Heterogeneity, Uncertainty and Learning: Semiparametric Identification and Estimation*

Jackson Bunting †　　Paul Diegert‡　　Arnaud Maurel§

February 14, 2024

Abstract

We provide semiparametric identification results for a broad class of learning models in which continuous outcomes depend on three types of unobservables: i) known heterogeneity, ii) initially unknown 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 identification of the outcome equation parameters and the distribution of the three types of unobservables, under the standard assumption that unknown heterogeneity and uncertainty are normally distributed. We also show that, absent known 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.

*First version: November 2022. We thank seminar participants at CREST-PSE, LMU Munich, TSE, UC Davis, UT Austin, conference participants at the 34th EC² conference, the 2024 NAWMES, 2023 IAAE, SETA and SOLE meetings, Elena Pastorino, Yuya Sasaki as well as Karun Adusumilli, Victor Aguirregabiria, Peter Arcidiacono, Stéphane Bonhomme, Xavier D'Haultfoeuille, Yuichi Kitamura, Mauricio Olivares, Chris Taber and Daniel Wilhelm for useful comments. We also thank Zhangchi Ma and Chinncy Qin for capable resesarch assistance.

†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), 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 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. 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 unobserved individual heterogeneity. Specifically, we consider throughout our analysis a potential outcome model where individual i's potential outcome in period t from assignment d is given by

$$Y_{i,t}(d) = X_{i,t}^{\intercal}\beta_{t,d} + (X_i^*)^{\intercal}\lambda_{t,d} + \epsilon_{i,t}(d), \tag{1}$$

where $X_{i,t}$ is a vector of explanatory variables associated with individual i in period t (including an intercept), X_i^* denotes a vector of latent individual effects (or factors), $\epsilon_{i,t}(d)$ is a transitory shock, and $(\beta_{t,d}^{\intercal}, \lambda_{t,d}^{\intercal})^{\intercal}$ is an unknown parameter vector. 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 X_i^* consists of two components: $X_{k,i}^*$, which are supposed to be known by the agent, and $X_{u,i}^*$ which are initially unknown but may be learned over time. We complement this potential 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 establish that the model is 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 transitory shocks from the outcome equations ($\epsilon_{i,t}(d)$), as well as the unknown heterogeneity component ($X_{u,i}^*$), are normally distributed. In contrast, the distribution of the known heterogeneity component ($X_{k,i}^*$) is left unspecified. From the observation that the distribution of 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.

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

We propose to estimate the model parameters using a sieve maximum likelihood estimator which we show to be consistent. We then focus on a general class of functionals of the model parameters, 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; Gong et al., 2019). We show that, under mild regularity conditions, the resulting estimators are consistent and asymptotically normal. We implement our sieve maximum likelihood estimator using a profile likelihood based procedure. Importantly for practical purposes, the resulting procedure only involves a modest computational cost. Monte Carlo simulation results further indicate that our estimator exhibits good finite-sample properties.

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 (see, e.g., Abbring and Campbell, 2005; Arcidiacono et al., 2023; Gong, 2019; Pastorino, 2022). A key distinction with most of the papers in this literature is that we only impose mild restrictions on the choice process. Importantly, we remain agnostic about how choices depend on individual beliefs about $X^*_{u,i}$, while allowing these beliefs to depend arbitrarily on past choices and realized outcomes. Particularly relevant for us is recent complementary work by Pastorino (2022), which establishes formal identification results in a different context of a two-sided learning model where workers and firms have imperfect information. Key to the strategy proposed in that paper is to leverage for identification particular mixture representations of selected one-dimensional outcomes. Related mixture representations also play an important role in our analysis.

Our paper also fits into a literature that focuses on the identification of Markovian dynamic discrete choice models in the presence of persistent unobserved heterogeneity (see, among others, Heckman and Navarro, 2007; Hu and Schennach, 2008; Kasahara and Shimotsu, 2009; Hu and Shum, 2012; Sasaki, 2015; Hu and Sasaki, 2018; Aguirregabiria et al., 2021; Bunting, 2022). Unlike these papers, we do not impose a Markov structure, since current beliefs and decisions are allowed to depend on the entire history of past outcomes and decisions.[1] More broadly, our analysis is related to the literature that deals with the identification of mixture models (see, for example, 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 unobserved individual- and time-specific effects, our paper also fits into the literature that examines 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). An important 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.

[1] Although our framework is more general, Bayesian learning models often naturally possess a first order Markov structure. There are, however, several additional significant differences between our paper and the listed literature. Notably, Hu and Shum (2012) focus on scalar unobserved heterogeneity, whereas the existence of multivariate unobserved heterogeneity is fundamental to our main setting. Beyond this, several of their assumptions may fail to hold in our setup. For instance, since the support of the latent beliefs is larger than the support of the choices, the requirement that the observed variables be invertible measurements of the latent variables (Hu and Shum, 2012, Assumption 2) will generally fail to hold.

Organization of the paper

The remainder of the paper is organized as follows. Section 2 introduces and discusses 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 and its finite-sample performances. Finally, Section 6 concludes. The appendix gathers all the proofs, additional material on the variance decompositions, the implementation of our estimator, and further Monte Carlo simulation results.

Notation: for a given random variable A, we denote by a its realization, $\mathcal{S}(A)$ indicates its support, F_A denotes its cumulative distribution function, $q_\alpha[A]$ its $\alpha \in [0,1]$ quantile, whereas f_A indicates its probability mass or density function. For any sequence $(a_1, a_2, \dots, a_S)$ and $s \leq S$, we let $a^s = (a_1, a_2, \dots, a_s)$. $A \perp\!\!\!\perp B \mid C$ indicates that A and B are statistically independent conditional on C. Finally, unless stated otherwise, we suppress the individual subscript i from all random variables in the remainder of the paper.

2 Set-up

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

$$Y_t(d) = X_t^\intercal \beta_{t,d} + X_k^* \lambda_{t,d}^k + (X_u^*)^\intercal \lambda_{t,d}^u + \epsilon_t(d), \tag{2}$$

where d represents a possible value of individual i's assignment in period t, $Y_t(d)$ is a scalar potential outcome variable associated with assignment d, X_t is a vector of observed explanatory variables, $X^* := (X_k^*, (X_u^*)^\intercal)^\intercal$ are unobserved (to the econometrician) factors, $(\beta_{t,d}^\intercal, \lambda_{t,d}^\intercal)^\intercal$ with $\lambda_{t,d} := (\lambda_{t,d}^k, (\lambda_{t,d}^u)^\intercal)^\intercal$ is an unknown parameter vector, and $\epsilon_t(d)$ is an idiosyncratic random shock. 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 (X_t) as well as on multiple dimensions of unobserved abilities (X^*), which may play different roles in different occupations (see, e.g., Hincapié, 2020; Arcidiacono et al., 2023). This setup is fairly general and can be applied in a wide range of contexts. For instance, $Y_t(d)$ may alternatively represent the potential log-quantity of a particular product sold by a firm in a given market d (see, e.g., Berman et al., 2019). This framework can also be used in the health context, where $Y_t(d)$ may correspond to a health outcome measure associated with a certain drug (e.g., CD4 cell counts associated with a particular HIV drug treatment, as in Chan and Hamilton, 2006), or to the body mass index associated with a certain type of diet.

Importantly, we allow for two distinct types of latent individual effects. Namely, X_k^* is assumed to be known by the agent, while X_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, $X_k^*\lambda_{t,d}^k + (X_u^*)^\intercal\lambda_{t,d}^u$. As the worker accumulates more experience, she may update her belief about X_u^*, and thus about the initially unknown portion of productivity in each of the possible occupations.

Turning to the choice and learning process, the key restriction that we place on an individual's assignment in period t (denoted as D_t) is that it does not directly depend on the unknown component of heterogeneity. Specifically, we assume that:

$$D_t \perp\!\!\!\perp X_u^* \mid X^t, Y^{t-1}, D^{t-1}, X_k^*. \tag{3}$$

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

This choice process accommodates a wide range of models that have been considered in the learning literature. In particular, this framework is consistent with a setup where agents are rational and Bayesian updaters, so that beliefs coincide with the true distribution of X_u^* conditional on their information set at a given point in time, which may include all realized variables and model parameters. Alternatively, this accommodates situations where individual decisions may not involve beliefs over the distribution of X_u^*, or depend instead on myopic beliefs that are formed based on the prior-period choice and outcome. This setup also allows for heterogeneous beliefs formation, where, for instance, some agents may have rational expectations about their unobserved characteristic X_u^*, while others may have biased (e.g. over-optimistic) beliefs.

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

$$h_t(d^t, x^t, y^{t-1}, x_k^*) \coloneqq \Pr(D_t = d \mid X^t = x^t, Y^{t-1} = y^{t-1}, D^{t-1} = d^{t-1}, X_k^* = x_k^*).$$

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 X_k^* - are identified. In empirical applications it is very common to impose some structure on the choice process. For example, in a dynamic discrete choice framework it is standard to assume that

$$D_t = \underset{d \in \mathcal{S}(D_t)}{\arg\max} \left\{ \overline{v}(d, X_t, X_k^*, S_t) + \eta_t(d) \right\},$$

where the conditional value function $\overline{v}$ is known up to a finite-dimensional vector of parameters, S_t are sufficient statistics for the conditional distribution of X_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 to identify $\overline{v}$ (see, e.g., Hotz and Miller, 1993; Aguirregabiria and Mira, 2010; Chiong et al., 2016), and then recover the primitives of the choice model (see, e.g., Arcidiacono et al., 2023).

Uncertainty and learning. A central feature of the model is the distinction between three forms of unobserved heterogeneity: (1) permanent heterogeneity that is known to the agent, X_k^*, (2) permanent heterogeneity that is initially unknown to the agent, X_u^*, and (3) transitory time-varying shocks, $\epsilon = \{\epsilon_t(d) : d \in \mathcal{S}(D_t),\ t = 1, 2, \ldots\}$. This provides a framework for quantifying the importance of uncertainty in outcomes. At $t = 1$, the variance in future outcomes can be decomposed into a component that depends on (X_u^*, ϵ) and a component that depends on X_k^*. Cunha et al. (2005) and Cunha and Heckman (2016) consider this decomposition in the context of educational choice, decomposing the variance in lifetime earnings into a component that is predictable when deciding to go to college and a component that is not.

In our framework, the importance of uncertainty can change over time as agents learn about X_u^* by observing realized outcomes and covariates, and use this information to self-select into different alternatives. We provide in Appendix B.2 a class of variance decomposition parameters which includes both the $t = 1$ decomposition as well as $t > 1$ decompositions that incorporate these learning and selection effects. These decompositions, which are identified from the model parameters, each provide different ways of quantifying the importance of uncertainty to future outcomes. After establishing identification of the model, we will pay special attention to estimation and inference of a broad class of functionals that encompasses these kinds of variance decompositions.

3 Identification

We first provide in Subsection 3.1 a high-level overview of the underlying reweighting scheme that plays an important role in both 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 Reweighting strategy

Key to the identification problem analyzed in this paper is how to recover the conditional distributions of potential outcomes (i.e., $f_{Y_t(d_t)|X_t,X^*}$ for each t and d_t) and selection probabilities (i.e., $f_{D_t|X_t,X^*_k}$ for each t), from the selected population distribution (i.e., f_{Y^T,D^T,X^T}) which is directly identified from the data.

We now provide intuition as to how one can leverage the structure imposed on the choice process to address the censored data problem. To illustrate, consider a simplified version of our model with a binary choice in each period (i.e., $\mathcal{S}(D_t) = \{0,1\}$) and without covariates. Let $D := \prod_{t=1}^{T} D_t$, $Y := (Y_1, \ldots, Y_T)$ and $Y(1) := (Y_1(1), \ldots, Y_T(1))$, and focus on identification of the distribution of the potential outcome $Y(1)$. By Bayes' rule, the relationship between the target and censored distributions can be characterized as follows:

$$f_{Y|D}(y|1)\frac{f_D(1)}{f_{D|Y(1)}(1|y)} = f_{Y(1)}(y)$$

where the conditional density $f_{Y|D}(y|1)$, which is directly identified from the data, is weighted by a selection adjustment term, $\frac{f_D(1)}{f_{D|Y(1)}(1|y)}$.

Our learning framework provides one strategy for identifying these selection weights. Let us first assume that all components of the latent effect are initially unknown. In a learning context where the decision makers' actions depend on beliefs over X^*, it is often natural to assume that beliefs depend only on past realized outcomes and choices, and that:

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

where the right hand side of Equation (4) is identified from the joint distribution of (D^t, Y^{t-1}) conditional on $D^{t-1} = 1$. Applying this reasoning recursively, it follows that $f_{D|Y(1)}(1|y)$ (and thus the selection weight) is identified as follows:

$$f_{D|Y(1)}(1 \mid y) = f_{D_T|Y^{T-1}(1),D^{T-1}}(1 \mid y^{T-1},1) f_{D_{T-1}|Y^{T-2}(1),D^{T-2}}(1 \mid y^{T-2},1) \cdots f_{D_1}(1).$$

We build on this idea when establishing in Section 3.3 identification of a version of the model we call *pure learning* (where $X^* = X_u^*$). The conditional independence restriction in Equation (4) will generally break down, however, when agents also possess persistent private information that affects their decision (i.e., X_k^*). We propose in Section 3.2 an identification strategy that can be used in such situations. A key and non-trivial additional step in this context is to show, relying on existing results from Bruni and Koch (1985), that maintaining a normality assumption commonly made in the learning literature is sufficient to identify the joint distribution of (Y^T, D^T, X_k^*) 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 baseline model discussed in Section 2. We first impose a form of conditional independence on $(\epsilon_t(d), D_t, X_t)$.

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

$$F_{\epsilon_t(d),D_t,X_t|Y^{t-1},D^{t-1},X^{t-1},X^*} = F_{\epsilon_t(d)} F_{D_t|X^t,Y^{t-1},D^{t-1},X_k^*} F_{X_t|Y^{t-1},D^{t-1},X^{t-1}}.$$

Furthermore, for any $d \in \mathcal{S}(D_1)$, $F_{\epsilon_1(d),D_1,X_1|X^*} = F_{\epsilon_1(d)} F_{D_1|X_1,X_k^*} F_{X_1|X^*}$.

Assumption KL1 imposes the potential outcome model in Equation (2) and contains three independence conditions. First, it implies that the additive transitory shock in the outcome equation ($\epsilon_t(d)$) is independent of all contemporaneous and lagged variables. This is closely related to the standard fixed effect assumption that dependence in outcomes across periods is due to the latent fixed effect (e.g., Freyberger (2018, Assumption N5) and Sasaki (2015, Restriction 2)). However, note that we allow for arbitrary within-period dependence between the additive shocks ($\epsilon_t(d)$ and $\epsilon_t(\tilde{d})$, for $d \neq \tilde{d}$). Second, the unknown factor (X_u^*) does not directly affect treatment assignments (D_t), a natural restriction discussed in Section 2. Third, we also impose that the transition of the control variables (X_t) does not directly depend on the time-invariant unobservables (X^*). Importantly, this does allow X_t to depend on X^* through past choices and outcomes. For instance, in the context of occupational choices, this restriction accommodates occupation-specific work experiences whose accumulation depends on X^* through past occupational choices.

Our second assumption KL2 imposes that the unknown component of the individual effect is drawn from a multivariate normal distribution, and that the random shock in the outcome equation is normally distributed too. This is a very frequent assumption in the learning literature, to which we return in Remark 2.

Assumption KL2. For all $(x_1, x_k^*) \in \mathcal{S}(X_1) \times \mathcal{S}(X_k^*)$, $X_u^* \mid (X_1, X_k^*) = (x_1, x_k^*) \sim N(0, \Sigma_u(x_1))$ and $\forall\ d \in \mathcal{S}(D_t)$, $\epsilon_t(d) \sim N(0, \sigma_{t,d}^2)$.

Assumption KL2 implies a Gaussian conjugate posterior distribution for X_u^*, which we summarize in Lemma 1. Importantly, neither this assumption nor Assumption KL1 place any restriction on the dependence between X_k^* and X_1.[2] To do so, define

[2]Lemma 1 and our main identification result would go through if one replaces the first part of Assumption KL2 with $X_u^* \mid (X_1 = x_1, X_k^* = x_k^*) \sim N(0, \Sigma_u(x_1, x_k))$ under appropriate regularity conditions on $x_k \mapsto \Sigma_u(x_1, x_k)$, including for each $x_k^* - \tilde{x}_k^* > 0$, $\Sigma_u(x_1, x_k^*) - \Sigma_u(x_1, \tilde{x}_k^*)$ is positive (or negative) semi-definite. For simplicity, we maintain the stronger Assumption KL2 when establishing identification in Theorem 1 below.

