Heterogeneity, Uncertainty and Learning: Semiparametric Identification and Estimation*

Jackson Bunting † Paul Diegert‡ Arnaud Maurel§

November 16, 2022

Abstract

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

*Preliminary version. We thank seminar participants at CREST-PSE, TSE and UC Davis as well as Xavier D'Haultfoeuille for useful comments.

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

‡Duke University, paul.diegert@duke.edu.

§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 widely used in economics. These models have received particular interest in various subfields in empirical microeconomics, including industrial organization and health (Coscelli and Shum, 2004; Crawford and Shum, 2005; Aguirregabiria and Jeon, 2020) as well as in labor economics and economics of education (Miller, 1984; Stange, 2012; Arcidiacono et al., 2016). Since the seminal work of Erdem and Keane (1996), learning models have also been popular in the marketing literature (see Ching et al., 2013, for a survey). However, while learning models are often structurally estimated, relatively little is known about the identification of this important class of models.

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

$$Y_{it}(d) = \alpha_t(d) + Z_{it}^{\intercal}\beta_t(d) + \lambda_{k,i}F_{kt}(d) + \lambda_{u,i}^{\intercal}F_{ut}(d) + \epsilon_{it}(d), \tag{1}$$

where Z_{it} is a vector of explanatory variables, $\theta := (\alpha(d), \beta_t(d), F_{kt}(d), F_{ut}(d))$ are unknown parameters, $\lambda_i = (\lambda_{k,i}, \lambda_{u,i}^{\intercal})^{\intercal}$ denotes a vector of latent individual effects, and $\epsilon_{it}(d)$ is an idiosyncratic shock. While these types of interactive fixed effects models have been the object of much interest in the econometrics literature, a key distinctive feature of our setup is the existence of two different types of individual effects, namely $\lambda_{k,i}$, which are supposed to be initially known by the agent, and $\lambda_{u,i}$, which are initially unknown, but may be learned over time. We complement this outcome model with a very flexible choice model, in which agent i's assignment in period t can depend arbitrarily on contemporaneous and lagged explanatory variables, assignments and

outcomes. The choice model encompasses most of the decision models that have been estimated in the learning literature.

We first show that the model is point-identified under two alternative sets of conditions. Our first and main identification result applies to a version of the learning where, consistent with most of the empirical Bayesian learning models, we assume that the idiosyncratic shocks from the outcome equations ($\epsilon_{it}(d)$), as well as the unknown heterogeneity component ($\lambda_{u,i}$), are normally distributed. The distribution of the known heterogeneity component ($\lambda_{k,i}$) is left unspecified. We then also show that a pure learning model with only one type of permanent unobserved heterogeneity ($\lambda_{u,i}$) actually remains point-identified without making any distributional assumption.

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

Our paper fits into a growing literature that examines the identification of dynamic discrete choice models in the presence of unobserved heterogeneity (Heckman and Navarro, 2007; Kasahara and Shimotsu, 2009; Arcidiacono and Miller, 2011; Hu and Shum, 2012; Sasaki, 2015; Bunting, 2022). Unlike these papers, we focus on a learning framework in which a portion of the permanent individual unobserved heterogeneity is initially unknown to the agents, so that decisions may depend on the unknown component ($\lambda_{u,i}$) only through the sequence of past outcomes. This asymmetry property plays an important role in our ability to address the deconvolution problem associated with the coexistence of both types of unobserved heterogeneity.

Also relevant for us is recent work by Pastorino (2022) and Gong (2019) which both consider the identification of specific types of learning models. Beyond the fact that Pastorino restricts her analysis to the specific context of workers' and firms' learning, there are two important differences relative to our paper. First, ability in that paper is restricted to be discrete, whereas we allow for both continuous and multivariate abilities. Second, and importantly, the outcomes which forms the basis of learning is assumed to depend on the learned ability only. In our setting, these outcomes may depend on both known and learned abilities. Our framework also differs from Gong (2019) in important ways. Notably, while we remain agnostic about how choices depend on agents' beliefs about the distribution of λ_u, Gong assumes that assignment depends on prior ability mean only. Gong further imposes significant restrictions on the updating rule, while we remain agnostic about how agents update their beliefs about λ_u.

As the outcome equation in our model involves interactions between unobserved individual- and time-specific effects, our paper also fits into the literature that deals with the identification and estimation of panel data models with interactive fixed effects (Bai, 2009; Gobillon and Magnac, 2016; Freyberger, 2018). Our analysis is most closely related to Freyberger (2018). A fundamental distinction though comes from the fact that Freyberger considers a selection-free environment. On the other hand, choices, along with the underlying selection issues, play a central role in our analysis.

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

Notation: Supp(A) *indicates the support of random variable* A. F_A *indicates the distribution function of random variable* A. *For any sequence* $(a_1, a_2, \ldots, a_S)$ *and* $s \leq$

S, we let $a^s = (a_1, a_2, \ldots, a_s)$. Upper case letters represent random variables, lower case represent realized values. $A \perp\!\!\!\perp B \mid C$ indicates that A and B are statistically independent conditional upon C.

2 Set-up

In this paper we consider a model where potential outcomes have an interactive fixed effect structure:

$$Y_{it}(d) = \alpha_t(d) + Z_{it}^{\intercal}\beta_t(d) + \lambda_{k,i}F_{kt}(d) + \lambda_{u,i}^{\intercal}F_{ut}(d) + \epsilon_{it}(d), \tag{2}$$

where d represents individual i's realized assignment in period t, $Y_{it}(d)$ is a scalar outcome variable, Z_{it} a vector of explanatory variables, $(\alpha(d), \beta_t(d), F_{kt}(d), F_{ut}(d))$ are unknown structural parameters, $\lambda_i = (\lambda_{k,i}, \lambda_{u,i}^{\intercal})^{\intercal}$ is the latent individual effect and $\epsilon_{it}(d)$ is an unobserved random variable. For example $Y_{it}(d)$ may represent wages in occupation d, which depend on multiple dimensions of unobserved abilities λ_i, which might differ in importance across occupations.

Importantly, we allow for two types of latent individual effects: $\lambda_{k,i}$ that is known by the agent, and $\lambda_{u,i}$ that is initially unknown but (possibly) learned over time. For example, a worker i's log-wage in occupation d at time t, $Y_{it}(d)$, may depend on their occupation specific productivity $\lambda_{k,i}F_{kt}(d) + \lambda_{u,i}^{\intercal}F_{ut}(d)$, part of which the worker becomes more certain of as they accumulate more experience.

The only restriction placed on an individual's assignment in period t is that it does not directly depend on the unknown component of latent heterogeneity. Specifically we assume that

$$D_{it} \perp\!\!\!\perp \lambda_{u,i} \mid Z_i^t, Y_i^{t-1}, D_i^{t-1}, \lambda_{k,i}. \tag{3}$$

The above conditional independence condition highlights the asymmetry between the two types of latent effect: assignment may directly depend on the known component of the latent effect $\lambda_{k,i}$, but not on the unknown component of the latent effect $\lambda_{u,i}$.

By allowing the assignment rule to depend arbitrarily on lagged variables, we remain agnostic about how assignments depend on agents' beliefs over $\lambda_{u,i}$, as well as about how agents form their beliefs about $\lambda_{u,i}$. For example, D_{it} may represent the outcome of a worker-firm matching process, which depends on the common prior probability distribution over the worker's latent abilities (Pastorino, 2022). To take another example, D_{it} may be the college major choice, which depends on the student's belief about their abilities in various fields of study (Arcidiacono et al., 2016). In one extreme, the beliefs of a rational Bayesian agent coincide with the objective distribution of $\lambda_{u,i}$ conditional upon their information set at time t, which may include all realized variables and model parameters. In the other extreme, an agent's assignment may not depend on beliefs over the distribution of $\lambda_{u,i}$. In addition to these two extremes, the flexible assignment rule allow for myopic or uninformative beliefs. Importantly, agent beliefs may be heterogeneous — for example, some agents may be rational Bayesian updaters, while others may be myopic.

Given the assignment rule, we define the conditional choice probability (CCP) function as

$$\bar{h}_t(d^t, z^t, y^{t-1}, v_k) \equiv \Pr(D_{it} = d \mid Z_i^t = z^t, Y_i^{t-1} = y^{t-1}, D_i^{t-1} = d^{t-1}, \lambda_{k,i} = v_k).$$

These CCPs play a central role in our identification analysis. In applications it is common to impose structure on the assignment rule. For example, in a model of school choice and labor supply, Arcidiacono et al. (2016) assume that

$$D_{it} = \arg \max_{\tilde{d} \in \text{Supp}(D_t)} \left\{ v_t(\tilde{d}, Z_{it}, \lambda_{k,i}, X_{it}) + \eta_{it}(\tilde{d}) \right\},$$

where v_t is known up to a finite dimensional parameter, X_{it} are sufficient statistics for the conditional distribution of $\lambda_{u,i}$ at time t, and η_{it} follows a known distribution. In what follows, we focus on point identification of the (latent) CCP functions $\bar{h} = (\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_T)$ and the outcome equation parameters, from which standard arguments can be applied to identify the primitives (see, e.g., Magnac and Thesmar, 2002).

3 Identification results

This section considers identification of the model of Section 2. We provide two sets of sufficient conditions for point identification. One set of conditions (Theorem 1) assume that ϵ_{it} belongs to the Gaussian family. The second set of conditions (Theorem 2) does not require parametric assumptions on the distribution of ϵ_{it}, but does assume all components of the individual effect are initially unknown by the agent.

3.1 Known and unknown heterogeneity

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

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

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

Assumption KL2 imposes that the learned individual effect is drawn from a multivariate normal distribution, and that the random shock in the outcome equation, which forms the basis to update beliefs over $\lambda_{u,i}$, is also normal.

