through if one replaces the first part of Assumption KL2 with $\lambda_u \mid (Z_1 = z_1, \lambda_k = v_k) \sim N(0, \Sigma_u(v_k, z_1))$ under some regularity conditions on $v_k \mapsto \Sigma_u(v_k, z_1)$, including for each $v_k - \tilde{v}_k > 0$, $\Sigma_u(z_1, v_k) - \Sigma_u(z_1, \tilde{v}_k)$ is positive (or negative) semi-definite. For simplicity, we maintain the stronger Assumption KL2 when establishing identification in Theorem 1 below.

This holds if the distribution of λ_k has discrete support although this obviously applies to a broader set of distributions. We return to this compactness condition in Remark 1 below. Part (C) requires that the distribution of $Y_t(d) \mid (Z_t, D_t, \lambda_k)$ is non-degenerate. Part (D) is a "rectangular" support assumption on λ_k. This assumption is typically satisfied in dynamic discrete choice models, as they generally impose a large support assumption on the random utility shocks. Finally, Part (E) imposes that there exists sufficient variation in Z_t conditional on D_t.

Assumption KL5. (A) For each d_t there are sequences $d^{t-1}, \tilde{d}^{t-1}$ such that $F_{ut}(d_t)^\intercal \Sigma_t \sum_{s=1}^{t-1} \left(F_{us}(d_s) \frac{F_{ks}(d_s)}{\sigma_s^2(d_s)} - F_{us}(\tilde{d}_s) \frac{F_{ks}(\tilde{d}_s)}{\sigma_s^2(\tilde{d}_s)} \right) \neq 0$. (B) For all d_t, $F_{kt}(d_t) \neq 0$. (C) For all d^t, $F_{kt}(d_t) - F_{ut}(d_t)^\intercal \Sigma_t \sum_{s=1}^{t-1} F_{us}(d_s) \frac{F_{ks}(d_s)}{\sigma_s^2(d_s)} \neq 0$. (D) For each (d_2, d_1), $F_{u2}(d_2)^\intercal \Sigma_2 F_{u1}(d_1) \frac{F_{k1}(d_1)}{\sigma_1^2(d_1)} \neq 0$ (E) There are sets $\{d_{2,i} \in \mathcal{S}(D_2) : i = 1, 2, \ldots, p\}$, $\{\tilde{d}_{2,i} \in \mathcal{S}(D_2) : i = 1, 2, \ldots, p\}$ which satisfy

$$
\begin{aligned}
&(F_{u2}(d_{2,1}) F_{u2}(d_{2,2}) \ldots F_{u2}(d_{2,p}))^{-\intercal} \operatorname{vec}(F_{k2}(d_{2,1}), \ldots, F_{k2}(d_{2,p})) \\
&\neq \left(F_{u2}(\tilde{d}_{2,1}) F_{u2}(\tilde{d}_{2,2}) \ldots F_{u2}(\tilde{d}_{2,p}) \right)^{-\intercal} \operatorname{vec}(F_{k2}(\tilde{d}_{2,1}), \ldots, F_{k2}(\tilde{d}_{2,p})).
\end{aligned}
$$

(F) For any d^T, $\{F_{ut}(d_t) : t = 1, \ldots, T\}$ is linearly independent.

Assumption KL5 is a regularity condition that basically ensures that the latent individual effect λ alters outcomes sufficiently differently across time and assignments. This condition is relatively mild as it primarily rules out knife-edge cases where the cumulative effect of different elements of the individual effect perfectly offset each other.[3] More specifically, Part (A) requires that the aggregate effect of λ_k on outcomes for choice d_t is different for at least two histories $(d^{t-1}, \tilde{d}^{t-1})$. Part (B) assumes that the direct effect of λ_k is non-zero in each period for each assignment. Part (C) states the aggregate effect of λ_k on outcomes must be non-zero—that is, that the direct effect $F_{kt}(d_t)$ is not perfectly offset by the effect mediated through previous choices. Part (D) ensures that there is a non-zero effect of previous choices in $t = 2$. Part (E) requires

[3]This type of assumption is similarly required in latent factor models without selection or learning in order to rule out degeneracies (see,e.g., Freyberger, 2018, Assumption L4).

that in $t = 2$ the relative effect of known and unknown λ changes across choices. In the special case where $\lambda_u \in \mathbb{R}$, the condition reduces to $\frac{F_{k2}(d_2)}{F_{u2}(d_2)} \neq \frac{F_{k2}(\tilde{d}_2)}{F_{u2}(\tilde{d}_2)}$, i.e., that the ratio of factor loadings is non-constant across assignments. More generally, for $\lambda_u \in \mathbb{R}^p$, this condition implies that, at least for $t = 2$, the set of assignments must contain at the minimum $p + 1$ elements. Finally, part (F) requires that the unknown factor affects each outcome via a different linear combination.

Define $v_k \mapsto f_{\lambda_k}(v_k, z_1)$ to be the distribution function of λ_k conditional upon the initial exogenous covariates $Z_1 = z_1$. We are now in a position to state our main identification result for the model parameters $\theta = \big((\alpha_t, \beta_t, F_{kt}, F_{ut}, \sigma_t)_{t=1}^T, \Sigma_u, h, f_{\lambda_k}\big) \in \Theta$.

Theorem 1. Suppose the distribution of $(Y_t, D_t, Z_t)_{t=1}^T$ is observed for $T = 2p + 1$ periods, and that Assumptions KL1-KL5 hold. Then θ is point identified.

The proof of this theorem relies on the normality of the error term $\epsilon_t(d)$. The first step is to show that Y_t is normally distributed conditional upon lagged outcomes Y^{t-1}, assignments D^t, covariates Z^t and the known component of the latent individual effect λ_k. This implies that that Y_t conditional upon (Y^{t-1}, D^t, Z^t) is a mixture distribution parameterized by λ_k. Then under the compact support and non-degeneracy assumptions (Assumptions KL4 (A)-(C)), one can apply a result from Bruni and Koch (1985) to identify the aforementioned mixture distribution up to an affine transformation of λ_k. Next, the normalization and regularity assumptions (Assumptions KL3-KL5) are used to pin down the affine transformation, leading to identification of the joint distribution of $(Y^T, D^T, Z^T, \lambda_k)$. Knowledge of this distribution identifies the components of the model related to the known component of the latent individual effect, namely $\big((\alpha_t, \beta_t, F_{kt})_{t=1}^T, h, f_{\lambda_k}\big)$. Thus it remains to disentangle the effect of the learned component (i.e., λ_u) and uncertainty (i.e., $\epsilon_t(d)$) in order to identify $\big((F_{ut}, \sigma_t)_{t=1}^T, \Sigma_u\big)$. To do so, we show that the joint distribution of (Y^T, D^T, Z^T) conditional upon λ_k, suitability weighted by the assignment probabilities, is a normal-weighted mixture of normal distributions. This observation leads to identification $\big((F_{ut}, \sigma_t)_{t=1}^T, \Sigma_u\big)$ from the second moments of the reweighted distribution. See Section B.1 for the formal argument.

Remark 1 (Compact support assumption). Assumption KL4 (B) imposes that the known component of the latent individual effect has bounded support. In applications, it is common to assume λ_k has finite support with known cardinality. Assumption KL4 (B) relaxes this assumption in the sense that the number of support points of λ_k need not be known a priori, and indeed may be infinite. The assumption that the support of the mixing distribution is compact plays an important role in establishing identification.[4]

Remark 2 (Normality of unknown factor). As summarized in Lemma 1, an important advantage of the normality assumptions (Assumption KL2) is the resulting conjugate prior with a tractable closed form. For this reason, these assumptions are very common in the applied literature. In the context of our analysis, the most important implication of these assumptions is to enable identification of the (latent) distribution of $Y_t \mid (\lambda_k, Y^{t-1}, D^t, Z^t)$ from variation in the realized outcome Y_t only. First, the normality assumptions on ϵ_t and λ_u lead to normality of $Y_t \mid (\lambda_k, Y^{t-1}, D^t, Z^t)$, using standard Bayesian arguments. It follows that, for any given $(Y^{t-1}, D^t, Z^t) = (y^{t-1}, d^t, z^t)$, the distribution of $Y_t \mid (Y^{t-1}, D^t, Z^t)$ is a mixture of normal distributions with mixture weights given by the distribution of $\lambda_k \mid (Y^{t-1}, D^t, Z^t)$. Identification then follows from existing results for mixtures of normal distribution (Bruni and Koch, 1985).

This discussion also highlights why we restrict λ_k to be a scalar random variable. Namely, that identification of its distribution is coming from variation in the scalar outcome variable Y_t. If a vector of outcomes were available—that is, if Y_t was vector-valued—then we expect our arguments to easily extend to multivariate λ_k.

Remark 3 (Invariance to normalization). The normalization assumption (Assumption KL3) is a true normalization in the sense that particular meaningful economic

[4]Compactness is used in particular to apply the Stone - Weierstrass approximation theorem, which is a central argument in Bruni and Koch (1985, Theorem 1).

parameters are invariant to the assumption. In particular, we can show that average and quantile structural functions are identified without the normalization assumption. To formalize this notion, define $C_{kt}(d) \equiv \lambda_k^\intercal F_{kt}(d)$, $C_{ut}(d) \equiv \lambda_u^\intercal F_{ut}(d)$ and let $Q_\alpha[X]$ be the α-quantile of the random variable X. Let $z \in \mathcal{S}(Z_t)$ and define the quantile structural functions

$$
\begin{aligned}
s_{1,t}(z,\alpha) =& \alpha_t(d) + z^\intercal \beta_t(d) + Q_\alpha[C_{kt}(d) + C_{ut}(d) + \epsilon_t(d)], \\
s_{2,t}(z,\alpha_1,\alpha_2,\alpha_3) =& \alpha_t(d) + z^\intercal \beta_t(d) + Q_{\alpha_1}[C_{kt}(d)] + Q_{\alpha_2}[C_{ut}(d)] + Q_{\alpha_3}[\epsilon_t(d)],
\end{aligned}
$$

and the average structural function as $s_{3,t}(z) = \alpha_t(d) + Z_t^\intercal \beta_t(d) + \int e F_{C_{kt}+C_{ut}+\epsilon_t}(e)de$.

In Appendix B.1 we prove the following corollary:

Corollary 1. Suppose the Assumptions KL1, KL4 and KL5 hold and that $(\lambda_u \mid Z_1 = z_1, \lambda_k = v_k) \sim N\left(\mu_u, \tilde{\Sigma}_u(z_1)\right)$ and $\epsilon_t(d) \sim N(c_t(d), \sigma_t(d)^2)$. Furthermore, suppose for some $(d_1, d_2, \ldots, d_p)$, $(F_{u1}(d_1)F_{u2}(d_2)\ldots F_{up}(d_p))$ is full rank. Then $s_{1,t}$, $s_{2,t}$ and $s_{3,t}$ are identified on the support of Z_t.

3.3 Pure learning model

This section considers a special case of the model of Section 2, in which all components of the latent individual effect are initially unknown to the decision making agent. That is, $\lambda = \lambda_u$. Without needing to distinguish initially known and unknown heterogeneity, a stronger identification result is achieved. In particular, no parametric restrictions on the distribution of the unobservables are required.

Assumption L1. For any $d \in \mathcal{S}(D_t)$,

$$
F_{\epsilon_t(d)D_tZ_t|Y^{t-1}D^{t-1}Z^{t-1}\lambda} = F_{\epsilon_t(d)}F_{D_t|Y^{t-1}D^{t-1}Z^t}F_{Z_t|Y^{t-1}D^{t-1}Z^{t-1}}.
$$

Assumption L1 adapts Assumption KL1 to reflect there is no initially-known component on latent unobserved heterogeneity.

Assumption L2. (A) The joint density of Y, λ and D, Z admits a bounded density with respect to the product measure of the Lebesgue measure on $\mathcal{S}(Y) \times \mathcal{S}(\lambda)$ and some dominating measure on $\mathcal{S}(D) \times \mathcal{S}(Z)$. All marginal and conditional densities are bounded. (B) $\lambda \mid Z_1$ has full support. (C) The characteristic function of $\epsilon_t(d)$ is non-vanishing, $\mathbb{E}[\epsilon_t] = 0$.

Assumption L2 substantially weakens Assumption KL2 by replacing the normality assumption with a full support assumption. Note that a full support assumption on $Y_t(d)$ is implied by Assumption KL2. Let $\lambda \in \mathbb{R}^p$.

Assumption L3. For some choice sequence $(d_t \colon t = 1, 2, \ldots, p)$, (A) $(F_1(d_1) \ldots F_p(d_p)) = I_{p \times p}$ and (B) $\alpha_t(d_t) = 0$ for each $t = 1, 2, \ldots, p$.

Assumption L4. (A) For each $(y^{t-1}, z^t) \in \mathcal{S}(Y^{t-1}, Z^t)$, $\Pr(D_t = d \mid Y^{t-1} = y^{t-1}, Z^t = z^t) > 0$ for all $d \in \mathcal{S}(D_t)$. (B) The variance-covariance matrix of $\lambda \mid Z_1$ is full rank. (C) The variance-covariance matrix of $(1, Z_t)$ conditional upon $D_t = d_t$ is non-singular.

Assumption L3 are normalization assumptions, which are standard in interactive fixed effect models. An alternative normalization could be placed on the expectation of λ conditional upon Z. Assumption L4 (A) is similar to Assumption KL4 (D). It requires that for each history (y^{t-1}, d^{t-1}, z^t), some units are assigned to $D_t = d_t$ for each $d_t \in \mathcal{S}(D_t)$. This assumption is satisfied in many standard parametric discrete choice models (e.g., Keane and Wolpin (1997)). At the cost of notational burden, this assumption could be weakened to hold for certain sequences of choices. In particular, that for each $d_t \in \mathcal{S}(D_t)$, there is a finite sequence of choice sequences whose first element is the choice sequence of Assumption L3 (A), whose adjacent elements are equal on at least p points of their domain, and whose final element maps t to d_t.

Assumption L5. For any d^T, $\{F_{ut}(d_t) : t = 1, \ldots, T\}$ are linearly independent.

Assumption L5 is a standard assumption in the interactive fixed effect literature (Assumption N6, Freyberger, 2018). Similar to Assumption KL5, it rules out degeneracies by ensuring that the outcome in each period $Y_t(d_t)$ depends on a distinct linear combination of λ_u.

We now define the conditional choice probability function

$$h_t(y^{t-1}, d^t, z^t) := \Pr(D_t = d_t \mid Y^{t-1} = y^{t-1}, D^{t-1} = d^{t-1}, Z^t = z^t),$$

and let $h = (h_1, h_2, \ldots, h_T)$. Since, unlike in Section 3.2, there is no latent variable that enters the CCP function, h is identified directly from the data. As in Section 3.2, we place very little structure on the learning process of decision making agents. This highlights that the core identification results do not rely on structure imposed on the belief formation process. However it is worth emphasizing that, should there be such structure, our identification results would enable identification of the belief formation process. To illustrate this, consider the case that the decision making agents are rational and Bayesian updaters and that the sufficient statistics for λ_u at time t are a known function of the information set and the model parameters. That is, that there is a known function g such that the sufficient statistics equal $g(Y^{t-1}, D^{t-1}, Z^{t-1}, \theta)$, where θ are the model parameters. In this case, identification of θ is sufficient for identification of the beliefs.

To state the identification result, let $\theta_1 = \big((\alpha_t, \beta_t, F_t,)_{t=1}^T\big)$ and define $f_\lambda(v_k, z_1)$ to be the distribution function of λ conditional upon the initial exogenous covariates. Finally, define $f_\epsilon = \big\{f_{\epsilon_t(d)} \colon d \in \mathcal{S}(D_t),\ t = 1, \ldots, T\big\}$. Then, the structural parameter is $\theta = (\theta_1, f_\lambda, f_\epsilon, h)$. The following theorem states that the preceding conditions are sufficient for point identification of θ.

Theorem 2. Suppose the distribution of $(Y_t, D_t, Z_t)_{t=1}^T$ is observed for $T = 2p + 1$ and that Assumptions L1-L5 hold. Then θ is point identified.

The key insight that enables identification of θ is that, under Assumption L1, this is a model of selection on observables. That is, although assignment probabilities depend on unobserved beliefs over λ, they do not depend on the unobserved factor λ itself. It follows that one can control for beliefs at time t by conditioning upon prior outcomes, choices and covariates. This in turn allows us to express the joint distribution of (Y^t, D^t, Z^t), suitably weighted by the assignment probabilities, as a mixture model over the potential outcomes $Y^t(d_t)$ conditional upon the latent factor λ and exogenous covariates Z. From here the arguments of Freyberger (2018) yield identification of the mixture and component distributions. See Section B.2 for details.

Remark 4 (Auxiliary measurements)*.* In some cases, additional unselected noisy measurements of known abilities are available. See, for instance, Cunha et al. (2005) and Heckman and Navarro (2007). With this additional data, sufficient conditions for identification of the distribution of the latent effect are well known in the literature (Hu and Schennach, 2008; Cunha et al., 2010). If the sufficient conditions are satisfied conditional on each $(Y_t, D_t, Z_t)_{t=1}^{T}$, then the joint distribution of $\left((Y_t, D_t, Z_t)_{t=1}^{T}, \lambda_k\right)$ is identified from the auxiliary measurements. From here, one can redefine $Z_t = (Z_t, \lambda_k)$ and the conditions of Theorem 2 are sufficient for distribution-free identification of the model with known and unknown heterogeneity.