(μ_t, Σ_t) recursively as follows. First, $(\mu_1, \Sigma_1) = (0, \Sigma_u(x_1))$. Second,

$$\Sigma_{t+1} = \left(\Sigma_t^{-1} + \lambda^u_{t,d_t}(\lambda^u_{t,d_t})^\intercal \sigma^{-2}_{t,d_t}\right)^{-1},$$
$$\mu_{t+1} = \Sigma_{t+1}\left(\Sigma_t^{-1}\mu_t + \lambda^u_{t,d_t}\frac{y_t - x_t^\intercal \beta_{t,d_t} - x_k^* \lambda^k_{t,d_t}}{\sigma^2_{t,d_t}}\right).$$

Lemma 1. Let Assumptions KL1 and KL2 hold. Then, for all $t \geq 2$, X_u^* conditional on $(Y^{t-1}, D^{t-1}, X^t, X_k^*) = (y^{t-1}, d^{t-1}, x^t, x_k^*)$ is distributed $N(\mu_t, \Sigma_t)$.

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

Assumption KL3. (A) For some $d \in \mathcal{S}(D_1)$, the element of $\beta_{1,d}$ associated with the constant term is zero, and $\lambda^k_{1,d} = 1$. (B) For some $d^p \in \mathcal{S}(D^p)$, $\left(\lambda^u_{1,d_1} \cdots \lambda^u_{p,d_p}\right) = I_{p\times p}$.

Assumption KL3 is a location-scale normalization on the finite dimensional parameters, which reflects the fact that the latent factors are only identified up to location and scale. This type of assumption is standard in interactive fixed effect models (Freyberger, 2018).

Finally, we impose in Assumptions KL4 and KL5 below several regularity conditions. We start with Assumption KL4, which places support restrictions on various objects of the model. In what follows, we let $\theta_1 := \left\{\{\beta_t, \lambda_t, \sigma_t^2\}_{t=1}^T, \Sigma_u(x_1)\right\} \in \Theta_1 \subset \mathbb{R}^{\dim \Theta_1}$, where $\{\beta_t, \lambda_t, \sigma_t^2\} := \{\beta_{t,d}, \lambda_{t,d}, \sigma^2_{t,d} \colon d \in \mathcal{S}(D_t)\}$.

Assumption KL4. (A) For each $x_1 \in \mathcal{S}(X_1)$, Θ_1 is a compact set. (B) $\mathcal{S}(X_k^*)$ is compact. (C) For each t and $d \in \mathcal{S}(D_t)$, $(\lambda^u_{t,d})^\intercal \Sigma_t \lambda^u_{t,d} + \sigma^2_{t,d} \neq 0$, $\sigma^2_{t,d} \neq 0$ and $\forall\ x_1 \in \mathcal{S}(X_1)$, $\Sigma_u(x_1)$ is non-singular. (D) For each y^{t-1}, d^t, x^t in their support, $\mathcal{S}(X_k^* \mid (Y^{t-1}, D^t, X^t) = (y^{t-1}, d^t, x^t)) = \mathcal{S}(X_k^*)$ and $Var(X_k^*) \neq 0$. (E) For each t and $d \in \mathcal{S}(D_t)$, $E[X_t X_t^\intercal \mid D_t = d]$ is non-singular. (F) For all t, $Var(D_t) \neq 0$.

Part (A) states the finite dimensional parameters θ_1 belong to a compact set. Part (B) imposes that the known latent factor X_k^* has compact support. This holds if the distribution of X_k^* has discrete support although this clearly applies to a broader set of distributions. We return to this compactness condition in Remark 1 below. Part (C) requires certain normally distributed random variables to have non-singleton support. Part (D) imposes a rectangular support condition and a non-degeneracy assumption on the distribution of X_k^*. These conditions are typically satisfied in dynamic discrete choice models with unobserved heterogeneity, which generally impose a large support assumption on the random utility shocks. Part (E) imposes that the support of X_t conditional on D_t is sufficiently rich. Finally, Part (F) imposes that the support of the choice variables contain at least two elements.

Next, Assumption KL5 below contains a set of regularity conditions that ensure that the latent individual effect X^* alters outcomes sufficiently differently across time and assignments.

Assumption KL5. (A) For each t and $d_t \in \mathcal{S}(D_t)$ there exists two sequences $(d^{t-1}, \tilde{d}^{t-1}) \in \mathcal{S}(D^{t-1})^2$ such that $(\lambda^u_{t,d_t})^\intercal \Sigma_t \sum_{s=1}^{t-1} \left(\lambda^u_{s,d_s} \frac{\lambda^k_{s,d_s}}{\sigma^2_{s,d_s}} - \lambda^u_{s,\tilde{d}_s} \frac{\lambda^k_{s,\tilde{d}_s}}{\sigma^2_{s,\tilde{d}_s}} \right) \neq 0$. (B) For all t and $d_t \in \mathcal{S}(D_t)$, $\lambda^k_{t,d_t} \neq 0$. (C) For all t and $d^t \in \mathcal{S}(D^t)$, $\lambda^k_{t,d_t} - (\lambda^u_{t,d_t})^\intercal \Sigma_t \sum_{s=1}^{t-1} \lambda^u_{s,d_s} \frac{\lambda^k_{s,d_s}}{\sigma^2_{s,d_s}} \neq 0$. (D) For all $d^2 \in \mathcal{S}(D^2)$, $(\lambda^u_{2,d_2})^\intercal \Sigma_2 \lambda^u_{1,d_1} \frac{\lambda^k_{1,d_1}}{\sigma^2_{1,d_1}} \neq 0$. (E) There exists $\{(d_{2,i}, \tilde{d}_{2,i}) \in \mathcal{S}(D_2)^2 : i = 1, 2, \ldots, p\}$ which satisfy

$$\left(\lambda^u_{2,d_{2,1}} \cdots \lambda^u_{2,d_{2,p}} \right)^{-\intercal} \mathrm{vec}(\lambda^k_{2,d_{2,1}}, \ldots, \lambda^k_{2,d_{2,p}}) \neq \left(\lambda^u_{2,\tilde{d}_{2,1}} \cdots \lambda^u_{2,\tilde{d}_{2,p}} \right)^{-\intercal} \mathrm{vec}(\lambda^k_{2,\tilde{d}_{2,1}}, \ldots, \lambda^k_{2,\tilde{d}_{2,p}}).$$

(F) For all $d^T \in \mathcal{S}(D^T)$, $\{\lambda^u_{t,d_t} : t = 1, \ldots, T\}$ is linearly independent.

This assumption is fairly mild as it primarily rules out knife-edge cases where the effect of different elements of permanent unobserved heterogeneity is exactly zero.[3] Part (A) requires that the aggregate effect of X_k^* on outcomes associated with 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 X_k^* is non-zero in each period and each assignment. Part (C) states the aggregate effect of X_k^* on outcomes must be non-zero—that is, that the direct effect

[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).

λ^k_{t,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 that for $t = 2$ the relative effect of known and unknown X^* changes across choices. In the special case where $X^*_u \in \mathbb{R}$ (i.e., $p = 1$), the condition reduces to $\frac{\lambda^k_{2,d_2}}{\lambda^u_{2,d_2}} \neq \frac{\lambda^k_{2,\tilde{d}_2}}{\lambda^u_{2,\tilde{d}_2}}$, i.e., that the ratio of factor loadings varies across some assignments. More generally, for $X^*_u \in \mathbb{R}^p$, this condition implies that, for $t = 2$, the set of assignments must contain at least $p + 1$ elements. Finally, Part (F) requires that the initially unknown factor affects each outcome via a different linear combination.

We are now in a position to state our main identification result. We denote by $\theta = \left\{\{\beta_t, \lambda_t, \sigma_t, g_t, h_t\}_{t=1}^T, \Sigma_u, F_{X^*_k,X_1}\right\} \in \Theta$ the model parameters, where $g_t := dF_{X_t|Y^{t-1},D^{t-1},X^{t-1}}$.

Theorem 1. Suppose the distribution of $(Y_t, D_t, X_t)_{t=1}^T$ 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, from Assumptions KL1 and KL2 and Lemma 1 that Y_t is normally distributed conditional on lagged outcomes Y^{t-1}, assignments D^t, covariates X^t and the known component of the latent individual effect, X^*_k. This implies that Y_t conditional on (Y^{t-1}, D^t, X^t) is a Gaussian mixture distribution parameterized by X^*_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 X^*_k. Next, the normalization and regularity assumptions (Assumptions KL3-KL5) are used to pin down the affine transformation, leading to identification of the distribution of (Y^T, D^T, X^T, X^*_k). Knowledge of this distribution identifies the components of the model related to the known component of the latent individual effect, namely $\left\{\{\beta_t, \lambda^k_t, h_t\}_{t=1}^T, F_{X^*_k,X_1}\right\}$. The final step is to disentangle the effect of the learned component (X^*_u) and uncertainty ($\epsilon_t(d)$) in order to identify $\left\{\{\lambda^u_t, \sigma^2_t\}_{t=1}^T, \Sigma_u\right\}$. This is done by showing that the joint distribution of (Y^T, D^T, X^T) conditional on X^*_k, suitability weighted by the assignment probabilities, is a normal-weighted mixture of normal distributions. This allows us to identify $\left\{\{\lambda_t^u, \sigma_t^2\}_{t=1}^T, \Sigma_u\right\}$ from the second moments of the reweighted distribution. We refer the interested reader to Section A.2 for a formal derivation.

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 X_k^* has finite support with known cardinality. Assumption KL4 (B) relaxes this restriction in the sense that the number of support points of X_k^* need not be known a priori, and indeed may be infinite.[4]

Remark 2 (**Normality of unknown factor**)**.** As summarized in Lemma 1, an important implication of the normality assumptions (Assumption KL2) is the resulting normal 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 though, the key implication of normality is rather to enable identification of the distribution of $Y_t \mid (Y^{t-1}, D^t, X^t, X_k^*,)$ from variation in the realized outcome Y_t only. Namely, under Assumption KL2, the distribution of $Y_t \mid (Y^{t-1}, D^t, X^t)$ is a mixture of normal distributions with mixture weights given by the distribution of $X_k^* \mid (Y^{t-1}, D^t, X^t)$. This allows us to establish identification by leveraging existing results for mixtures of normal distributions (Bruni and Koch, 1985).[5]

Remark 3 (**Role of covariates**)**.** Inspection of the proof shows that the covariates X_t do not actually play any role in the identification of the parameters θ, beyond $\{\beta_t : t = 1, \ldots, T\}$. In particular, one can easily adapt the proof to establish identification for a more flexible specification where X_t enters the outcome equation through an additive nonparametric shifter. We maintain linearity throughout for estimation

[4] Compactness is used in particular to apply the Stone-Weierstrass approximation theorem, which plays an important role in the identification proof of Bruni and Koch (1985, Theorem 1).

[5] That identification of the distribution of X_k^* arises from variation in the scalar outcome variable Y_t highlights why we restrict X_k^* to be a scalar random variable. If Y_t was vector-valued instead, then we expect that our arguments would easily extend to allow for a multivariate X_k^*.

precision and to preserve tractability.

Remark 4 (**Invariance to normalization**). The normalization assumption (Assumption KL3) is a true normalization in the sense that particular meaningful economic parameters are invariant to the assumption. Specifically, we can show that this is the case of the average and quantile structural functions. To formalize this notion, define $C^k_{t,d} \coloneqq X^*_k \lambda^k_{t,d}$, $C^u_{t,d} \coloneqq (X^*_u)^\intercal \lambda^u_{t,d}$ and let $Q_\alpha[X]$ be the α-quantile of a random variable X. Let $x \in \mathcal{S}(X_t)$ and define the quantile structural functions associated with the potential outcomes $Y_t(d_t)$ as follows:

$$\begin{aligned} s_{1,t}(x,\alpha) =& x^\intercal \beta_{t,d_t} + Q_\alpha[C^k_{t,d_t} + C^u_{t,d_t} + \epsilon_t(d_t)], \\ s_{2,t}(x,\alpha_1,\alpha_2,\alpha_3) =& x^\intercal \beta_{t,d_t} + Q_{\alpha_1}[C^k_{t,d_t}] + Q_{\alpha_2}[C^u_{t,d_t}] + Q_{\alpha_3}[\epsilon_t(d_t)], \end{aligned}$$

and the average structural function as $s_{3,t}(x) = x^\intercal \beta_{t,d_t} + \int u dF_{C^k_{t,d_t}+C^u_{t,d_t}+\epsilon_t(d_t)}(u)$. In Appendix B.1 we prove the following corollary:

Corollary 1. Suppose the Assumptions KL1, KL4 and KL5 hold and that for each $(x_1, x^*_k) \in \mathcal{S}(X_1) \times \mathcal{S}(X^*_k)$, $X^*_u \mid (X_1, X^*_k) = (x_1, x^*_k) \sim N(\mu_u, \Sigma_u(x_1))$ and for all t and $d \in \mathcal{S}(D_t)$, $\epsilon_t(d) \sim N(c_{t,d}, \sigma^2_{t,d})$. Furthermore, suppose for some $d^p \in \mathcal{S}(D^p)$, $(\lambda^u_{1,d_1} \cdots \lambda^u_{p,d_p})$ is full rank. Then $s_{1,t}(x,\cdot)$, $s_{2,t}(x,\cdot,\cdot,\cdot)$ and $s_{3,t}(x)$ are identified for all x on the support of X_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 maker ($X^* = X^*_u$). 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. We establish identification in this model under Assumptions L1-L5 stated below.

Assumption L1. For all t and $d \in \mathcal{S}(D_t)$, $Y_t(d) = X_t^\intercal \beta_{t,d} + (X^*)^\intercal \lambda_{t,d} + \epsilon_t(d)$. For any $t \geq 2$ and $d \in \mathcal{S}(D_t)$,

$$F_{\epsilon_t(d), D_t, X_t | Y^{t-1}, D^{t-1}, X^{t-1}, X^*} = F_{\epsilon_t(d)} F_{D_t | Y^{t-1}, D^{t-1}, X^t} F_{X_t | Y^{t-1}, D^{t-1}, X^{t-1}}.$$

Furthermore, for any $d \in \mathcal{S}(D_1)$, $F_{\epsilon_1(d), D_1, X_1 | X^*} = F_{\epsilon_1(d)} F_{D_1 | X_1} F_{X_1 | X^*}$.

Assumption L1 adapts Assumption KL1 to reflect that there is no initially known component of unobserved heterogeneity.

Assumption L2. (A) The joint density of (Y, X^*) and (D, X) admits a bounded density with respect to the product measure of the Lebesgue measure on $\mathcal{S}(Y) \times \mathcal{S}(X^*)$ and some dominating measure on $\mathcal{S}(D) \times \mathcal{S}(X)$. All marginal and conditional densities are bounded. (B) For each $x_1 \in \mathcal{S}(X_1)$, $X^* \mid X_1 = x_1$ has full support. (C) For each t and $d \in \mathcal{S}(D_t)$, the characteristic function of $\epsilon_t(d)$ is non-vanishing, and $E[\epsilon_t] = 0$.

Assumption L2 substantially weakens Assumption KL2 by replacing the normality assumption with a full support assumption. Let $X^* \in \mathbb{R}^p$.

Assumption L3. For some $d^p \in \mathcal{S}(D^p)$, (A) $\left(\lambda_{1,d_1} \cdots \lambda_{p,d_p}\right) = I_{p \times p}$ and (B) the element of β_{t,d_t} associated with the constant component of X_t is zero.

Assumption L4. (A) For each $(y^{t-1}, x^t) \in \mathcal{S}(Y^{t-1}, X^t)$, $\Pr(D_t = d \mid Y^{t-1} = y^{t-1}, X^t = x^t) > 0$ for all $d \in \mathcal{S}(D_t)$. (B) For each $x_1 \in \mathcal{S}(X_1)$, the variance-covariance matrix of $X^* \mid X_1 = x_1$ is full rank. (C) For each t and $d \in \mathcal{S}(D_t)$, the variance-covariance matrix of X_t conditional on $D_t = d$ is non-singular.

Assumption L3 are normalization assumptions, which are standard in interactive fixed effect models. Assumption L4 (A) is similar to Assumption KL4 (D). It requires that for each history (y^{t-1}, d^{t-1}, x^t), some units are assigned to $D_t = d_t$ for each $d_t \in \mathcal{S}(D_t)$. This assumption is typically satisfied in parametric dynamic discrete choice models (see, e.g., Keane and Wolpin, 1997 and Blundell, 2017 for a survey).

At the cost of increased notational burden, this assumption could be weakened to hold for certain sequences of choices only.

Assumption L5. For any $d^T \in \mathcal{S}(D^T)$, $\{\lambda^u_{t,d_t} : t = 1, \dots, T\}$ areS linearly independent.

Assumption L5 is a standard assumption in the interactive fixed effect literature (see, e.g., 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 X^*_u.

We now define the period t conditional choice probability function as $h_t(y^{t-1}, d^t, x^t) := \Pr(D_t = d_t \mid Y^{t-1} = y^{t-1}, D^{t-1} = d^{t-1}, X^t = x^t)$. In this pure learning environment, the CCP function does not depend on any latent variable and is thus identified directly from the data. As in Section 3.2, our identification result (Theorem 2 below) does not rely on a particular structure imposed on the belief formation process. However, should there be such structure, our identification result would enable identification of the belief formation process. To illustrate this, consider a situation where agents are rational and Bayesian updaters, and where beliefs about X^*_u at time t are a known function of the information set and the model parameters. That is, there is a known function s such that beliefs are given by $s(Y^{t-1}, D^{t-1}, X^{t-1}, \theta)$, where θ are the model parameters. In this case, identification of θ is sufficient for identification of the beliefs.