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

This assumption leads to a specific functional form for the posterior distribution, namely the Gaussian conjugate distribution. We summarize this in Lemma 1.

Lemma 1. Define (E_t, Σ_t) recursively as follows. First, $(E_1, \Sigma_1) = (0, \Sigma_u(Z_1, \lambda_k))$. Second,

$$\begin{aligned}
\Sigma_{t+1} &= \left(\Sigma_t^{-1} + F_{ut}(D_t)F_{ut}(D_t)^\intercal \sigma_t^{-2}(D_t)\right)^{-1} \\
E_{t+1} &= \Sigma_{t+1}\left(\Sigma_t^{-1}E_t + F_{ut}(D_t)\frac{Y_{it} - \alpha_t(D_t) - Z_{it}^\intercal\beta_t(D_t) - \lambda_{k,i}F_{kt}(D_t)}{\sigma_t^2(D_t)}\right).
\end{aligned}$$

Then, under Assumption KL2, $\lambda_u \mid (D^{t-1}, Y^{t-1}, Z^t, \lambda_k) \sim N(E_t, \Sigma_t)$.

Since λ_u conditional upon $(Y^{t-1}, D^{t-1}, Z^t, \lambda_k)$ is normal with mean E_t and variance-covariance matrix Σ_t, it follows that (E_t, Σ_t) are sufficient statistics for $\lambda_{u,i}$ at time t. Notice (E_t, Σ_t) are a deterministic function of $(D^{t-1}, Y^{t-1}, Z^t, \lambda_k)$ and $\theta_1 = \left((\alpha_t, \beta_t, F_{kt}, F_{ut}, \sigma_t)_{t=1}^T, \Sigma_u\right) \in \Theta_1$. Furthermore, we can express (E_t, Σ_t) non-recursively as:

$$\begin{aligned}
\Sigma_{t+1} =& \left(\Sigma_u^{-1}(Z_1, \lambda_k) + \sum_{s=1}^{t} F_{us}(D_s) F_{us}(D_s)^\intercal \sigma_s^{-2}(D_s)\right)^{-1} \\
E_{t+1} =& \Sigma_{t+1} \left(\sum_{s=1}^{t} F_{us} \frac{Y_{is} - \alpha_s(D_s) - Z_{is}^\intercal \beta_s(D_s) - \lambda_{k,i} F_{ks}(D_s)}{\sigma_s^2(D_s)}\right)
\end{aligned}$$

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

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

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

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

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

(F) Any $p \times p$ submatrix of $(F_{u1}(d_1) F_{u2}(d_2) \ldots F_{uT}(d_T))$ has full rank.

Assumption KL3 is a normalization on the finite dimensional parameters. This type of assumption is standard in interactive fixed effect models (Freyberger, 2018), since no scale assumption is placed on the distribution of the unknown fixed effects. For example, it may be possible to replace Assumption KL3(A) by a zero mean assumption on the latent individual effect λ_i. Assumption KL3(B) is required since the latent effect is inherently scale free: multiplying the latent effect by a scalar and dividing the coefficient by the same scalar are observationally equivalent.

Assumption KL4 places support restrictions on various objects of the model. Part (A) states the finite dimensional parameters belong to a compact set. Part (B) imposes that $\lambda_{k,i}$ has compact support. This would be satisfied if the distribution of $\lambda_{k,i}$ has discrete support, although this applies, of course, to a broader set of distributions. Part (C) requires that the distribution of $(Y_{it}(d)|Z_{it}, \lambda_{k,i}, D_t = d)$ is non-degenerate. Part (D) is a 'rectangular' support assumption on $\lambda_{k,i}$. It states that given each history (Y^{t-1}, D^{t-1}, Z^t) there are some v_k in the support of $\lambda_{k,i}$ that are assigned to $D_t = d_t$. This will be satisfied by any standard dynamic discrete choice model, due to the large support assumption on the random utility shocks. Finally, part (E) imposes sufficient variation in Z_t.

Assumption KL5 is a regularity condition that ensures that the latent individual effect λ_i alters outcomes sufficiently differently across time and assignments. In broad terms, it rules out "knife-edge" cases where the cumulative effect of different elements of the individual effect perfectly offset each other. This type of assumption is similarly required in latent factor models without selection or learning (Freyberger, 2018, Assumption L4) to rule out degeneracies. In this sense, Assumption KL5 can be viewed as a generalization of a standard assumption in linear factor models to models with selection and learning. Part (A) requires that the aggregate effect of $\lambda_{k,i}$ on outcomes for choice d_t is different for at least two histories $(d^{t-1}, \tilde{d}^{t-1})$. Part (B) assumes that the direct effect of $\lambda_{k,i}$ is non-zero in each period for each assignment. Part (C) states

the aggregate effect of $\lambda_{k,i}$ on outcomes must be non-zero—that is, that the direct effect $F_{kt}(d_t)$ is not perfectly offset by the effect mediated through previous choices. Part (D) ensures that there is a non-zero effect of previous choices in $t = 2$. Finally, Part (E) requires that in $t = 2$ the relative effect of known and unknown λ_i changes across choices. In the case that $\lambda_{u,i} \in \mathbb{R}$, it reduces to $\frac{F_{k2}(d_2)}{F_{u2}(d_2)} \neq \frac{F_{k2}(\tilde{d}_2)}{F_{u2}(\tilde{d}_2)}$—that the ratio of factor loadings is non-constant across assignments. Notice that it implies that, at least in $t = 2$, the set of assignments must contain at least $p+1$ elements for $\lambda_{u,i} \in \mathbb{R}^p$.

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

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

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

suitability weighted by the assignment probabilities, is a normal-weighted mixture of normals. This observation leads to identification $\left((F_{ut}, \sigma_t)_{t=1}^T, \Sigma_u\right)$ from the second moments of the reweighted distribution.

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 $\lambda_{k,i}$ has finite support with known cardinality. Assumption KL4(B) relaxes this assumption in the sense that the number of support points of $\lambda_{k,i}$ need not be known a priori.

Remark 2 (Normality of unknown factor). As summarized in Lemma 1, an important advantage of the normality assumptions (Assumption KL2) is the resulting conjugate prior with a tractable closed form. For this reason, these assumptions are very common in the applied literature. In our identification result, the most important implication of these assumptions is to enable identification of the (latent) distribution of $Y_{it} \mid \left(\lambda_k, Y_i^{t-1}, D_i^t, Z_i^t\right)$ from variation in Y_{it} only. First, the normality assumptions on ϵ_t and λ_u lead to normality of $Y_{it} \mid \left(\lambda_k, Y_i^{t-1}, D_i^t, Z_i^t\right)$ by standard Bayesian arguments. For each fixed $(Y_i^{t-1}, D_i^t, Z_i^t) = (y^{t-1}, d^t, z^t)$ this is a mixture of normal distributions weighted by the (continuous) distribution of λ_k conditional upon $(Y_i^{t-1}, D_i^t, Z_i^t)$. Then classical continuous mixture of normals arguments yield identification.

This discussion also highlights why we restrict $\lambda_{k,i}$ to be a scalar random variable. Namely, that identification of its distribution is coming from variation in the scalar outcome variable Y_{it}. If a vector of outcomes were available—that is, if Y_{it} was vector-valued—then our arguments would easily extend to multivariate $\lambda_{k,i}$.

Remark 3 (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. In particular, we can show that average and quantile structural functions are identified without the normalization assumption. To formalize this notion, define $C_{kt}(d) \equiv \lambda_k' F_{kt}(d)$, $C_{ut}(d) \equiv \lambda_u' F_{ut}(d)$ and let $Q_\alpha[X]$ be

the α-quantile of the random variable X. Let $z \in \text{Supp}(Z_{it})$ and define the quantile structural functions

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

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

In Appendix A.1 we prove the following corollary:

Corollary 1. Suppose the Assumptions KL1, KL4 and KL5 and that $(\lambda_u \mid Z_1 = z_1, \lambda_k = v_k) \sim N(\mu_u, \Sigma_u(z_1, v_k))$ and $\epsilon_t(d) \sim N(c_t(d), \sigma_t(d)^2)$. Furthermore, suppose for some $(d_1, d_2, \ldots, d_p)$, $F_{k1}(d_1) \neq 0$ and $F_p = (F_{u1}(d_1)F_{u2}(d_2)\ldots F_{up}(d_p))$ is full rank. Then $s_{1,t}$, $s_{2,t}$ and $s_{3,t}$ are identified on the support of Z_t.

3.2 Pure learning model

This section considers a special case of the model of Section 2, in which all components of the latent individual effect are initially unknown to the decision making agent. That is, $\lambda_i = \lambda_{u,i}$. 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.

Suppose $\lambda \in \mathbb{R}^p$. The required assumptions are as follows:

Assumption L1. $\epsilon_1, \ldots, \epsilon_T, \nu_1, \ldots, \nu_T, \lambda$ are mutually independent conditional upon Z.

Assumption L2. (A) The joint PDF of Y, λ conditional upon Z is bounded and continuous, as are all marginal and conditional densities. (B) $\lambda \mid Z$ has full support. (C) The characteristic function of $\epsilon_t(d)$ is non-vanishing, $\mathbb{E}[\epsilon_t | Z, \lambda] = 0$.

Assumption L1 weakens Assumptions KL1 and KL2 by relaxing the restriction that Z_t be first-order Markov. Assumption L2 places a full support assumption on $Y_{it}(d)$, which is implied by Assumption KL2.