4 Estimation

We propose to estimate the model parameters via sieve maximum likelihood. Let $W_i = (Y_{it}, D_{it}, Z_{it} \colon t = 1, \ldots, T)$ and $\theta^* \in \Theta$ be the true value of the parameters. In the following we focus on the model of Section 3.2, although similar conditions could be presented for the model of Section 3.3. The log-likelihood contribution of $W_i = w$

is

$$\ell(w;\theta) = \log \int \prod_{t=1}^{T} \Bigg(\frac{1}{\sigma_t(d_t)} \phi_1 \left(\frac{y_t - \alpha_t(d_t) - z_t^{\intercal}\beta_t(d_t) - v_k F_{kt}(d_t) - v_u^{\intercal} F_{ut}(d_t)}{\sigma_t(d_t)} \right) \\ \times h_t(y^{t-1}, d^t, z^t, v_k) \Bigg) \times \prod_{t=1}^{T-1} g_t(z_{t+1}; y^t, d^t, z^t) \\ \times \frac{1}{\sqrt{|\Sigma_u(z_1)|}} \phi_p \left(\Sigma_u^{-\frac{1}{2}}(z_1)\, v_u \right) \times f_{\lambda_k}(v_k; z_1)\, dv \tag{5}$$

where ϕ_s is the probability distribution function of the standard multivariate normal distribution with s components, g_t is the distribution of Z_{t+1} conditional upon $(Y^t, D^t, Z^t) = (y^t, d^t, z^t)$. There are four components of the likelihood function: the outcomes, the assignment probabilities, the distribution of the covariates, and the distribution of latent factors $(\lambda_u^{\intercal}, \lambda_k)^{\intercal}$.

To estimate θ, let Θ_n be a finite dimensional sieve space that serves as an approximation to Θ. The sieve maximum-likelihood estimator for θ^* is defined as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(w_i;\hat{\theta}) \geq \sup_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} \ell(w_i;\theta) - o_p(1/n) \tag{6}$$

The following result states that under standard conditions (stated in Appendix C.1), if our model is identified, $\hat{\theta}$ is consistent for θ^*,

Theorem 3. Let $(Y_{it}, D_{it}, Z_{it} \colon t = 1, \ldots, T)_{i=1}^{n}$ be i.i.d. data where $T \geq 2p+1$ and Assumptions KL1-KL5 and Assumptions E1-E5 hold. Then $\hat{\theta}$ as defined in Equation (6) is consistent for θ^*.

Researchers are often interested in functionals of the model parameters, such that the variance decompositions discussed in section 2.1. The variance decompositions (7), (8), (9), and (10) involve both the finite dimensional parameters of the model as well as the distribution of λ_k and the CCPs. Therefore, many of the existing results on inference on the finite dimensional parameters of a semiparametric model (e.g. Ai and Chen, 2003) do not directly apply to this setting.

Instead, we next provide an inference result for a plug-in estimator of a more general class of functionals of the model parameters. For a functional f, under a set of smoothness and regularity conditions similar to those given in Chen and Liao (2014), we show establish that the plug-in estimator $f(\hat{\theta})$ has an asymptotically normal distribution and characterize its asymptotic variance.

Theorem 4. Let $(Y_{it}, D_{it}, Z_{it} \colon t = 1, \ldots, T)_{i=1}^{n}$ be i.i.d. data where $T \geq 2p+1$ and that Assumptions KL1-KL5 and Assumptions E1-E13 hold. Then $\sqrt{n}\frac{f(\hat{\theta})-f(\theta^*)}{\|v_n^*\|} \underset{d}{\to} \mathcal{N}(0,1)$ where v_n^* is the sieve Riesz representer of $f(\theta)$ and $\|\cdot\|$ is defined in Equation (13) in the online appendix.

The asymptotic variance and rate of convergence of the plug in sieve estimator depend on the $\|v_n^*\|$. For *regular* functionals, $\|v_n^*\|$ converges to a constant, which implies that the plug-in estimator has a root-n convergence rate. Note, however, that Theorem 4 also allows for the possibility that the sieve variance v_n^* may diverge–that is, that f is an *irregular* functional. In either case, consistent estimators for the sieve variance are available (Chen and Liao, 2014, Section 3).

We leave it to future work to derive primitive conditions under which functionals such as the variances decompositions discussed in section 2.1 satisfy the high level conditions of Theorem 4. However, we do provide in Appendix C.2 lower level conditions under which Theorem 4 holds.

5 Implementation

To implement the sieve maximum likelihood estimation developed in the previous section, first partition θ into $\{F_{\lambda_k|Z_1}\}$ and $\theta_1 := \theta \setminus \{F_{\lambda_k|Z_1}\}$. Integrating over λ_u in (5), we obtain $\ell(w;\theta) = \log \int f(w, v_k; \theta_1) dF_{\lambda_k|Z_1}$ where,

$$f(w, v_k; \theta_1) := \frac{1}{\sqrt{|V(w, v_k; \theta)|}} \phi_1\left(m(w, v_k; \theta)^T V(w, v_k; \theta)^{-1} m(w, v_k; \theta)\right) \times \prod_{t=1}^{T} h_t(d^t, z^t, y^{t-1}, v_k) \times \prod_{t=1}^{T-1} g_t(z_{t+1} \mid z_t, y_t, d_t)$$

where $m(w, v_k; \theta)$ and $V(w, v_k; \theta)$ are the T-dimensional vector and $T \times T$ matrix giving the expected mean and variance of Y^T conditional on $(D^T, Z^T, \lambda_k) = (d^T, z^T, v_k)$. They are defined as follows. Let $F_u(w) = \left[F_{u1}(d_t) \cdots F_{uT}(d_T)\right]$. Then,

$$\begin{aligned}
[m(w, v_k; \theta)]_t &= \alpha_t(d_t) + \beta_t^\intercal z_t + F_k(d_t) v_k, \\
V(w, v_k; \theta) &= F_u(w)^T \Sigma_u F_u(w) + \operatorname{diag}(\sigma_1^2(d_1), \ldots, \sigma_T^2(d_T)),
\end{aligned}$$

There are three non-parametric objects in the likelihood function: h, g, and $F_{\lambda_k|Z_1}$. These can be estimated non-parametrically using a sieve space, or the researcher can impose a parametric form on any . The choice of a model or sieve spaces for h and g are typically context specific. For $F_{\lambda_k|Z_1}$, we propose using a sieve space closely related to estimator discussed in Koenker and Mizera (2014) and Fox et al. (2016). Assume that Z_1 has finite support, $(z_1, \ldots, z_R)$. For each n, fix a grid of support points for λ_k, $\mathcal{S}_n = \{\bar{v}_{1n}, \ldots, \bar{v}_{q_n n}\}$, for some finite q_n. Then, we can use the sieve space for $F_{\lambda_k|Z_1}$:

$$\mathcal{F}_n = \left\{ (v; z_r) \mapsto \sum_{s=1}^{q_n} \omega_{sr} \mathbf{1}\{v \leq \bar{v}_{sn}\} \middle| \omega \in \Delta^R(q_n) \right\}$$

where $\Delta^k(m) = \{x \in \mathbb{R}^{m \times k} : x_{ij} \geq 0, \sum_{s=1}^k x_{is} = 1, 1 \leq i \leq m, 1 \leq j \leq k\}$ is the k-product m-dimensional simplexes. The space $\mathcal{F}_n$ is simply the space of conditional distributions with support contained in $\mathcal{S}_n$. If $\mathcal{S}_n$ becomes dense in $\mathbb{R}$ and the number of terms grows at a suitable rate, this sieve space satisfies the conditions of theorems 3 and 4. This is the sieve space studied in Fox et al. (2016), who show that in several related settings, the rate of $O(n^{-1}\log(n))$ is sufficient for consistency.

This sieve space is useful to make computation tractable while estimating $F_{\lambda_k|Z_1}$ nonparametrically by splitting the optimization problem into two steps. Define the *profile log likelihood* for θ_1 as,

$$\sum_{i=1}^n \tilde{\ell}(w_i; \theta_1) := \max_{F_{\lambda_k|Z_1} \in \mathcal{F}_n} \sum_{i=1}^n \ell(w_i; \theta_1, F_{\lambda_k|Z_1})$$

For each $F \in \mathcal{F}_n$, there exists $\omega \in \Delta^R(q_n)$ such that $\ell(w;\theta_1,F) = \log\int f(w,v;\theta_1)dF(v;z_1) = \log\sum_{s=1}^{q_n}\omega_{sr}\sum_{r=1}^{R}1(z=z_r)f(w,\bar{v}_{sr};\theta_1)$. Notice that fixing θ_1, $\sum_{r=1}^{R}1(z=z_r)f(w,\bar{v}_{sr};\theta_1)$ is constant, so this maximization problem amounts to maximizing a convex objective function in ω subject to the linear constraints, $\omega \in \Delta^R(q_n)$. This inner problem can solved efficiently and reliably using standard software for convex optimization. For example, the algorithm proposed in Kim et al. (2020), is specialized for this setting and implemented in the R package *mixsqp*.

To solve the original maximum likelihood problem, therefore, we simply maximize the profile likelihood function over θ_1. Separating the maximization problem into this inner and outer maximization significantly reduces the dimensionality of the problem, without restricting the distribution of λ_k.

In order to efficiently solve the the outer likelihood maximization problem, it is useful to be able to calculate the gradient of the profile likelihood function with respect to θ_1. This involves differentiating through the solution to the inner optimization problem. In appendix C.3, we provide details of how this derivative is calculated, and in the accompanying R package, we provide code to calculate the derivative of the profile likelihood in this setup.

6 Monte Carlo simulations

In this section, we present results from Monte Carlo simulations which illustrate the computational tractablity and finite-sample performance of the proposed estimator.

The data generating process (DGP) used in the simulations is based on the model in Section 3.2 with both known and unknown heterogeneity. We include two time-invariant covariates, one continuous and one discrete, which are independent of λ_i. Assignment probabilities are derived from a model in which agents maximize the following utility function,

$$u_t(d,\lambda_{k,i},Y_i^{t-1},Z_i,D_i^{t-1}) = \rho E(Y_t(d)|\lambda_{k,i},Y_i^{t-1},Z_i,D_i^{t-1}) + \rho\gamma\mathbf{1}(D_{t,i}=2)\lambda_{k,i} + \nu_{i,t}(d),$$

where $\{\nu_{i,t}(d) : t = 1, 2, 3, d = 1, 2\}$ are mutually independent with an Extreme Value Type 1 distribution. This utility function puts a weight on the expected outcome of their choice and another term which depends on λ_k. This additional term can reflect biased beliefs, heterogeneity in preferences, or a combination of both. λ_k is distributed as a mixture of truncated normal random variables. The parameter values used in the simulations are reported in Appendix C.4.

Table 1: Average computational times by sample size

	N = 250	N = 500	N = 1000	N = 2000	N = 4000
Computational Time	0:24	0:31	0:55	2:15	3:32

We perform a Monte Carlo experiment, estimating parameters of the model with 200 simulations and sample sizes of $250, 500, 1,000, 2,000$ and $4,000$. We use the sieve MLE estimator described in Section 4, maintaining the parametric structure on the assignment probabilities but estimating F_{λ_k} nonparametrically using the sieve space described in Section 5. The number of support points in the estimated distributions, q_n, grows at a rate of $n^{1/3}$, from 62 to 158. There are 32 parameters in θ_1.

With this specfication, computation remains tractable for these sample sizes. Average computational times to solve the profile MLE problem reported in Table 1 run from half a minute to around three and half minutes.

Table 2: Bias and Variance ($\times 1,000$) of Finite Parameter Estimators

	N = 250		N = 500		N = 1000		N = 2000		N = 4000	
	sq bias	var	sq bias	var	sq bias	var	sq bias	var	sq bias	var
$\alpha_1(2)$	71.722	87.915	34.056	60.969	12.915	47.132	0.729	19.025	0.042	5.697
$\alpha_2(1)$	0.148	27.977	0.264	12.384	0.116	7.391	0.002	2.878	0.010	1.376
$\alpha_2(2)$	73.516	108.963	34.175	74.416	12.407	57.187	0.463	25.797	0.027	8.115
$\alpha_3(1)$	0.005	36.562	0.455	13.824	0.197	5.311	0.000	2.237	0.013	0.963
$\alpha_3(2)$	47.840	163.156	32.091	82.423	12.025	62.314	0.593	25.979	0.037	7.322
$\beta_{z1,1}(1)$	0.513	10.084	0.399	5.220	0.137	3.172	0.016	1.486	0.000	0.721
$\beta_{z1,1}(2)$	0.852	15.221	0.304	6.753	0.045	3.349	0.007	1.744	0.002	0.801
$\beta_{z2,1}(1)$	0.837	16.296	0.659	7.859	0.392	4.464	0.038	1.849	0.006	0.803
$\beta_{z2,1}(2)$	1.378	20.814	0.601	12.059	0.093	5.618	0.003	2.687	0.006	1.215
$\beta_{z3,1}(1)$	0.408	9.298	0.243	3.879	0.156	1.886	0.028	1.030	0.007	0.569
$\beta_{z3,1}(2)$	0.383	19.188	0.402	9.105	0.083	4.201	0.010	2.096	0.002	0.861
$\beta_{z1,2}(1)$	0.606	58.909	0.362	23.238	0.361	11.159	0.027	4.772	0.004	2.295
$\beta_{z1,2}(2)$	0.187	46.657	0.218	25.403	0.022	11.158	0.002	5.124	0.010	2.611
$\beta_{z2,2}(1)$	0.005	40.411	0.001	19.844	0.001	9.046	0.002	4.347	0.041	2.476
$\beta_{z2,2}(2)$	0.038	57.755	0.055	26.567	0.000	12.370	0.000	6.761	0.012	3.295
$\beta_{z3,2}(1)$	0.495	40.189	0.079	19.943	0.017	7.637	0.000	3.943	0.022	2.046
$\beta_{z3,2}(2)$	0.102	65.654	0.332	32.107	0.014	15.179	0.024	7.111	0.000	3.438
$F_{k1}(1)$	2.746	27.521	1.701	12.890	0.624	7.268	0.009	3.684	0.001	1.468
$F_{k2}(1)$	1.148	25.977	0.558	10.825	0.226	4.777	0.004	2.594	0.000	1.089
$F_{k2}(2)$	0.869	10.978	0.254	5.825	0.073	2.654	0.007	1.383	0.000	0.743
$F_{k3}(1)$	3.986	33.663	0.873	13.719	0.178	5.683	0.003	3.069	0.000	1.330
$F_{k3}(2)$	5.702	36.861	0.674	12.556	0.224	5.305	0.011	2.408	0.005	1.080
$F_{u1}(2)$	0.979	13.945	0.306	4.733	0.170	2.438	0.015	1.330	0.000	0.605
$F_{u2}(1)$	0.040	8.317	0.027	5.139	0.036	1.947	0.014	1.003	0.002	0.481
$F_{u2}(2)$	1.478	14.880	0.494	6.218	0.130	3.324	0.009	1.522	0.004	0.643
$F_{u3}(1)$	0.446	9.912	0.093	5.003	0.062	2.187	0.030	0.968	0.023	0.469
$F_{u3}(2)$	0.106	21.919	0.101	8.903	0.112	4.148	0.004	2.140	0.005	0.936
$\sigma^2(1)$	0.453	2.477	0.091	1.241	0.030	0.672	0.007	0.298	0.001	0.136
$\sigma^2(2)$	1.228	4.449	0.242	2.237	0.027	1.058	0.016	0.701	0.008	0.332
σ_u^2	0.017	72.902	0.052	41.174	0.043	17.906	0.012	9.337	0.010	4.335
γ	3.791	103.555	0.562	34.226	0.192	18.098	0.357	10.611	0.043	5.779
ρ	3.317	121.851	0.165	51.065	0.031	25.777	0.012	11.329	0.077	4.724

All calculations are based on 200 Monte Carlo simulations of the DGP described in the main text. Squared bias and variance of finite parameter estimates are multiplied times $1,000$

The squared bias and variance of the sieve estimator of the finite parameter sub-vector, θ_1, are presented in Table 2. (Note that all values in Table 2 are multiplied by

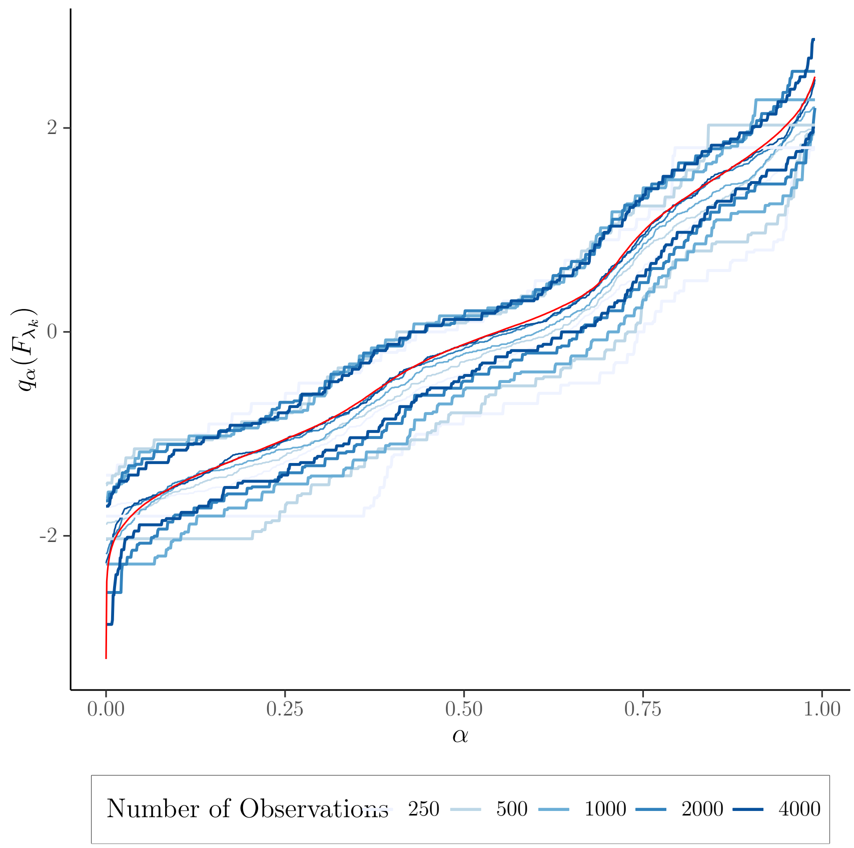
1,000.) For each of the parameters, the bias becomes negligible for all parameters relative to variance as sample size grows. For most parameter execpt the intercepts, this bias is small even for small sample sizes. The variance declines at a rate consistent with $\sqrt{n}$ convergence of the mean squared error. This is consistent with Theorem 4 since the functional mapping the parameter space to a single element of θ_1 is known to be a regular functional.