We now turn to our identification result. Define $f_{\epsilon_t} = \left\{ f_{\epsilon_t(d)} \colon d \in \mathcal{S}(D_t) \right\}$. Let the model parameter vector be $\theta = \left\{ \{\beta_t, \lambda_t, f_{\epsilon_t}, g_t, h_t\}_{t=1}^{T}, \Sigma_u, F_{X^*_k, X_1} \right\} \in \Theta$. The following theorem states that the previous conditions are sufficient for point identification of θ.

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

Key to this result is a simple but powerful insight, namely that, under Assumption L1, this pure learning model is a model of selection on observables. That is, although assignment probabilities depend on unobserved beliefs over X^*, they do not depend on the unobserved factor X^* itself. It follows that one can control for beliefs at time t by conditioning on prior outcomes, choices and covariates. This, in turn, allows us to express the joint distribution of (Y^t, D^t, X^t), suitably weighted by the assignment probabilities, as a mixture over the potential outcomes $Y^t(d_t)$, conditional on the latent factor X^* and exogenous covariates X. From here, the arguments of Freyberger (2018) yield identification of the mixture and component distributions. See Section A.3 for a formal proof.

Remark 5 (Auxiliary measurements). In some cases, additional unselected noisy measurements of known heterogeneity factors are available. This includes, in particular, the Armed Services Vocational Aptitude Battery (ASVAB) ability measures that are available in the National Longitudinal Survey of Youth panels. See, among many others, Cunha et al. (2005), Cunha et al. (2010) and Ashworth et al. (2021). With such auxiliary 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 these conditions are satisfied conditional on each $(Y_t, D_t, X_t)_{t=1}^T$, then the joint distribution of $\left((Y_t, D_t, X_t)_{t=1}^T, X_k^*\right)$ is identified from the auxiliary measurements. From here, one can redefine X_t as (X_t, X_k^*), and Theorem 2 then yields distribution-free identification of the model with both known and unknown heterogeneity.

4 Estimation

We propose to estimate the model parameters via sieve maximum likelihood. We let $W_i = (Y_{i,t}, D_{i,t}, X_{i,t} \colon t = 1, \ldots, T)$ and $\theta^* \in \Theta$ be the true value of the parameters. In the following we focus on the model of Section 3.2 with both known and unknown heterogeneity.[6] Under the conditions of Theorem 1, the log-likelihood contribution of

[6] While we focus on this specification, analogous conditions could be derived for the pure learning model considered in Section 3.3.

$W_i = w$ is given by:

$$\ell(w;\theta) = \log \int\int \prod_{t=1}^{T} \frac{1}{\sigma_t(d_t)} \phi_1\left(\frac{y_t - x_t^\intercal \beta_t(d_t) - x_k^* \lambda_{t,d_t}^k - (x_u^*)^\intercal \lambda_{t,d_t}^u}{\sigma_t(d_t)}\right)$$
$$\times \prod_{t=1}^{T} h_t(d^t, x^t, y^{t-1}, x_k^*) \times \prod_{t=1}^{T-1} g_t(x_{t+1}; y^t, d^t, x^t) dF_{X_1}(x_1)$$
$$\times \frac{1}{\sqrt{|\Sigma_u(x_1)|}} \phi_p\left(\Sigma_u^{-\frac{1}{2}}(x_1)\, x_u^*\right) \times dx_u^* dF_{X_k^*|X_1}(x_k^*, x_1) \tag{5}$$

where ϕ_s is the probability distribution function of the standard multivariate normal distribution with s components, g_t is the distribution of X_{t+1} conditional on $(Y^t, D^t, X^t) = (y^t, d^t, x^t)$. There are four components of the likelihood function, which are associated with the outcomes, the assignment probabilities, the distribution of the covariates, and the joint distribution of (X_1, X^*), respectively.

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^*, $\hat{\theta}$, 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 Assumptions KL1-KL5 under which θ^* is identified, and additional standard conditions (stated in Appendix B.3.1), $\hat{\theta}$ is a consistent estimator for θ^*.

Theorem 3. Let $(W_i)_{i=1}^n$ 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^*.

In practice, researchers are often interested in functionals of the model parameters, such as the variance decompositions discussed in Section 2 and Appendix B.2. These decompositions involve both the finite dimensional parameters of the model, as well as the distribution of X_k^* and the CCPs. We provide in Theorem 4 below an inference result for a plug-in estimator of a general class of functionals of the model parameters, which include those defined in Appendix B.2. For a functional f, under a set of smoothness and regularity conditions similar to those given in Chen and Liao (2014), we show that the plug-in estimator $f(\hat{\theta})$ has an asymptotically normal distribution and characterize its asymptotic variance.

Theorem 4. Let $(W_i)_{i=1}^n$ be i.i.d. data where $T \geq 2p+1$ and Assumptions KL1-KL5 and E1-E13 hold. Then $\sqrt{n}\frac{f(\hat{\theta})-f(\theta^*)}{\|v_n^*\|} \underset{d}{\rightarrow} N(0,1)$ where v_n^* is the sieve Riesz representer of $f(\theta)$ and $\|\cdot\|$ is defined in Equation (15) in Appendix B.3.2.

The rate of convergence of the plug-in sieve estimator depends on the behavior of the sieve variance $\|v_n^*\|$ as n diverges. Note that Theorem 4 does not require that $\|v_n^*\|$ is convergent. That is, Theorem 4 still applies in cases where the parameter of interest is an irregular (i.e., not $\sqrt{n}$ estimable) functional. In either case, consistent estimators for the sieve variance of certain functionals are available (Chen and Liao, 2014, Section 3).[7]

5 Implementation and Monte Carlo simulations

In this section we show how the sieve MLE estimator introduced in Section 4 can be tractably implemented, and then perform a Monte Carlo experiment illustrating the good finite sample performance of the estimator.

5.1 Implementation

We propose an implementation method combining a profiling approach that exploits the parametric components of our model, with a convenient choice of sieve space. Notice first that by integrating out X_u^* in Equation (5), we obtain $\ell(w;\theta) =$

[7] We leave it to future work to derive primitive conditions under which functionals such as the variances decompositions discussed in Section 2 satisfy the high level conditions of Theorem 4.

$\log \int \ell^c(w, x_k^*; \theta^c) dF_{X_k^*|X_1}(x_k^*; x_1)$ with

$$\ell^c(w, x_k^*; \theta^c) := \frac{1}{\sqrt{|V(w, x_k^*; \theta^c)|}} \phi_T \left(V(w, x_k^*; \theta_1^c)^{-\frac{1}{2}} (y^T - m(w, x_k^*; \theta^c)) \right) \times \prod_{t=1}^{T} h_t(d^t, x^t, y^{t-1}, x_k^*) \times \prod_{t=1}^{T-1} g_t(x_{t+1}; y^t, d^t, x^t) dF_{X_1}(x_1),$$

where $m(w, x_k^*; \theta^c) = (\beta_{1,d_1} \cdots \beta_{T,d_T})^\intercal x + \left(\lambda_{1,d_1}^k \cdots \lambda_{T,d_T}^k\right)^\intercal x_k^*$, $V(w, x_k^*; \theta^c) = \left(\lambda_{1,d_1}^u \cdots \lambda_{T,d_T}^u\right)^\intercal \Sigma_u(x_1) \left(\lambda_{1,d_1}^u \cdots \lambda_{T,d_T}^u\right) + \text{diag}(\sigma_{1,d_1}^2, \ldots, \sigma_{T,d_T}^2)$, and θ^c denotes the parameter vector excluding $F_{X_k^*|X_1}$. The above re-expression of the likelihood function embodies two insights. First, although the 'complete' likelihood function ℓ^c is itself an integral over the missing data X_u^*, within our model this integral has the convenient analytical expression described above. Second, the ℓ^c function does not depend on the distribution of the missing data X_k^*, which enables a profiling approach to forming the maximum likelihood estimator.

To explain our profiling approach, suppose for simplicity that $X_k^* \perp\!\!\!\perp X_1$.[8] The profile likelihood approach boils down to solving Equation (6) as

$$\max_{\theta \in \Theta_n} \sum_{i=1}^{n} \ell(w_i, \theta) = \max_{\theta^c \in \Theta_n^c} \sum_{i=1}^{n} \log \int \ell^c(w_i, x_k^*; \theta^c) d[F(\theta^c)](x_k^*),$$

where $F(\theta^c) = \arg\max_{F \in \mathcal{F}_n} \sum_{i=1}^{n} \log \int \ell^c(w_i, x_k^*; \theta^c) dF(x_k^*)$, and $\mathcal{F}_n$ and Θ_n^c are a sieve spaces for $F_{X_k^*}$ and θ^c, respectively. As the non-parametric objects in θ^c are often context specific (for example, g_t may be estimated in a first step, or h_t may be a parametric choice model), we focus on the choice of $\mathcal{F}_n$. Namely, we propose using a sieve space closely related to the estimator discussed in Koenker and Mizera (2014) and Fox et al. (2016). For each n, let us fix a grid of support for X_k^* with $q_n < \infty$

[8] We assume this simply for clarity of exposition. In the general case, a sieve space for $(X_k^*|X_1)$ can be constructed similarly as the cross product of unit simplexes over a grid of $\mathcal{S}(X_1)$.

points, $\mathcal{S}_n = \{\bar{x}^*_{n,1}, \ldots, \bar{x}^*_{n,q_n}\}$. We can then use the following sieve space,

$$\mathcal{F}_n = \left\{ x^* \mapsto \sum_{s=1}^{q_n} \omega_s \mathbf{1}\{x^* \le \bar{x}^*_{n,s}\} \;\middle|\; \omega \in \Delta(q_n) \right\}$$

where $\Delta(m)$ is the $(m-1)$-dimensional unit simplex. Notice that $\mathcal{F}_n$ is the space of distributions with support contained in $\mathcal{S}_n$. As long as the support points are chosen so that $\mathcal{S}_n$ becomes dense in $\mathbb{R}$ and the number of points grows at a suitable rate, this sieve space satisfies the conditions of Theorems 3 and 4.

Importantly for practical purposes, this sieve space turns out to be particularly convenient computationally. To see this, note that under the sieve space $\mathcal{F}_n$ considered above,

$$dF(\theta^c) = \arg\max_{\omega \in \Delta(q_n)} \sum_{i=1}^{n} \log \sum_{s=1}^{q_n} \omega_s \; \ell^c(w_i, \bar{x}^*_{n,s}; \theta^c).$$

Thus the profile step reduces to a convex programming problem. This problem can be solved very efficiently and reliably using recent convex optimization algorithms available in standard softwares. For example the algorithm proposed in Kim et al. (2020) is specialized for this setting and readily implemented in the R package *mixsqp*. This allows us to calculate the profile log likelihood so the full MLE problem can be solved by maximizing this function in θ^c.[9]

5.2 Monte Carlo simulations

Next, we present results from Monte Carlo simulations which illustrate the computational tractability and finite-sample performance of the proposed estimator. We focus here on a specification with a parametric assignment model. In Appendix B.4.3 we consider a specification with a nonparametric assignment model, and show that the estimator achieves similar performance.

[9] In Appendix B.4.1 we show how the gradient of the profile log likelihood function can be calculated implicitly, making it feasible to use first order optimization algorithms to maximize the profile log likelihood function over θ^c efficiently.

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, $X = (X_1, X_2)$, where X_1 has a standard normal distribution and X_2 as a Bernoulli distribution with equal weights. We assume that X_1 and X_2 are independent from each other, and from X^*.

Assignment probabilities are derived from a model in which agents maximize the following expected utility function,

$$\upsilon_t(d, X_k^*, Y^{t-1}, X, D^{t-1}) = \rho E(Y_t(d)|X_k^*, Y^{t-1}, X, D^{t-1}) + \rho\kappa\mathbf{1}(d=2)X_k^* + \nu_t(d),$$

where $Y_t(d) = \alpha_{t,d} + X_1\gamma_{t,d}^{(1)} + X_2\gamma_{t,d}^{(2)} + X_k^*\lambda_{t,d}^k + X_u^*\lambda_{t,d}^u + \epsilon_t(d)$, where $\epsilon_t(d) \sim N(0, \sigma_d^2)$, and $\{\nu_t(d) : t = 1, 2, 3, d = 1, 2\}$ are exogenous and mutually independent with a standard Extreme Value Type 1 distribution. ρ is a scale parameter which affects the relative weight of preference shocks compared to systematic preferences. κ reflects heterogeneity in preferences and/or beliefs that allows X_k^* to affect choices beyond its impact on the expectation of $Y_t(d)$. We assume $X_u^* \sim N(0, \sigma_u^2)$ with $\sigma_u^2 = 1.5$. Finally, X_k^* is distributed following a finite mixture of three truncated normal distributions, with means $(-1.2, 0, 1.5)$, variances $(0.2, 0.1, 0.3)$, and mixing weights $(0.4, 0.3, 0.3)$.[10] The parameter values used in the simulations are reported in Appendix B.4.2. This expected utility function puts a weight on the expected choice-specific potential outcomes, and add another term which depends on X_k^*. This additional term can reflect biased beliefs, heterogeneity in preferences, or a combination of both.

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 $F_{X_k^*}$ nonparametrically using the sieve space described in Section 5.1.[11] The sieve is chosen to have $6n^{1/3}$ uniformly spaced support

[10] Each component distribution is truncated at the third standard deviation of its distribution.

[11] Since X_1 is independent of X_k^*, $F_{X_k^*|X_1} = F_{X_k^*}$.

points.[12]

With this implementation method, computation remains highly tractable for all the sample sizes considered in these simulations. Average computational times to evaluate the maximum likelihood estimator are reported in Table 1 below. Run times increase with sample size from less than half a minute (for $n = 250$), to around three and half minutes for our largest sample size ($n = 4,000$).

	$n = 250$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 4{,}000$
Time (seconds)	24	31	55	135	212

Table 1: Time to compute the estimator. Computational times were obtained using an Intel Core i9-12900K CPU, and are computed as the average over 200 simulations.

The squared bias and variance of the sieve estimator of the finite dimensional parameters are presented in Table 2 below. (Note that all values in this table are multiplied by 1,000.) For each of the parameters, the bias becomes negligible relative to the variance as sample size grows. The variance also declines with sample size, as expected given the consistency of our estimators, at a rate consistent with $\sqrt{n}$-convergence of the mean squared error. Overall most of the parameters are precisely estimated for realistic sample sizes $n \geq 2,000$.

[12] This rate of growth is consistent with the rate conditions of Theorem 4, in particular Assumptions E6 and E7. To contain the unknown bounded support of X_k^*, the grid is chosen to have minimum and maximum values at $(-0.7n^{1/6}, 0.7n^{1/6})$.

	n = 250		n = 500		n = 1,000		n = 2,000		n = 4,000	
	Bias^2	Var	Bias^2	Var	Bias^2	Var	Bias^2	Var	Bias^2	Var
$\alpha_{1,2}$	71.72	87.92	34.06	60.97	12.91	47.13	0.73	19.02	0.04	5.70
$\alpha_{2,1}$	0.15	27.98	0.26	12.38	0.12	7.39	0.00	2.88	0.01	1.38
$\alpha_{2,2}$	73.52	108.96	34.18	74.42	12.41	57.19	0.46	25.80	0.03	8.11
$\alpha_{3,1}$	0.01	36.56	0.45	13.82	0.20	5.31	0.00	2.24	0.01	0.96
$\alpha_{3,2}$	47.84	163.16	32.09	82.42	12.03	62.31	0.59	25.98	0.04	7.32
$\gamma_{1,1}^{(1)}$	0.51	10.08	0.40	5.22	0.14	3.17	0.02	1.49	0.00	0.72
$\gamma_{1,2}^{(1)}$	0.85	15.22	0.30	6.75	0.05	3.35	0.01	1.74	0.00	0.80
$\gamma_{2,1}^{(1)}$	0.84	16.30	0.66	7.86	0.39	4.46	0.04	1.85	0.01	0.80
$\gamma_{2,2}^{(1)}$	1.38	20.81	0.60	12.06	0.09	5.62	0.00	2.69	0.01	1.21
$\gamma_{3,1}^{(1)}$	0.41	9.30	0.24	3.88	0.16	1.89	0.03	1.03	0.01	0.57
$\gamma_{3,2}^{(1)}$	0.38	19.19	0.40	9.11	0.08	4.20	0.01	2.10	0.00	0.86
$\gamma_{1,1}^{(2)}$	0.61	58.91	0.36	23.24	0.36	11.16	0.03	4.77	0.00	2.29
$\gamma_{1,2}^{(2)}$	0.19	46.66	0.22	25.40	0.02	11.16	0.00	5.12	0.01	2.61
$\gamma_{2,1}^{(2)}$	0.01	40.41	0.00	19.84	0.00	9.05	0.00	4.35	0.04	2.48
$\gamma_{2,2}^{(2)}$	0.04	57.76	0.05	26.57	0.00	12.37	0.00	6.76	0.01	3.29
$\gamma_{3,1}^{(2)}$	0.50	40.19	0.08	19.94	0.02	7.64	0.00	3.94	0.02	2.05
$\gamma_{3,2}^{(2)}$	0.10	65.65	0.33	32.11	0.01	15.18	0.02	7.11	0.00	3.44
$\lambda_{1,1}^{k}$	2.75	27.52	1.70	12.89	0.62	7.27	0.01	3.68	0.00	1.47
$\lambda_{2,1}^{k}$	1.15	25.98	0.56	10.83	0.23	4.78	0.00	2.59	0.00	1.09
$\lambda_{2,2}^{k}$	0.87	10.98	0.25	5.82	0.07	2.65	0.01	1.38	0.00	0.74
$\lambda_{3,1}^{k}$	3.99	33.66	0.87	13.72	0.18	5.68	0.00	3.07	0.00	1.33
$\lambda_{3,2}^{k}$	5.70	36.86	0.67	12.56	0.22	5.30	0.01	2.41	0.01	1.08
$\lambda_{1,2}^{u}$	0.98	13.94	0.31	4.73	0.17	2.44	0.01	1.33	0.00	0.61
$\lambda_{2,1}^{u}$	0.04	8.32	0.03	5.14	0.04	1.95	0.01	1.00	0.00	0.48
$\lambda_{2,2}^{u}$	1.48	14.88	0.49	6.22	0.13	3.32	0.01	1.52	0.00	0.64
$\lambda_{3,1}^{u}$	0.45	9.91	0.09	5.00	0.06	2.19	0.03	0.97	0.02	0.47
$\lambda_{3,2}^{u}$	0.11	21.92	0.10	8.90	0.11	4.15	0.00	2.14	0.01	0.94
$\sigma^2(1)$	0.45	2.48	0.09	1.24	0.03	0.67	0.01	0.30	0.00	0.14
$\sigma^2(2)$	1.23	4.45	0.24	2.24	0.03	1.06	0.02	0.70	0.01	0.33
σ_u^2	0.02	72.90	0.05	41.17	0.04	17.91	0.01	9.34	0.01	4.33

Table 2: Simulation results for estimation of finite dimensional parameters. 'Bias^2' and 'Var' refer to the average empirical squared bias and variance scaled by 1,000, respectively, computed over 200 simulations.