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

Assumption L4. (A) $f_{Y^{t-1},Z^t,D^t}(y^{t-1}, z^t, d^t) > 0$ for all t. (B) The variance-covariance matrix of $\lambda \mid Z$ is full rank.

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

Assumption L5. Any $p \times p$ sub-matrix of $F(d) = (F_1(d_1)F_2(d_2) \ldots F_T(d_T))$ is full rank.

Assumption L5 is a standard assumption in the interactive fixed effect literature (Freyberger, 2018). Similar to the more general 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 $\lambda_{u,i}$.

We now define the conditional choice probability function

$$\bar{h}_t(d^t, z^t, y^{t-1}) \equiv \Pr(D_{it} = d_t \mid Z_i^t = z^t, Y_i^{t-1} = y^{t-1}, D_i^{t-1} = d^{t-1}),$$

and let $\bar{h} = \left(\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_T\right)$. Compared to Section 3.1, there is no latent variable that enters $\bar{h}$. Therefore, $\bar{h}$ is identified directly from the data. As in Section 3.1, we place very little structure on the learning process of decision making agents. This highlights that the core identification results do not rely on structure imposed on

the belief formation process. However it is worth emphasizing that, should there be such structure, our identification results would enable identification of the belief formation process. To illustrate this, consider the case that the decision making agents are rational Bayesians and that the sufficient statistics for λ_u at time t are a known function of the information set and the model parameters. That is, that there is a known function g such that the sufficient statistics equal $g(Y_i^{t-1}, D_i^{t-1}, Z_i^{t-1}, \theta)$, where θ are the model parameters. In this case, identification of θ is sufficient for identification of the beliefs.

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

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

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

Remark 4 (Auxiliary selection-free measurements). In some cases, additional unselected noisy measurements of known abilities are available. See, for instance, Cunha et al. (2005) and Heckman and Navarro (2007). With this additional data, sufficient

conditions for identification of the distribution of the latent effect are well known in the literature (Hu and Schennach, 2008; Cunha et al., 2010). If the sufficient conditions are satisfied conditional on each $(Y_t, D_t, X_t)_{t=1}^T$, then the joint distribution of $\left((Y_t, D_t, Z_t)_{t=1}^T, \lambda_k\right)$ is identified from the additional outcome variables. From here, one can redefine $Z_t = (Z_t, \lambda_k)$ and the conditions of Theorem 2 are sufficient for distribution-free identification of the model with known and unknown heterogeneity.

4 Estimation

We propose to estimate the model parameters via sieve maximum likelihood. Let $W_i = (D_{it}, Z_{it}, Y_{it})_{t=1}^T$ and $\theta^* \in \Theta$ be the true value of the parameters. We focus in the following on the model of Section 3.1, although similar conditions could be presented for the model of Section 3.2. The log-likelihood contribution of $W_i = w$ is

$$\begin{aligned}\ell(w;\theta) = &\log \int \prod_{t=1}^{T} \left(\frac{1}{\sigma_t\left(d_t\right)} \phi_1 \left(\frac{y_t - \alpha_t\left(d_t\right) - \beta_t\left(d_t\right)^T z_t - F_{ut}\left(d_t\right) v_u - F_{kt}\left(d_t\right) v_k}{\sigma_t\left(d_t\right)} \right) \right. \\ &\left. \times \bar{h}_t(d^t, z^t, y^{t-1}, v_k) \right) \times \prod_{t=1}^{T-1} g_t(z_{t+1} \mid z_t, y_t, d_t) \\ &\times \frac{1}{\sqrt{\left|\Sigma_u\left(z_1, v_k\right)\right|}} \phi_p \left(\Sigma_u^{-\frac{1}{2}}\left(z_1, v_k\right) v_u \right) \times dF_{\lambda_k}\left(v_k; z_1\right) dv_u\end{aligned}$$

where ϕ_s is the probability distribution function of the standard multivariate normal distribution with s components, g_t is the Markov kernel for z_{t+1}. There are four components of the likelihood function: the outcomes, the assignment probabilities, the Markov kernel of the covariates, and the distribution of latent factors (λ_u, λ_k).

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

$$\frac{1}{n}\sum_{i=1}^{n} \ell(w_i;\hat{\theta}) \geq \sup_{\theta \in \Theta_n} \frac{1}{n}\sum_{i=1}^{n} \ell(w_i;\theta) - o_p(1/n) \tag{4}$$

The following result states that under standard conditions (stated in the Appendix) $\hat{\theta}$ is consistent for θ^*.

Theorem 3. Suppose the distribution of $(Y_t, D_t, Z_t)_{t=1}^T$ is observed for $T \geq 2p+1$ and that Assumptions KL1-KL5 and Assumptions E1-E6 hold. Then $\hat{\theta}$ as defined in Equation (4) is consistent for θ^*.

In practice, researchers are often interested in particular low-dimensional functions of the model parameters. We begin by introducing the general class of functionals we consider for estimation, before turning to an illustration in the context of variance decomposition a la Cunha et al. (2005). We let $I_t \subseteq \times_{s=t+1}^T \text{Supp}(D_s)$, $\omega = \{\omega_i \colon i \in I_t\}$ be a user chosen subset of future choices and weights which could depend on θ. Then define the function f_1 which maps (θ, w, v_k) to $\mathbb{R}$ as

$$f_1\left(\mathrm{E}\left[\sum_{i\in I_t}\omega_i Y_{t_i}(i) \mid W^t = w, \lambda_k = v_k\right], \mathrm{Var}\left[\sum_{i\in I_t}\omega_i Y_{t_i}(i) \mid W^t = w, \lambda_k = v_k\right]\right), \quad (5)$$

where t_i denotes the time period to which $i \in I_t$ belongs. Now let $d\mu(w, v_k)$ be a user chosen measure on $\text{Supp}(W_t) \times \text{Supp}(\lambda_k)$. We define the functional of θ, which we propose to estimate by plug-in sieve MLE, as

$$f(\theta) = \int f_1(\theta, w, v_k) dF_{W^t,\lambda_k}(w, v_k).$$

Notice the measure $dF_{W^t,\lambda}$ also depends on θ.

This class of functionals encompass several economically meaningful objects, and can be used in particular to decompose earnings variance between predictable and unpredictable components. This question has attracted much interest in labor economics, where a growing literature aims to quantify the relative importance of uncertainty and unobserved heterogeneity in lifetime earnings (see, e.g. Cunha et al., 2005; Cunha and Heckman, 2008, 2016; Gong et al., 2019).

We illustrate this application with a simple model where agents make an one-time educational decision in period $t = t_0$, which determines their sector of employment

for their entire career. For a discount value ρ, the present value of lifetime earnings is:

$$\widetilde{Y}_{t_0}(d) = \sum_{t=t_0}^{T} \frac{Y_t(d)}{(1+\rho)^{t-t_0}}$$

where the predictable component of the present value of lifetime earnings is given by, denoting by $\mathcal{I}_{t_0}$ the agent's information set at time $t = t_0$:

$$E(\widetilde{Y}_{t_0}(d)|\mathcal{I}_{t_0})$$

where we assume that $\mathcal{I}_{t_0} = \{\lambda_k, W^{t_0}\}$ with $W^t = (Y^{t-1}, D^{t-1}, Z^t)$. The variance of the predictable and unpredictableTioarss2 components of $\widetilde{Y}_{t_0}(d)$ are then given by:

$$\sigma^2_{k,t_0}(d) = \int \left(E(\widetilde{Y}_{t_0}(d)|\mathcal{I}_{t_0}) - E(\widetilde{Y}_{t_0}(d)) \right)^2 dF_{\lambda_k, W^{t_0}}(x^*_k, w^{t_0})$$

$$\sigma^2_{u,t_0}(d) = \int \mathrm{Var}(\widetilde{Y}_{t_0}(d)|\mathcal{I}_{t_0}) dF_{\lambda_k, W^{t_0}}(x^*_k, w^{t_0})$$

The share of predictable relative to unpredictable earnings variance evolves over time as agents learn and update their beliefs about λ_u.

Theorem 4 shows that, under mild regularity conditions, the plug-in sieve MLE estimator for the functional $f(\theta)$ is root-n consistent and asymptotically normal.

Theorem 4. Suppose the distribution of $(Y_t, D_t, Z_t)_{t=1}^T$ is observed for $T \geq 2p+1$ and that Assumptions KL1-KL5 and Assumptions E1-E13 hold.Then $\sqrt{n}\frac{f(\hat{\theta})-f(\theta^*)}{\|v_n^*\|} \underset{d}{\to} \mathcal{N}(0,1)$

Note that Theorem 4 allows for the possibility that the sieve variance v_n^* may diverge– that is, that f is an *irregular* functional. In either case, consistent estimators are readily available for the sieve variance (Chen and Liao, 2014, Section 3).

5 Implementation

To implement the sieve maximum likelihood estimation developed in the previous section, first notice that we can write the likelihood as follows:

$$\ell(w;\theta) = \log \int \frac{1}{|2\pi V(w,v_k;\theta)|^{T/2}} \phi\left(m(w,v_k;\theta)^T V(w,v_k;\theta)^{-1} m(w,v_k;\theta)\right) \times \prod_{t=1}^{T} \bar{h}_t(d^t, z^t, y^{t-1}, v_k) \times \prod_{t=1}^{T-1} g_t(z_{t+1} \mid z_t, y_t, d_t) dF_{\lambda_k}(v_k; z_1)$$

where $m(w, v_k; \theta)$ and $V(w, v_k; \theta)$ are the T-dimensional vector and $T \times T$ matrix giving the expected mean and variance of Y^T conditional on $(D^T, Z^T, \lambda_k) = (d^T, z^T, v_k)$. They are defined as follows. $m(w, v_k; \theta) = (m_1(w, v_k; \theta), \ldots, m_T(w, v_k; \theta))^T$, where,

$$m_t(w, v_k; \theta) = \alpha_t(d_t) + \beta_t^\intercal z_t + F_k(d_t) v_k,$$

and,

$$V(w, v_k; \theta) = F_u(w)^T \Sigma_u F_u(w) + \text{diag}(d_1, \ldots, d_T)$$

where $F_u(w)$ is the $p \times T$,

$$F_u(w) = \Big[F_{u1}(d_t) \cdots F_{uT}(d_T) \Big]$$

There are three non-parametric objects in the likelihood function, $\bar{h}$, g, and F_{λ_k}.