To present results for the nonparametric estimator of the distribution of known unobserved heterogeneity F_{λ_k}, we focus on the quantiles of F_{λ_k}. Let $q_\alpha(F)$ be the α quantile of a random variable with the distribution F. For each value of $\alpha \in [0,1]$, we calculate the mean and the 5th and 95th percentile of the simulated distribution of the estimator of $q_\alpha(f_{\lambda_k})$. The results are presented in Figure 1. The red line shows the CDF of the true distribution of λ_k, while the blue lines that closely follow the red line are the mean of the simulated distribution of the quantile estimators for each sample size. Darker blue lines represent larger sample sizes. The blue lines above and below the CDF are the 5th and 95th percentiles of the simulated distribution of the quantile estimators.

The results indicate that the bias of the quantile estimators becomes negligible in moderate sample sizes. The estimator broadly captures the shape of the true distribution of λ_k, and also appears to converge toward the true distribution as the sample size grows. We do not provide a formal result on the rate of convergence of this parameter, but we expect this nonparametric estimator to converge at rate a slower than $\sqrt{n}$. At a sample size of $n = 4,000$, the simulated distribution of this estimator is still relatively disperse.

Finally, Table 3 presents the variance and squared bias of the plug-in functional estimator for the functionals discussed in section 2.1. We focus on the decomposition of variance of lifetime earnings into forecastable and non-forecastable components at time zero. Under the decomposition, and setting the discount rate at .5, for each

Figure 1: Quantiles of Estimator of λ_k: 95% Coverage Intervals



Note: The red line shows the true distribution of λ_k. The blue lines show the mean, and the 5th and 95th percentiles of the simulated distribution of the estimate of $q_\alpha(f_{\lambda_k})$.

choice sequence $d = (d_1, d_2, d_3)$, we calculate two functionals:

$$\text{unknown}: \qquad \sigma_u^2 \sum_{1 \leq t_1, t_2 \leq 3} (.95)^{t_1+t_2-2} F_{ut_1}(d_{t_1}) F_{ut_2}(d_{t_2}) + \sum_{1 \leq t \leq 3} (.95)^{2t-2} \sigma_t^2(d_t)$$

$$\text{known}: \qquad \mathbb{V}(\lambda_k) \sum_{1 \leq t_1, t_2 \leq 3} (.95)^{t_1+t_2-2} F_{kt_1}(d_{t_1}) F_{kt_2}(d_{t_2})$$

For moderate sample sizes, the squared bias is small relative to the variance, and like the estimators of θ_1, the variance appears to be consistent with a $\sqrt{r}$ convergence rate.

Table 3: Bias and Variance of the Period 0 Lifetime Earnings Estimators

		N = 250		N = 500		N = 1000		N = 2000		N = 4000	
		sq bias	var	sq bias	var	sq bias	var	sq bias	var	sq bias	var
(1, 1, 1)	known	0.007	0.994	0.000	0.451	0.001	0.214	0.002	0.136	0.001	0.067
(1, 1, 1)	unknown	0.001	3.060	0.001	1.512	0.000	0.677	0.001	0.332	0.000	0.154
(1, 1, 2)	known	0.003	1.456	0.012	0.697	0.005	0.380	0.001	0.225	0.001	0.095
(1, 1, 2)	unknown	0.002	2.324	0.000	1.134	0.001	0.516	0.000	0.268	0.000	0.117
(1, 2, 1)	known	0.313	1.774	0.130	0.931	0.041	0.530	0.000	0.282	0.000	0.111
(1, 2, 1)	unknown	0.031	1.723	0.003	0.853	0.001	0.368	0.000	0.192	0.000	0.087
(1, 2, 2)	known	0.218	3.134	0.173	1.533	0.047	0.876	0.000	0.412	0.000	0.155
(1, 2, 2)	unknown	0.011	1.201	0.006	0.599	0.003	0.284	0.000	0.154	0.000	0.062
(2, 1, 1)	known	0.235	1.492	0.065	0.819	0.018	0.363	0.000	0.223	0.000	0.097
(2, 1, 1)	unknown	0.028	1.755	0.003	0.850	0.007	0.363	0.002	0.165	0.001	0.084
(2, 1, 2)	known	0.143	2.432	0.075	1.125	0.021	0.560	0.000	0.324	0.001	0.137
(2, 1, 2)	unknown	0.009	1.228	0.007	0.595	0.010	0.265	0.001	0.132	0.000	0.066
(2, 2, 1)	known	1.051	3.036	0.295	1.556	0.072	0.734	0.002	0.380	0.000	0.174
(2, 2, 1)	unknown	0.103	1.097	0.019	0.450	0.011	0.191	0.002	0.095	0.001	0.046
(2, 2, 2)	known	0.482	5.843	0.247	2.771	0.032	1.562	0.003	0.761	0.000	0.330
(2, 2, 2)	unknown	0.062	0.786	0.027	0.318	0.015	0.170	0.001	0.088	0.000	0.040

All calculations are based on 200 Monte Carlo simulations of the DGP described in the main text.

7 Conclusion

We provide new identification results for a general class of learning models, that encompasses many of the models that have been considered in the applied literature. We consider an environment where the researcher has access to panel data on choices and realized outcomes only. As such, our results are widely applicable, including in frequent environments where one does not have access to elicited beliefs data or auxiliary selection-free measurements. We show that the model is point-identified under two alternative sets of conditions. Our first set of conditions applies to a version of the learning where we assume that the idiosyncratic shocks from the outcome equations are normally distributed, a restriction that is very commonly imposed in empirical Bayesian learning models. We also show that normality can be relaxed in the case of a pure learning model, and establish identification for this class of models.

We then derive a sieve MLE estimator for the model parameters and a particular class of functionals, which includes as a leading special cases the predictable and unpredictable outcome variances. Notably, these variances can in turn be used to evaluate the relative importance of uncertainty versus heterogeneity in lifecycle earnings variability (Cunha et al., 2005). Under certain regularity conditions, the resulting estimators are consistent and asymptotically normal. Importantly for practical purposes, the profile likelihood based estimation procedure proposed in this paper can be implemented at a modest computational cost.

Bibliography

Abbring, J. and Campbell, J. (2005), A firm's first year, Technical report, Tinbergen Institute Discussion Paper 05-046/3.

Ackerberg, D. A. (2003), 'Advertising, learning, and consumer choice in experience good markets: an empirical examination', *International Economic Review* **44**(3), 1007–1040.

Aguirregabiria, V., Gu, J. and Luo, Y. (2021), 'Sufficient statistics for unobserved heterogeneity in structural dynamic logit models', *Journal of Econometrics* **223**(2), 280–311.

Aguirregabiria, V. and Jeon, J. (2020), 'Firms' beliefs and learning: Models, identification, and empirical evidence', *Review of Industrial Organization* **56**, 203–235.

Ai, C. and Chen, X. (2003), 'Efficient estimation of models with conditional moment restrictions containing unknown functions', *Econometrica* **71**(6), 1795–1843.

Antonovics, K. and Golan, L. (2012), 'Experimentation and job choice', *Journal of Labor Economics* **30**(2), 333–366.

Arcidiacono, P. (2004), 'Ability sorting and the returns to college major', *Journal of Econometrics* **121**(1-2), 343–375.

Arcidiacono, P., Aucejo, E., Maurel, A. and Ransom, T. (2023), College attrition and the dynamics of information revelation, Technical report, Duke University.

Arcidiacono, P. and Miller, R. A. (2011), 'Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity', *Econometrica* **79**(6), 1823–1867.

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279.

Bruni, C. and Koch, G. (1985), 'Identifiability of continuous mixtures of unknown gaussian distributions', *The Annals of Probability* pp. 1341–1357.

Bunting, J. (2022), 'Continuous permanent unobserved heterogeneity in dynamic discrete choice models', *arXiv preprint arXiv:2202.03960* .

Chan, T. Y. and Hamilton, B. H. (2006), 'Learning, private information, and the economic evaluation of randomized experiments', *Journal of Political Economy* **114**(6), 997–1040.

Chen, X. (2007), 'Large sample sieve estimation of semi-nonparametric models', *Handbook of econometrics* **6**, 5549–5632.

Chen, X. and Liao, Z. (2014), 'Sieve m inference on irregular parameters', *Journal of Econometrics* **182**(1), 70–86.

Chen, X., Liao, Z. and Sun, Y. (2014), 'Sieve inference on possibly misspecified semi-nonparametric time series models', *Journal of Econometrics* **178**, 639–658.

Ching, A. T., Erdem, T. and Keane, M. P. (2013), 'Learning models: An assessment of progress, challenges, and new developments', *Marketing Science* **32**(6), 913–938.

Compiani, G. and Kitamura, Y. (2016), 'Using mixtures in econometric models: a brief review and some new results', *Econometrics Journal* **19**(3), C95–C127.

Coscelli, A. and Shum, M. (2004), 'An empirical model of learning and patient spillovers in new drug entry', *Journal of Econometrics* **122**(2), 213–246.

Crawford, G. and Shum, M. (2005), 'Uncertainty and learning in pharmaceutical demand', *Econometrica* **73**(4), 1137–1173.

Cunha, F. and Heckman, J. J. (2016), 'Decomposing trends in inequality in earnings into forecastable and uncertain components', *Journal of Labor Economics* **34**(S2), S31–S65.

Cunha, F., Heckman, J. J. and Navarro, S. (2005), 'Separating uncertainty from heterogeneity in life cycle earnings', *Oxford Economic Papers* **57**(2), 191–261.

Cunha, F., Heckman, J. J. and Schennach, S. M. (2010), 'Estimating the technology of cognitive and noncognitive skill formation', *Econometrica* **78**(3), 883–931.

D'Haultfoeuille, X. (2011), 'On the completeness condition in nonparametric instrumental problems', *Econometric Theory* **27**(3), 460–471.

Erdem, T. and Keane, M. P. (1996), 'Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods', *Marketing Science* **15**(1), 1–20.

Fox, J. T., il Kim, K. and Yang, C. (2016), 'A simple nonparametric approach to estimating the distribution of random coefficients in structural models', *Journal of Econometrics* **195**(2), 236–254.

Freyberger, J. (2018), 'Non-parametric panel data models with interactive fixed effects', *The Review of Economic Studies* **85**(3), 1824–1851.

Gobillon, L. and Magnac, T. (2016), 'Regional policy evaluation: Interactive fixed effects and synthetic controls', *The Review of Economics and Statistics* **98**(3), 535–551.

Gong, Y. (2019), Signal-based learning models without the rational expectations assumption: Identification and counterfactuals, Technical report, Technical report, mimeo.

Heckman, J. J. and Navarro, S. (2007), 'Dynamic discrete choice and dynamic treatment effects', *Journal of Econometrics* **136**(2), 341–396.

Hincapié, A. (2020), 'Entrepreneurship over the life cycle: Where are the young entrepreneurs?', *International Economic Review* **61**(2), 617–681.

Hu, Y. and Schennach, S. M. (2008), 'Instrumental variable treatment of nonclassical measurement error models', *Econometrica* **76**(1), 195–216.

Hu, Y. and Shum, M. (2012), 'Nonparametric identification of dynamic models with unobserved state variables', *Journal of Econometrics* **171**(1), 32–44.

Huggett, M., Ventura, G. and Yaron, A. (2011), 'Sources of lifetime inequality', *American Economic Review* **101**(7), 2923–2954.

Kasahara, H. and Shimotsu, K. (2009), 'Nonparametric identification of finite mixture models of dynamic discrete choices', *Econometrica* **77**(1), 135–175.

Keane, M. P. and Wolpin, K. I. (1997), 'The career decisions of young men', *Journal of political Economy* **105**(3), 473–522.

Kim, Y., Carbonetto, P., Stephens, M. and Anitescu, M. (2020), 'A fast algorithm for maximum likelihood estimation of mixture proportions using sequential quadratic programming', *Journal of Computational and Graphical Statistics* **29**(2), 261–273.

Kinsler, J. and Pavan, R. (2021), 'Local distortions in parental beliefs over child skill', *Journal of Political Economy* **129**(1), 81–100.

Kitamura, Y. and Laage, L. (2018), 'Nonparametric analysis of finite mixtures', *arXiv preprint arXiv:1811.02727v1* .

Koenker, R. and Mizera, I. (2014), 'Convex optimization, shape constraints, compound decisions, and empirical bayes rules', *Journal of the American Statistical Association* **109**(506), 674–685.

Magnac, T. and Thesmar, D. (2002), 'Identifying dynamic discrete decision processes', *Econometrica* **70**(2), 801–816.

Miller, R. A. (1984), 'Job matching and occupational choice', *Journal of Political Economy* **92**(6), 1086–1120.

Pastorino, E. (2015), 'Job matching within and across firms', *International Economic Review* **56**(2), 647–671.

Pastorino, E. (2022), 'Careers in firms: The role of learning about ability and human capital acquisition'. Forthcoming in the *Journal of Political Economy.*

Sasaki, Y. (2015), 'Heterogeneity and selection in dynamic panel data', *Journal of Econometrics* **188**(1), 236–249.

Sasaki, Y. and Hu, Y. (2018), 'Heterogeneity and selection in dynamic panel data', *Econometric Theory* **34**(1), 166–185.

Stange, K. M. (2012), 'An empirical investigation of the option value of college enrollment', *American Economic Journal: Applied Economics* **4**(1), 49–84.

Stinebrickner, T. and Stinebrickner, R. (2012), 'Learning about academic ability and the college dropout decision', *Journal of Labor Economics* **30**(4), 707–748.

Thomas, J. (2019), 'The signal quality of grades across academic fields', *Journal of Applied Econometrics* **34**(4), 566–587.

Yang, N. (2020), 'Learning in retail entry', *International Journal of Research in Marketing* **37**(2), 336–355.

Zafar, B. (2011), 'How do college students form expectations?', *Journal of Labor Economics* **29**(2), 301–348.

A Variance decompositions

We therefore consider a class of variance decomposition parameters which includes both the $t = 1$ orthogonal decomposition as well as other decompositions that incorporate these learning and selection effects.

To define this class of parameters, consider a weighted sum of potential outcomes, $Y(\omega, d) = \sum_t \omega_t Y_t(d_t)$ for a sequence of choices $d = \{d_t : t \leq T\}$ and weights, $\omega = \{\omega_t : t \leq T\}$. Lifetime earnings as defined in Cunha and Heckman are a special case of $Y(\omega, d)$, with $\omega_t = 1(t \geq t_0)(1-\rho)^{t_0 - t}$, for some discount rate $\rho < 1$.

Next, define the agent's information set, $\mathcal{I}_t = \{D_s, Y_s : s < t\} \cup \{Z_s : s \leq t\} \cup \{\lambda_k\}$. Restricting attention to weighted sums where $\omega_s = 0$ for $s \leq t$, the variance of $Y(\omega, d)$ conditional on $\mathcal{I}_t$ can be understood as the variance that is due to the agent's uncertainty about a particular weighted sum of future potential outcomes at period t. We refer to this as the *posterior variance*, because this is derived from the posterior distribution of λ_u after performing a Bayesian update with the information up to $\mathcal{I}_t$.