Next, we present results for the nonparametric estimator of the distribution of known unobserved heterogeneity X_k^*, focusing on its quantiles $q_\alpha[X_k^*]$. 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 $q_\alpha[X_k^*]$. The results are presented in Figure 1 below. The red line shows the quantile function of the true distribution of X_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 quantile function are the 95th and 5th percentiles of the simulated distribution of the quantile estimators.

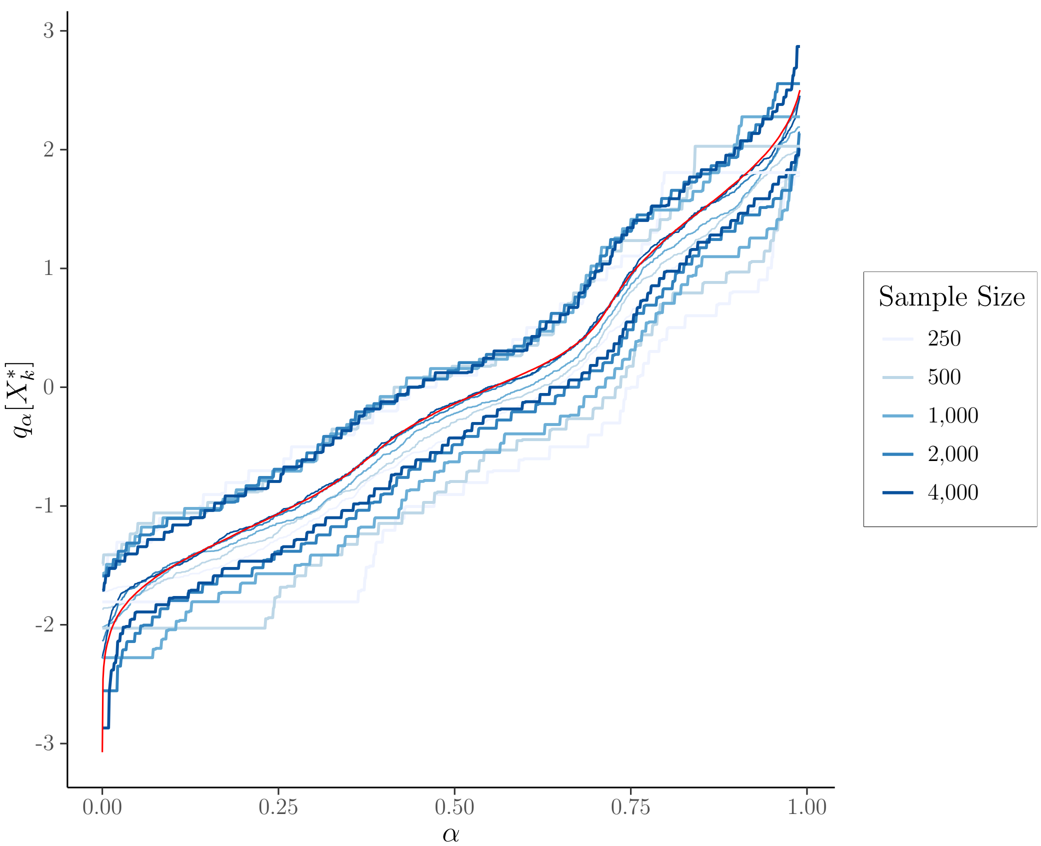

Figure 1: Quantiles of the estimator of $q_\alpha[X_k^*]$. The red line shows the true distribution of X_k^*. The blue lines show the mean, and the 5th and 95th percentiles of the simulated distribution of the estimator of $q_\alpha[X_k^*]$ for each sample size.

The results indicate that the bias of the quantile estimators becomes negligible in moderate sample sizes. The estimator also broadly captures the shape of the true distribution of X_k^*. Besides, and even though the simulated distribution is still relatively disperse for the sample sizes we consider in these simulations, the estimator also appears to converge toward the true distribution as the sample size grows.

Finally, we conclude this section by considering the plug-in estimator for one of the functionals discussed in Section 2 and Appendix B.2. Namely, we focus on the decomposition of the present value of a stream of outcomes into known and unknown components at $t = 1$. Setting the discount rate equal to 0.95, the variance of the unknown and known components corresponding to the two terms in Equation (10) in Appendix B.2 are, for a given choice sequence d^3,[13]

$$\begin{aligned} V_{d^3}^u &:= \sigma_u^2 \sum_{1 \le t_1, t_2 \le 3} (.95)^{t_1+t_2-2} \lambda_{t_1,d_{t_1}}^u \lambda_{t_2,d_{t_2}}^u + \sum_{1 \le t \le 3} (.95)^{2t-2} \sigma_{d_t}^2, \\ V_{d^3}^k &:= \mathrm{Var}(X_k^*) \sum_{1 \le t_1, t_2 \le 3} (.95)^{t_1+t_2-2} \lambda_{t_1,d_{t_1}}^k \lambda_{t_2,d_{t_2}}^k. \end{aligned} \tag{7}$$

We estimate these functionals, which involve both the finite dimensional parameters and $F_{X_k^*}$, using the plug-in estimator described in Section 4. The results are presented in Table 3. For moderately small sample sizes starting with $n = 500$, the squared bias is generally negligibly small relative to the variance. Besides, variance (and MSE) decrease with the sample sizes, at a rate that appears to be consistent with a $\sqrt{n}$-convergence rate.

[13] The sum of these two terms is the variance of $\sum_{t=1}^{3} (.95)^{1-t} Y_t(d_t)$, which is the present value of $(Y_1(d_1), Y_2(d_2), Y_3(d_3))$ at period 1. This is a special case of the class of weighted sums of potential outcomes considered in Appendix B.2, where the weights are $\omega_t = (.95)^{1-t}$, and the choice sequence is d^3. The two terms correspond to the two terms of Equation (10) with ω_t defined as above.

Parameter	n = 250		n = 500		n = 1,000		n = 2,000		n = 4,000	
	Bias^2	Var	Bias^2	Var	Bias^2	Var	Bias^2	Var	Bias^2	Var
$V^k_{(1,1,1)}$	0.01	0.99	0.00	0.45	0.00	0.21	0.00	0.14	0.00	0.07
$V^u_{(1,1,1)}$	0.00	3.06	0.00	1.51	0.00	0.68	0.00	0.33	0.00	0.15
$V^k_{(1,1,2)}$	0.00	1.46	0.01	0.70	0.00	0.38	0.00	0.23	0.00	0.09
$V^u_{(1,1,2)}$	0.00	2.32	0.00	1.13	0.00	0.52	0.00	0.27	0.00	0.12
$V^k_{(1,2,1)}$	0.32	1.77	0.13	0.93	0.04	0.53	0.00	0.28	0.00	0.11
$V^u_{(1,2,1)}$	0.03	1.72	0.00	0.85	0.00	0.37	0.00	0.19	0.00	0.09
$V^k_{(1,2,2)}$	0.21	3.13	0.16	1.53	0.05	0.88	0.00	0.41	0.00	0.15
$V^u_{(1,2,2)}$	0.01	1.20	0.01	0.60	0.00	0.28	0.00	0.15	0.00	0.06
$V^k_{(2,1,1)}$	0.24	1.49	0.07	0.82	0.02	0.36	0.00	0.22	0.00	0.10
$V^u_{(2,1,1)}$	0.03	1.75	0.00	0.85	0.01	0.36	0.00	0.16	0.00	0.08
$V^k_{(2,1,2)}$	0.15	2.43	0.08	1.13	0.03	0.56	0.00	0.32	0.00	0.14
$V^u_{(2,1,2)}$	0.01	1.23	0.01	0.60	0.01	0.27	0.00	0.13	0.00	0.07
$V^k_{(2,2,1)}$	1.00	3.04	0.30	1.56	0.07	0.73	0.00	0.38	0.00	0.17
$V^u_{(2,2,1)}$	0.10	1.10	0.02	0.45	0.01	0.19	0.00	0.09	0.00	0.05
$V^k_{(2,2,2)}$	0.45	5.84	0.21	2.77	0.04	1.56	0.00	0.76	0.00	0.33
$V^u_{(2,2,2)}$	0.06	0.79	0.03	0.32	0.01	0.17	0.00	0.09	0.00	0.04

Table 3: Simulation results for estimation of $V^p_{d^3}$ for $p = k, u$ as defined in Equation (7). 'Bias^2' and 'Var' refer to the average empirical squared bias and variance respectively, computed over 200 simulations.

6 Conclusion

We provide new identification results for a general class of learning models, that encompasses many of the setups that have been considered in the applied literature. We focus on a context where the researcher has access to a short panel of choices and realized outcomes only. As such, our approach is 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 apply to a setup with both known and unknown unobserved heterogeneity. We show that the model is identified under the assumption that the idiosyncratic shocks from the outcome equations and the unknown heterogeneity components are normally distributed, a very frequent restriction in empirical Bayesian learning models. We also show that normality can be relaxed in the case of a pure learning model, while preserving point-identification for this class of models.

We then derive a sieve MLE estimator for the model parameters and a particular class of functionals. The latter includes as special cases the predictable and unpredictable outcome variances, which can in turn be used to evaluate the relative importance of uncertainty versus heterogeneity in life-cycle earnings variability (Cunha et al., 2005). Under appropriate regularity conditions, the resulting estimators are consistent and asymptotically normal. Importantly for practical purposes, we devise a profile likelihood-based procedure that allows us to implement our estimator 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.

Aguirregabiria, V. and Mira, P. (2010), 'Dynamic discrete choice structural models: A survey', *Journal of Econometrics* **156**(1), 38–67.

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. IZA Discussion Paper No. 16585.

Ashworth, J., Hotz, V. J., Maurel, A. and Ransom, T. (2021), ‘Changes across cohorts in wage returns to schooling and early work experiences’, *Journal of Labor Economics* **39**(4), 931–964.

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

Berman, N., Rebeyrol, V. and Vicard, V. (2019), ‘Demand learning and firm dynamics: evidence from exporters’, *Review of Economics and Statistics* **101**(1), 91–106.

Blundell, R. (2017), ‘What have we learned from structural models?’, *American Economic Review* **107**(5), 287–292.

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.

Chiong, K., Galichon, A. and Shum, M. (2016), 'Duality in dynamic discrete-choice models', *Quantitative Economics* **7**(1), 83–115.

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, University of Western Ontario.

Gong, Y., Stinebrickner, T. and Stinebrickner, R. (2019), 'Uncertainty about future income: Initial beliefs and resolution during college', *Quantitative Economics* **10**(2), 607–641.

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.

Hotz, V. J. and Miller, R. A. (1993), 'Conditional choice probabilities and the estimation of dynamic models', *Review of Economic Studies* **60**(3), 497–529.

Hu, Y. and Sasaki, Y. (2018), 'Closed-form identification of dynamic discrete choice models with proxies for unobserved state variables', *Econometric Theory* **34**(1), 166–185.

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.

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.

Shen, X. and Wong, W. H. (1994), 'Convergence rate of sieve estimates', *The Annals of Statistics* pp. 580–615.

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 Proofs for identification section

In this section, we let ϕ denote the standard normal p.d.f.

A.1 Proof of Lemma 1

Proof. We proceed inductively. First, by Assumption KL2 and the definition of (μ_1, Σ_1), $X_u^* \mid (X_1, X_k^*) = (x_1, x_k^*) \sim N(\mu_1, \Sigma_1)$. Second, for $t \geq 1$ suppose $X_u^* \mid (Y^{t-1}, D^{t-1}, X^t, X_k^*) = (y^{t-1}, d^{t-1}, x^t, x_k^*) \sim N(\mu_t, \Sigma_t)$. Then

$$
\begin{aligned}
&f_{X_u^*|Y^t,D^t,X^{t+1},X_k^*}(x_u^*; y^t, d^t, x^{t+1}, x_k^*) \\
&\propto_{(1)} f_{X_u^*|Y^{t-1},D^{t-1},X^t,X_k^*}(x_u^*; y^{t-1}, d^{t-1}, x^t, x_k^*) \\
&\quad \times f_{Y_t,D_t,X_{t+1}|Y^{t-1},D^{t-1},X^t,X^*}(y_t, d_t, x_{t+1}; y^{t-1}, d^{t-1}, x^t, x^*) \\
&\propto_{(2)} f_{X_u^*|Y^{t-1},D^{t-1},X^t,X_k^*}(x_u^*; y^{t-1}, d^{t-1}, x^t, x_k^*) f_{Y_t(d_t)|X_t,X^*}(y_t; x_t, x^*) \\
&\propto_{(3)} \exp\left(-\frac{1}{2}(x_u^* - \mu_t)^\intercal \Sigma_t^{-1}(x_u^* - \mu_t)\right) \phi\left(\frac{y_t - x_t^\intercal \beta_{t,d_t} - x_k^* \lambda_{t,d_t}^k - (x_u^*)^\intercal \lambda_{t,d_t}^u}{\sigma_{t,d_t}}\right) \\
&\propto \exp\left(-\frac{1}{2}(x_u^* - \mu_t)^\intercal \Sigma_t^{-1}(x_u^* - \mu_t)\right) \\
&\quad \times \exp\Bigg(-\frac{1}{2}\left(x_u^* - \lambda_{t,d_t}^u \left((\lambda_{t,d_t}^u)^\intercal \lambda_{t,d_t}^u\right)^{-1}(y_t - x_t^\intercal \beta_{t,d_t} - x_k^* \lambda_{t,d_t}^k)\right)^\intercal \\
&\quad \times \frac{\lambda_{t,d_t}^u (\lambda_{t,d_t}^u)^\intercal}{\sigma_{t,d_t}^2}\left(x_u^* - \lambda_{t,d_t}^u \left((\lambda_{t,d_t}^u)^\intercal \lambda_{t,d_t}^u\right)^{-1}(y_t - x_t^\intercal \beta_{t,d_t} - x_k^* \lambda_{t,d_t}^k)\right)\Bigg) \\
&=_{(4)} \exp\left(-\frac{1}{2}(x_u^* - \mu_{t+1})^\intercal \Sigma_{t+1}^{-1}(x_u^* - \mu_{t+1})\right).
\end{aligned}
$$

Display (1) follows from Bayes' theorem. Display (2) holds since Assumption KL1 has the following three implications: first $X_{t+1} \perp\!\!\!\perp X^* \mid (Y^t, D^t, X^t)$; second $\epsilon_t(d_t) \perp\!\!\!\perp (Y^{t-1}, D^t, X^t, X^*) \implies \epsilon_t(d_t) \perp\!\!\!\perp (Y^{t-1}, D^t, X^{t-1}) \mid (X_t, X^*) \implies Y_t(d_t) \perp\!\!\!\perp (Y^{t-1}, D^t, X^{t-1}) \mid (X_t, X^*)$; third $D_t \perp\!\!\!\perp X_u^* \mid (Y^{t-1}, D^{t-1}, X^t, X_k^*)$. Display (3) holds from the induction assumption and Assumptions KL1 and KL2. Display (4) follows from the definitions in Lemma 1. □

A.2 Proof of Theorem 1

The proof of Theorem 1 uses the following lemmas.