The choice of sieve spaces for $\bar{h}$ and g are typically context specific. For F_{λ_k}, we propose using the sieve estimator discussed in Koenker and Mizera (2014) and closely related to the estimator in Fox et al. (2016). In particular, for each n, fix a grid of support points for λ_k, $\mathcal{S}_n = \{\bar{v}_{1n}, \ldots, \bar{v}_{q_n n}\}$, for some finite q_n. Then, we can use the sieve space for F_{λ_k}:

$$\mathcal{F}_n = \left\{ v \mapsto \sum_{s=1}^{q_n} \omega_s \mathbf{1}\{v \leq \bar{v}_{sn}\} \,\middle|\, \sum_{s}^{q_n} \omega_s = 1 \right\}$$

We can show that if $q_n \to \infty$ the space of distribution functions is dense in $\{\mathcal{F}_n\}_n$.

Koenker and Mizera (2014) note that fixing the other parameters, maximizing $\{\omega_s : s \leq q_n\}$ is a convex optimization problem that can solved efficiently and reliably

using standard software for convex optimization. In practice, we use the algorithm proposed in Kim et al. (2020), which is specialized for this setting, and implemented in the R package "mixsqp".

Given the efficiency of solving this problem, it is useful to define the *profile likelihood* function. Partition θ into $\{F_{\lambda_k}\}$ and $\theta_1 = \theta_1 = \theta \setminus \{F_{\lambda_k}\}$, then:

$$\sum_{i=1}^{n} \ell(w; \theta_1) = \max_{F_{\lambda_k}} \sum_{i=1}^{n} \ell(w; \theta_1, F_{\lambda_k})$$

To solve the original maximum likelihood problem, therefore, we simply maximize the profile likelihood function over θ_1. Separating the maximization problem into this inner and outer maximization significantly reduces the dimensionality of the problem, while allowing λ_k to have a very flexible distribution. Importantly, the inner maximization problem over F_{λ_k} can be solved very quickly and efficiently despite the potentially high dimensionality of $\{\omega_s : s \leq q_n\}$.

6 Monte Carlo Simulations

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

The data generating process (DGP) used in our simulations has both known and unknown heterogeneity, no covariates, and $\text{supp}(\lambda_{u,i}) = \mathbb{R}$. With this specification, the potential outcomes equation is:

$$Y_t(d) = \alpha_t(d) + F_{ut}(d)\lambda_u(d) + F_{kt}(d)\lambda_k + U_t(d)$$

Recall that under the assumptions of Theorem 1, the marginal distribution of λ_k, F_{λ_k}, and the CCP function, $\bar{h}$, are point identified without further functional form assumptions or assumptions on the primitive choice process generating $\bar{h}$. We adopt a specification in which agents choose an option to maximize utility in each period.

The utility that individual i derives from choice d in period t is,

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

where $\{\nu_{i,t}(d) : t = 1, 2, 3, d = 1, 2\}$ are mutually independent with an Extreme Value Type 1 distribution. This assumption embeds several features that are common in learning models: (1) there is positive selection on the expected outcome, and (2) individuals have rational expectations and use their past outcomes and choices to form their expectations about future outcomes. Since λ_k enters linearly in the utility and in the conditional expectation term, this specification also allows agents to have permanent unobserved heterogeneity which affects both their choices and their learning process.

A common approach to estimating models of this form is to assume that λ_k has a discrete distribution, typically with a small number of support points. The conditions of Theorem 1, however, impose the much weaker condition that the support of λ_k is compact. To illustrate the performance of our estimator when λ_k is not drawn from a low-dimensional discrete distribution, we use a mixture of Gaussian random variables with three components where the mean and variance are both allowed to be mixture-component specific. The resulting distribution is then truncated to satisfy the assumption that λ_k has compact support.

We perform a Monte Carlo experiment, estimating the model with 400 simulations and sample sizes of $250, 500, 1,000, 2,000$ and $4,000$. We use the sieve MLE estimator described in Section 5, with the number of support points in the estimated distributions growing at a rate of $n^{1/3}$, from 62 to 158. This is a pretty high dimensional problem. With $n = 4,000$, the inner convex optimization problem has $4,000$ constraints and 158 parameters, and the outer non-convex optimization problem has 20 parameters. Importantly though, using the computational approach described earlier, the MLE problem can be solved in under 3 minutes using a 4-core CPU. Computation times ranged from approximately 30 seconds to 180 seconds depending on sample size.

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

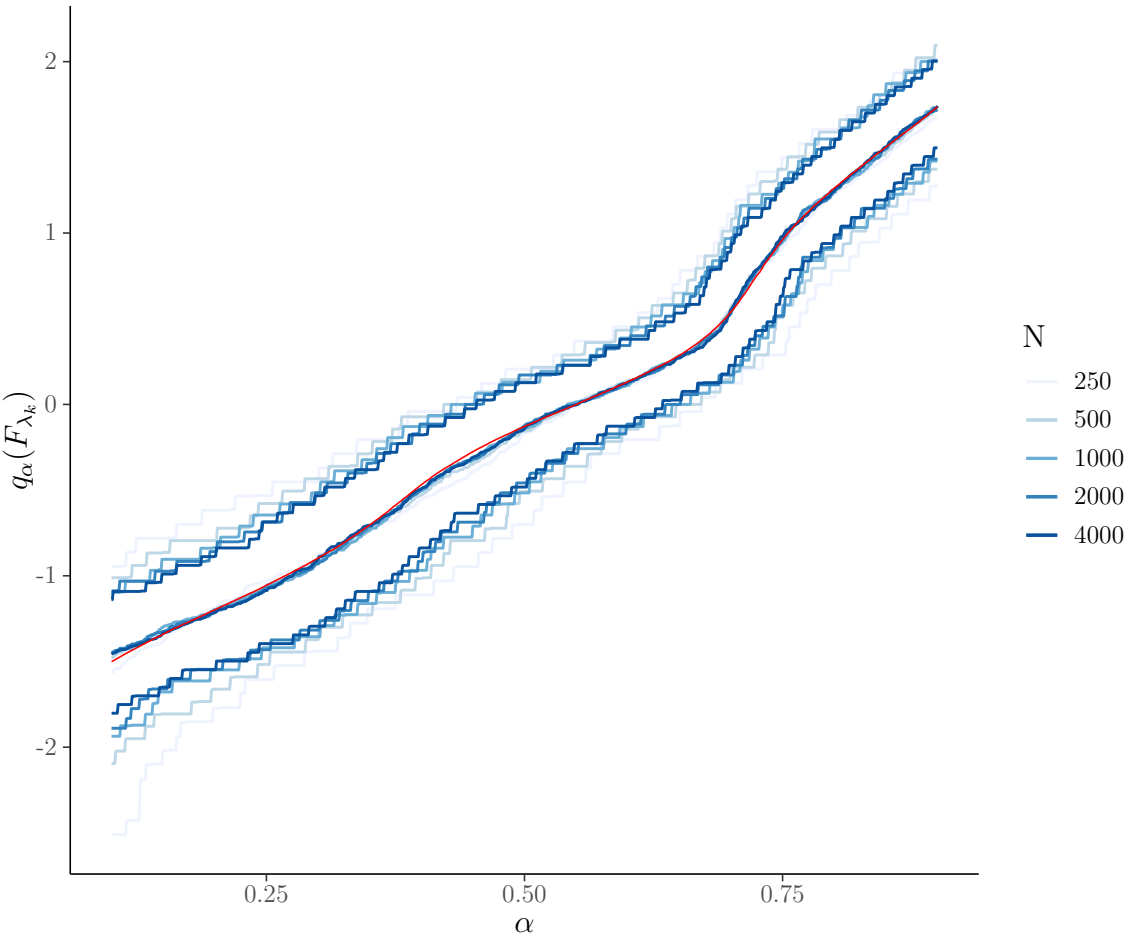
	N = 250		N = 500		N = 1000		N = 2000		N = 4000	
	sq bias	var	sq bias	var	sq bias	var	sq bias	var	sq bias	var
$\alpha_1(2)$	0.746	6.953	0.056	2.311	0.017	0.763	0.006	0.276	0.001	0.134
$\alpha_2(1)$	0.037	6.060	0.000	2.274	0.001	0.752	0.000	0.265	0.000	0.112
$\alpha_2(2)$	0.001	8.545	0.001	2.617	0.008	0.849	0.001	0.321	0.001	0.118
$\alpha_3(1)$	0.007	5.236	0.002	2.291	0.002	0.762	0.001	0.293	0.000	0.122
$\alpha_3(2)$	0.052	8.169	0.010	2.950	0.006	0.950	0.002	0.312	0.003	0.120
$F_{u1}(2)$	0.011	10.801	0.044	3.707	0.019	1.307	0.010	0.381	0.004	0.158
$F_{u2}(1)$	0.019	6.508	0.023	3.056	0.001	1.067	0.000	0.363	0.000	0.158
$F_{u2}(2)$	0.183	13.906	0.159	4.276	0.039	1.287	0.016	0.368	0.008	0.152
$F_{u3}(1)$	0.132	5.932	0.093	3.077	0.068	1.035	0.019	0.362	0.002	0.120
$F_{u3}(2)$	0.879	13.157	0.299	5.717	0.096	1.504	0.027	0.496	0.013	0.190
$F_{k1}(1)$	0.006	5.126	0.000	2.262	0.006	0.871	0.008	0.301	0.003	0.118
$F_{k2}(1)$	0.125	5.589	0.053	2.600	0.021	0.960	0.008	0.305	0.003	0.122
$F_{k2}(2)$	0.943	7.306	0.080	2.860	0.031	0.929	0.008	0.284	0.002	0.135
$F_{k3}(1)$	0.001	6.225	0.001	2.528	0.002	0.786	0.000	0.289	0.000	0.115
$F_{k3}(2)$	0.587	6.793	0.043	2.650	0.000	0.818	0.002	0.323	0.001	0.136
γ	0.016	2.773	0.000	1.308	0.000	0.403	0.000	0.138	0.000	0.057
$\sigma^2(1)$	0.037	0.341	0.011	0.144	0.003	0.057	0.001	0.018	0.000	0.007
$\sigma^2(2)$	0.064	0.599	0.024	0.249	0.005	0.066	0.001	0.020	0.000	0.008
σ_u^2	0.019	2.853	0.004	1.049	0.002	0.317	0.000	0.097	0.000	0.043
ρ	0.035	14.565	0.004	6.103	0.002	1.855	0.001	0.697	0.000	0.308