In particular, Lemma 1 the next section implies that the posterior variance has the form, $\mathbb{V}\left(Y(\omega, d) | \mathcal{I}_t\right) = V^u(D^{t-1}, Z^t; \omega, d)$, where,

$$V^u(D^t, Z; \omega, d) := \sum_{t_1, t_2 > t} \omega_{t_1} \omega_{t_2} F_{ut_t}(d_{t_1})^T \Sigma_t(D^{t-1}, Z_1) F_{ut_2}(d_{t_2}) + \sum_{t_1 > t} \omega_{t_1}^2 \sigma_{t_1}^2(d_{t_1})$$

where $\Sigma_t(D^{t-1}, Z_1)$ is the posterior variance of λ_u as written in Lemma 1.

At $t = 1$, the following variance decomposition provides a natural way to quantify the relative importance of this source of variance:

$$\mathbb{V}(Y(\omega, d) | Z_1 = z) = V^u(\emptyset, z; \omega, d) + \sum_{t_1, t_2 \geq 1} \omega_{t_1} \omega_{t_2} F_{kt_1}(d_{t_1}) F_{kt_2}(d_{t_2}) \mathbb{V}(\lambda_k | Z_1 = z) \quad (7)$$

This corresponds to the decomposition in Cunha and Heckman (2016) and has the simple interpretation that the first term is the portion of variance in the lifetime earnings that is due to uncertainty and the second part is due to heterogeneity.

For $t > 1$, the analysis is more complicated. For any $t > 1$, $V^u(D^{t-1}, z; \omega, d) < V^u(\emptyset, z; \omega, d)$, because the realized outcomes are informative about the λ_u, but agents also select D^{t-1} based on their private information and the sequence of observed outcomes. There are several different possible ways of quantifying the relative importance of uncertainty which treat selection differently. The following are three alternatives:

$$\mathbb{V}(Y(\omega, d)|D^t = d, Z_1 = z) = V^u(d^t, z; \omega, d) + \mathbb{V}(\mathbb{E}(Y(\omega, d)|\mathcal{I}_t)|D^t = d^t, Z = z) \tag{8}$$

$$\mathbb{V}(Y(\omega, d)|Z_1 = z) = \mathbb{E}(V^u(D^t, z; \omega, d)) + \mathbb{V}(\mathbb{E}(Y(\omega, d)|\mathcal{I}_t)|Z = z) \tag{9}$$

$$\mathbb{V}(Y(\omega, d)|Z_1 = z) = V^u(d^t, z; \omega, d) + \tilde{\mathbb{V}}(\tilde{\mathbb{E}}(Y(\omega, d)|\mathcal{I}_t)|Z = z) \tag{10}$$

Decomposition (8) compares the variance of uncertainty to the total variance conditional on choosing the sequence d^t. This is a natural parameter to consider, but it does not disentangle selection from learning. In particular, the ratio, $V^u(d^t, z; \omega, d)/\mathbb{V}(Y(\omega, d)|D^t = d^t, |Z_1 = z)$ reflects both the effect of learning in the numerator and selection in the denominator.

Decomposition (9) compares the total variance $Y(\omega, d)$ to the expected posterior varince of $Y(\omega, d)$ after t periods. The expectation of $V^u(d^t, z; \omega, d)$ can be understood as the uncertainty that a randomly choosen person would have in period t after observering their outcomes and endogenously choosing actions based on that information and their private information.

Finally decomposition (10) is based on a counterfactual distribution. Here $\tilde{\mathbb{E}}$ and $\tilde{\mathbb{V}}$ represent the expectation and variance in a counterfactual distribution where D_s does not depend on $\mathcal{I}_s$ for $s \leq t$. This decomposition compares the variance in $Y(\omega, d)$ which is due to uncertainty vs. known heterogeneity among people randomly assigned to the choice sequence d^t.

B Identification proofs and auxiliary results

In this section, we use the following notations: ϕ denotes the standard normal p.d.f.; $\mathcal{S}(X)$ represents the support of a random variable X.

B.1 Proofs for Section 3.2

Proof of Lemma 1. We proceed inductively. First, by Assumption KL2 and the definition of (E_1, Σ_1), $\lambda_u \mid (Z_1, \lambda_k) \sim \mathcal{N}(E_1, \Sigma_1)$. Second, for $t > 1$ suppose $\lambda_u \mid (Y^{t-2}, D^{t-2}, Z^{t-1}) \sim \mathcal{N}(E_{t-1}, \Sigma_{t-1})$ and that

$$
\begin{aligned}
&f_{\lambda_u|Y^{t-1}D^{t-1}Z^t\lambda_k}(v_u; y^{t-1}, d^{t-1}, z^t, v_k)\\
&\propto_{(1)} f_{\lambda_u|Y^{t-2}D^{t-2}Z^{t-1}\lambda_k}(v_u; y^{t-2}, d^{t-2}, z^{t-1}, v_k)\\
&\quad \times f_{Y_{t-1}D_{t-1}Z_t|Y^{t-2}D^{t-2}Z^{t-1}\lambda}(y_{t-1}, d_{t-1}, z_t; y^{t-2}, d^{t-2}, z^{t-1}, v)\\
&= f_{\lambda_u|Y^{t-2}D^{t-2}Z^{t-1}\lambda_k}(v_u; y^{t-2}, d^{t-2}, z^{t-1}, v_k) f_{Z_t|Y^{t-1}D^{t-1}Z^{t-1}\lambda}(z_t; y^{t-1}, d^{t-1}, z^{t-1}, v)\\
&\quad \times f_{Y_{t-1}(d_{t-1})|Y^{t-2}D^{t-1}Z^{t-1}\lambda}(y_{t-1}; y^{t-2}, d^{t-1}, z^{t-1}, v) f_{D_{t-1}|Y^{t-2}D^{t-2}Z^{t-1}\lambda}(d_{t-1}; y^{t-2}, d^{t-2}, z^{t-1}, v)\\
&\propto_{(2)} f_{\lambda_u|Y^{t-2}D^{t-2}Z^{t-1}\lambda_k}(v_u; y^{t-2}, d^{t-2}, z^{t-1}, v_k) f_{Y_{t-1}(d_{t-1})|Z_{t-1}\lambda}(y_{t-1}; z_{t-1}, v)\\
&\propto_{(3)} \exp\left(-\frac{1}{2}(v_u - E_t)^\intercal \Sigma_t^{-1}(v_u - E_t)\right) \phi\left(\frac{y_t - \alpha_t(d_t) - z_t^\intercal \beta_t(d_t) - v_k F_{kt}(d_t) - v_u^\intercal F_{ut}(d_t)}{\sigma_t(d_t)}\right)\\
&\propto \exp\left(-\frac{1}{2}(v_u - E_t)^\intercal \Sigma_t^{-1}(v_u - E_t)\right)\\
&\quad \times \exp\Bigg(-\frac{1}{2}\left(v_u - F_{ut}(d_t)\left(F_{ut}(d_t)^\intercal F_{ut}(d_t)\right)^{-1}\left(y_t - \alpha_t(d_t) - z_t^\intercal \beta_t(d_t) - v_k F_{kt}(d_t)\right)\right)^\intercal\\
&\quad \times \frac{F_{ut}(d_t) F_{ut}(d_t)^\intercal}{\sigma_t^2(d_t)}\left(v_u - F_{ut}(d_t)\left(F_{ut}(d_t)^\intercal F_{ut}(d_t)\right)^{-1}\left(y_t - \alpha_t(d_t) - z_t^\intercal \beta_t(d_t) - v_k F_{kt}(d_t)\right)\right)\Bigg)\\
&=_{(4)} \exp\left(-\frac{1}{2}(v_u - E_{t+1})^\intercal \Sigma_{t+1}^{-1}(v_u - E_{t+1})\right).
\end{aligned}
$$

Display (1) follows from Bayes' theorem. Display (2) holds since Assumption KL1 has the following three implications: first $Z_t \perp\!\!\!\perp \lambda \mid (Y^{t-1}, D^{t-1}, Z^{t-1})$; second $\epsilon_{t-1}(d_{t-1}) \perp\!\!\!\perp (Y^{t-2}, D^{t-1}, Z^{t-1}, \lambda) \Rightarrow \epsilon_{t-1}(d_{t-1}) \perp\!\!\!\perp (Y^{t-2}, D^{t-1}, Z^{t-2},) \mid (Z_{t-1}, \lambda) \Rightarrow Y_{t-1}(d_{t-1}) \perp\!\!\!\perp (Y^{t-2}, D^{t-1}, Z^{t-2}) \mid (Z_{t-1}, \lambda)$; third $D_{t-1} \perp\!\!\!\perp \lambda_u \mid (Y^{t-2}, D^{t-2}, Z^{t-1}, \lambda_k)$.

Display (3) holds from the induction assumption and Assumptions KL1 and KL2. Display (4) follows from the definitions in Lemma 1. $\square$

The following corollary of Lemma 1 will be used in the proof to Theorem 1.

Corollary 2. Let Assumptions KL1 and KL2 hold. Then Y_t conditional upon $(Y^{t-1}, D^t, Z^t, \lambda_k) = (y^{t-1}, d^t, z^t, v_k)$ is distributed

$$N\left(\alpha_t(d_t) + z_t^\intercal \beta_t(d_t) + \lambda_k F_{kt}(d_t) + E_t^\intercal F_{ut}(d_t), \quad F_{ut}(d_t)^\intercal \Sigma_t F_{ut}(d_t) + \sigma_t^2(d_t)\right)$$

Proof of Corollary 2. For $t > 1$,

$$\begin{aligned}
& f_{Y_t|Y^{t-1}D^tZ^t\lambda_k}(y_t; y^{t-1}, d^t, z^t, v_k) \\
&= \int f_{Y_t(d_t)|Y^{t-1}D^tZ^t\lambda}(y_t; y^{t-1}, d^t, z^t, v) f_{\lambda_u|Y^{t-1}D^tZ^t\lambda_k}(v_u; y^{t-1}, d^t, z^t, v_k) dv_u \\
&=_{(1)} \int f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) f_{\lambda_u|Y^{t-1}D^{t-1}Z^t\lambda_k}(v_u; y^{t-1}, d^{t-1}, z^t, v_k) dv_u \\
&\propto_{(2)} \int \phi\left(\frac{y_t - \alpha_t(d_t) - z_t^\intercal \beta_t(d_t) - v_k F_{kt}(d_t) - v_u^\intercal F_{ut}(d_t)}{\sigma_t(d_t)}\right) \\
&\times \exp\left((v_u - E_t)^\intercal \Sigma_t^{-1}(v_u - E_t)\right) dv_u \\
&= \phi\left(\frac{y_t - \alpha_t(d_t) - z_t^\intercal \beta_t(d_t) - v_k F_{kt}(d_t) - E_t^\intercal F_{ut}(d_t)}{\sqrt{F_{ut}^\intercal(d_t)\Sigma_t F_{ut}(d_t) + \sigma_t^2(d_t)}}\right)
\end{aligned}$$

Equality (1) holds because Assumption KL1 implies $Y_t(d_t) \perp\!\!\!\perp (Z^{t-1}, D^t, Y^{t-1} \mid Z_t, \lambda)$ and $D_t \perp\!\!\!\perp \lambda_u \mid (\lambda_k, Z^t, Y^{t-1}, D^{t-1})$. Equality (2) holds because Assumption KL1 and KL2 imply Lemma 1 and $\epsilon_t(d) \mid (Z_t, \lambda) \sim N(0, \sigma_t(d)^2)$. A similar argument applies for $t = 1$. $\square$

Proof of Theorem 1. The proof is in three parts. First, we use Corollary 2 and Bruni and Koch (1985) to identify the distribution of $Y_t \mid (D^t, Y^{t-1}, Z^t, \lambda_k)$ up to an affine transformation of λ_k. The second part is to use the normalization (Assumption KL3) to show that the affine transformation is the identity function. Finally, we use identification of the distribution of $(Y^t, D^t, Z^t, \lambda_k)$ to identify the distribution of (Y^t, D^t, Z^t, λ).

Part 1. By Corollary 2, $f_{Y_t|Y^{t-1}D^tZ^t}(y_t; y^{t-1}, d^t, z^t) =$

$$\int f_{Y_t|Y^{t-1}D^tZ^t\lambda_k}(y_t; y^{t-1}, d^t, z^t, v_k) df_{\lambda_k|Y^{t-1}D^tZ^t}(v_k; y^{t-1}, d^t, z^t) dv_k.$$

I.e., the nonparametrically identified $f_{Y_t|Y^{t-1}D^tZ^t}(y_t; y^{t-1}, d^t, z^t)$ is a mixture of Gaussians. To identify the component and mixture distributions, we will apply Bruni and Koch (1985, Theorem 3). First, for any t, d^t and z_1, define

$$\Lambda = \left\{ v_k \mapsto (\alpha_t(d_t) + z_t^\intercal \beta(d_t) + v_k \mu_1(\theta^t) + \mu_2(\theta^t), \sigma(\theta^t)) \colon \theta^t \in \Theta^t \right\},$$

where $\theta^t = (\alpha^t, \beta^t, F_k^t, F_u^t, \sigma^t, \Sigma_u)$, Θ^t is the corresponding subset of Θ, and

$$\begin{aligned}
\mu_1(\theta^t) &= \left(F_{kt}(d_t) - F_{ut}^\intercal(d_t)\Sigma_t \sum_{s=1}^{t-1} F_{us}(d_s) \frac{F_{ks}(d_s)}{\sigma_s^2(d_s)} \right), \\
\mu_2(\theta^t) &= F_{ut}(d_t)^\intercal \Sigma_t \sum_{s=1}^{t-1} F_{us}(d_s) \frac{y_{is} - \alpha_s(d_s) - z_{is}^\intercal \beta_s(d_s)}{\sigma_s^2(d_s)}, \\
\sigma(\theta^t) &= F_{ut}(d_t)^\intercal \Sigma_t F_{ut}(d_t) + \sigma_t^2(d_t).
\end{aligned}$$

For example, for $t = 1$, $\sigma(\theta^1) = F_{u1}(d_1)^\intercal \Sigma_u(z_1) F_{u1}(d_1) + \sigma_1^2(d_1)$. Notice that $\lambda_k F_{kt}(d_t) + E_t^\intercal F_{ut}(d_t) = \lambda_k \mu_1(\theta^t) + \mu_2(\theta^t)$. Under Assumptions KL4(A,B,C) and KL5(C), 4, $\Lambda \subset \Lambda_4$ where Λ_4 is defined in Bruni and Koch (1985, p. 1344). Thus Bruni and Koch (1985, Theorem 3) applies and

$$\left((\alpha_t(d_t) + z_t^\intercal \beta(d_t) + \pi(v_k)\mu_1(\theta^t) + \mu_2(\theta^t),\ \sigma(\theta^t),\ df_{\lambda_k|Y^{t-1}D^tZ^t}(\pi(v_k); y^{t-1}, d^t, z^t) \right) \tag{11}$$

is identified with π an unknown non-constant affine function which may depend on the history (y^{t-1}, d^t, z^t).

To conclude this part, we show that if π is identity for each history (d^s, y^{s-1}, z^s) $s = 1, 2, \ldots, t$, then $f_{Y^tD^tZ^t\lambda_k}(y^t, d^t, z^t, v_k)$ is point identified (i.e., $(\alpha_t(d_t), \beta_t(d_t), \mu_1(\theta^t), \mu_2(\theta^t), \sigma(\theta^t), df_{\lambda_k|Y^{t-1}D^tZ^t}(v_k; y^{t-1}, d^t, z^t))$ is point identified). For $t = 1$, as $(\mu_1(\theta^1), \mu_2(\theta^1)) = (F_{k1}(d_1), 0)$ identification follows immediately from display (11) and Assumption KL4(E).

Now suppose $(\alpha_s(d_s), \beta_s(d_s), \mu_1(\theta^s), \mu_2(\theta^s), \sigma(\theta^s), df_{\lambda_k|Y^{s-1}D^sZ^s}(v_k; y^{s-1}, d^s, z^s))$ is point identified for each $s < t$. From equation (11),

$$\left((\alpha_t(d_t) + z_t^\intercal \beta(d_t) + v_k \mu_1(\theta^t) + \mu_2(\theta^t),\ \sigma(\theta^t),\ df_{\lambda_k|Y^{t-1}D^tZ^t}(v_k; y^{t-1}, d^t, z^t)\right)$$

is identified for every (d^t, y^{t-1}, z^t). $\mu_1(\theta^t)$ is identified from variation in v_k. Assumption KL4(E) implies identification of

$$\left(\alpha_t(d_t) + \mu_2(\theta^t),\ \beta(d_t)\right).$$

Then $\mu_2(\theta^t) = \sum_{s=1}^{t-1} (y_s - \alpha_s(d_s) - z_s^\intercal \beta_s(d_s)) \frac{\partial}{\partial y_s} (\alpha_t(d_t) + \mu_2(\theta^t))$, from which follows identification of $\alpha_t(d_t)$.