Lemma 2. Let Assumptions KL1 and KL2 hold. Then Y_t conditional on $(Y^{t-1}, D^t, X^t, X_k^*) = (y^{t-1}, d^t, x^t, x_k^*)$ is distributed

$$N\left(x_t^\intercal \beta_{t,d_t} + x_k^* \lambda_{t,d_t}^k + \mu_t^\intercal \lambda_{t,d_t}^u,\ \ (\lambda_{t,d_t}^u)^\intercal \Sigma_t \lambda_{t,d_t}^u + \sigma_{t,d_t}^2\right).$$

Proof. For $t > 1$,

$$\begin{aligned}
&f_{Y_t|Y^{t-1},D^t,X^t,X_k^*}(y_t; y^{t-1}, d^t, x^t, x_k^*) \\
&= \int f_{Y_t(d_t)|Y^{t-1},D^t,X^t,X^*}(y_t; y^{t-1}, d^t, x^t, x^*) f_{X_u^*|Y^{t-1},D^t,X^t,X_k^*}(x_u^*; y^{t-1}, d^t, x^t, x_k^*) dx_u^* \\
&=_{(1)} \int f_{Y_t(d_t)|X_t,X^*}(y_t; x_t, x^*) f_{X_u^*|Y^{t-1},D^{t-1},X^t,X_k^*}(x_u^*; y^{t-1}, d^{t-1}, x^t, x_k^*) dx_u^* \\
&\propto_{(2)} \int \phi\left(\frac{y_t - x_t^\intercal \beta_{t,d_t} - x_k^* \lambda_{t,d_t}^k - (x_u^*)^\intercal \lambda_{t,d_t}^u}{\sigma_{t,d_t}}\right) \exp\left((x_u^* - \mu_t)^\intercal \Sigma_t^{-1}(x_u^* - \mu_t)\right) dx_u^* \\
&= \phi\left(\frac{y_t - x_t^\intercal \beta_{t,d_t} - x_k^* \lambda_{t,d_t}^k - \mu_t^\intercal \lambda_{t,d_t}^u}{\sqrt{(\lambda_{t,d_t}^u)^\intercal \Sigma_t \lambda_{t,d_t}^u + \sigma_{t,d_t}^2}}\right)
\end{aligned}$$

Display (1) holds because Assumption KL1 implies $Y_t(d_t) \perp\!\!\!\perp (Y^{t-1}, D^t, X^{t-1}) \mid (X_t, X^*)$ and $D_t \perp\!\!\!\perp X_u^* \mid (Y^{t-1}, D^{t-1}, X^t, X_k^*)$. Display (2) holds because Assumption KL1 and KL2 imply Lemma 1 and $\epsilon_t(d) \mid (X_t, X^*) \sim N(0, \sigma_{t,d}^2)$. A similar argument applies for $t = 1$. □

For the following results, it is useful to notice that, for $t \geq 1$,

$$\begin{aligned}
\Sigma_{t+1} &= \left(\Sigma_u^{-1}(x_1) + \sum_{s=1}^{t} \sigma_{s,d_s}^{-2} \lambda_{s,d_s}^u (\lambda_{s,d_s}^u)^\intercal\right)^{-1}, \\
\mu_{t+1} &= \Sigma_{t+1}\left(\sum_{s=1}^{t} \lambda_{s,d_s}^u \frac{y_s - x_s^\intercal \beta_{s,d_s} - x_k^* \lambda_{s,d_s}^k}{\sigma_{s,d_s}^2}\right).
\end{aligned}$$

Lemma 3. Let Assumptions KL1, KL2, KL4 (A,B,C) and KL5 (C) hold. Then, for each $(y^{t-1}, d^t, x^t) \in \mathcal{S}((Y^{t-1}, D^t, X^t))$ there exists an affine function π such that, for all $y_t \in \mathcal{S}(Y_t)$, $F_{Y^t, D^t, X^t, X_k^*}(y^t, d^t, x^t, \pi(x_k^*))$ is identified.

Proof. Fix $(y^{t-1}, d^t, x^t) \in \mathcal{S}((Y^{t-1}, D^t, X^t))$. Since $f_{Y_t|Y^{t-1}, D^t, X^t}(y_t; y^{t-1}, d^t, x^t) =$

$$\int f_{Y_t|Y^{t-1}, D^t, X^t, X_k^*}(y_t; y^{t-1}, d^t, x^t, x_k^*) dF_{X_k^*|Y^{t-1}, D^t, X^t}(x_k^*; y^{t-1}, d^t, x^t),$$

Lemma 2 implies $f_{Y_t|Y^{t-1}, D^t, X^t}(y_t; y^{t-1}, d^t, x^t)$ is a mixture of normal random variables. To identify the component and mixture distributions, we apply Bruni and Koch (1985, Theorem 3). First, for any t and $(y^{t-1}, d^t, x^t) \in \mathcal{S}((Y^{t-1}, D^t, X^t))$, define $\Lambda :=$

$$\left\{x_k^* \mapsto \left(x_t^\intercal \beta_{t,d_t} + x_k^*(\lambda_{t,d_t}^k + (\mu_t^k)^\intercal \lambda_{t,d_t}^u) + (\mu_t^u)^\intercal \lambda_{t,d_t}^u,\ (\lambda_{t,d_t}^u)^\intercal \Sigma_t \lambda_{t,d_t}^u + \sigma_{t,d_t}^2\right) : \theta^t \in \Theta^t\right\},$$

where $\theta^t := \left\{\{\beta_{s,d_s}, \lambda_{s,d_s}^k, \lambda_{s,d_s}^u, \sigma_{s,d_s}^2 : s = 1, \ldots, t\}, \Sigma_u(x_1)\right\}$, Θ^t is the corresponding subset of Θ, and $\mu_t = \mu_t^k x_k^* + \mu_t^u$ for all x_k^*. I.e., $\mu_1^k = \mu_1^u = 0$ and for $t > 1$,

$$\mu_t^k := -\Sigma_t \sum_{s=1}^{t-1} \lambda_{s,d_s}^u \frac{\lambda_{s,d_s}^k}{\sigma_{s,d_s}^2}, \qquad \mu_t^u := \Sigma_t \sum_{s=1}^{t-1} \lambda_{s,d_s}^u \frac{y_{is} - x_{is}^\intercal \beta_{s,d_s}}{\sigma_{s,d_s}^2}.$$

Under Assumptions KL4 (A,B,C) and KL5 (C), $\Lambda \subset \Lambda_4$ where Λ_4 is defined in Bruni and Koch (1985, p. 1344). Thus Bruni and Koch (1985, Theorem 3) applies and

$$\left\{x_t^\intercal \beta_{t,d_t} + \pi(x_k^*)(\lambda_{t,d_t}^k + (\mu_t^k)^\intercal \lambda_{t,d_t}^u) + (\mu_t^u)^\intercal \lambda_{t,d_t}^u,\ (\lambda_{t,d_t}^u)^\intercal \Sigma_t \lambda_{t,d_t}^u + \sigma_{t,d_t}^2\right\}$$

and $F_{X_k^*|Y^{t-1}, D^t, X^t}(\pi(x_k^*); y^{t-1}, d^t, x^t)$ are identified with $\pi(x_k^*) = \pi_0 + \pi_1 x_k^*$. □

Lemma 4. Let Assumptions KL1, KL2, KL3 (A), KL4 and KL5 (C) hold. Then $\mathcal{S}(X_k^*)$ is identified from $F_{Y_1, D_1, X_1}(y_1, d_1, x_1)$.

Proof. In this proof, it will be useful to denote $\beta_{1,d} = (\alpha_{1,d}, \gamma_{1,d}^\intercal)^\intercal$, where $\alpha_{1,d}$ is the coefficient on the constant term in X_1.

For any $x_1 \in \mathcal{S}(X_1)$ and $d \in \mathcal{S}(D_1)$, Lemma 3 implies

$$\left\{x_1^\intercal \beta_{1,d} + (\pi_0 + \pi_1 x_k^*)\lambda_{1,d}^k,\ (\lambda_{1,d}^u)^\intercal \Sigma_1(x_1)\lambda_{1,d}^u + \sigma_{1,d}^2,\ F_{X_k^*|D_1,X_1}(\pi_0 + \pi_1 x_k^*; d, x_1)\right\}$$

is identified. Set $d \in \mathcal{S}(D_1)$ as in Assumption KL3 (A). We now show $(\pi_0, \pi_1) = (0, 1)$.[14] By Assumption KL4 (D), $\exists\, x_k^* \neq \tilde{x}_k^*$ such that $dF_{X_k^*|D_1,X_1}(\pi_0+\pi_1 x_k^*; d, x_1) > 0$ and $dF_{X_k^*|D_1,X_1}(\pi_0 + \pi_1 \tilde{x}_k^*; d, x_1) > 0$. Then by Assumption KL3 (A), $1 = \lambda_{1,d}^k = \frac{(x_1^\intercal \beta_{1,d}+(\pi_0+\pi_1 x_k^*)\lambda_{1,d}^k)-(x_1^\intercal \beta_{1,d}+(\pi_0+\pi_1 \tilde{x}_k^*)\lambda_{1,d}^k)}{x_k^*-\tilde{x}_k^*} = \pi_1$. Thus $x_1^\intercal \beta_{1,d}+\pi_0$ is identified by $(x_1^\intercal \beta_{1,d}+(\pi_0 + x_k^*)) - x_k^*$. If $\exists\, x_1, \tilde{x}_1 \in \mathcal{S}(X_1)$ such that their respective π_0 differ, then $\mathcal{S}(X_k^* \mid X_1 = x_1, D_1 = d) \neq \mathcal{S}(X_k^* \mid X_1 = \tilde{x}_1, D_1 = d)$, which contradicts Assumption KL4 (D). Therefore $(\alpha_{1,d} + \pi_0, \gamma_{1,d}^\intercal)^\intercal = E[X_1 X_1^\intercal | D_1 = d]^{-1} E[X_1\,(X_1^\intercal \beta_{1,d} + \pi_0) \mid D_1 = d]$, which exists by Assumption KL4 (E). Finally, by Assumption KL3 (A), $0 = \alpha_{1,d} = (x_1^\intercal \beta_{1,d} + \pi_0) - x_1^\intercal(\alpha_{1,d}, \gamma_{1,d}^\intercal)^\intercal = \pi_0$. To conclude, by Assumption KL4 (D), $\mathcal{S}(X_k^*) = \mathcal{S}(X_k^* \mid D_1 = d_1, X_1 = x_1)$. □

Lemma 5. Under the assumptions in Theorem 1, $F_{Y^T,D^T,X^T,X_k^*}(y^T, d^T, x^T, x_k^*)$ is identified on its support.

Proof. For any t and $(y^{t-1}, d^t, x^t) \in \mathcal{S}((Y^{t-1}, D^t, X^t))$, it follows from Lemma 3 that $dF_{X_k^*|Y^{t-1},D^t,X^t}(\pi(x_k^*); y^{t-1}, d^t, x^t)$ is identified. Then since $\mathcal{S}(X_k^*)$ is known by Lemma 4, Assumption KL4 (D) implies $\mathcal{S}(X_k^*) =$

$$dF^{-1}_{X_k^*|Y^{t-1},D^t,X^t}(\cdot; y^{t-1}, d^t, x^t)[\mathbb{R}_+] = (dF_{X_k^*|Y^{t-1},D^t,X^t}(\cdot; y^{t-1}, d^t, x^t) \circ \pi)^{-1}[\mathbb{R}_+],$$

where $R_+ = \{x \in \mathbb{R}\colon x > 0\}$. Then, since π is bijective, $\pi[\mathcal{S}(X_k^*)] = \mathcal{S}(X_k^*)$. The only affine functions that satisfy this identity are $\pi(x_k^*) = x_k^*$ and $\pi(x_k^*) = \sup \mathcal{S}(X_k^*) + \inf \mathcal{S}(X_k^*) - x_k^*$. To conclude the proof, we need to rule out the second function.

To proceed, let μ_t^k and μ_t^u be defined as in the proof to Lemma 3, and, for any $1 \leq s < t$, let $\tilde{\mu}_{t,s}(d^{t-1}) \coloneqq \Sigma_t \frac{\lambda_{s,d_s}^u}{\sigma_{s,d_s}^2}$. Now note that by Lemma 3 and Assumption KL4,

[14] Recall from Lemma 3 that the affine function π may depend on the history (y^{t-1}, d^t, x^t). In this lemma we show that the affine function is the identity for one particular choice history.

for any t and $d^t \in \mathcal{S}(D^t)$, $jc_t(d^t) = \lambda^k_{t,d_t} + (\mu^k_t)^\intercal \lambda^u_{t,d_t}$ with $j \in \{-1, 1\}$ unknown and $c_t(d^t) := \frac{(x_t^\intercal \beta_{t,d_t} + \pi(x^*_k)\lambda^k_{t,d_t} + \mu_t^\intercal \lambda^u_{t,d_t}) - (x_t^\intercal \beta_{t,d_t} + \pi(\tilde{x}^*_k)\lambda^k_{t,d_t} + \mu_t^\intercal \lambda^u_{t,d_t})}{x^*_k - \tilde{x}^*_k}$ known. In addition, for any $1 \le s < t$, $\frac{\partial}{\partial y_s}(x_t^\intercal \beta_{t,d_t} + \pi(x^*_k)\lambda^k_{t,d_t} + \mu_t^\intercal \lambda^u_{t,d_t}) = (\lambda^u_{t,d_t})^\intercal \tilde{\mu}_{t,s}(d^{t-1})$.

The proof is inductive. First consider $t = 1$. Applying the above argument to the sequences $\{\tilde{d}_1, (d_1, d_2), (\tilde{d}_1, d_2)\}$ for $d_1 \in \mathcal{S}(D_1)$ as in Assumption KL3 (A), $\tilde{d}_1 \in \mathcal{S}(D_1) \setminus \{d_1\}$, and $d_2 \in \mathcal{S}(D_2)$, yields identification of $j_1 c_1(\tilde{d}_1)$, $j_{d_2} c_2((d_1, d_2))$ $(\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(d_1)$, $\tilde{j}_{d_2} c_2((\tilde{d}_1, d_2))$, and $(\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(\tilde{d}_1)$ with $(j_1, \tilde{j}_{d_2}, j_{d_2}) \in \{-1, 1\}^3$ unknown. Since $\lambda^k_{1,d_1} = 1$, $j_1 c_1(\tilde{d}_1) = \lambda^k_{1,\tilde{d}_1}$, $j_{d_2} c_2((d_1, d_2)) = \lambda^k_{2,d_2} - (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(d_1)$, and $\tilde{j}_{d_2} c_2((\tilde{d}_1, d_2)) = (\lambda^k_{2,d_2} - (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(\tilde{d}_1)\lambda^k_{1,\tilde{d}_1})$, it must be that

$$(\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(d_1) + j_{d_2} c_2((d_1, d_2)) = (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(\tilde{d}_1) j_1 c_1(\tilde{d}_1) + \tilde{j}_{d_2} c_2((\tilde{d}_1, d_2)). \tag{8}$$

We use this identity to show $(j_1, \tilde{j}_{d_2}, j_{d_2}) = (1, 1, 1)$. Suppose $j_{d_2} = 1$. It is straightforward to show that Equation (8) implies:

$$\begin{aligned}
(j_1, \tilde{j}_{d_2}) = (-1, -1) &\implies \lambda^k_{2,d_2} = 0, \\
(j_1, \tilde{j}_{d_2}) = (1, -1) &\implies \lambda^k_{2,d_2} - (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(\tilde{d}_1)\lambda^k_{1,\tilde{d}_1} = 0, \\
(j_1, \tilde{j}_{d_2}) = (-1, 1) &\implies (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(\tilde{d}_1)\lambda^k_{1,\tilde{d}_1} = 0,
\end{aligned}$$

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

$$\begin{aligned}
(j_1, \tilde{j}_{d_2}) = (1, 1) &\implies \lambda^k_{2,d_2} - (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(d_1)\lambda^k_{1,d_1} = 0, \\
(j_1, \tilde{j}_{d_2}) = (-1, -1) &\implies (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(d_1)\lambda^k_{1,d_1} = 0, \\
(j_1, \tilde{j}_{d_2}) = (1, -1) &\implies (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(\tilde{d}_1)\lambda^k_{1,\tilde{d}_1} - (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(d_1)\lambda^k_{1,d_1} = 0, \\
(j_1, \tilde{j}_{d_2}) = (-1, 1) &\implies \lambda^k_{2,d_2} - (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(\tilde{d}_1)\lambda^k_{1,\tilde{d}_1} - (\lambda^u_{2,d_2})^\intercal \tilde{\mu}_{2,1}(d_1)\lambda^k_{1,d_1} = 0.
\end{aligned}$$

The first three implications contradict Assumptions KL5 (C), (D) and (A), respectively. To conclude, for each $d \in \{d_{2,i}, \tilde{d}_{2,i} \in \mathcal{S}(D_2) : i = 1, 2, \ldots, p\}$ of Assumption

KL5 (E), by considering the sequences $\{(d_1, d), (\tilde{d}_1, d)\}$, $j_d c_2((d_1, d))$ and $\tilde{j}_d c_2((\tilde{d}_1, d))$ are identified with $(j_d, \tilde{j}_d) \in \{(-1, 1), (1, 1)\}$. Since $\lambda^k_{1,\tilde{d}_1} \neq 0$ by Assumption KL5 (B), for the sign of $\lambda^k_{1,\tilde{d}_1}$ to be constant across sequences, we can rule out all signs except $\left(j_1, (j_{d_{2,i}}, \tilde{j}_{d_{2,i}}, j_{\tilde{d}_{2,i}}, \tilde{j}_{\tilde{d}_{2,i}} : i = 1, \ldots, p)\right) \in \{(1, (1, 1, 1, 1)^p), (-1, (-1, 1, -1, 1)^p)\}$. If $\left(j_1, (j_{d_{2,i}}, \tilde{j}_{d_{2,i}}, j_{\tilde{d}_{2,i}}, \tilde{j}_{\tilde{d}_{2,i}} : i = 1, \ldots, p)\right) = (-1, (-1, 1, -1, 1)^p)$, then

$$\begin{aligned} 0 &= \text{vec}\left(\lambda^k_{2,d_{2,1}}, \ldots, \lambda^k_{2,d_{2,p}}\right) - \left(\lambda^u_{2,d_{2,1}} \cdots \lambda^u_{2,d_{2,p}}\right)^\intercal \left(\tilde{\mu}_{2,1}(\tilde{d}_1)\lambda^k_{1,\tilde{d}_1} + \tilde{\mu}_{2,1}(d_1)\lambda^k_{1,d_1}\right) \\ &= \text{vec}\left(\lambda^k_{2,\tilde{d}_{2,1}}, \ldots, \lambda^k_{2,\tilde{d}_{2,p}}\right) - \left(\lambda^u_{2,\tilde{d}_{2,1}} \cdots \lambda^u_{2,\tilde{d}_{2,p}}\right)^\intercal \left(\tilde{\mu}_{2,1}(\tilde{d}_1)\lambda^k_{1,\tilde{d}_1} + \tilde{\mu}_{2,1}(d_1)\lambda^k_{1,d_1}\right), \end{aligned}$$

which contradicts Assumption KL5 (E).