All calculations are based on 400 Monte Carlo simulations of the DGP described in the main text. Squared bias and variance of finite parameter estimates are multiplied times $1,000$

Starting with the finite parameters, $\{\alpha_t, F_{ut}, F_{kt}, \gamma, \sigma, \rho\}$, Table 1 shows squared bias and variance of the parameter estimates under the simulated distribution. (Note that all values in Table 1 are multiplied by $1,000$.) For each of the parameters, the bias tends to be small even for small sample sizes. In particular, the squared bias tends to be negligible relative to the variance. Besides, the variance declines at a rate consistent with $\sqrt{n}$ convergence of the mean squared error.

To present results for the nonparametric estimator of the distribution of known unobserved heterogeneity F_{λ_k}, we focus on the quantiles of F_{λ_k}. Let $q_\alpha(F)$ be the α quantile of a random variable with the distribution F. For each value of $\alpha \in [0, 1]$,

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



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

we calculate the mean and the 5th and 95th percentile of the simulated distribution of the estimator of $q_\alpha(F_{\lambda_k})$. The results are illustrated in Figure 1. The red line is the CDF of the true distribution of λ_k, while the blue lines that closely follow the red line are the mean of the simulated distribution of the quantile estimators for each sample size. Darker blue lines represent larger sample sizes. The blue lines above and below the CDF are the 5th and 95th percentiles of the simulated distribution of the quantile estimators.

The results indicate that the bias of the quantile estimators are negligible even at

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

	N = 250		N = 500		N = 1000		N = 2000		N = 4000	
	sq bias	var	sq bias	var	sq bias	var	sq bias	var	sq bias	var
$\tilde{\sigma}_{u,1}(1)$	0.028	3.550	0.049	1.712	0.006	0.652	0.000	0.242	0.001	0.109
$\tilde{\sigma}_{u,2}(1)$	0.050	6.383	0.022	4.431	0.015	1.675	0.023	0.573	0.009	0.251
$\tilde{\sigma}_{u,3}(1)$	0.000	6.660	0.001	3.824	0.017	1.192	0.007	0.432	0.002	0.160
$\tilde{\sigma}_{k,1}(1)$	0.111	3.514	0.029	1.231	0.008	0.389	0.002	0.114	0.000	0.048
$\tilde{\sigma}_{k,2}(1)$	0.074	2.697	0.007	1.167	0.003	0.443	0.010	0.144	0.014	0.067
$\tilde{\sigma}_{k,3}(1)$	0.076	1.082	0.011	0.466	0.002	0.173	0.000	0.067	0.004	0.026

All calculations are based on 400 Monte Carlo simulations of the DGP described in the main text. Squared bias and variance of finite parameter estimates are multiplied times $1,000$

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

The results in Table 1 confirms that despite the slower convergence of the nonparametric estimator of F_{λ_k}, the finite parameters converge at the usual parametric rate. Theorem 4 shows that the parametric convergence rate is attained for a broader class of linear functionals of the model parameters, including the parameters that decompose variance in potential outcomes into uncertainty and known heterogeneity. We now turn to assessing the performance of estimators of this kind of parameter. For simplicity, we focus here on the variance of potential outcomes in one period rather than a sum of potential outcomes. In particular, we consider the following parameters, which are single-period analogues of the parameters $(\sigma^2_{k,t_0}(d), \sigma^2_{u,t_0}(d))$ as defined

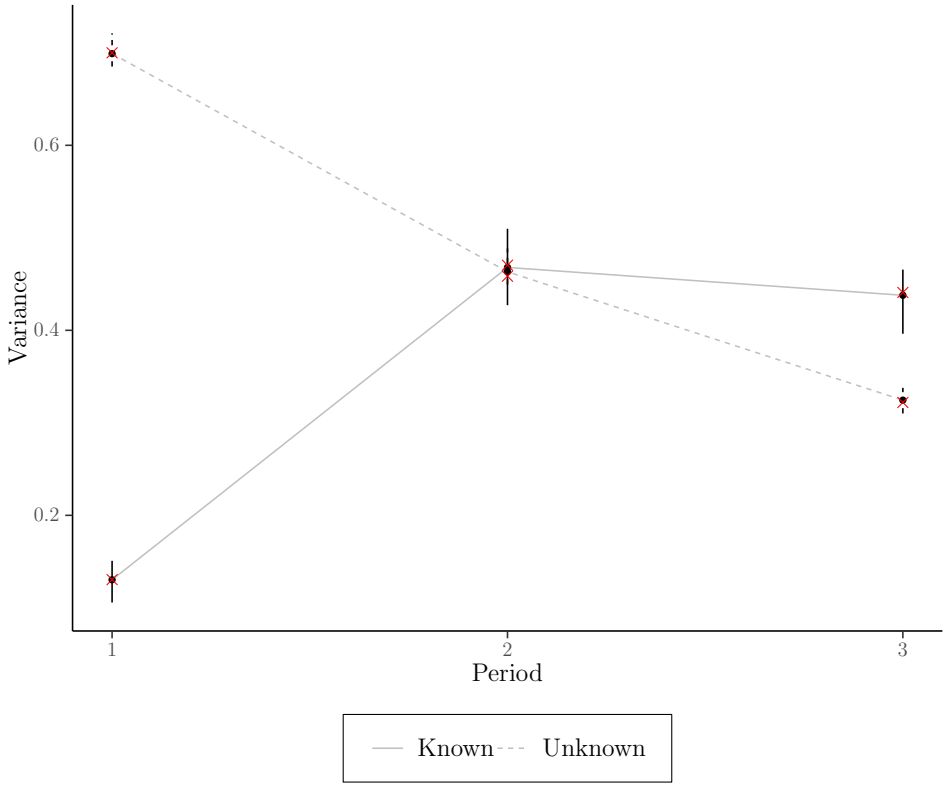
in Section 4,

$$\tilde{\sigma}^2_{k,t}(d) = \int \left(E(Y_t(d)|\mathcal{I}_t) - E(\widetilde{Y}_t(d)) \right)^2 dF_{\lambda_k, W^{t_0}}(x^*_k, w^t),$$
$$\tilde{\sigma}^2_{u,t}(d) = \int \mathrm{Var}(Y_t(d)|\mathcal{I}_t) dF_{\lambda_k, W^t}(x^*_k, w^t).$$

Table 2 reports the squared bias and variance of the parameter estimates for each of these two parameters for $t = 1, 2, 3$ and $d = 1$. These results are similar to the results for the finite parameters of the model. The bias tends to be small relative to the variance, and the variance declines at a rate consistent with $\sqrt{n}$ consistency. Consistent with Theorem 4, these results indicate that this class of parameters can be estimated at a parametric rate, even though they are functionals of all the parameters, including F_{λ_k}, which is estimated nonparametrically.

To explore the magnitude of the estimation error of these parameters, Figure 2 shows the estimates of the variance decomposition for $d = 1$. The vertical bars in Figure 2 show the 5th and 95th percentiles of the simulated distributions of the estimators. The pattern over time in the variance decomposition reflects the learning dynamics in the model: uncertainty declines over time so the unknown component of variance declines relative to the known component. In particular, with a sample size of $2,000$, the estimation error is small compared to the magnitude of these dynamics.

Figure 2: Variance Decomposition Estimators, $N = 2,000$



Note: The red crosses show the true parameters. Vertical bars show the range between the 5th and 95th percentiles of the simulated distributions of the estimators. The dashed line shows the trajectory of variance in potential outcomes that is unknown to individuals when they choose an option. The solid line shows the remaining variance in potential outcomes, which is known to individuals.

7 Conclusion

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

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

Bibliography

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

Arcidiacono, P., Aucejo, E., Maurel, A. and Ransom, T. (2016), College attrition

and the dynamics of information revelation, Technical report, National Bureau of Economic Research.

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

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

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

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

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

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.

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. (2008), 'A new framework for the analysis of inequality', *Macroeconomic Dynamics* **12**(S2), 315–354.

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.

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

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

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

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

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

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.

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.

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.

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

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

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

Pastorino, E. (2022), Careers in firms: the role of learning and human capital, Technical report, Technical report, mimeo.