Part 2. To show the affine function π is identity, we proceed in three steps. First, we show π is identity for the normalized choice $D_1 = d_1$, which provides identification of $\mathcal{S}(\lambda_k)$. Second, we use knowledge of $\mathcal{S}(\lambda_k)$ to prove the affine function must satisfy $|\frac{\partial}{\partial v}\pi(v)| = 1$ for any history (y^{t-1}, d^t, z^t). Third, we use restrictions on the panel dimension to conclude π is identity for each history (y^{t-1}, d^t, z^t).

First, Let $t = 1$ and d_1 as in Assumption KL3(A), then since $\mu_1(\theta^1) = F_{k1}(d_1)$ and $\mu_2(\theta^1) = 0$, from Part 1 we have identified:

$$\left(z_1^\intercal \beta(d_1) + \pi(v_k),\ \sigma(\theta^1),\ df_{\lambda_k|D^1Z^1}(\pi(v_k); d^1, z^1)\right),$$

with $\pi(v_k) = \pi_0 + \pi_1 v_k$. Since $F_{k1}(d_1) = 1$, $\pi_1 = 1$. We now show $\pi_0 = 0$. First notice that π_0 does not depend on (d_1, z_1) since the support of $\lambda_k \mid (D_1 = d_1, Z_1 = z_1)$ is the same for each (d_1, z_1). Now suppose that for any z_1, $z_1^\intercal \beta(d_1) + \pi_0 = z_1^\intercal \tilde{\beta}(d_1) + \tilde{\pi}_0$. In particular for $\tilde{z}_1 \neq z_1$, $(z_1 - \tilde{z}_1)^\intercal(\beta(d_1) - \tilde{\beta}(d_1)) = 0$. By Assumption KL4(E), we conclude $\beta(d_1) - \tilde{\beta}(d_1) = 0$. This in conjunction with $\alpha_1(d_1) = 0$ gives $\pi_0 = \tilde{\pi}_0 = 0$.

Second, for each fixed (y^{t-1}, d^t, z^t), $df_{\lambda_k|Y^{t-1}D^tZ^t}(\pi(v_k); y^{t-1}, d^t, z^t)$ is identified from part 1. Then, by Assumption KL4(D),

$$\mathcal{S}(\lambda_k) = df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+] = (df_{\lambda_k|Y^{t-1}D^tZ^t} \circ \pi)^{-1}[\mathbb{R}_+],$$

where $R_+ = \{x \in \mathbb{R} \colon x > 0\}$. And since π is bijective,

$$(\pi \circ df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t})[\mathbb{R}_+] = df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+].$$

In particular

$$\pi(\sup dF^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+]) = \sup df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+],$$
$$\pi(\inf df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+]) = \inf df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+].$$

The only affine functions that satisfy these identities are $\pi^+(v) = v$ and $\pi^-(v) = (\bar{v} + \underline{v}) - v$ for $\underline{v} = \inf df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+]$ and $\bar{v} = \sup df^{-1}_{\lambda_k|Y^{t-1}D^tZ^t}[\mathbb{R}_+]$.

Third, it remains to show that $\pi = \pi^+$. To do so, it will be useful to define:

$$\tilde{\mu}_{ts}(d^{t-1}) = \Sigma_t \frac{F_{us}(d_s)}{\sigma_s^2(d_s)}$$

It will also be useful to denote $\mu_j(d^t) = \mu_j(\theta^t)$, to emphasize the dependence of μ_j on d^t. Then notice $\mu_1(d^t) = F_{kt}(d_t) - F_{ut}(d_t) \sum_{s=1}^{t-1} \tilde{\mu}_{ts}(d^{t-1}) F_{ks}(d_s)$ and $\mu_2(d^t) = F_{ut}(d_t)^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{ts}(d^{t-1}) \left(Y_{is} - \alpha_s(d_s) - Z_{is}^\intercal \beta_s(d_s)\right)$.

The proof is inductive. First consider $t = 1$. From Assumption KL3(A), $F_{k1}(d_1) = 1$. For $\tilde{d}_1 \neq d_1$, given part 1 and $\frac{\partial}{\partial x}|\pi(x)| = 1$, $F_{k1}(\tilde{d}_1)$ is identified up to sign as $\frac{\partial}{\partial v_k}\left(\alpha_1(\tilde{d}_1) + z_1^\intercal \beta(\tilde{d}_1) + F_{k1}(\tilde{d}_1)\pi(v_k)\right)$. Similarly, for $d^2 = (d_2, d_1)$, $\mu_1(d^2) = F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1)$ is identified up to sign and $F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1)$ are identified as $\frac{\partial}{\partial x}\left(\alpha_2(d_2) + \beta(d_2)' z_2 + \pi(v_k)\mu_1(d^2) + \mu_2(d^2)\right)$ for $x = v_k$ and $x = y_1$, respectively (since $\mu_2(d^2) = F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1)(y_1 - \alpha_1(d_1) - z_1^\intercal \beta_1(d_1))$ and $\mu_1(d^2)$ does not depend on y_1). Repeating this argument for the choice sequence $(\tilde{d}_1, d_2)$ yields identification of $(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1))$ up to sign and $F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1)$.

Summarizing, we have identification of $F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1)$, $F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1)$, and $(-1)^{j_1} F_{k1}(\tilde{d}_1)$, $(-1)^{j_{d_2}}(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1))$, and $(-1)^{\tilde{j}_{d_2}}(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1))$ with $(j_1, \tilde{j}_{d_2}, j_{d_2}) \in \{0,1\}^3$. We show only the correct choice of sign will satisfy

$$\begin{aligned} &F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1)(-1)^{j_1} F_{k1}(\tilde{d}_1) + (-1)^{\tilde{j}_{d_2}}(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1)) \\ &= F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1) + (-1)^{j_{d_2}}(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1)). \end{aligned}$$

Suppose $j_{d_2} = 0$. It is straightforward to show the following implications:

$$(j_1, \tilde{j}_{d_2}) = (1,1) \implies F_{k2}(d_2) = 0,$$
$$(j_1, \tilde{j}_{d_2}) = (0,1) \implies F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) = 0,$$
$$(j_1, \tilde{j}_{d_2}) = (1,0) \implies F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1)) = 0.$$

The three implications contradict Assumptions KL5 (B), (C) and (D), respectively. Now suppose $j_{d_2} = 1$, then

$$(j_1, \tilde{j}_{d_2}) = (0,0) \implies F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1) = 0,$$
$$(j_1, \tilde{j}_{d_2}) = (1,1) \implies F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1) = 0,$$
$$(j_1, \tilde{j}_{d_2}) = (0,1) \implies F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1) = 0,$$
$$(j_1, \tilde{j}_{d_2}) = (1,0) \implies F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1) = 0.$$

The first three implications contradict Assumptions KL5 (C), (D) and (A), respectively. Finally, we show that the final equality contradicts Assumption KL5 (E). For each $d \in \{d_{2,i} \in \mathcal{S}(D_2) : i = 1, 2, \ldots, p\} \cup \{\tilde{d}_{2,i} \in \mathcal{S}(D_2) : i = 1, 2, \ldots, p\}$ of Assumption KL5 (E), by considering the sequences (d_1, d), $(\tilde{d}_1, d)$, $(-1)^{j_d}(F_{k2}(d) - F_{u2}(d)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1))$ and $(-1)^{\tilde{j}_d}(F_{k2}(d) - F_{u2}(d)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1))$ is identified with $(j_d, \tilde{j}_d) \in \{(1,0), (0,0)\}$. Since $F_{k1}(\tilde{d}_1) \neq 0$ by Assumption KL5 (B), for the sign of $F_{k1}(\tilde{d}_1)$ to be constant across sequences, we can rule out all signs except $\left(j_1, (j_{d_{2,i}}, \tilde{j}_{d_{2,i}}, j_{\tilde{d}_{2,i}}, \tilde{j}_{\tilde{d}_{2,i}} : i = 1, \ldots, p)\right) \in \{(0, (0,0,0,0)^p), (1, (1,0,1,0)^p)\}$. If $\left(j_1, (j_{d_{2,i}}, \tilde{j}_{d_{2,i}}, j_{\tilde{d}_{2,i}}, \tilde{j}_{\tilde{d}_{2,i}} : i = 1, \ldots, p)\right) = (1, (1,0,1,0)^p)$, then

$$0 = \operatorname{vec}\left(F_{k2}(d_{2,1}), \ldots, F_{k2}(d_{2,k})\right) - \left(F_{u2}(d_{2,1}) \ldots F_{u2}(d_{2,k})\right)^\intercal \left(\tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) + \tilde{\mu}_{21}(d^1) F_{k1}(d_1)\right)$$
$$= \operatorname{vec}\left(F_{k2}(\tilde{d}_{2,1}), \ldots, F_{k2}(\tilde{d}_{2,k})\right) - \left(F_{u2}(\tilde{d}_{2,1}) \ldots F_{u2}(\tilde{d}_{2,k})\right)^\intercal \left(\tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) + \tilde{\mu}_{21}(d^1) F_{k1}(d_1)\right),$$

which contradicts Assumption KL5(E).

For the induction step, suppose π is identity for each history (d^s, y^{s-1}, z^s) $s = 1, \ldots, t-1$ and consider choice sequences $d^{t-1} = ((d^{t-2})^\intercal, d_{t-1})^\intercal$ and $\tilde{d}^{t-1} = ((d^{t-2})^\intercal, \tilde{d}_{t-1})^\intercal$ for $d_{t-1} \neq \tilde{d}_{t-1}$.

From part 1, $(\alpha_t(d_t)+\beta_t(d_t)'z_t+\mu_1(d^{t-2},d,d_t)\pi(v_k)+\mu_2(d^{t-2},d,d_t))$ is identified for $d = d_{t-1},\tilde{d}_{t-1}$ and $\pi \in \{\pi^+,\pi^-\}$. By the preceding arguments, $(F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(d^{t-1})F_{ks}(d_s))$, $F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(\tilde{d}^{t-1})F_{ks}(\tilde{d}_s)$. $(-1)^{j_1}\left(F_{kt}(d_t)-F_{ut}(d_t)^\intercal\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(d^{t-1})F_{ks}(d_s)\right)$, and $(-1)^{j_2}\left(F_{kt}(d_t)-F_{ut}(d_t)^\intercal\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(\tilde{d}^{t-1})F_{ks}(\tilde{d}_s)\right)$ is identified with $(j_1,j_2)\in\{0,1\}^2$

As before we show that only that only $(j_1,j_2)=(0,0)$ is consistent with the identity

$$
\begin{aligned}
&(-1)^{j_1}\left(F_{kt}(d_t)-F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(d^{t-1})F_{ks}(d_s)\right)+F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(d^{t-1})F_{ks}(d_s)\\
&=(-1)^{j_2}\left(F_{kt}(d_t)-F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(\tilde{d}^{t-1})F_{ks}(\tilde{d}_s)\right)+F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(\tilde{d}^{t-1})F_{ks}(\tilde{d}_s)
\end{aligned}
$$

For this, consider

$$
\begin{aligned}
(j_1,j_2)=(0,1) &\implies \left(F_{kt}(d_t)-F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(\tilde{d}^{t-1})F_{ks}(\tilde{d}_s)\right)=0,\\
(j_1,j_2)=(1,0) &\implies \left(F_{kt}(d_t)-F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(d^{t-1})F_{ks}(d_s)\right)=0,\\
(j_1,j_2)=(1,1) &\implies F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(d^{t-1})F_{ks}(d_s)-F_{ut}(d_t)\sum_{s=1}^{t-1}\tilde{\mu}_{ts}(\tilde{d}^{t-1})F_{ks}(\tilde{d}_s)=0,
\end{aligned}
$$

which contradict Assumptions KL5 (C), (C) and (A), respectively. Thus π is the identity function for the history (d^t,y^{t-1},z^t).

Part 3. From parts 1 and 2, $f_{Y^TD^TZ^T\lambda_k}$, and thus h, is identified. First,

$$
\begin{aligned}
&f_{Y^TD^TZ^T\lambda_k}\left(y^T,d^T,z^T,v_k\right)\\
&=\int f_{Y^T(d^T)D^TZ^T\lambda}\left(y^T,d^T,z^T,v\right)dv_u\\
&=\int f_{Y_T(d_T)|Z_T,\lambda}\left(y_T;z_T,v\right)f_{D_T|Y^{T-1}D^{T-1}Z^T\lambda_k}(d_T;y^{T-1},d^{T-1},z^T,v_k)\\
&\quad\times f_{Z_T|Y^{T-1}D^{T-1}Z^{T-1}}(z_T;y^{T-1},d^{T-1},z^{T-1})\ldots f_{Y_1(d_1)|Z_1\lambda}\left(y_1;z_1,v\right)\\
&\quad\times f_{D_1|Z_1\lambda_k}(d_1;z_1,v_k)f_{\lambda_u|Z_1\lambda_k}(v_u;z_1,v_k)f_{Z_1\lambda_k}(z_1,v_k)dv_u.
\end{aligned}
$$

This implies that on the support of $f_{Y^TD^TZ^T\lambda_k}$,

$$\frac{f_{Y^TD^TZ^T\lambda_k}\left(y^T,d^T,z^T,v_k\right)}{f_{D_1Z_1\lambda_k}(d_1,z_1,v_k)\prod_{t=2}^T f_{D_tZ_t|Y^{t-1}D^{t-1}Z^{t-1}\lambda_k}(d_t;y^{t-1},d^{t-1},z^t,v_k)}$$
$$=\int\prod_{t=1}^T f_{Y_t(d_t)|Z_t\lambda}\left(y_t;z_t,v\right)f_{\lambda_u|\lambda_kZ_1}(v_u;v_k,z_1)dv_u.$$

I.e., the function is equal to the probability density function of a jointly normal random variable with mean

$$\left(\alpha_t(d_t)+z_t^\intercal\beta_t(d_t)+v_kF_{kt}(d_t)\right)_{t=1}^T,$$

and covariance matrix

$$F_u(d)^\intercal\Sigma_u(z_1)F_u(d)+\operatorname{diag}\left(\sigma_t^2(d_t)\colon t=1,\ldots,T\right),$$

where $F_u(d)=(F_{u1}(d_1)F_{u2}(d_2)\ldots F_{uT}(d_T))$. From parts 1 and 2, the components of the mean function are identified. The components of the covariance matrix are identified under Assumptions KL3 (B) and KL5 (F). □

Proof of Corollary 1. Fix $(d_1,d_2,\ldots,d_p)$ as in the statement and define $F_u=(F_{u1}(d_1)F_{u2}(d_2)\ldots F_{up}(d_p))$, $\tilde{\lambda}_u=F_u^\intercal(\lambda_u-\mu_u)$, $\tilde{\epsilon}_t(d)=\epsilon_t(d)-c_t(d)$, $\tilde{\lambda}_k=b+F_{k1}(d_1)\lambda_k$ where $b=\alpha_1(d_1)+F_{u1}(d_1)^\intercal\mu_u+c_1(d_1)$. Finally, define $\tilde{F}_{kt}(d_t)=F_{k1}(d_1)^{-1}F_{kt}(d_t)$, $\tilde{F}_{ut}(d_t)=F_u^{-1}F_{ut}(d_t)$, and $\tilde{\alpha}_t(d)=\alpha_t(d)-\tilde{F}_{kt}(d)b+F_{ut}(d)^\intercal\mu_u+c_t(d)$. We then have that

$$Y_t(d)=\tilde{\alpha}_t(d)+Z_t^\intercal\beta_t(d)+\tilde{\lambda}_u^\intercal\tilde{F}_{ut}(d)+\tilde{\lambda}_k\tilde{F}_{kt}(d)+\tilde{\epsilon}_t(d),$$

$E[\tilde{\epsilon}_t(d)]=0$ and $E[\tilde{\lambda}_u\mid Z_1=z,\lambda_k=v_k]=0$ so that the reparameterized model satisfies Assumption KL2 (with $\Sigma_u(Z_1)=F_u^\intercal\tilde{\Sigma}_u(Z_1)F_u$). Also, $\tilde{F}_{k1}(d_1)=1$, $\tilde{\alpha}_1(d_1)=0$ and $\tilde{F}_p\equiv\left(\tilde{F}_{u1}(d_1)\tilde{F}_{u2}(d_2)\ldots\tilde{F}_{up}(d_p)\right)=I_{p\times p}$ so the reparameterized model satisfies Assumption KL3. By Theorem 1, $\tilde{\theta}=\left((\tilde{\alpha}_t,\beta_t,\tilde{F}_{kt},\tilde{F}_{ut},\sigma_t^2)_{t=1}^T,\Sigma_u,\tilde{h},F_{\tilde{\lambda}_k}\right)$ is identified, where $\tilde{h}$ and $F_{\tilde{\lambda}_k}$ are the CCPs and conditional distribution of $\tilde{\lambda}_k$, respectively.