For the induction step, suppose π is identity for each history (y^{s-1}, d^s, x^s), $s = 1, \ldots, t-1$, and let $d^t, \tilde{d}^t \in \mathcal{S}(D^t)$ satisfy $d_t = \tilde{d}_t$ and $d_{t-1} \neq \tilde{d}_{t-1}$. By the preceding arguments, $j_1 c_t(d^t)$, $j_2 c_t(\tilde{d}^t)$ with $(j_1, j_2) \in \{-1, 1\}^2$, and, for each $s < t$, $(\lambda^u_{t,d_t})^\intercal \tilde{\mu}_{t,s}(d^{t-1})$ and $(\lambda^u_{t,d_t})^\intercal \tilde{\mu}_{t,s}(\tilde{d}^{t-1})$ are identified. Since $\lambda^k_{s,d}$ is identified for any $s < t$ and $d \in \mathcal{S}(D_s)$, $j_1 c_t(d^t) = \lambda^k_{t,d_t} - (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(d^{t-1})\lambda^k_{s,d_s}$ and $j_2 c_t(\tilde{d}^t) = \lambda^k_{t,d_t} - (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(d^{t-1})\lambda^k_{s,\tilde{d}_s}$, it must be that

$$j_1 c_t(d^t) + (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(d^{t-1})\lambda^k_{s,d_s} = j_2 c_t(\tilde{d}^t) + (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(\tilde{d}^{t-1})\lambda^k_{s,\tilde{d}_s}. \quad (9)$$

We use this identity to show $(j_1, j_2) = (1, 1)$. Consider

$$\begin{aligned} (j_1, j_2) = (1, -1) &\implies \left(\lambda^k_{t,d_t} - (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(\tilde{d}^{t-1})\lambda^k_{s,\tilde{d}_s}\right) = 0, \\ (j_1, j_2) = (-1, 1) &\implies \left(\lambda^k_{t,d_t} - (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(d^{t-1})\lambda^k_{s,d_s}\right) = 0, \\ (j_1, j_2) = (-1, -1) &\implies (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(d^{t-1})\lambda^k_{s,d_s} - (\lambda^u_{t,d_t})^\intercal \sum_{s=1}^{t-1} \tilde{\mu}_{t,s}(\tilde{d}^{t-1})\lambda^k_{s,\tilde{d}_s} = 0, \end{aligned}$$

which contradict Assumptions KL5 (C), (C) and (A), respectively. □

Proof of Theorem 1. By Lemma 5, f_{Y^T,D^T,X^T,X_k^*}, and thus h_t, is identified. First,

$$\begin{aligned}
& f_{Y^T,D^T,X^T,X_k^*}\left(y^T,d^T,x^T,x_k^*\right) \\
&= \int f_{Y^T(d^T),D^T,X^T,X^*}\left(y^T,d^T,x^T,x^*\right)dx_u^* \\
&= \int f_{Y_T(d_T)|X_T,X^*}\left(y_T;x_T,x^*\right) f_{D_T|Y^{T-1},D^{T-1},X^T,X_k^*}(d_T;y^{T-1},d^{T-1},x^T,x_k^*) \\
&\quad \times f_{X_T|Y^{T-1},D^{T-1},X^{T-1}}(x_T;y^{T-1},d^{T-1},x^{T-1})\dots f_{Y_1(d_1)|X_1,X^*}\left(y_1;x_1,x^*\right) \\
&\quad \times f_{D_1|X_1,X_k^*}(d_1;x_1,x_k^*)f_{X_u^*|X_1,X_k^*}(x_u^*;x_1,x_k^*)f_{X_1,X_k^*}(x_1,x_k^*)dx_u^*.
\end{aligned}$$

This implies that on the support of f_{Y^T,D^T,X^T,X_k^*},

$$\begin{aligned}
& \frac{f_{Y^T,D^T,X^T,X_k^*}\left(y^T,d^T,x^T,x_k^*\right)}{f_{D_1,X_1,X_k^*}(d_1,x_1,x_k^*)\prod_{t=2}^T f_{D_t,X_t|Y^{t-1},D^{t-1},X^{t-1},X_k^*}(d_t,x_t;y^{t-1},d^{t-1},x^{t-1},x_k^*)} \\
&= \int \prod_{t=1}^T f_{Y_t(d_t)|X_t,X^*}\left(y_t;x_t,x^*\right) f_{X_u^*|X_k^*,X_1}(x_u^*;x_k^*,x_1)dx_u^*.
\end{aligned}$$

The function is equal to the probability density function of a jointly normal random variable with mean

$$\left(x_t^\intercal \beta_{t,d_t} + x_k^* \lambda_{t,d_t}^k\right)_{t=1}^T,$$

and covariance matrix

$$(\lambda_d^u)^\intercal \Sigma_u(x_1)\lambda_d^u + \operatorname{diag}\left(\sigma_{t,d_t}^2 : t=1,\dots,T\right),$$

where $\lambda_d^u = \left(\lambda_{1,d_1}^u \cdots \lambda_{T,d_T}^u\right)$. By Assumptions KL4 (D) and (E), the components of the mean function are identified. The components of the covariance matrix are identified under Assumptions KL3 (B) and KL5 (F). □

A.3 Proof of Theorem 2

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 $x \in \mathcal{S}(X)$ and $d^T \in \mathcal{S}(D^T)$ 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 $[L_{123} m](w_1) =$

$$\int \frac{f_{Y,D,X}(y,d,x)}{f_{D_1,X_1}(d_1,x_1) \prod_{t=2}^{T} f_{D_t,X_t|Y^{t-1},D^{t-1},X^{t-1}}(d_t, x_t; y^{t-1}, d^{t-1}, x^{t-1})} m(w_3) dw_3,$$

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

$$L_{1X^*} : \mathcal{L} \to \mathcal{L} \qquad [L_{1X^*} m](w_1) = \int \prod_{t=1}^{p} f_{Y_t(d_t)|X_t,X^*}(y_t; x_t, x^*) m(x^*) dx^*,$$

$$L_{X^*3} : \mathcal{L}_{W_3} \to \mathcal{L} \qquad [L_{X^*3} m](x^*) = \int \prod_{t=p+2}^{T} f_{Y_t(d_t)|X_t,X^*}(y_t; x_t, x^*) f_{X^*|X_1}(x^*; x_1) m(w_1) dw_1,$$

$$D_{X^*} : \mathcal{L}_{X^*} \to \mathcal{L}_{X^*} \qquad [D_{X^*} m](x^*) = f_{Y_{p+1}(d_{p+1})|X_{p+1},X^*}(y_{p+1}; x_{p+1}, x^*) m(x^*).$$

The following derivation shows that $L_{123} = L_{1X^*} D_{X^*} L_{X^*3}$. First,

$$\begin{aligned} f_{Y,D,X}(y,d,x) &= \int f_{Y,D,X,X^*}(y,d,x,x^*) dx^* \\ &= \int f_{Y_T(d_T)|X_T,X^*}(y_T; x_T, x^*) f_{D_T,X_T|Y^{T-1},D^{T-1},X^{T-1}}(d_T, x_T; y^{T-1}, d^{T-1}, x^{T-1}) \\ &\quad \times f_{Y_{T-1}(d_{T-1})|X_{T-1},X^*}(y_{T-1}; x_{t-1}, x^*) \ldots f_{D_1,X_1}(d_1, x_1) f_{X^*|X_1}(x^*; x_1) dx^*. \end{aligned}$$

Then, by Assumption L4 (A),

$$\frac{f_{Y,D,X}(y,d,x)}{f_{D_1,X_1}(d_1,x_1)\prod_{t=2}^{T} f_{D_t,X_t|Y^{t-1},D^{t-1},X^{t-1}}(d_t,x_t;y^{t-1},d^{t-1},x^{t-1})}$$
$$= \int \prod_{t=1}^{T} f_{Y_t(d_t)|X_t,X^*}(y_t;x_t,x^*) f_{X^*|X_1}(x^*;x_1)dx^*,$$

and therefore it follows that

$$\begin{aligned}
[L_{123}m](w_1) =& \int \left(\int \prod_{t=1}^{T} f_{Y_t(d_t)|X_t,X^*}(y_t;x_t,x^*) f_{X^*|X_t}(x^*;x_t)dx^* \right) m(w_3)dw_3 \\
=& \int \prod_{t=1}^{p+1} f_{Y_t(d_t)|X_t,X^*}(y_t;x_t,x^*) \left(\int \prod_{t=p+2}^{T} f_{Y_t(d_t)|X_t,X^*}(y_t;x_t,x^*) f_{X^*|X_t}(x^*)m(w_3)dw_3 \right) dx^* \\
=& \int \prod_{t=1}^{p} f_{Y_t(d_t)|X_t,X^*}(y_t;x_t,x^*) \left(f_{Y_{p+1}(d_{p+1})|X_{p+1},X^*}(y_{p+1};x_{p+1},x^*)[L_{X^*3}m](x^*) \right) dx^* \\
=& \int \int \prod_{t=1}^{p} f_{Y_t(d_t)|X_t,X^*}(y_t;x_t,x^*)[D_{X^*}L_{X^*3}m](x^*)dx^* \\
=&[L_{1X^*}D_{X^*}L_{X^*3}m](w_1),
\end{aligned}$$

and $L_{123} = L_{1X^*}D_{X^*}L_{X^*3}$. Similarly, $L_{13} = L_{1X^*}L_{X^*3}$.

From here, Assumptions L1, L2, L3, L4 (B), and L5 imply the arguments of Theorem 1 Freyberger (2018) apply, so that λ_{t,d_t}, $f_{Y_t(d_t)|X_t,X^*}(\cdot;x_t,\cdot)$ and $f_{X^*|X_1}(\cdot;x_1)$ are identified for each t for the given (d_t,x).[15] Given identification of $f_{Y_t(d_t)|X_t,X^*}(\cdot;x_t,\cdot)$ for each $x_t \in \mathcal{S}(X_t)$ and λ_{t,d_t}, Assumption L4 (C) implies identification of β_{t,d_t} and thus $f_{\epsilon_t(d_t)}$.

Next, given an arbitrary t and d_t, define $\tilde{d}$ by replacing the t-th element of d with d_t. Then consider a permutation $(1,2,\ldots,T) \mapsto (t_1,t_2,\ldots,t_T)$ such that $t \mapsto t_1$ and

[15]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 $f_{Y_t(d_t)|X_t,X^*}(y;x,x^*) = f_{\epsilon_t(d_t)}(y - x^\intercal\beta_{t,d_t} - (x^*)^\intercal\lambda_t)$, Freyberger (2018, Lemma 1) and D'Haultfoeuille (2011) may be applied with minor modifications.

define $\tilde{W}_1 = (Y_{t_1}, Y_{t_2}, \ldots, Y_{t_p})$, $\tilde{W}_2 = (Y_{t_{p+1}}, Y_{t_{p+1}}, \ldots, Y_{t_T})$,

$$\tilde{L}_{2X^*} : \mathcal{L} \to \mathcal{L} \qquad [\tilde{L}_{2X^*} m](\tilde{w}_2) = \int \prod_{i=p+1}^{T} f_{Y_{t_i}(d_{t_i})|X_{t_i},X^*}(y_{t_i}; x_{t_i}, x^*) f_{X^*|X_1}(x^*; x_1) m(x^*) dx^*,$$

$$\tilde{L}_{X^*1} : \mathcal{L}_{\tilde{W}_1} \to \mathcal{L} \qquad [\tilde{L}_{X^*1} m](x^*) = \int \prod_{i=1}^{p} f_{Y_{t_i}(d_{t_i})|X_{t_i},X^*}(y_{t_i}; x_{t_i}, x^*) m(\tilde{w}_1) d\tilde{w}_1,$$

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

$$[\tilde{L}_{21} m](\tilde{w}_2) = \int \frac{f_{Y,D,X}(y,d,x)}{f_{D_1,X_1}(d_1,x_1) \prod_{t=2}^{T} f_{D_t,X_t|Y^{t-1},D^{t-1},X^{t-1}}(d_t, x_t; y^{t-1}, d^{t-1}, x^{t-1})} m(\tilde{w}_1) d\tilde{w}_1.$$

As before, $\tilde{L}_{21} = \tilde{L}_{2X^*} \tilde{L}_{X^*1}$. Since $\tilde{L}_{2X^*}$ and $\tilde{L}_{21}$ are identified and injective, $\tilde{L}_{X^*1}$ is identified by $\tilde{L}_{2X^*}^{-1} \tilde{L}_{21} = \tilde{L}_{X^*1}$ and thus $\beta_{t,d_t}, \lambda_{t,d_t}, f_{\epsilon(d_t)}$. □

B Online Appendix

B.1 Proof of Corollary 1

In this proof we denote $\beta_{t,d} = (\alpha_{t,d}, \gamma_{t,d}^\intercal)^\intercal$, where $\alpha_{t,d}$ is the coefficient on the constant term in X_t. Fix d^p as in the statement and define $\lambda_u = \left(\lambda_{1,d_1}^u \cdots \lambda_{p,d_p}^u\right)$, $\tilde{X}_u^* = \lambda_u^\intercal (X_u^* - \mu_u)$, $\tilde{\epsilon}_t(d) = \epsilon_t(d) - c_{t,d}$, $\tilde{X}_k^* = b + \lambda_{1,d_1}^k X_k^*$ where $b = \alpha_{1,d_1} + \mu_u^\intercal \lambda_{1,d_1}^u + c_{1,d_1}$. Finally, define $\tilde{\lambda}_{t,d_t}^k = (\lambda_{1,d_1}^k)^{-1}\lambda_{t,d_t}^k$, $\tilde{\lambda}_{t,d_t}^u = \lambda_u^{-1}\lambda_{t,d_t}^u$, and $\tilde{\alpha}_{t,d_t} = \alpha_{t,d_t} - \tilde{\lambda}_{t,d_t}^k b + \mu_u^\intercal \lambda_{t,d_t}^u + c_{t,d_t}$. We then have that

$$Y_t(d_t) = X_t^\intercal \left(\tilde{\alpha}_{t,d_t}, \gamma_{t,d_t}^\intercal\right)^\intercal + (\tilde{X}_u^*)^\intercal \tilde{\lambda}_{t,d_t}^u + \tilde{X}_k^* \tilde{\lambda}_{t,d_t}^k + \tilde{\epsilon}_t(d_t),$$