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

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

A Identification proofs and auxiliary results

A.1 Proofs for Section 3.1

Proof of Lemma 1. We proceed inductively. First, by Assumption KL2 and the definition of (E_1, Σ_1), $\lambda_u|(Z_1 = z_1, \lambda_k = v_k) \sim \mathcal{N}(E_1, \Sigma_1)$. Second, for $t > 1$ suppose $\lambda_u|(Y^{t-2}, D^{t-2}, Z^{t-1}) \sim \mathcal{N}(E_{t-1}, \Sigma_{t-1})$ and consider the following argument:

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

Display (1) follows from Bayes' theorem. Display (2) holds since Assumption KL1 has the following three implications: first $(Z_t \perp\!\!\!\perp \lambda \mid Z^{t-1}, Y^{t-1}, D^{t-1})$; second $(\epsilon_{t-1}(d_{t-1}) \perp\!\!\!\perp Z^{t-1}, Y^{t-2}, D^{t-1}, \lambda) \Rightarrow (\epsilon_{t-1}(d_{t-1}) \perp\!\!\!\perp Z^{t-2}, Y^{t-2}, D^{t-1} \mid Z_{t-1}, \lambda) \Rightarrow (Y_{t-1}(d_{t-1}) \perp\!\!\!\perp Z^{t-2}, Y^{t-2}, D^{t-1} \mid Z_{t-1}, \lambda)$; third $(D_{t-1} \perp\!\!\!\perp \lambda_u \mid Z^{t-1}, Y^{t-2}, D^{t-2}, \lambda_k)$. Display (3) holds from the induction assumption and Assumptions KL1 and KL2. Display (4) follows from the definitions in Lemma 1. □

Lemma 2. Let Assumptions KL1 and KL2 hold. Then Y_t conditional upon

$(D^t, Y^{t-1}, Z^t, \lambda_k) = (d^t, y^{t-1}, z^t, v_k)$ is distributed

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

Proof of Lemma 2. We show that the distribution of Y_t conditional upon $(D^t, Y^{t-1}, Z^t, \lambda_k)$ is distributed according to

$$N\left(\alpha_t(D_t) + \beta(D_t)' Z_t + F_{kt}(D_t)\lambda_k + F_{ut}(D_t)^\intercal E_t,\ F_{ut}(D_t)^\intercal \Sigma_t F_{ut}(D_t) + \sigma_t^2(D_t)\right).$$

First consider $t = 1$. In this case

$$\begin{aligned}
&f_{Y_1|D_1 Z_1 \lambda_k}(y_1; d_1, z_1, v_k) \\
&= f_{Y_1(d_1)|D_1 Z_1 \lambda_k}(y_1; d_1, z_1, v_k) \\
&= \int f_{Y_1(d_1)\lambda_u|D_1 Z_1 \lambda_k}(y_1, v_u; d_1, z_1, v_k) dv_u \\
&= \int f_{Y_1(d_1)|D_1 Z_1 \lambda}(y_1; d_1, z_1, v) f_{\lambda_u|D_1 Z_1 \lambda_k}(v_u; d_1, z_1, v_k) dv_u \\
&=_{(1)} \int f_{Y_1(d_1)|Z_1 \lambda}(y_1; z_1, v) f_{\lambda_u|Z_1 \lambda_k}(v_u; z_1, v_k) dv_u \\
&=_{(2)} \int \frac{1}{\sigma_1(d_1)} \phi\left(\frac{y_1 - \alpha_1(d_1) - z_1^\intercal \beta_1(d_1) - v_k F_{k1}(d_1) - v_u F_{u1}(d_1)}{\sigma_1(d_1)}\right) \\
&\times (2\pi)^{p/2} \det \Sigma_1^{-1/2} \exp\left(-1/2 (v_u - E_1)^\intercal \Sigma_1^{-1} (v_u - E_1)\right) dv_u \\
&= \frac{1}{\sqrt{F_{u1}(d_1)^\intercal \Sigma_1 F_{u1}(d_1) + \sigma_1^2(d_1)}} \phi\left(\frac{y_1 - \alpha_1(d_1) - z_1^\intercal \beta_1(d_1) - v_k F_{k1}(d_1) - E_1^\intercal F_{u1}(d_1)}{\sqrt{F_{u1}(d_1)^\intercal \Sigma_1 F_{u1}(d_1) + \sigma_1^2(d_1)}}\right)
\end{aligned}$$

Equality (1) holds since Assumption KL1 implies $(Y_1(d_1) \perp\!\!\!\perp D_1 \mid Z_1, \lambda)$ and $(D_1 \perp\!\!\!\perp \lambda_u \mid \lambda_k, Z_1)$, as argued in the proof to Lemma 1. Equality (2) holds because Assumption KL1 and KL2 imply $\epsilon_1(d) \mid (Z_1, \lambda) \sim N(0, \sigma_1(d)^2)$ in addition to Assumption 4. Notice that $(E_1, \Sigma_1) = (0, \Sigma_u(v_k, z_1))$.

Now consider $t > 1$. By Lemma 1, $\lambda_u \mid (D^{t-1}, Y^{t-1}, Z^{t-1}, \lambda_k)$ is distributed $N(E_t, \Sigma_t)$.

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

Equality (1) holds because Assumption KL1 implies $(Y_t(d_t) \perp\!\!\!\perp Z^{t-1}, D^t, Y^{t-1} \mid Z_t, \lambda)$ and $(D_t \perp\!\!\!\perp \lambda_u \mid \lambda_k, Z^t, Y^{t-1}, D^{t-1})$, as argued in the proof to Lemma 1. Equality (2) holds because Assumption KL1 and KL2 imply $\epsilon_t(d) \mid (Z_t, \lambda) \sim N(0, \sigma_t(d)^2)$.

□

Proof of Theorem 1. The proof is in three parts. First, we show that $(Y_t \mid D^t, Y^{t-1}, Z^t, \lambda_k)$ is normally distributed and apply Bruni and Koch (1985, Theorem 3) to identify its distribution up to an affine transformation of λ_k. The second part uses the normalization (Assumption KL3) to show that the affine transformation is the identity function. The final part uses identification of the distribution of $(Y^t, D^t, Z^t, \lambda_k)$ to identify the distribution of (Y^t, D^t, Z^t, λ).

Part 1: Identification of $f_{Y^tD^tZ^t\lambda_k}(y^t, d^t, z^t, \pi(v_k))$ for an unknown affine function π

In Lemma 2, we show that $(Y_t \mid D^t, Y^{t-1}, Z^t, \lambda_k)$ is distributed according to

$$
N\left(\alpha_t(D_t) + \beta(D_t)' Z_t + F_{kt}(D_t)\lambda_k + F_{ut}(D_t)^\intercal E_t,\ F_{ut}(D_t)^\intercal \Sigma_t F_{ut}(D_t) + \sigma_t^2(D_t)\right).
$$

It follows that

$$f_{Y_t|D^tY^{t-1}Z^t}(y_t; d^t, y^{t-1}, z^t) =$$
$$\int f_{Y_t|D^tY^{t-1}Z^t\lambda_k}(y_t; d^t, y^{t-1}, z^t, v_k) dF_{\lambda_k|D^tY^{t-1}Z^t}(v_k; d^t, y^{t-1}, z^t) dv_k$$

is a Gaussian mixture. To identify the component and mixture distributions, we will apply Bruni and Koch (1985, Theorem 3). First, define the set Λ as

$$\left\{(\alpha_t(d_t) + \beta(d_t)'z_1 + v_k\mu_1(\theta^t) + \mu_2(\theta^t), \sigma(v_k, \theta^t)) \colon \theta^t \in \Theta^t\right\},$$

where $\theta^t = \text{vec}(\alpha^t, \beta^t, F_k^t, F_u^t, \sigma^t, \Sigma_u)$ and

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

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

$$\left((\alpha_t(d_t) + \beta(d_t)'z_1 + \pi(v_k)\mu_1(\theta^t) + \mu_2(\theta^t),\ \sigma(\pi(v_k), \theta^t),\ dF_{\lambda_k|D^tY^{t-1}Z^t}(\pi(v_k); d^t, y^{t-1}, z^t)\right) \tag{6}$$

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

To conclude Part 1, we show that if $\pi = I$ for each history (d^s, y^{s-1}, z^s) $s = 1, 2, \ldots, t$, then $f_{Y^t,D^t,Z^t,\lambda_k}(y^t, d^t, z^t, v_k)$ is point identified. That is, $(\alpha_t(d_t), \beta_t(d_t), \mu_1(\theta^t), \mu_2(\theta^t), \sigma(v_k, \theta^t), dF_{\lambda_k|D^tY^{t-1}Z^t}(v_k; d^t, y^{t-1}, z^t))$ is point identified. For $t = 1$, as $(\mu_1(\theta^1), \mu_2(\theta^1)) = (F_{k1}(d_1), 0)$ identification follows immediately from equation (6) and Assumption KL4(E). Now suppose

$(\alpha_s(d_s), \beta_s(d_s), \mu_1(\theta^s), \mu_2(\theta^s), \sigma(v_k, \theta^s), dF_{\lambda_k|D^sY^{s-1}Z^s}(v_k; d^s, y^{s-1}, z^s))$ is point identified for each $s < t$. From equation (6) and $\pi = I$,

$$\left((\alpha_t(d_t) + \beta(d_t)'z_1 + v_k\mu_1(\theta^t) + \mu_2(\theta^t),\ \sigma(v_k, \theta^t),\ dF_{\lambda_k|D^tY^{t-1}Z^t}(v_k; d^t, y^{t-1}, z^t)\right)$$

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

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

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

Part 2: Showing π is the identity function

In this part we use the normalization assumption to prove the affine function π is identity. First, we show $\pi = I$ for the normalized choice d_1, which provides identification of the support of λ_k. Second, we use knowledge of $\text{Supp}(\lambda_k)$ to prove the affine function must satisfy $|\frac{\partial}{\partial v}\pi(v)| = 1$ for any history (d^t, y^{t-1}, z^t). Third, we use restrictions on the panel dimension to conclude $\pi = I$ for each history (d^t, y^{t-1}, z^t).