This, in turn, implies the identification of the distribution of C_{jt} for $j = k, u$. Finally,

$$
\begin{aligned}
&\tilde{\alpha}_t + z^\intercal \beta_t + Q_\alpha[\tilde{C}_{kt} + \tilde{C}_{ut} + \tilde{\epsilon}] \\
=&\alpha_t - \tilde{F}_{kt} b + F_{ut}\mu_u + c_t + z^\intercal \beta_t + Q_\alpha[\tilde{C}_{kt} + \tilde{C}_{ut} + \tilde{\epsilon}] \\
=&\alpha_t - \tilde{F}_{kt} b + F_{ut}\mu_u + c_t + z^\intercal \beta_t + Q_\alpha[C_{kt} + \tilde{F}_{kt} b + C_{ut} - F_{ut}^\intercal \mu_u + \epsilon_t - c_t] \\
=&\alpha_t + z^\intercal \beta_t + Q_\alpha[C_{kt} + C_{ut} + \epsilon_t]
\end{aligned}
$$

□

B.2 Proofs for Section 3.3

In this section denote $\mathcal{L} = \{m\colon \mathbb{R}^k \to \mathbb{R} : \sup_{a\in\mathbb{R}^k} |m(a)| < \infty, \int |m(a)|da < \infty\}$ and $\mathcal{L}_A = \{m\colon \mathbb{R}^k \to \mathbb{R} : \sup_{a\in\mathbb{R}^k} |m(a)| < \infty,\ \int |m(a)| f_A(a)da < \infty\}$ for a random variable A with p.d.f. f_A.

Proof. Let $z \in \mathcal{S}(Z)$ be given and fix a choice sequence $d = (d_1, d_2, \ldots, d_T)$ whose first p elements satisfy Assumption L3, and define $W_1 = (Y_1, \ldots, Y_p)$, $W_2 = Y_{p+1}$ and $W_3 = (Y_{p+2}, \ldots, Y_T)$. Let $L_{123} : \mathcal{L}_{W_3} \to \mathcal{L}$ and $L_{13} : \mathcal{L}_{W_3} \to \mathcal{L}$ be defined as $[L_{123}m](w_1) =$

$$\int \frac{f_{YDZ}(y, d, z)}{f_{D_1 Z_1}(d_1, z_1) \prod_{t=2}^{T} f_{D_t Z_t | Y^{t-1} D^{t-1} Z^{t-1}}(d_t, z_t; y^{t-1}, d^{t-1}, z^{t-1})} m(w_3) dw_3,$$

and $[L_{13}m](w_1) = \int [L_{123}m](w_1) dw_2$. In addition, define

$$
\begin{aligned}
L_{1\lambda} : \mathcal{L} \to \mathcal{L} \qquad & [L_{1\lambda}m](w_1) = \int \prod_{t=1}^{p} f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) m(v) dv, \\
L_{\lambda 3} : \mathcal{L}_{W_3} \to \mathcal{L} \qquad & [L_{\lambda 1}m](v) = \int \prod_{t=p+2}^{T} f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) f_{\lambda|Z_1}(v; z_1) m(w_1) dw_1, \\
D_\lambda : \mathcal{L}_\Lambda \to \mathcal{L}_\Lambda \qquad & [D_\lambda m](v) = f_{Y_{p+1}(d_{p+1})|Z_{p+1}\lambda}(y_{p+1}; z_{p+1}, v) m(v).
\end{aligned}
$$

The following derivation shows $L_{123} = L_{1\lambda} D_\lambda L_{\lambda 3}$. First,

$$\begin{aligned} f_{YDZ}(y,d,z) &= \int f_{YDZ\lambda}(y,d,z,v)dv \\ &= \int f_{Y_T(d_T)|Z_T\lambda}(y_T; z_T, v) f_{D_T Z_T|Y^{T-1}D^{T-1}Z^{T-1}}(d_T, z_T; y^{T-1}, d^{T-1}, z^{T-1}) \\ &\quad \times f_{Y_{T-1}(d_{T-1})|Z_{T-1}\lambda}(y_{T-1}; z_{t-1}, v) \dots f_{D_1 Z_1}(d_1, z_1) f_{\lambda|Z_1}(v; z_1)dv. \end{aligned}$$

Then, by Assumption L4 (A),

$$\begin{aligned} &\frac{f_{YDZ}(y,d,z)}{f_{D_1 Z_1}(d_1,z_1)\prod_{t=2}^T f_{D_t Z_t|Y^{t-1}D^{t-1}Z^{t-1}}(d_t, z_t; y^{t-1}, d^{t-1}, z^{t-1})} \\ &= \int \prod_{t=1}^T f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) f_{\lambda|Z_1}(v; z_1)dv, \end{aligned}$$

and therefore that

$$\begin{aligned} [L_{123}m](w_1) &= \int \left(\int \prod_{t=1}^T f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) f_{\lambda|Z_t}(v; z_t)dv \right) m(w_3)dw_3 \\ &= \int \prod_{t=1}^{p+1} f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) \left(\int \prod_{t=p+2}^{T} f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) f_{\lambda|Z_t}(v) m(w_3)dw_3 \right) dv \\ &= \int \prod_{t=1}^{p} f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v) \left(f_{Y_{p+1}(d_{p+1})|Z_{p+1}\lambda}(y_{p+1}; z_{p+1}, v)[L_{\lambda 3}m](v) \right) dv \\ &= \int \int \prod_{t=1}^{p} f_{Y_t(d_t)|Z_t\lambda}(y_t; z_t, v)[D_\lambda L_{\lambda 3}m](v)dv \\ &= [L_{1\lambda} D_\lambda L_{\lambda 3} m](w_1), \end{aligned}$$

and $L_{123} = L_{1\lambda} D_\lambda L_{\lambda 3}$. Similarly, $L_{13} = L_{1\lambda} L_{\lambda 3}$.

From here, Assumptions L1, L2, L3, L4 (B), and L5 imply the arguments of Theorem 1 Freyberger (2018) apply[5], so that $F_t(d_t)$, $f_{Y_t(d_t)|Z_t\lambda}(\cdot; z_t, \cdot)$ and $f_{\lambda|Z_1}(\cdot; z_1)$

[5]The listed assumptions imply the assumptions of Freyberger (2018, Theorem 1) with the primary exception of Assumption L1 that differs from Assumption N5 in Freyberger (2018) by allowing period t variables to impact the evolution of period t' covariates for $t' > t$. However, since Assumption L1 implies $f_{Y_t(d_t)|Z_t\lambda}(y; z, v) = f_{\epsilon_t(d_t)}(y - \alpha_t(d_t) - \beta_t(d_t)^\intercal z - F_t^\intercal v)$, Freyberger (2018, Lemma 1) and D'Haultfoeuille (2011) may be applied with small modification.

are identified for each t for the given (d_t, z). Given identification of $f_{Y_t(d_t)|Z_t\lambda}(\cdot; z_t, \cdot)$ for each $z_t \in \mathcal{S}(Z_t)$ and $F_t(d_t)$, Assumption L4 (C) implies identification of $\alpha_t(d_t)$ and $\beta_t(d_t)$ and thus $f_{\epsilon_t(d_t)}$.

Next, given an arbitrary t and d_t, define $\tilde{d}$ by replacing the tth element of d with d_t. Then let ρ be a permutation $(1, 2, \ldots, T) \mapsto (t_1, t_2, \ldots, t_T)$ such that $t \mapsto t_1$ and define $\tilde{W}_1 = (Y_{t_1}, Y_{t_2}, \ldots, Y_{t_p})$, $\tilde{W}_2 = (Y_{t_{p+1}}, Y_{t_{p+1}}, \ldots, Y_{t_T})$,

$$\tilde{L}_{2\lambda} : \mathcal{L} \to \mathcal{L} \qquad [\tilde{L}_{2\lambda} m](\tilde{w}_2) = \int \prod_{i=p+1}^{T} f_{Y_{t_i}(d_{t_i})|Z_{t_i}\lambda}(y_{t_i}; z_{t_i}, v) f_{\lambda|Z_1}(v; z_1) m(v) dv,$$

$$\tilde{L}_{\lambda 1} : \mathcal{L}_{\tilde{W}_1} \to \mathcal{L} \qquad [\tilde{L}_{\lambda 1} m](v) = \int \prod_{i=1}^{p} f_{Y_{t_i}(d_{t_i})|Z_{t_i}\lambda}(y_{t_i}; z_{t_i}, v) m(\tilde{w}_1) d\tilde{w}_1,$$

and $\tilde{L}_{21} : \mathcal{L}_{\tilde{W}_1} \to \mathcal{L}$ as

$$[\tilde{L}_{21} m](\tilde{w}_2) = \int \frac{f_{YDZ}(y, d, z)}{f_{D_1 Z_1}(d_1, z_1) \prod_{t=2}^{T} f_{D_t Z_t | Y^{t-1} D^{t-1} Z^{t-1}}(d_t, z_t; y^{t-1}, d^{t-1}, z^{t-1})} m(\tilde{w}_1) d\tilde{w}_1.$$

As before, $\tilde{L}_{21} = \tilde{L}_{2\lambda} \tilde{L}_{\lambda 1}$. Since $\tilde{L}_{2\lambda}$ and $\tilde{L}_{21}$ are identified and injective, $\tilde{L}_{\lambda 1}$ is identified by $\tilde{L}_{2\lambda}^{-1} \tilde{L}_{21} = \tilde{L}_{\lambda 1}$ and thus $\alpha_t(d_t), \beta_t(d_t), F_t(d_t), f_{\epsilon(d_t)}$. $\square$

C Estimation appendix

C.1 Consistency of sieve MLE

In this section we introduce conditions for the sieve maximum likelihood estimator defined in Equation (6) to be consistent for the true model parameters. We begin by imposing smoothness restrictions on the unknown functions. To do so, given $\gamma > 0$, $\omega \geq 0$ and $\mathcal{X}$ a subset of a Euclidean space, let $\Lambda^{\lambda}(\mathcal{X})$ denote a Hölder space equipped with the Hölder norm $\|h\|_{\Lambda^\gamma}$ (that is, for k the largest integer smaller than γ, $\Lambda^{\lambda}(\mathcal{X})$ is a space of functions $h\colon \mathcal{X} \to \mathbf{R}$ having at least k continuous derivatives, the kth of which is Hölder continuous with exponent $\gamma - k$). Then define a weighted Hölder ball with radius $c \in (0, \infty)$ as $\Lambda_c^{\gamma,\omega}(\mathcal{X}) = \{h \in \Lambda^\gamma(\mathcal{X})\colon \|h(\cdot)[1 + \|\cdot\|_E^2]^{-\omega}\|_{\Lambda^\gamma} \leq c\}$, where $\|\cdot\|_E$ is the Euclidean norm.

Without loss of generality, suppose the CCP function $h_t(d^t, z^t, y^{t-1}, v_k)$ depends on (d^t, z^t, y^{t-1}) via some measureable vector-valued function $(d^t, z^t, y^{t-1}) \mapsto j_t$ which is known up to $\big((\alpha_{st}, \beta_{st}, F_{kst}, F_{ust}, \sigma_{st})_{st=1}^T, \Sigma_u\big)$. This is without loss of generality since the function may be identity. Other examples include rational learning where $j_t \in \mathbb{R}^{p(p+3)/2+2}$ includes sufficient statistics for λ_u, and a sort of myopia where $j_t \in \mathbb{R}^{3+2}$ depends on the history only via the previous period $(d_{t-1}, z_{t-1}, y_{t-1})$. Write $J_t = (J_{1,t}^\intercal, J_{2,t}^\intercal)^\intercal$ and $Z_t = (Z_{1,t}^\intercal, Z_{2,t}^\intercal)^\intercal$ where $J_{1,t}, Z_{1,t}$ are continuous random variables and $J_{2,t}, Z_{2,t}$ are random variables with finite support and, with some abuse of notation, redefine the CCP function as $h_t(j_{1,t}, j_{2,t}, v_k)$. Define

$$\begin{aligned}
\mathcal{H}_t &= \Lambda_c^{\gamma_1,\omega_1}\left(\mathcal{S}(\lambda_k) \times \mathcal{S}(J_{1,t})\right), \\
\mathcal{F} &= \{f \colon \mathcal{S}(\lambda_k, Z_{1,1}) \to \mathbb{R} \big| f(\cdot, z_1) \text{ is càdlàg }, f(v, \cdot) \in \Lambda_c^{\gamma_2,\omega_2}(\mathcal{S}(Z_{1,1}))\} \\
\mathcal{G}_t &= \Lambda_c^{\gamma_3,\omega_3}\left(\mathcal{S}(Z_{1,t+1}) \times \mathcal{S}(Y_t) \times \mathcal{S}(Z_{1,t})\right).
\end{aligned}$$

The use of a weighted Holder space enables us to allow the support of the continuous random variables to be unbounded. Though not required for consistency, Assumption E6 places restrictions on $(\gamma_1, \gamma_2, \gamma_3)$, the parameters that govern the smoothness of the function classes. Next, to simplify notation we make the following assumption which strengthens Assumption KL1:

Assumption E1. For any t, $F_{Z_{t+1}|Y^tD^tZ^t} = F_{Z_{t+1}|Y_tD_tZ_t}$.

Define $k_{1,t} = |\mathcal{S}(J_{2,t})|$, $k_2 = |\mathcal{S}(Z_{2,1})|$, and $k_{3,t} = |\mathcal{S}(Z_{2,t+1}, D_t, Z_{2,t})|$. Notice that $\Theta = \Theta_1 \times \mathcal{H}_1^{k_{1,1}} \times \cdots \times \mathcal{H}_T^{k_{1,T}} \times \mathcal{F}^{k_2} \times \mathcal{G}_1^{k_{3,1}} \times \cdots \times \mathcal{G}_{T-1}^{k_{3,T-1}}$ and we denote an element of Θ as $\theta = (\theta_1, h_1, \ldots, h_T, f_\lambda, g_1, \ldots, g_{T-1})$. Define the norms on $\mathcal{H}_t^{k_{1,t}}$, $\mathcal{F}^{k_2}$ and $\mathcal{G}_t^{k_{3,t}}$ as follows:

$$\begin{aligned}
\|h_t\|_{\infty,\omega_1} &= \sup_{j_2 \in \mathcal{S}(J_{2,t})} \|h_t(\cdot, j_2, \cdot)[1 + \|\cdot\|_E^2]^{-\omega_1}\|_\infty, \\
\|f_\lambda\|_{\infty,\omega_2} &= \sup_{z_2 \in \mathcal{S}(Z_{2,1})} \|f_\lambda\left(\cdot, (\cdot, z_2)\right)[1 + \|\cdot\|_E^2]^{-\omega_2}\|_\infty, \\
\|g_t\|_{\infty,\omega_3} &= \sup_{(z_2', d, z_2) \in \mathcal{S}(Z_{2,t+1}, D_t, Z_{2,t})} \|g_t\left((\cdot, z_2'); \cdot, d, (\cdot, z_2)\right)[1 + \|\cdot\|_E^2]^{-\omega_3}\|_\infty,
\end{aligned}$$

where $\|\cdot\|_\infty$ is the uniform norm. Finally, define a metric d on Θ as

$$d(\theta,\tilde{\theta}) = \|\theta_1 - \tilde{\theta}_1\|_E + \sum_{t=1}^{T} \|h_t - \tilde{h}_t\|_{\infty,\tilde{\omega}_1} + \|f_\lambda - \tilde{f}_\lambda\|_{\infty,\tilde{\omega}_2} + \sum_{t=1}^{T-1} \|g_t - \tilde{g}_t\|_{\infty,\tilde{\omega}_3},$$

for scalars $\tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3$. Now, let $\mathcal{H}_{n,t}$, $\mathcal{F}_n$ and $\mathcal{G}_{n,t}$ be sieve spaces for $\mathcal{H}_t$, $\mathcal{F}$ and $\mathcal{G}_t$ respectively. Then $\Theta_n = \Theta_1 \times \mathcal{H}_{n,1}^{k_{1,1}} \times \dots \mathcal{H}_{n,T}^{k_{1,T}} \times \mathcal{F}_n^{k_2} \times \mathcal{G}_{n,1}^{k_{3,1}} \times \cdots \times \mathcal{G}_{n,T-1}^{k_{3,T-1}}$ and

$$\frac{1}{n}\sum_{i=1}^{n} \ell(w_i;\hat{\theta}) \geq \sup_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} \ell(w_i;\theta) - o_p(1/n).$$

Assumption E2. $\theta^* \in \Theta$ and (Θ, d) is compact.

Assumption E3. For each $n \geq 1$, $\Theta_n \subseteq \Theta_{n+1} \subseteq \Theta$ and Θ_n is compact under d. As $n \to \infty$, $\min_{\theta\in\Theta_n} d(\theta,\theta_0) \to 0$.

Assumption E4. $\mathrm{E}[\ell(W_i,\theta)]$ is continuous at $\theta = \theta^*$

Assumption E5.

(i) For each n, $\mathrm{E}[\sup_{\theta\in\Theta_n} \ell(W_i,\theta)|]$ is finite.

(ii) There is a non-zero $s < \infty$ and integrable random variable $g(W_i)$ such that $\forall\ \theta, \tilde{\theta} \in \Theta_n$, $d(\theta,\tilde{\theta}) < \delta \implies |\ell(W_i,\theta) - \ell(W_i,\tilde{\theta})| \leq \delta^s g(W_i)$.

(iii) For all $\delta > 0$, $\log N(\delta^{1/s}, \Theta_n, d) = o(n)$.