$E[\tilde{\epsilon}_t(d_t)] = 0$ and $E[\tilde{X}_u^* \mid X_1 = x, X_k^* = x_k^*] = 0$ so that the reparameterized model satisfies Assumption KL2 (with $\tilde{\Sigma}_u(x_1) = \lambda_u^\intercal \Sigma_u(x_1)\lambda_u$). Also, $\tilde{\lambda}_{1,d_1}^k = 1$, $\tilde{\alpha}_{1,d_1} = 0$ and $\left(\tilde{\lambda}_{1,d_1}^u \cdots \tilde{\lambda}_{p,d_p}^u\right) = I_{p\times p}$ so the reparameterized model satisfies Assumption KL3. By Theorem 1, $\tilde{\theta} = \left\{\{\tilde{\alpha}_{t,d_t}, \gamma_{t,d_t}, \tilde{\lambda}_{t,d_t}^k, \tilde{\lambda}_{t,d_t}^u, \sigma_{t,d_t}^2, g_t, \tilde{h}_t\}_{t=1}^T, \tilde{\Sigma}_u, F_{\tilde{X}_k^* X_1}\right\}$ is identified, where $\tilde{h}_t$ and $F_{\tilde{X}_k^* X_1}$ are the CCPs and distribution of $(\tilde{X}_k^*, X_1)$, respectively. This, in turn, implies the identification of the distribution of C_{t,d_t}^j for $j = k, u$. Finally,

$$\begin{aligned}
&x^\intercal \left(\tilde{\alpha}_{t,d_t}, \gamma_{t,d_t}^\intercal\right)^\intercal + Q_\alpha[\tilde{C}_{t,d_t}^k + \tilde{C}_{t,d_t}^u + \tilde{\epsilon}_t(d_t)] \\
=& x^\intercal \beta_{t,d_t} - \tilde{\lambda}_{t,d_t}^k b + \mu_u^\intercal \lambda_{t,d_t}^u + c_{t,d_t} + Q_\alpha[\tilde{C}_{t,d_t}^k + \tilde{C}_{t,d_t}^u + \tilde{\epsilon}_t(d_t)] \\
=& x^\intercal \beta_{t,d_t} - \tilde{\lambda}_{t,d_t}^k b + \mu_u^\intercal \lambda_{t,d_t}^u + c_{t,d_t} + Q_\alpha[C_{t,d_t}^k + \tilde{\lambda}_{t,d_t}^k b + C_{t,d_t}^u - \mu_u^\intercal \lambda_{t,d_t}^u + \epsilon_t(d_t) - c_{t,d_t}] \\
=& x^\intercal \beta_{t,d_t} + Q_\alpha[C_{t,d_t}^k + C_{t,d_t}^u + \epsilon_t(d_t)].
\end{aligned}$$

B.2 Variance decompositions

As discussed in Section 2, an important class of parameters in learning models are terms that decompose the variance of potential outcomes into components that are predictable and unpredictable given the agents' information. These parameters can be expressed as functionals of the finite- and infinite-dimensional components of the model parameters. Section 4 provides general inference results, which can be applied to a plug-in sieve MLE estimator of these parameters. In this section, we define these parameters and discuss their relevance to quantifying the importance of uncertainty and learning.

To define this class of parameters, consider a weighted sum of potential outcomes, $Y(\omega^T, d^T) = \sum_t \omega_t Y_t(d_t)$ for a sequence of choices d^T and weights, ω^T. Cunha and Heckman (2016) consider a special case of this parameter in the context of an educational choice model. In particular, they consider the present value of lifetime earnings, which is defined as $Y(\omega^T, d^T)$, with $\omega_t = \mathbb{1}(t \geq t_0)(1-\rho)^{t_0 - t}$, for some discount rate $0 \leq \rho < 1$.

Next, define the agent's information set as $\mathcal{I}_t = \{Y^{t-1}, D^{t-1}, X^t, X_k^*\}$ for $t > 1$ and $\mathcal{I}_1 = \{X_1, X_k^*\}$. Restricting attention to weighted sums where $\omega_s = 0$ for $s < t$, the variance of $Y(\omega^T, d^T)$ conditional on $\mathcal{I}_t$ can be understood as the variance that is due to the agent's uncertainty over $Y(\omega^T, d^T)$ given their information up to period t. We refer to this as the *posterior variance*, because this is derived from the posterior distribution of X_u^* after performing a Bayesian update with the information in $\mathcal{I}_t$.

In its full generality, the model allows for endogeneity in X_t as the transition probabilities depend on past choices and outcomes. Therefore, the posterior variance of $Y(\omega^T, d^T)$ includes terms which reflects uncertainty about the future realizations of X_t conditional on X_k^*. In order to focus on uncertainty over X^*, we abstract from this by assuming that the covariates are not time varying, which we denote as X.[16]

In particular, with this restriction on the covariates, Lemma 1 implies that the posterior variance, which we denote as $V_t^u(X, D^{t-1}; \omega^T, d^T) \coloneqq \text{Var}\big(Y(\omega^T, d^T) \mid \mathcal{I}_t\big)$,

[16]When the covariates are time varying and transitions depend on (D^{t-t}, Y^{t-1}), the posterior variance will include the covariances between future realizations of X_t and between X_t and X_u^* conditional on the information set. These terms reflect another channel through which unobserved heterogeneity is related to the agents' uncertainty. In this case, the plug-in estimator of the posterior variance will involve other infinite dimension parameters of the model (e.g., $f_{D_t|X^t, Y^{t-1}, D^{t-1}, X_k^*}$).

has the form

$$V_t^u(X, D^{t-1}; \omega^T, d^T) := \sum_{t_1,t_2 \geq t} \omega_{t_1}\omega_{t_2}(\lambda^u_{t_1,d_{t_1}})^\intercal \Sigma_t \lambda^u_{t_2,d_{t_2}} + \sum_{t_1 \geq t} \omega_{t_1}^2 \sigma^2_{t_1,d_{t_1}}$$

for $t > 1$ where Σ_t is the posterior variance of X_u^* as written in Lemma 1.[17] When $t = 1$, D^{t-1} is empty so we write $V_1^u(X; \omega^T, d^T) := \text{Var}\left(Y(\omega^T, d^T) \mid \mathcal{I}_1\right)$.

At $t = 1$, the following variance decomposition provides a natural way to quantify the relative importance of uncertainty in potential outcomes,

$$\text{Var}(Y(\omega^T, d^T) \mid X) = V_1^u(X; \omega^T, d^T) + \sum_{t_1,t_2 \geq 1} \omega_{t_1}\omega_{t_2}\lambda^k_{t_1,d_{t_1}}\lambda^k_{t_2,d_{t_2}}\text{Var}(X_k^* \mid X) \quad (10)$$

This corresponds to the decomposition in Cunha and Heckman (2016) and in that context, 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 privately known heterogeneity.

For $t > 1$, the analysis is more complicated. For any $t > 1$, $V_t^u(X, D^{t-1}; \omega^T, d^T) < V_1^u(X; \omega^T, d^T)$, because the realized outcomes are informative about X_u^*. Agents also select d^{t-1} based on their private information (X_k^*), which induces a selected distribution of X_k^* (i.e., conditional on $(X, Y^{t-1}, D^{t-1}) = (x, y^{t-1}, d^{t-1})$). Given these contributions of learning and selection to variance of $Y(\omega^T, d^T)$, there are several possible ways of quantifying the relative importance of uncertainty. The following are three alternative decompositions, which express total variance (conditional on some subset of observables) as the sum of a term that reflects uncertainty and another

[17] Note that Σ_t depends on certain components of $\mathcal{I}_t$.

reflecting variance induced by private information (X_k^*),

$$\begin{aligned}\mathrm{Var}(Y(\omega^T, d^T) \mid D^{t-1} = d^{t-1}, X = x) = V_t^u(d^{t-1}, x; \omega^T, d^T) \\ + \mathrm{Var}(E(Y(\omega^T, d^T) \mid \mathcal{I}_t) \mid D^{t-1} = d^{t-1}, X = x),\end{aligned} \tag{11}$$

$$\mathrm{Var}(Y(\omega^T, d^T) \mid X = x) = E(V_t^u(D^{t-1}, x; \omega^T, d^T)) + \mathrm{Var}(E(Y(\omega^T, d^T) \mid \mathcal{I}_t) \mid X = x), \tag{12}$$

$$\mathrm{Var}(Y(\omega^T, d^T) \mid X = x) = V_t^u(d^{t-1}, x; \omega^T, d^T) + \tilde{\mathrm{Var}}(\tilde{E}(Y(\omega^T, d^T) \mid \mathcal{I}_t) \mid X = x). \tag{13}$$

Decomposition (11) compares the variance of uncertainty to the total variance conditional on choosing the sequence d^t. These are natural parameters to consider, but the ratio, $V_t^u(d^t, x; \omega^T, d^T)/\mathrm{Var}(Y(\omega^T, d^T) \mid D^t = d^t, X = x)$ reflects both the effect of learning in the numerator and selection in the denominator.

Decomposition (12) compares the total variance $Y(\omega^T, d^T)$ to the expected posterior variance of $Y(\omega^T, d^T)$ after t periods. The expectation of $V^u(D^t, x; \omega^T, d^T)$ can be understood as the uncertainty that a randomly chosen person would have in period t after observing their outcomes and endogenously choosing actions based on that information and their private information.

Finally decomposition (13) is based on a counterfactual distribution. Here $\tilde{E}$ and $\tilde{\mathrm{Var}}$ represent the expectation and variance in a counterfactual distribution where D^t is assigned randomly. This decomposition compares the variance in $Y(\omega^T, d^T)$ which is due to uncertainty vs. known heterogeneity among people randomly assigned to the choice sequence d^t.

B.3 Appendix to estimation section

B.3.1 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 parameter $\theta^* \in \Theta$. 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 $\Lambda^\gamma(\mathcal{X})$ denote a Hölder space equipped with the Hölder norm $\|h\|_{\Lambda^\gamma}$ (that is, for k the largest integer smaller than γ, $\Lambda^\gamma(\mathcal{X})$ is a space of functions $h\colon \mathcal{X} \to \mathbb{R}$ 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 $h_t(d^t, x^t, y^{t-1}, x_k^*)$ depends on (d^t, x^t, y^{t-1}) via some measurable vector-valued function $(d^t, x^t, y^{t-1}) \mapsto j_t$ which is known up to $\big((\beta_s, \lambda_s, \sigma_s)_{s=1}^T, \Sigma_u(x_1)\big)$. 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 X_u^* (i.e, the mean and variance), and a sort of myopia where $j_t \in \mathbb{R}^{3+2}$ depends on the history only via the previous period $(d_{t-1}, x_{t-1}, y_{t-1})$. Write $J_t = (J_{1,t}^\intercal, J_{2,t}^\intercal)^\intercal$ and $X_t = (X_{1,t}^\intercal, X_{2,t}^\intercal)^\intercal$ where $J_{1,t}, X_{1,t}$ are continuous random variables and $J_{2,t}, X_{2,t}$ are random variables with finite support and, with some abuse of notation, redefine the CCP function as $h_t(j_{1,t}, j_{2,t}, x_k^*)$. Define

$$\begin{aligned}
\mathcal{H}_t &= \Lambda_c^{\gamma_1,\omega_1}\left(\mathcal{S}(X_k^*) \times \mathcal{S}(J_{1,t})\right), \\
\mathcal{F} &= \{f\colon \mathcal{S}(X_k^*, X_{1,1}) \to \mathbb{R} \big| f(\cdot, x_1) \text{ is càdlàg }, f(x_k^*, \cdot) \in \Lambda_c^{\gamma_2,\omega_2}(\mathcal{S}(X_{1,1}))\} \\
\mathcal{G}_t &= \Lambda_c^{\gamma_3,\omega_3}\left(\mathcal{S}(X_{1,t+1}) \times \mathcal{S}(Y_t) \times \mathcal{S}(X_{1,t})\right).
\end{aligned}$$

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, $F_{X_{t+1}|Y^t,D^t,X^t} = F_{X_{t+1}|Y_t,D_t,X_t}$, and $F_{X_U^*|X_1} = F_{X_U^*}$.

Define $k_{1,t} = |\mathcal{S}(J_{2,t})|$, $k_2 = |\mathcal{S}(X_{2,1})|$, and $k_{3,t} = |\mathcal{S}((X_{2,t+1}, D_t, X_{2,t}))|$. 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 $\theta = (\theta_1, h_1, \ldots, h_T, f_{X^*}, g_1, \ldots, g_{T-1})$. Define the norms on $\mathcal{H}_t^{k_{1,t}}$, $\mathcal{F}^{k_2}$ and $\mathcal{G}_t^{k_{3,t}}$ as follows:

$$\begin{aligned}
\|h_t\|_{\infty,\omega_1} &= \sup_{j_2 \in \mathcal{S}(J_{2,t})} \|h_t(\cdot, j_2, \cdot)[1 + \|\cdot\|_E^2]^{-\omega_1}\|_\infty, \\
\|f_{X^*}\|_{\infty,\omega_2} &= \sup_{x_2 \in \mathcal{S}(X_{2,1})} \|f_{X^*}(\cdot, (\cdot, x_2))\,[1 + \|\cdot\|_E^2]^{-\omega_2}\|_\infty, \\
\|g_t\|_{\infty,\omega_3} &= \sup_{(x_2', d, x_2) \in \mathcal{S}(X_{2,t+1}, D_t, X_{2,t})} \|g_t\left((\cdot, x_2'); \cdot, d, (\cdot, x_2)\right)[1 + \|\cdot\|_E^2]^{-\omega_3}\|_\infty,
\end{aligned}$$

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

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

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 $\Theta_n = \Theta_1 \times \mathcal{H}_{n,1}^{k_{1,1}} \times \ldots \mathcal{H}_{n,T}^{k_{1,T}} \times \mathcal{F}_n^{k_2} \times \mathcal{G}_{n,1}^{k_{3,1}} \times \cdots \times \mathcal{G}_{n,T-1}^{k_{3,T-1}}$ 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. $\mathrm{E}[\ell(W, \theta)]$ is continuous at $\theta = \theta^*$

Assumption E5.

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

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

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

The identification assumptions imply $\theta^* = \arg\max_{\theta \in \Theta} \mathrm{E}[\ell(W, \theta)]$ and for all $\theta \in \Theta \backslash \{\theta^*\}$, $\mathrm{E}[\ell(W, \theta^*)] \geq \mathrm{E}[\ell(W, \theta)]$. By assuming compactness of Θ, we ensure that θ^* is a well-separated maximum of $\mathrm{E}[\ell(W, \theta)]$. 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 similar to Condition 3.5M in Chen (2007).

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

B.3.2 Plug-in sieve estimator

We first assume a linear sieve space and limit its complexity.

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

Assumption E6 controls the rate at which the number of sieve terms grow. To achieve this, part (i) of Assumption E6 requires that the nonparametric functions have adequate smoothness. In applied work, one may focus on discrete X_t and posit a parametric model for h_t, in which case the above restrictions are milder.

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

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

$$\langle \theta_1 - \theta^*, \theta_2 - \theta^* \rangle = -\frac{\partial^2}{\partial \tau_1 \partial \tau_2} E\left[\ell\left(W, \theta^* + \tau_1\left(\theta_1 - \theta^*\right) + \tau_2\left(\theta_2 - \theta^*\right)\right)\right]|_{\tau_1=0,\tau_2=0}, \tag{14}$$

for $\theta_1, \theta_2 \in \Theta$. the corresponding norm for $\theta \in \Theta$ is

$$\left\|\theta - \theta^*\right\|^2 := -\left.\frac{\partial^2}{\partial \tau^2} E\left[\ell\left(W, \theta^* + \tau\left(\theta - \theta^*\right)\right)\right]\right|_{\tau=0}. \tag{15}$$

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

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

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

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

with $E\left(\left[U\left(W\right)\right]^\gamma\right) \leqslant C_2$ 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 $(Y_{i,t}, D_{i,t}, X_{i,t} \colon t = 1, \ldots, T)_{i=1}^n$ 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

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

and $\mathcal{N}_n := \mathcal{N}_0 \cap \Theta_n$. 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 $(\theta - \theta^*)$:

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

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 $0 \in \mathbb{R}$. (i) 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 $\tau \in \mathcal{T}$; (iii) 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 $\langle \cdot, \cdot \rangle$.[18] 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. (i) $v \mapsto \frac{\partial f(\theta^*)}{\partial \theta}[v]$ is a linear functional. (ii) 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 $\left|\frac{\partial f(\theta^*)}{\partial \theta}[\theta_n^* - \theta^*]\right| / \|v_n^*\| = o(n^{-1/2})$. (iii) $\sup_{\theta \in \mathcal{N}_0} \frac{\left|f(\theta) - f(\theta^*) - \frac{\partial f(\theta^*)}{\partial \theta}[\theta - \theta^*]\right|}{\|v_n^*\|} = o(n^{-1/2})$.

Assumption E11 imposes some restrictions on the functional of interest $\theta \mapsto f(\theta)$. Part (i) imposes that the directional derivative is a linear functional, a mild condition that is satisfied by our examples in Section 4. Part (ii) is a restriction on the growth rate of the dimension of the sieve space. Part (iii) 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 $u_n^* := \frac{v_n^*}{\|v_n^*\|}$, $\varepsilon_n = o\left(n^{-1/2}\right)$ and $\mu_n\{g(\boldsymbol{W})\} := n^{-1} \sum_{i=1}^n \left[g\left(W_i\right) - \mathrm{E}[g\left(W_i\right)]\right]$ denote the centered empirical process indexed by the function g.