Part 2.1: Identifying π for $D_1 = d_1$.

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

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

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

Part 2.2: Restricting π to have modulus derivative equal to one.

Fix $(Y^t, D^t, Z^t) = (y^t, d^t, z^t)$. From Part 1, we have identification of

$$\left((\alpha_t(d_t) + \beta(d_t)'z_1 + \pi(v_k)\mu_1(\theta^t) + \mu_2(\theta^t), \sigma(\theta^t), dF_{\lambda_k|D^tY^{t-1}Z^t}(\pi(v_k); d^t, y^{t-1}, z^t)\right).$$

In this part we use the known support of λ_k to prove the modulus of the derivative of π is unity. First, consider that by Assumption KL4(D),

$$\text{Supp}(\lambda_k) = dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+] = (dF_{\lambda_k|D^tY^{t-1}Z^t} \circ \pi)^{-1}[\mathbb{R}_+]$$

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

$$(\pi \circ dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t})[\mathbb{R}_+] = dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+].$$

In particular

$$\pi(\sup dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+]) = \sup dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+]$$
$$\pi(\inf dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+]) = \inf dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+]$$

The only affine functions that satisfy these identities are $\pi^+(v) = v$ and $\pi^-(v) = (\bar{v} + \underline{v}) - v$ for $\underline{v} = \inf dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+]$ and $\bar{v} = \sup dF^{-1}_{\lambda_k|D^tY^{t-1}Z^t}[\mathbb{R}_+]$. It remains to show that $\pi = \pi^+$.

Part 2.3: Concluding $\pi = I$.

For this part, it will be useful to define:

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

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

The proof is inductive. First consider $t = 1$. From Assumption KL3(A), $F_{k1}(d_1) = 1$. For $\tilde{d}_1 \neq d_1$, from Part 1 we have identified

$$\left(\alpha_1(\tilde{d}_1) + \beta(\tilde{d}_1)'z_1 + F_{k1}(\tilde{d}_1)\pi(v_k)\right)$$

And from Part 2.2, we conclude that $F_{k1}(\tilde{d}_1)$ is identified up to sign. For $d^2 = (d_2, d_1)$, by Part 1 we identify

$$\left(\alpha_2(d_2) + \beta(d_2)' z_2 + \pi(v_k)\mu_1(d^2) + \mu_2(d^2)\right). \tag{7}$$

From Part 2.2, we conclude that $\mu_1(d^2) = F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1)$ is identified up to sign. And since $\mu_2(d^2) = F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1)(y_1 - \alpha_1(d_1) - z_1^\intercal \beta_1(d_1))$ and $\mu_1(d^2)$ does not depend on y_1, we can identify $F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1)$ by taking the derivative of (7) with respect to y_1.

Repeating this argument for the choice sequence $(\tilde{d}_1, d_2)$ yields identification of $(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1))$ up to sign and $F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1)$.

Summarizing, we have identification of the set

$$\left\{(-1)^{j_1} F_{k1}(\tilde{d}_1),\ (-1)^{j_{d_2}}(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1)),\ (-1)^{\tilde{j}_{d_2}}(F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1))\right\},$$

with $j = (j_1, j_{d_2}, \tilde{j}_{d_2}) \in \{0,1\}^3$. We show only the correct choice of sign will satisfy

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

First, suppose $\tilde{j}_{d_2} = 1$. It is straightforward to show the following implications

$$\begin{aligned}
(j_1, j_{d_2}) = (0,0) &\ \Rightarrow\ F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) =_{(1)} 0, \\
(j_1, j_{d_2}) = (0,1) &\ \Rightarrow\ F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^{'}) F_{k1}(d_1) =_{(2)} 0, \\
(j_1, j_{d_2}) = (1,0) &\ \Rightarrow\ F_{k2}(d_2) =_{(3)} 0, \\
(j_1, j_{d_2}) = (1,1) &\ \Rightarrow\ F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^{'}) F_{k1}(d_1) =_{(4)} 0.
\end{aligned}$$

Equalities (1), (2), (3) and (4) contradict Assumptions KL5 (C), (A), (B) and (D) respectively. Similarly,

$$\begin{aligned}
(j_1, j_{d_2}, \tilde{j}_{d_2}) = (1,0,0) &\ \Rightarrow\ F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1)) = 0, \\
(j_1, j_{d_2}, \tilde{j}_{d_2}) = (0,1,0) &\ \Rightarrow\ F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^1) F_{k1}(d_1) = 0, \\
(j_1, j_{d_2}, \tilde{j}_{d_2}) = (1,1,0) &\ \Rightarrow\ F_{k2}(d_2) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(\tilde{d}^1) F_{k1}(\tilde{d}_1) - F_{u2}(d_2)^\intercal \tilde{\mu}_{21}(d^{'}) F_{k1}(d_1) = 0.
\end{aligned}$$

The first two equalities contradict Assumptions KL5 (D) and (C) respectively. To conclude, we show the third equality contradicts Assumption KL5(E). For each $d \in \{d_{2,i} : i = 1, 2, \ldots, k\} \cup \{\tilde{d}_{2,i} : i = 1, 2, \ldots, k\}$ from Assumption KL5(E), consider the sequences (d_1, d), $(\tilde{d}_1, d)$. The above argument applies and we identify

$$\left\{(-1)^{j_d}(F_{k2}(d) - F_{u2}(d)^\intercal \tilde{\mu}_{21}(d^1)F_{k1}(d_1)),\ (-1)^{\tilde{j}_d}(F_{k2}(d) - F_{u2}(d)^\intercal \tilde{\mu}_{21}(\tilde{d}^1)F_{k1}(\tilde{d}_1))\right\},$$

for $\left(j_1, (j_{d_{2,i}}, \tilde{j}_{d_{2,i}}, j_{\tilde{d}_{2,i}}, \tilde{j}_{\tilde{d}_{2,i}} : i = 1, \ldots, k)\right) \in \left\{\left(0, (0,0,0,0)^k\right), \left(1, (1,0,1,0)^k\right)\right\}$. If $\left(j_1, (j_{d_{2,i}}, \tilde{j}_{d_{2,i}}, j_{\tilde{d}_{2,i}}, \tilde{j}_{\tilde{d}_{2,i}} : i = 1, \ldots, k)\right) = \left(1, (1,0,1,0)^k\right)$, then

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

which contradicts Assumption KL5(E).

For the induction step, suppose $\pi = I$ for each history (d^s, y^{s-1}, z^s) $s = 1, \ldots, t-1$ and consider choice sequences $d^{t-1} = (d^{t-2}, d_{t-1})$ and $\tilde{d}^{t-1} = (d^{t-2}, \tilde{d}_{t-1})$ for $d_{t-1} \neq \tilde{d}_{t-1}$. From part 1, we have identification of

$$\left(\alpha_t(d_t) + \beta_t(d_t)' z_t + \mu_1(d^{t-2}, d, d_t)\pi(v_k) + \mu_2(d^{t-2}, d, d_t)\right),$$

for $d = d_{t-1}, \tilde{d}_{t-1}$. By the previous arguments, we identify

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

with $(j_1, j_2) \in \{0,1\}^2$ in addition to $(F_{ut}(d_t)\sum_{s=1}^{t-1} \tilde{\mu}_{ts}(d^{t-1})F_{ks}(d_s))$, $F_{ut}(d_t)\sum_{s=1}^{t-1} \tilde{\mu}_{ts}(\tilde{d}^{t-1})F_{ks}(\tilde{d}_s)$. As before we show that only that only $(j_1, j_2) = (0,0)$ is consistent with the identity

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

To see this, consider following implications:

$$(j_1, j_2) = (0,1) \;\Rightarrow\; \left(F_{kt}(d_t) - F_{ut}(d_t) \sum_{s=1}^{t-1} \tilde{\mu}_{ts}(\tilde{d}^{t-1}) F_{ks}(\tilde{d}_s) \right) = 0,$$

$$(j_1, j_2) = (1,0) \;\Rightarrow\; \left(F_{kt}(d_t) - F_{ut}(d_t) \sum_{s=1}^{t-1} \tilde{\mu}_{ts}(d^{t-1}) F_{ks}(d_s) \right) = 0,$$

$$(j_1, j_2) = (1,1) \;\Rightarrow\; F_{ut}(d_t) \sum_{s=1}^{t-1} \tilde{\mu}_{ts}(d^{t-1}) F_{ks}(d_s) - F_{ut}(d_t) \sum_{s=1}^{t-1} \tilde{\mu}_{ts}(\tilde{d}^{t-1}) F_{ks}(\tilde{d}_s) = 0$$

The first two equalities contradict Assumption KL5(C), and the final equality contradicts Assumption KL5(A). Thus π is the identity function for the history (d^t, y^{t-1}, z^t).