The identification assumptions imply $\theta^* = \arg\max_{\theta\in\Theta} \mathrm{E}[\ell(W_i,\theta)]$ and for all $\theta \in \Theta \setminus \{\theta^*\}$, $\mathrm{E}[\ell(W_i,\theta^*)] \geq \mathrm{E}[\ell(W_i,\theta)]$. By assuming compactness of Θ, we ensure that θ^* is a well-separated maximum of $\mathrm{E}[\ell(W_i,\theta)]$. Assumption E3 requires the sieve space Θ_n to be a good approximation to Θ. Assumption E4 requires the population criterion to be continuous. Finally, Assumption E5 is related to Condition 3.5 in Chen (2007).

Theorem 3 follows from Remark 3.3 in Chen (2007), so its proof is omitted.

C.2 Plug-in sieve estimator

We assume a linear sieve space and limit its complexity.

Assumption E6. (1) $\mathcal{H}_{n,t}$, $\mathcal{F}_n$ and $\mathcal{G}_{n,t}$ are linear sieves of length $M_{Hn,t}$, M_{Fn} and $M_{Gn,t}$ respectively, where $M_{Hn,t} = O(n^{\frac{1}{2\gamma_1/(1+\dim(J_{1,t}))+1}})$, $M_{Fn} = O(n^{\frac{1}{2\gamma_2/(1+\dim(Z_{1,1}))+1}})$, and $M_{Gn,t} = O(n^{\frac{1}{2\gamma_3/(\dim(Z_{1,t+1})+1+\dim(Z_{1,t}))+1}})$. (2) $\min\left\{\frac{\gamma_1}{1+\dim(J_{1,t})}, \frac{\gamma_2}{1+\dim(Z_{1,1})}, \frac{\gamma_3}{\dim(Z_{1,t+1})+1+\dim(Z_{1,t})}\right\} > 1/2$.

Assumption E6 controls the rate at which the number of sieve terms grow. To achieve this, part (2) of Assumption E6 requires that the CCP functions have adequate smoothness. In applications, it is common to assume a parametric model for h_t, in which case the above curse-of-dimensionality is avoided.

The next assumption strengthens E3 and ensures the number of sieve terms grows sufficiently fast.

Assumption E7. $\min_{\theta\in\Theta_n} d(\theta,\theta^*) = o(n^{-1/4})$.

Assume ℓ is pathwise differentiable and define an inner product on Θ as

$$\langle\theta_1-\theta^*,\theta_2-\theta^*\rangle = -\frac{\partial^2}{\partial\tau_1\partial\tau_2}E\left[\ell\left(W_i,\theta^*+\tau_1\left(\theta_1-\theta^*\right)+\tau_2\left(\theta_2-\theta^*\right)\right)\right]|_{\tau_1=0,\tau_2=0}, \tag{12}$$

with the corresponding norm for $\theta\in\Theta$ as

$$\|\theta-\theta^*\|^2 \equiv -\frac{\partial^2}{\partial\tau^2}E\left[\ell\left(W_i,\theta^*+\tau\left(\theta-\theta^*\right)\right)\right]\Big|_{\tau=0}. \tag{13}$$

Assumption E8. There is $C_1>0$ such that for all small $\varepsilon>0$

$$\sup_{\{\theta\in\Theta_n:\|\theta-\theta^*\|\leqslant\varepsilon\}} \mathrm{Var}\left(\ell\left(W_i,\theta\right)-\ell\left(W_i,\theta^*\right)\right) \leqslant C_1\varepsilon^2$$

Assumption E9. For any $\delta>0$, there exists a constant $s\in(0,2)$ such that

$$\sup_{\{\theta\in\Theta_n:\|\theta-\theta^*\|\leqslant\delta\}} \left|\ell\left(W_i,\theta\right)-\ell\left(W_i,\theta^*\right)\right| \leqslant \delta^s U\left(W_i\right)$$

with $E\left(\left[U\left(W_i\right)\right]^\gamma\right)\leqslant C_2$ for some $\gamma\geqslant 2$.

The following theorem is now a consequence of Theorem 3.2 in Chen (2007) or Theorem 1 in Shen and Wong (1994).

Theorem 5. Let $(Y_{it}, D_{it}, Z_{it} \colon t = 1, \ldots, T)_{i=1}^{n}$ be i.i.d. data where $T \geq 2p+1$ and Assumptions KL1-KL5 and Assumptions E1-E9 hold. Then $\|\hat{\theta} - \theta^*\| = o_p(n^{-1/4})$.

Given the preceding result, we focus on a a shrinking neighborhood of θ^*. Let

$$\mathcal{N}_0 \equiv \left\{\theta \in \Theta \colon \|\theta - \theta^*\| = o(n^{-1/4}),\ d(\theta, \theta^*) = o(1)\right\},$$

and $\mathcal{N}_n \equiv \mathcal{N}_0 \cap \Theta_n$. Define $\theta_n^* = \operatorname{argmin}_{\theta \in \mathcal{N}_n} \|\theta - \theta^*\|$. Let $\mathcal{V}$ denote the closed (under $\|\cdot\|$) linear span of $\mathcal{N}_0$ centered at θ^*, and define $\mathcal{V}_n$ as the analogous closure of $\mathcal{N}_n$.

Then we define a linear approximation to $\ell(W, \theta) - \ell(W, \theta^*)$ as the directional derivative of ℓ at (W, θ^*) in the direction $(\theta - \theta^*)$:

$$\frac{\partial \ell\left(W, \theta^*\right)}{\partial \theta}[\theta - \theta^*] \equiv \left.\frac{\partial \ell\left(W, \theta^* + \tau(\theta - \theta^*)\right)}{\partial \tau}\right|_{\tau=0}.$$

Likewise, let $\frac{\partial f(\theta^*)}{\partial \theta}[v] = \left.\frac{\partial f(\theta^* + \tau v)}{\partial \tau}\right|_{\tau=0}$ for any $v \in \mathcal{V}$.

Assumption E10. Let $\mathcal{T}$ be an epsilon ball about $0 \in \mathbb{R}$. (1) for all $\theta \in \mathcal{N}_0$ and W, the derivative $\partial \ell\left(W, \theta^* + \tau(\theta - \theta^*)\right)/\partial \tau$ exists for all $\tau \in \mathcal{T}$; (ii) for all $\theta \in \mathcal{N}_0$, $\mathrm{E}\left[\ell\left(W, \theta^* + \tau\left(\theta - \theta^*\right)\right)\right]$ is finite for each $\tau \in \mathcal{T}$; (3) for all $\theta \in \mathcal{N}_0$, $\mathrm{E}\left[\sup_{\tau \in \mathcal{T}} \left|\frac{\partial}{\partial \tau} \ell\left(W, \theta^* + \tau\left[\theta - \theta^*\right]\right)\right|\right] < \infty$.

Assumption E10 provides sufficient conditions for the set $\mathcal{V}$ to be a Hilbert space under $\langle \cdot, \cdot \rangle$[6]. Define v_n^* to be the Riesz representer of $\frac{\partial f(\theta^*)}{\partial \theta}[\cdot]$ on $\mathcal{V}_n$, which exists under Assumption E11.

Assumption E11. (1) $v \mapsto \frac{\partial f(\theta^*)}{\partial \theta}[v]$ is a linear functional. (2) If $\lim_{n \to \infty} \|v_n^*\|$ is finite then $\|v_n^* - v^*\| \times \|\theta_n^* - \theta^*\| = o(n^{-1/2})$ where v^* is the limit of v_n^*. Otherwise $\left|\frac{\partial f(\theta^*)}{\partial \theta}[\theta_n^* - \theta^*]\right| / \|v_n^*\| = o(n^{-1/2})$. (3) $\sup_{\theta \in \mathcal{N}_0} \frac{\left|f(\theta) - f(\theta^*) - \frac{\partial f(\theta^*)}{\partial \theta}[\theta - \theta^*]\right|}{\|v_n^*\|} = o(n^{-1/2})$.

[6]See Chen and Liao (2014, p. 642).

Assumption E11 imposes some restrictions on the functional of interest $\theta \mapsto f(\theta)$. Part (1) imposes that the directional derivative is a linear functional, a mild condition that is satisfied by our examples in Section 4. Part (2) is a restriction on the growth rate of the dimension of the sieve space. Part (3) restricts the linear approximation error of $f(\cdot)$ in a neighborhood of θ^*, for which sufficient conditions could be stated in terms of the smoothness of $f(\cdot)$ and the growth rate of the dimension of the sieve space. See Chen et al. (2014) for further discussion.

Let $u_n^* \coloneqq \frac{v_n^*}{\|v_n^*\|}$ and $\varepsilon_n = o\left(n^{-1/2}\right)$. Let $\mu_n\{g(W)\} \coloneqq n^{-1}\sum_{i=1}^n \left[g\left(W_i\right) - \mathrm{E}[g\left(W_i\right)]\right]$ denote the centered empirical process indexed by the function g.

Assumption E12. (1) $\mu_n\{\frac{\partial \ell(W,\theta^*)}{\partial\theta}[v]\}$ is linear in $v \in \mathcal{V}$.

$$\sup_{\theta\in\mathcal{N}_n} \mu_n\left\{\ell\left(W, \theta \pm \varepsilon_n u_n^*\right) - \ell(W,\theta) - \frac{\partial \ell\left(W,\theta^*\right)}{\partial\theta}\left[\pm\varepsilon_n u_n^*\right]\right\} = O_p\left(\varepsilon_n^2\right)$$

For some positive sequence $\eta_n \to 0$,

$$\sup_{\theta\in\mathcal{N}_n}\left|E\left[\ell(W,\theta) - \ell\left(W,\theta \pm \varepsilon_n u_n^*\right)\right] - \frac{\left\|\theta \pm \varepsilon_n u_n^* - \theta^*\right\|^2 - \left\|\theta - \theta^*\right\|^2}{2}\left(1 + O\left(\eta_n\right)\right)\right| = O\left(\varepsilon_n^2\right)$$

Assumption E13. $\sqrt{n}\mu_n\left\{\frac{\partial \ell(W,\theta^*)}{\partial\theta}\left[u_n^*\right]\right\} \to_d N(0,1)$

Theorem 4 is a direct application of Lemma 2.1 in Chen and Liao (2014) so its proof is omitted.

C.2.1 Sieve Riesz representer for variance decomposition

Here we discuss whether the following functional $f(\theta) = g_1(\theta_1)V(\lambda_k) + g_2(\theta_1)$, with g_i, $i = 1,2$ known real valued functions, may or may not satisfy Assumption E11. For simplicity set $\theta = (\theta_1, \theta_2)$ with $\theta_1 \in \mathbb{R}^{\dim(\theta_1)}$ and $f_{\lambda_k} = \theta_2$. Then $f(\theta) = g_1(\theta_1)\int x^2 d\theta_2(x) + g_2(\theta_1)$ (where $\int$ is understood as the Riemann–Stieltjes integral) and

$$\begin{aligned}
\frac{\partial f(\theta)}{\partial \tau}[v] &= \frac{\partial f(\theta+\tau v)}{\partial \tau}\Big|_{\tau=0} \\
&= \frac{\partial}{\partial \tau}\left(g_1(\theta_1+\tau v_1)\int x^2 d[\theta_2+\tau v_2](x) + g_2(\theta_1+\tau v_1)\right)\Bigg|_{\tau=0} \\
&= v_1 \frac{\partial}{\partial \theta_1} g_1(\theta_1)\int x^2 d\theta_2(x) + g_1(\theta_1)\int x^2 dv_2(x) + v_1\frac{\partial}{\partial \theta_1} g_2(\theta_1).
\end{aligned}$$

For E11 (1), $v \mapsto \frac{\partial f(\theta^*+\tau v)}{\partial \tau}$ is plainly a linear map. For the remaining parts of E11, it will be useful to derive the the sieve Riesz representer. We proceed following Section 3.3.1. in Chen et al. (2014). Note $\frac{\partial f(\theta^*)}{\partial \theta_2}[P_{k_n}(\cdot)] = \frac{\partial}{\partial \tau} f((\theta_1^*, \theta_2^*+\tau P_{k_n}))|_{\tau=0}$ where P_{k_n} are the k_n sieve terms for sample size n. Then

$$\frac{\partial f(\theta^*)}{\partial \theta_2}[v(\cdot)] = g_1(\theta_1^*)\int x^2 dv(x)$$

An element of our choice of sieve space takes the form $f(x) = \sum_{j=1}^{k_n}\beta_j[\max\{x - t_j, 0\}]^0 =^* \sum_{j=1}^{k_n}\beta_j 1\{x \geq t_j\}$,[7] so that an element of the vector P_{k_n} is $1\{x \geq t_j\}$. Therefore

$$\frac{\partial f(\theta^*)}{\partial \theta_2}[1\{t_j \leq \cdot\}] = g_1(\theta_1^*)\int x^2 d1\{x \geq t_j\} = g_1(\theta_1^*)t_j^2.$$

It then follows that

$$\|\frac{\partial f(\theta^*)}{\partial \theta_2}[P_{k_n}(\cdot)]\|_E^2 = g_1(\theta_1^*)^2\sum_{j=1}^{k_n} t_j^4 \longrightarrow \infty,$$

and our functional is thus irregular by remark 3.2 in Chen et al. (2014)[8]. Since the support of λ_k is bounded, t_j is bounded as k_n grows, so $\|v_n^*\|^2 \asymp \|\frac{\partial f(\theta^*)}{\partial \theta_2}[P_{k_n}(\cdot)]\|_E^2 \asymp k_n$. As a result of this discussion, under the additional regularity condition of the above footnote and with the first-order spline sieve, we can provide the following 'lower-level' replacement of E11 for the functional considered in this remark:

Assumption E11†. $\max\left\{\left|\frac{\partial f(\theta^*)}{\partial \theta}[\theta_n^* - \theta^*]\right|, \sup_{\theta\in\mathcal{N}_0}\left|f(\theta) - f(\theta^*) - \frac{\partial f(\theta^*)}{\partial \theta}[\theta - \theta^*]\right|\right\} = o((k_n n)^{-1/2})$.

[7] The asterisk is since we use the convention $0^0 = 0$.

[8] This is under the additional condition that the sieve information matrix (i.e., $E[-\frac{\partial \ell}{\partial\theta\partial\theta'}]$ is non-singular).

C.3 Implicit Differentiation

Here we show how to calculate derivative the profile likelihood function. Let $\mathcal{N}(r)$ be the set of observations with $z_i = z^{(r)}$, and let $\omega_{\cdot,r} = (\omega_{1r}, \ldots, \omega_{q_n,r})$. The log likelihood for this set of observations is,

$$\mathcal{L}_r(\omega_{\cdot,r}, \theta_1) = \sum_{i \in \mathcal{N}(r)} \ell(w_i; \theta_1, \omega_{\cdot,r}) = \sum_{i \in \mathcal{N}(r)} \log \sum_{s=1}^{q_n} \omega_{sr} f(w_i, \bar{v}_{s,r}; \theta_1)$$

Let $\omega^*_{\cdot,r}(\theta)$ be the solution that maximizes $\mathcal{L}_r(\omega_{\cdot,r}, \theta_1)$ with respect to $\omega_{\cdot,r}$ subject to the constraint that $\omega_{\cdot,r} \in \Delta(q_n)$, and let $\mathcal{L}^*_r(\theta_1) = \mathcal{L}(\omega^*_{\cdot,r}(\theta_1), \theta_1)$ be the profile likelihood function. The gradient of the profile likelihood function is,

$$\frac{d\mathcal{L}^*_r(\theta_1)}{d\theta_1} = \sum_{i \in \mathcal{N}(r)} \frac{\partial \ell(w_i; \theta_1, \omega^*_{\cdot,r}(\theta_1))}{\partial \theta_1} + \frac{\partial \ell(w_i; \theta_1, \omega^*_{\cdot,r}(\theta_1))}{\partial \omega^*_{\cdot,r}(\theta_1)} \frac{d\omega^*_{\cdot,r}(\theta_1)}{d\theta_1}$$

The derivatives of the likelihood function can be calcaluated directly. The derivative $\frac{d\omega^*_{\cdot,r}(\theta_1)}{d\theta_1}$ is defined implicitly by the KKT conditions of the inner optimization step. We next show how to calculate it.

Proposition 3.3 in Kim et al. (2020) shows that for each θ_1, maximizing $\mathcal{L}_r(\omega_{\cdot,r}, \theta_1)$ subject to $\omega_{\cdot,r} \in \Delta(q_n)$ is equivalent to maximizing $\mathcal{L}_r(\omega_{\cdot,r}, \theta_1) + \sum_{s=1}^{q_n} \omega_{s,r}$ subject to $\omega_{s,r} \geq 0$ for all s.