[18] See Chen et al. (2014, p. 642).

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

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

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

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

B.4 Appendix to implementation and Monte Carlo simulations section

B.4.1 Implicit differentiation

For implementing the estimator, it can be useful to input the gradient of the objective function. In this section, we show how our profiling approach and choice of sieve space simplify this task. Recall that in Section 5.1, the profile log likelihood function with our proposed sieve space for $F_{X_k^*}$ is

$$\ell^p(\theta^c) := \sum_{i=1}^{n} \log \sum_{s=1}^{q_n} \omega_s(\theta^c)\ \ell^c(w_i, \bar{x}_{n,s}^*; \theta^c),$$

where $\omega(\theta^c) = \arg\max_{\omega\in\Delta(q_n)} \sum_{i=1}^{n} \log \sum_{s=1}^{q_n} \omega_s\ \ell^c(w_i, \bar{x}_{n,s}^*; \theta^c)$ is the solution to the inner problem for a fixed θ^c. Given an analytical expression for $\ell^c(w_i, x_k^*; \theta^c)$[19], the challenge of computing the gradient of $\ell^p(\theta^c)$ reduces to finding the Jacobian of $\omega(\theta^c)$ (i.e., $\frac{\partial}{\partial(\theta^c)^\intercal}\omega(\theta^c)$), which is defined implicitly by the Karush-Kuhn-Tucker (KKT)

[19]Given the analytical expression for ℓ^c, we use the software Google JAX to compute the derivative via autodifferention.

conditions of the inner optimization problem. In the following, we derive an analytical expression for $\frac{\partial}{\partial(\theta^c)^\intercal}\omega(\theta^c)$ in terms of $\ell^c(w_i, x_k^*; \theta^c)$, $\frac{\partial}{\partial\theta^c}\ell^c(w_i, x_k^*; \theta^c)$, and $\omega(\theta^c)$.

Proposition 3.3 in Kim et al. (2020) shows that $\omega(\theta^c)$ can be equivalently expressed as $\arg\max_{\omega\geq 0}\{\sum_{i=1}^n \log \sum_{s=1}^{q_n} \omega_s\, \ell^c(w_i, \bar{x}_{n,s}^*; \theta^c) + \sum_{s=1}^{q_n} \omega_s\}$, where $\omega \geq 0$ means $\omega_s \geq 0$ for all $s = 1, \ldots, q_n$. Letting $\lambda \in \mathbb{R}^{q_n}$ be the dual parameter corresponding to the constraint $\omega \geq 0$, and $\ell_i^c(\theta^c) \coloneqq (\ell^c(w_i, \bar{x}_{n,s}^*; \theta^c) \colon s = 1, \ldots, q_n)$, the equality constraints in the KKT conditions of this problem are,

$$0_{2q_n\times 1} = \begin{pmatrix} \sum_{i=1}^n \frac{1}{\omega^\intercal \ell_i^c(\theta^c)}\ell_i^c(\theta^c) + 1_{q_n} + \lambda \\ \lambda \circ \omega \end{pmatrix},$$

where $\circ$ is the Hadamard product. By definition, these constraints are identically zero for all θ^c, so under an implicit function theorem, $\frac{d}{d(\theta^c)^\intercal}\omega(\theta^c) = -G_1(\theta^c)^{-1}G_2(\theta^c)$,[20] where

$$G_1(\theta^c) = \begin{pmatrix} \sum_{i=1}^n \frac{1}{(\omega(\theta^c)^\intercal \ell_i^c(\theta^c))^2}\ell_i^c(\theta^c)(\ell_i^c(\theta^c)^\intercal & I_{q_n\times q_n} \\ \operatorname{diag}(\lambda(\theta^c)) & \operatorname{diag}(\omega(\theta^c)) \end{pmatrix},$$

and

$$G_2(\theta^c) = \begin{pmatrix} \sum_{i=1}^n \left(\frac{\frac{\partial}{\partial(\theta^c)^\intercal}\ell_i^c(\theta^c)}{\omega(\theta^c)^\intercal \ell_i^c(\theta^c)} - \frac{\ell_i^c(\theta^c)\omega(\theta^c)^\intercal \frac{\partial}{\partial(\theta^c)^\intercal}\ell_i^c(\theta^c)}{(\omega(\theta^c)^\intercal \ell_i^c(\theta^c))^2} \right) \\ 0_{q_n\times \dim(\theta^c)} \end{pmatrix}$$

Finally, note that the KKT conditions imply that $\lambda(\theta^c) = -1_{q_n} - \sum_{i=1}^n \frac{\ell_i^c(\theta^c)}{\omega(\theta^c)^\intercal \ell_i^c(\theta^c)}$.

B.4.2 Details on DGP

This section gives further details on the DGP used for Monte Carlo simulations discussed in Section 5.2. The values of the finite parameters used in the DGP are given in the table below.

[20] G_1 and G_2 are the partial derivatives of right hand side of the previous equation with respect to (ω, λ) and θ^c respectively, evaluated at $\omega(\theta^c)$ and $\lambda(\theta^c)$.

$\alpha_{1,1} = 0$	$\gamma^{(1)}_{1,1} = -0.5$	$\gamma^{(2)}_{1,1} = -0.58$	$\lambda^{u}_{1,1} = 1$	$\lambda^{k}_{1,1} = 0.3$
$\alpha_{2,1} = 0.1$	$\gamma^{(1)}_{2,1} = -0.8$	$\gamma^{(2)}_{2,1} = -0.83$	$\lambda^{u}_{2,1} = 1.05$	$\lambda^{k}_{2,1} = 0.35$
$\alpha_{3,1} = 0.2$	$\gamma^{(1)}_{3,1} = 0.12$	$\gamma^{(2)}_{3,1} = -0.83$	$\lambda^{u}_{3,1} = 1.01$	$\lambda^{k}_{3,1} = 0.33$
$\sigma^2_1 = 0.5$				
$\alpha_{1,2} = -0.1$	$\gamma^{(1)}_{1,2} = 0.13$	$\gamma^{(2)}_{1,2} = 0.71$	$\lambda^{u}_{1,2} = 0.4$	$\lambda^{k}_{1,2} = 1$
$\alpha_{2,2} = -0.22$	$\gamma^{(1)}_{2,2} = 0.89$	$\gamma^{(2)}_{2,2} = -0.36$	$\lambda^{u}_{2,2} = 0.36$	$\lambda^{k}_{2,2} = 1.05$
$\alpha_{3,2} = -0.33$	$\gamma^{(1)}_{3,2} = 0.32$	$\gamma^{(2)}_{3,2} = -0.36$	$\lambda^{u}_{3,2} = 0.44$	$\lambda^{k}_{3,2} = 1.02$
$\sigma^2_2 = 0.7$				
$\sigma^2_u = 1.5$	$\rho = 2.0$	$\kappa = 0.5$		

Table 4: Finite parameter values

B.4.3 DGP with risk aversion

In this section, we present results from an alternative DGP in which agents maximize their expected utility in each period which incorporates risk aversion, through constant relative risk aversion (CRRA) preferences, and subjective (possibly biased) beliefs. The expected utility that individual i derives from choice d in period t is given by:

$$v_{i,t}(d) := \mathcal{E}_{i,t}\left(\frac{Y_{i,t}(d)^{1-\chi}}{1-\chi}\right) + \eta_{i,t}(d)$$

where $\mathcal{E}_{i,t}$ denotes the expectation under individual i's subjective beliefs over $X^*_{u,i}$, given the information up to period t. $\eta_{i,t}(d)$ are independent preference shocks, which are supposed to follow an Extreme Value Type 1 distribution.

We assume that individuals' subjective beliefs over $X^*_{u,i}$ in time t are distributed $N(\mu_{i,t} + \delta X^*_{k,i}, \Sigma_{i,t})$ where $\mu_{i,t}, \Sigma_{i,t}$ are the correct posterior mean and variance of $X^*_{u,i}$ 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, $X^*_{k,i}$.

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

$$v_{i,t}(d) = \frac{\exp\left(\mu_{i,t}(d)(1-\chi) + \frac{1}{2}\sigma_{i,t}(d)(1-\chi)^2\right)}{1-\chi} + \eta_{i,t}(d) \tag{16}$$

where $\mu_{i,t}(d)$ ($\sigma_{i,t}(d)$) denote the subjective mean (variance) of $\log(Y_{i,t}(d))$.

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

$$\mathcal{P}_q(X_k^*) \otimes \mathcal{P}_q(X_1) \otimes \cdots \otimes \mathcal{P}_q(Y_1) \otimes \mathcal{P}_q(\delta_1) \otimes \cdots \otimes \mathcal{P}_q(Y_{t-1}) \otimes \mathcal{P}_q(\delta_{t-1}).$$

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.

The alternative approach we consider here is to use the following approximation

$$v_{i,t}(d) = \varphi\left(\sum_{h \in \mathcal{D}^{t-1}} 1(D^{t-1} = h)(\pi_{t,h,d,0} + \pi_{t,h,d,1}^\intercal X + \pi_{t,h,d,2} X_k^* + \pi_{t,h,d,3}^\intercal Y_i^{t-1})\right)$$

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 $n = 4,000$. Our choice of approximation is motivated by the fact that under Lemma 1 and Equation 16, there is a set of π parameters such that this equality holds, with $\varphi(\cdot) = \frac{1}{1-\chi}\exp(\cdot)$.

The finite parameters are the same as in our baseline simulations considered in Section 5.2, with the added risk aversion parameter χ, which we set to 1.5. X^* and X are generated from the same distributions as in the DGP considered in Section 5.2.

With the additional π parameters to estimate, the θ^c 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 $n = 500$. The large number of parameters to estimate in θ^c results in longer but still manageable computational times, which are reported in Table 5.

	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 4{,}000$
Time (minutes)	3	7.5	19.5	56

Table 5: Time to compute the estimator: DGP with risk aversion. Computational times were obtained using an Intel Core i9-12900K CPU, and are computed as the average over 200 simulations.

The results of the Monte Carlo simulations are presented in Table 6 and Figure 2. Despite the increased complexity of the model, our estimation procedure exhibits similar finite sample performance to the DGP considered in Section 5.2.

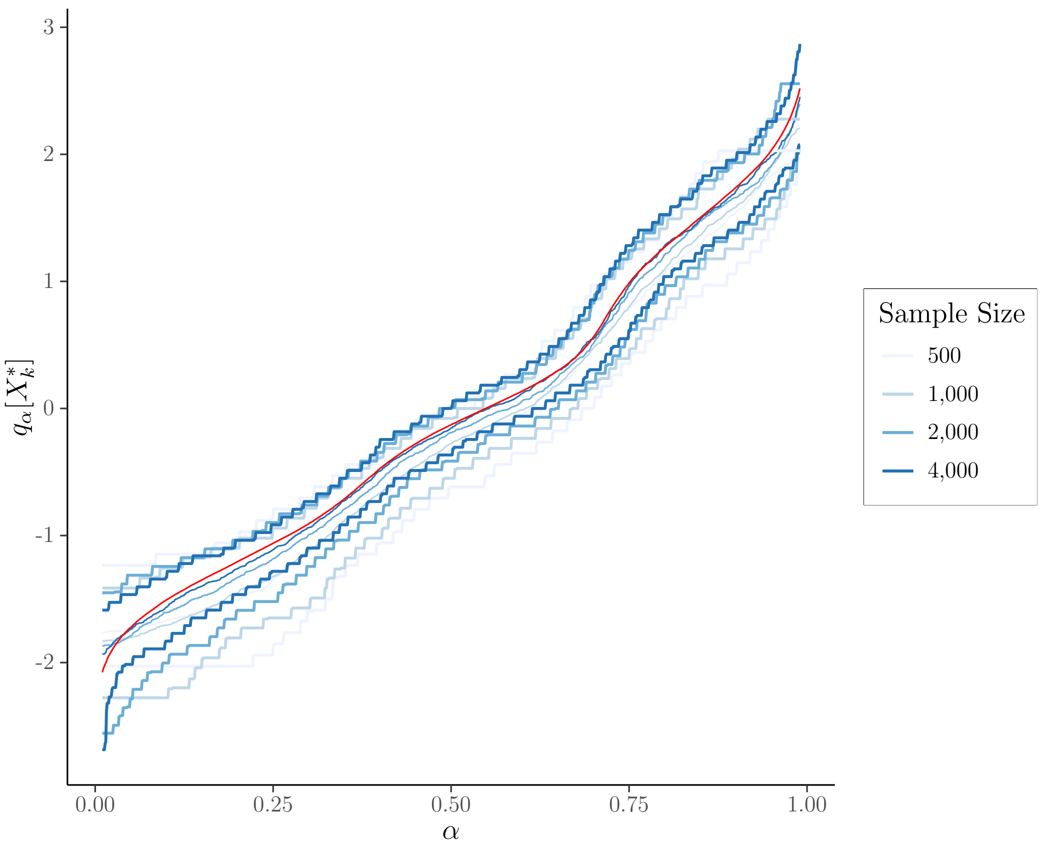

Figure 2: Quantiles of Estimator of $q_\alpha[X_k^*]$ under DGP with risk aversion. The red line shows the true distribution of X_k^*. The blue lines show the mean, and the 5th and 95th percentiles of the simulated distribution of the estimator of $q_\alpha[X_k^*]$ for each sample size.

	n = 500		n = 1,000		n = 2,000		n = 4,000	
	Bias^2	Var	Bias^2	Var	Bias^2	Var	Bias^2	Var
$\alpha_{1,2}$	66.15	38.25	18.40	20.20	3.97	12.19	0.05	7.69
$\alpha_{2,1}$	0.17	28.07	0.05	12.99	0.08	5.50	0.05	2.10
$\alpha_{2,2}$	69.24	42.16	18.40	23.49	3.25	14.33	0.00	9.20
$\alpha_{3,1}$	1.29	24.63	0.07	9.98	0.00	4.73	0.00	1.83
$\alpha_{3,2}$	68.62	42.86	23.69	21.80	3.41	13.62	0.01	8.28
$\gamma_{1,1}^{(1)}$	0.08	6.61	0.05	3.30	0.01	1.72	0.02	0.95
$\gamma_{1,2}^{(1)}$	0.12	8.29	0.09	3.55	0.02	1.64	0.01	0.78
$\gamma_{2,1}^{(1)}$	0.03	7.69	0.08	3.81	0.04	2.11	0.02	1.08
$\gamma_{2,2}^{(1)}$	0.21	9.49	0.25	4.13	0.06	2.18	0.03	0.79
$\gamma_{3,1}^{(1)}$	0.14	5.52	0.03	2.52	0.01	1.38	0.02	0.72
$\gamma_{3,2}^{(1)}$	0.08	9.43	0.11	4.03	0.03	1.84	0.02	0.83
$\gamma_{1,1}^{(2)}$	1.65	35.50	0.00	12.36	0.22	5.58	0.01	2.75
$\gamma_{1,2}^{(2)}$	0.09	28.70	0.09	11.52	0.16	6.99	0.06	3.19
$\gamma_{2,1}^{(2)}$	1.47	31.77	0.00	12.37	0.06	5.50	0.03	2.79
$\gamma_{2,2}^{(2)}$	0.08	28.45	0.11	13.67	0.23	7.50	0.11	3.25
$\gamma_{3,1}^{(2)}$	0.73	25.40	0.02	11.07	0.13	4.71	0.01	2.65
$\gamma_{3,2}^{(2)}$	0.17	29.53	0.00	14.60	0.16	7.89	0.09	3.35
$\lambda_{1,1}^{k}$	0.34	20.38	1.18	6.84	0.02	4.11	0.01	1.71
$\lambda_{2,1}^{k}$	0.18	21.01	2.41	9.54	0.42	5.21	0.09	1.91
$\lambda_{2,2}^{k}$	0.18	9.49	0.00	3.31	0.01	1.60	0.01	0.80
$\lambda_{3,1}^{k}$	0.45	17.32	1.53	8.13	0.15	4.25	0.01	1.53
$\lambda_{3,2}^{k}$	0.03	10.43	0.21	3.97	0.01	2.22	0.01	1.10
$\lambda_{1,2}^{u}$	0.11	6.31	0.03	2.65	0.00	1.23	0.00	0.52
$\lambda_{2,1}^{u}$	0.05	3.54	0.04	1.41	0.01	0.78	0.01	0.43
$\lambda_{2,2}^{u}$	0.09	8.36	0.01	3.61	0.00	1.65	0.01	0.69
$\lambda_{3,1}^{u}$	0.06	3.89	0.02	1.44	0.01	0.60	0.00	0.33
$\lambda_{3,2}^{u}$	0.35	9.16	0.15	4.34	0.00	1.90	0.01	0.87
$\sigma^2(1)$	0.15	0.68	0.01	0.36	0.01	0.17	0.00	0.07
$\sigma^2(2)$	0.06	0.24	0.00	0.15	0.00	0.07	0.00	0.03
σ_u^2	1.38	19.53	0.02	6.64	0.01	3.74	0.00	1.83

Table 6: Simulation results for estimation of finite dimensional parameters. 'Bias^2' and 'Var' refer to the average empirical squared bias and variance scaled by 1,000, respectively, computed over 200 simulations.