The equality constraints in the KKT conditions of this equivalent problem are $G_r(\omega_{\cdot,r}, \lambda_r; \theta_1) = 0$ where $\lambda_r \in \mathcal{R}^{q_n}$ are the dual parameters, and,

$$G_r(\omega_{\cdot,r}, \lambda_r, \theta_1) = \begin{bmatrix} \sum_{i \in \mathcal{N}(r)} \frac{f(w_i; \bar{v}_{\cdot,r}; \theta_1)}{\sum_{s=1}^{q_n} \omega_{s,r} f(w_i, \bar{v}_{\cdot,r}; \theta_1)} + \iota_{q_n} + \operatorname{diag}(\lambda_r) \\ \lambda_r \circ \omega_{\cdot,r} \end{bmatrix}$$

Since G_r is constant along the solution $(\omega^*_{\cdot,r}, \lambda^*)(\theta_1) := (\omega^*_{\cdot,r}(\theta_1), \lambda^*(\theta_1))$, for all θ_1, we have,

$$0 = \frac{\partial G_r(\omega^*(\theta_1), \lambda^*(\theta_1), \theta_1)}{\partial(\omega^*_{\cdot,r}(\theta_1), \lambda^*_r(\theta_1))} \frac{d(\omega^*_{\cdot,r}, \lambda^*_r)(\theta_1)}{d\theta_1} + \frac{\partial G_r(\omega^*(\theta_1), \lambda^*(\theta_1), \theta_1)}{\partial \theta_1}.$$

This implicitly defines $\frac{d\omega^*_{\cdot,r}(\theta_1)}{d\theta_1}$ assuming that the partial derivate of G_r with respect to its first two arguments is invertible.

The partial derivatives of G_r can be calculated as follows,

$$\frac{\partial G_r(\omega_{\cdot,r},\lambda_r,\theta_1)}{\partial(\omega_{\cdot,r},\lambda_r)} = \begin{bmatrix} -\sum_{i=1}^{n}\frac{f(\omega_i;\bar{v}_{\cdot,r};\theta_1)f(\omega_i;\bar{v}_{\cdot,r};\theta_1)^T}{\left(\sum_{s=1}^{q_n}\omega_{sr}f(\omega_i;\bar{v}_{\cdot,r};\theta_1)\right)^2} & I \\ \mathrm{diag}(\lambda_r) & \mathrm{diag}(\omega_{\cdot,r}) \end{bmatrix}$$

$$\frac{\partial G_r(\omega_{\cdot,r},\lambda_r,\theta_1)}{\partial\theta_1} = \begin{bmatrix} -\sum_{i=1}^{n}\frac{\nabla_{\theta_1}f(\omega_i;\bar{v}_{\cdot,r};\theta_1)\sum_{s=1}^{q_n}\omega_{sr}f(w_i,\bar{v}_{s,r};\theta_1)-f(w_i;\bar{v}_{\cdot,r};\theta_1)\sum_{s}^{q_n}\omega_{sr}\nabla_{\theta_1}f(w_i,\bar{v}_{rs};\theta_1)}{\left(\sum_{s=1}^{q_n}\omega_{sr}f(\omega_i;\bar{v}_{s,r};\theta_1)\right)^2} \\ 0 \end{bmatrix}$$

Summing over r, we obtain the derivative of the full profile likelihood function, $\frac{d\mathcal{L}^*(\theta_1)}{d\theta_1} = \sum_r \frac{d\mathcal{L}_r^*(\theta_1)}{d\theta_1}$

Finally, note that from the KKT conditions, $G_r(\omega_{\cdot,r},\lambda_r,\theta_1)=0$ imply that $\lambda_r^* = -\left(1+\sum_{i=1}^{n}\frac{f(w_i;\bar{v}_{sr};\theta_1)}{\sum_{s'=1}^{q_n}\omega_{s'r}^*f(w_i;\bar{v}_{s'r;\theta_1})}\right)$. Therefore, it is possible to calculate this gradient even if the values of $\lambda_r^*(\theta_1)$ available from the maximization procedure used.

C.4 Details on DGP

This section gives further details on the DGP used for Monte Carlo simulations discussed in Section 6. The values of the finite parameters used in the DGP are given in the table below.

Table 4: Finite parameter values

$\alpha_1(1)=0$	$\beta_{z1,1}(1)=-0.5$	$\beta_{z1,2}(1)=-0.58$	$F_{u1}(1)=1$	$F_{k1}(1)=0.3$
$\alpha_2(1)=0.1$	$\beta_{z2,1}(1)=-0.8$	$\beta_{z2,2}(1)=-0.83$	$F_{u2}(1)=1.05$	$F_{k2}(1)=0.35$
$\alpha_3(1)=0.2$	$\beta_{z3,1}(1)=0.12$	$\beta_{z3,2}(1)=-0.83$	$F_{u3}(1)=1.01$	$F_{k3}(1)=0.33$
$\sigma^2(1)=0.5$				
$\alpha_1(2)=-0.1$	$\beta_{z1,1}(2)=0.13$	$\beta_{z1,2}(2)=0.71$	$F_{u1}(2)=0.4$	$F_{k1}(2)=1$
$\alpha_2(2)=-0.22$	$\beta_{z2,1}(2)=0.89$	$\beta_{z2,2}(2)=-0.36$	$F_{u2}(2)=0.36$	$F_{k2}(2)=1.05$
$\alpha_3(2)=-0.33$	$\beta_{z3,1}(2)=0.32$	$\beta_{z3,2}(2)=-0.36$	$F_{u3}(2)=0.44$	$F_{k3}(2)=1.02$
$\sigma^2(2)=0.7$				
$\sigma_u^2=1.5$	$\rho=2.0$	$\gamma=0.5$		

The distribution of λ_k is a finite mixture of three truncated normal distributions.

The means of the component distributions are $(-1.2, 0, 1.5)$ and the variances of the component distributions are $(.2, .1, .3)$, and the weights of the mixing distribution are $(.4, .3, .3)$. Each component distribution is truncated at the third standard deviation of its distribution.

The distribution of the covariates $Z = (Z_1, Z_2)$ is as follows: Z_1 has a standard normal distribution and Z_2 as a Bernoulli distribution with equal weights. We assume that Z_1 and Z_2 are independent from each other and from λ

C.5 DGP with risk aversion

In this section, we present results from an alternative DGP where the choice problem incorporates risk aversion and biased beliefs. We adopt a specification in which agents maximize utility function in each period which incorporates risk aversion and subjective (possibly biased) beliefs. The utility that individual i derives from choice d in period t is,

$$U_{i,t}(d) := \mathbb{E}_{it}\left(\frac{Y_{it}(d)^{1-\gamma}}{1-\gamma}\right) + \eta_i(d)$$

where, E_{it} is the expectation under individual i's subjective beliefs over $\lambda_{u,i}$ given the information up to period t. $\{\eta_{i,t}(d)\}$ are mutually independent with an Extreme Value Type 1 distribution. This specification has a constant relative risk aversion (CRRA) functional form with an additive choice shock.

We assume that individuals' subjective beliefs over $\lambda_{u,i}$ are:

$$\lambda_{u,it} \sim N(E_{i,t} + F_b(d)\lambda_{k,i}, \Sigma_{i,t})$$

where $E_{i,t}, \Sigma_{i,t}$ are the correct posterior mean and variance of $\lambda_{u,i}$ given the information up to period $t-1$. This subjective belief process allows agents to have biased beliefs that can be correlated with the known part of their unobserved heterogeneity, λ_k.

Under this specification, expected utility has the following analytical solution,

$$U_{i,t}(d_t) = \frac{\exp\left(\mu_{i,t}(d_t)(1-\gamma) + \frac{1}{2}\sigma_{i,t}(d_t)(1-\gamma)^2\right)}{1-\gamma} + \eta_{i,t}(d_t) \tag{14}$$

where the i-specific mean and variance are,

$$\mu_{i,t}(d) = \alpha_t(d_t) + \beta_t(d)'Z_i + F_{kt}(d_t)\lambda_{k,i} + F_{ut}(d_t)'(E_{i,t} + F_b(d_t)\lambda_{k,i})$$
$$\sigma_{i,t}(d) = F_{ut}(d_t)'\Sigma_{i,t}F_{ut}(d) + \sigma_t^2(d_t).$$

Lemma 1, implies that this can written as $U_{i,t}(d) = u_{i,t}(d) + \eta_{i,t}(d)$, with,

$$u_{i,t}(d) = \frac{1}{1-\gamma}\exp\left(\sum_{h\in\mathcal{D}^{t-1}} 1(D^{t-1} = h)(\pi_{t,h,d,0} + \pi'_{t,h,d,1}Z + \pi_{t,h,d,2}\lambda_k + \pi'_{t,h,d,3}Y_i^{t-1})\right)$$

where the coefficients are derived from (14) and from lemma 1. The linear index in the previous equation has $|\mathcal{D}|^{t-1} * (t + \dim(Z) - 1)$ terms. Hence, with $T = 3$, $|\mathcal{D}| = 2$, and $\dim(Z_t) = 2$, this amounts to a total of 76 coefficients in π.

A naive approach to estimate $u_{i,t}(d)$ nonparametrically, would be to use a tensor product of polynomials $(\lambda_k, Z, Y^{t-1}, D^{t-1})$ as the sieve space. That is, for a univariate random variable Z, let $\mathcal{P}_q(Z) = \mathrm{sp}(\{1, Z, \ldots, Z^q\})$. Assume D_t is binary, and let $\delta_t = 1(D_t = 1)$, then the sieve space is,

$$\mathcal{P}_q(\lambda_k) \otimes \mathcal{P}_q(Z_1) \otimes \cdots \otimes \mathcal{P}_q(Y_1) \otimes \mathcal{P}_q(\delta_1) \otimes \cdots \otimes \mathcal{P}_q(Y_{t-1}) \otimes \mathcal{P}_q(\delta_{t-1}).$$

This sieve space is simply the linear span of the basis functions,

$$\left\{\lambda_k^{j_1} Z_1^{j_2} Z_2^{j_3} \prod_{1\leq s<t} Y_s^{2+2s}\delta_s^{3+2s} : 0 \leq j_s \leq q, s = 1, \ldots, 3 + 2(t-1)\right\}$$

For an q-order polynomial, the number of terms would be $(q+1)^3 + (q+1)^5 + (q+1)^7$, which grows very quickly in practical terms. (To include third-order terms, i.e., $q = 3$, this would involve $17,472$ basis functions).

An alternative approach we consider here is to instead assume that we know that the researcher is willint to assume that

$$u_{i,t(d)} = \varphi\left(\sum_{h\in\mathcal{D}^{t-1}} 1(D^{t-1} = h)(\pi_{t,h,d,0} + \pi'_{t,h,d,1}Z + \pi_{t,h,d,2}\lambda_k + \pi'_{t,h,d,3}Y_i^{t-1})\right)$$

for some unknown function φ. Since the argument of φ is scalar-valued, this means that the nonparametric estimation problem is greatly simplified to estimating a scalar-valued function. For this we use the sieve space of polynomials, with the order growing at the rate of $n^{1/3}$ with 3 terms with $n = 500$ and 6 terms for $n = 4,000$.

The finite parameters are the same as in Table **??**, with the added risk aversion parameter γ, which we set to 1.5. λ and Z are generated the same as in that DGP.

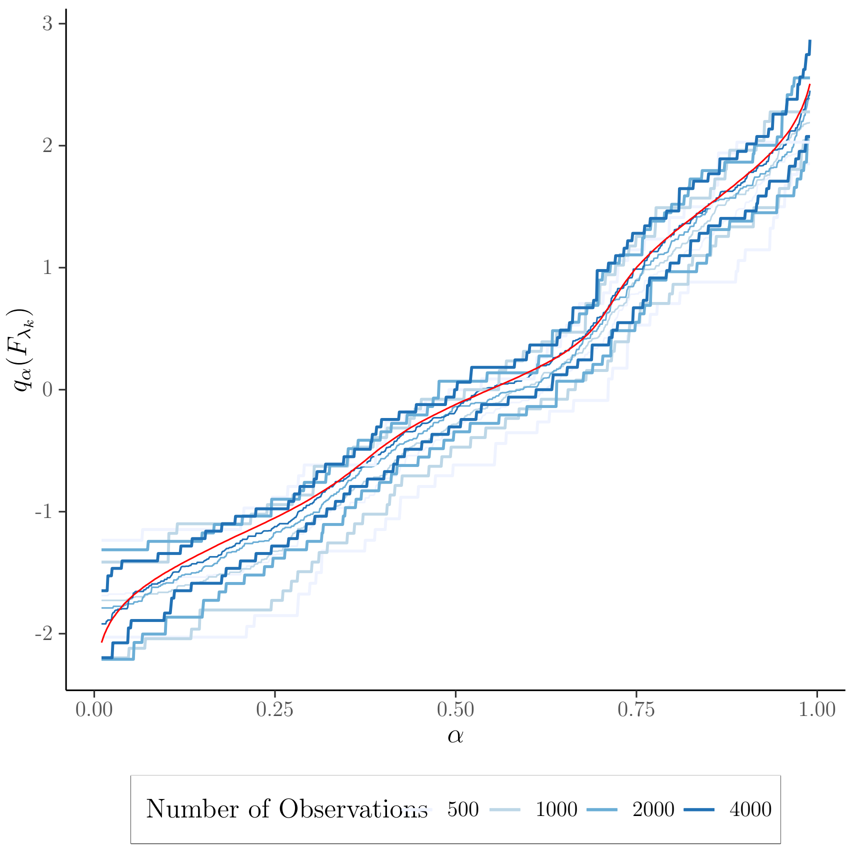
With the additional π parameters to estimate, the θ_1 has a total of 103 parameters. Given this large number of parameters to estimate, we expect $n = 250$ to be too small a sample size to perform well, and begin the Monte Carlo simulations with a sample size of $n = 500$.

Table 5: Bias and Variance ($\times 1,000$) of Finite Parameter Estimators: DGP with Risk Aversion

	N = 500		N = 1000		N = 2000		N = 4000	
	sq bias	var	sq bias	var	sq bias	var	sq bias	var
$\alpha_1(2)$	58.444	45.007	13.495	17.952	3.892	18.241	0.011	5.015
$\alpha_2(1)$	0.155	18.224	0.031	9.468	0.049	4.342	0.007	1.215
$\alpha_2(2)$	58.712	40.678	12.966	19.647	1.656	21.692	0.373	7.563
$\alpha_3(1)$	3.700	28.283	0.102	6.040	0.019	3.987	0.024	1.210
$\alpha_3(2)$	66.751	35.834	22.634	21.400	3.018	16.469	0.027	6.707
$\beta_{1,1}(1)$	0.268	5.490	0.002	3.218	0.031	1.606	0.017	0.724
$\beta_{1,1}(2)$	0.379	5.207	0.342	3.462	0.131	1.475	0.026	0.818
$\beta_{2,1}(1)$	0.031	6.463	0.003	4.190	0.004	1.905	0.000	1.097
$\beta_{2,1}(2)$	0.279	7.294	0.485	3.842	0.221	1.891	0.034	0.730
$\beta_{3,1}(1)$	0.449	5.089	0.011	3.202	0.006	1.305	0.011	0.728
$\beta_{3,1}(2)$	0.192	6.269	0.371	3.830	0.049	1.618	0.008	0.855
$\beta_{1,2}(1)$	0.047	31.005	0.206	10.995	0.112	4.917	0.093	2.757
$\beta_{1,2}(2)$	1.122	25.134	0.074	10.377	0.305	6.969	0.091	3.012
$\beta_{2,2}(1)$	0.020	26.534	0.654	10.006	0.038	4.102	0.028	2.648
$\beta_{2,2}(2)$	0.434	27.052	0.002	14.966	0.628	8.992	0.387	2.961
$\beta_{3,2}(1)$	0.011	23.634	0.118	9.195	0.194	4.541	0.000	2.585
$\beta_{3,2}(2)$	0.296	24.200	0.012	14.879	0.634	10.490	0.050	3.340
$F_{k1}(1)$	0.000	20.762	1.497	7.116	0.103	2.797	0.002	1.175
$F_{k2}(1)$	0.063	9.571	2.413	9.610	0.822	4.292	0.144	1.725
$F_{k2}(2)$	0.268	4.907	0.008	2.631	0.248	2.035	0.051	0.842
$F_{k3}(1)$	1.224	18.830	1.460	9.878	0.266	3.991	0.054	1.446
$F_{k3}(2)$	0.021	12.969	0.683	3.869	0.022	2.054	0.001	0.998
$F_{u1}(2)$	0.582	5.416	0.223	2.282	0.108	1.271	0.034	0.637
$F_{u2}(1)$	0.320	4.044	0.024	1.464	0.064	0.511	0.057	0.336
$F_{u2}(2)$	0.664	6.729	0.300	3.981	0.022	1.169	0.000	0.699
$F_{u3}(1)$	0.001	4.267	0.017	1.480	0.066	0.632	0.048	0.294
$F_{u3}(2)$	2.357	6.900	1.241	3.281	0.059	1.369	0.002	1.070
$\sigma^2(1)$	0.036	0.912	0.006	0.276	0.006	0.206	0.001	0.081
$\sigma^2(2)$	0.097	0.260	0.002	0.159	0.001	0.057	0.000	0.024
σ_u^2	3.062	12.636	0.252	4.785	0.119	3.063	0.324	1.800

All calculations are based on 200 Monte Carlo simulations of the DGP described in the main text. Squared bias and variance of finite parameter estimates are multiplied times $1,000$

Figure 2: Quantiles of Estimator of λ_k: 95% Coverage Intervals: DGP with Risk Aversion



Note: The red line shows the true distribution of λ_k. The blue lines show the mean, and the 5th and 95th percentiles of the simulated distribution of the estimate of $q_\alpha(f_{\lambda_k})$.