Part 3: Identification of $f_{Y^t D^t Z^t \lambda}(y^t, d^t, z^t, v_k, v_u)$

From Parts 1 and 2, $f_{Y^T D^T Z^T|\lambda_k}$ is identified. First,

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

This implies

$$\begin{aligned}
&\frac{f_{Y^T D^T Z^T|\lambda_k, Z_1}\left(y^T, d^T, z^T; v_k, z_1\right)}{f_{D_1|\lambda_k Z_1}(d_1; v_k, z_1) \prod_{t=2}^{T} f_{D_t|Y^{t-1} D^{t-1} \lambda_k Z^t}(d_t; y^{t-1}, d^{t-1}, v_k, z^t) f_{Z_t|Y_{t-1} D_{t-1} Z_{t-1}}(z_t; y_{t-1}, d_{t-1}, z_{t-1})} \\
=&\int \prod_{t=1}^{T} f_{Y_t^{d_t}|\lambda, Z_t}\left(y_t; v, z_t\right) f_{\lambda_u|\lambda_k Z_1}(v_u; v_k, z_1) dv_u
\end{aligned}$$

This is a normal-weighted mixture of normals. In particular, the function is equal to the pdf of a joint normal with mean

$$\left(\alpha_t(d_t) + Z_{it}^{\intercal} \beta_t(d_t) + v_k F_{kt}(d_t)\right)_{t=1}^{T}$$

and covariance matrix

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

where $F_u(d) = (F_{u1}(d_1)F_{u2}(d_2)\dots F_{uT}(d_T))$. From Parts 1 and 2, the components of the mean function are identified. The components of the covariance matrix are identified under the normalization assumption (Assumption KL3(B)) and the rank condition on $F_u(d)$ (Assumption KL5(F)).

□

Proof of Corollary 1. Fix $(d_1, d_2, \dots, d_p)$ and define $F_p = (F_{u1}(d_1)F_{u2}(d_2)\dots F_{up}(d_p))$, $\tilde{\lambda}_u = F_p^\intercal(\lambda_u - \mu_u)$, $\tilde{\epsilon}_t(d) = \epsilon_t(d) - c_t(d)$, $\tilde{\lambda}_k = b + F_{k1}(d_1)\lambda_k$ where $b = \alpha_1(d_1) + F_{1u}(d_1)^\intercal\mu_u + c_1(d_1)$. Finally, define $\tilde{F}_{kt}(d_t) = F_{k1}(d_1)^{-1}F_{kt}(d_t)$ and $\tilde{F}_{ut}(d_t) = F_p^{-1}F_{ut}(d_t)$.

and $\tilde{\alpha}_t(d) = \alpha_t(d) - \tilde{F}_{kt}(d)b + F_{ut}(d)^\intercal\mu_u + c_t(d)$. We then have that

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

$E[\tilde{\epsilon}_t(d)] = 0$ and $E[\tilde{\lambda}_u \mid Z_1 = z, \lambda_k = v_k] = 0$ so that the reparameterized model satisfies KL2. Also, $\tilde{F}_{k1}(d_1) = 1$, $\tilde{\alpha}_1(d_1) = 0$ and $\tilde{F}_p \equiv \left(\tilde{F}_{u1}(d_1)\tilde{F}_{u2}(d_2)\dots\tilde{F}_{up}(d_p)\right) = I_p$ so the reparameterized model satisfies KL3. By Theorem 1, $\tilde{\theta} = \left((\tilde{\alpha}_t, \beta_t, \tilde{F}_{kt}, \tilde{F}_{ut}, \sigma_t)_{t=1}^T, \Sigma_u, \tilde{\bar{h}}, F_{\tilde{\lambda}_k}\right)$ is identified, which imply the identification of the distribution of C_{jt} $(j = k, u)$. Finally,

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

□

A.2 Proofs for Section 3.2

Proof. Throughout this proof, let $f_{A|B}$ denote the conditional PDF of a random variable $A \mid B$, and f_A denote the marginal PDF of random variable A. The

notation $A \perp\!\!\!\perp B \mid C$ indicates that A is independent of B conditional upon C. Also, let $\mathcal{L} = \{m\colon \mathbb{R}^k \to \mathbb{R} : \sup_{a\in\mathbb{R}^k} |m(a)| < \infty, \int |m(a)|da < \infty\}$ 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.

Fix a choice sequence $d = (d_1, d_2, \ldots, d_T)$ whose first p elements satisfy Assumption L3, and define $W_1 = (Y_1, \ldots, Y_p)$, $W_2 = Y_{p+1}$ and $W_3 = (Y_{p+2}, \ldots, Y_T)$. Now define the following operators:

$$
\begin{aligned}
L_{123} : \mathcal{L}_{W_3} \to \mathcal{L} \qquad & [L_{123}m](w_1) = \int \frac{f_{YD|Z}(y,d;z)}{\prod_{t=2}^T f_{D_t|Y^{t-1}D^{t-1}Z}(d_t; y^{t-1}, d^{t-1}, z) f_{D_1|Z}(d_1;z)} m(w_3)dw_3 \\
L_{13} : \mathcal{L}_{W_3} \to \mathcal{L} \qquad & [L_{13}m](w_1) = \int\int \frac{f_{Y^TD^T|Z}(y^t,d^t;z)}{\prod_{t=2}^T f_{D_t|Y^{t-1}D^{t-1}Z}(d_t; y^{t-1}, d^{t-1}, z) f_{D_1|Z}(d_1;z)} dy_{p+1} m(w_3)dw_3 \\
L_{1\lambda} : \mathcal{L} \to \mathcal{L} \qquad & [L_{1\lambda}m](w_1) = \int \prod_{t=1}^p f_{Y_t(d_t)|Z\lambda}(y_t; z, v) m(v)dv \\
L_{\lambda 3} : \mathcal{L}_{W_3} \to \mathcal{L} \qquad & [L_{\lambda 1}m](v) = \int \prod_{t=p+2}^T f_{Y_t(d_t)|Z\lambda}(y_t; z, v) f_{\lambda|Z}(v) m(w_1)dw_1 \\
D_\lambda : \mathcal{L}_\Lambda \to \mathcal{L}_\Lambda \qquad & [D_\lambda m](v) = f_{Y_{p+1}(d_{p+1})|Z\lambda}(y_{p+1}; z, v) m(v)
\end{aligned}
$$

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

$$
\begin{aligned}
f_{YD|Z}(y,d;z) &= \int f_{YD\lambda|Z}(y,d,v;z)dv \\
&= \int f_{Y_T|DY^{T-1}Z\lambda}(y_T; d, y^{T-1}, z, v) f_{D_T|Y^{T-1}D^{T-1}Z\lambda}(d_T; y^{T-1}, d^{T-1}, z, v) \\
&\quad \times f_{Y_{T-1}|D^{T-1}Y^{T-2}Z\lambda}(y_{T-1}; d^{T-1}, y^{T-2}, z, v) \ldots f_{\lambda|Z}(v;z)d\lambda \\
&= \int f_{Y_T(d_T)|D^TY^{T-1}Z\lambda}(y_T; d^T, y^{T-1}, z, v) f_{D_T|Y^{T-1}D^{T-1}Z\lambda}(d_T; y^{T-1}, d^{T-1}, z, v) \\
&\quad \times f_{Y_{T-1}(d_{T-1})|D^{T-1}Y^{T-2}Z\lambda}(y_{T-1}; d^{T-1}, y^{T-1}, z, v) \ldots f_{\lambda|Z}(v;z)d\lambda \\
&= \int f_{Y_T(d_T)|Z\lambda}(y_T; z, v) f_{D_T|Y^{T-1}D^{T-1}Z}(d_T; y^{T-1}, d^{T-1}, z) \\
&\quad \times f_{Y_{T-1}(D_{T-1})|Z\lambda}(y_{T-1}; z, v) \ldots f_{\lambda|Z}(v;z)d\lambda
\end{aligned}
$$

The second equality holds by the law of total probability, the third holds by definition of $Y_t(d)$. The final equality holds since Assumption L1 has the following two implica-

tions: first, $(\epsilon_t \perp\!\!\!\perp \eta^t, \epsilon^{t-1}, \lambda \mid Z) \Rightarrow (\epsilon_t \perp\!\!\!\perp \eta^t, \epsilon^{t-1} \mid Z, \lambda) \Rightarrow (Y_t \perp\!\!\!\perp D^t, Y^{t-1} \mid Z, \lambda)$; and second, $\eta_t \perp\!\!\!\perp \eta^{t-1}, \epsilon^{t-1}\lambda \mid Z \Rightarrow (\eta_t \perp\!\!\!\perp D^{t-1}, Y^{t-1}\lambda \mid Z) \Rightarrow (\eta_t \perp\!\!\!\perp \lambda \mid Z, D^{t-1}, Y^{t-1}) \Rightarrow (D_t \perp\!\!\!\perp \lambda \mid Z, D^{t-1}, Y^{t-1})$. From this, it follows by Assumption L4(A) that

$$\frac{f_{YD|Z}(y,d;z)}{\prod_{t=2}^{T} f_{D_t|Y^{t-1}D^{t-1}Z}(d_t; y^{t-1}, d^{t-1}, z) f_{D_1|Z}(d_1; z)} = \int \prod_{t=1}^{T} f_{Y_t(d_t)|Z\lambda}(y_t; z, v) f_{\lambda|Z}(v; z) dv. \tag{8}$$

And therefore that

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

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

From here, under Assumptions L1, L2, L3, L4(B), and L5 imply the conditions of Freyberger (2018, Theorem 1) are satisfied, so that $f_{Y_t(d_t)|Z,\lambda}$ and $f_{\lambda|Z}$. From Assumptions L2 (C) and L4 (C), $\alpha_t(d_t), \beta_t(d_t)$ and $F_t(d_t)$ are identified.

Next, given an arbitrary t and d_t, define $\tilde{d}$ by replacing the tth element of d with d_t. Then let ρ be a permutation $(1, 2, \ldots, T) \mapsto (t_1, t_2, \ldots, t_T)$ such that $t \mapsto t_1$ and define $V_1 = (Y_{t_1}, Y_{t_2}, \ldots, Y_{t_p})$ and $V_2 = (Y_{t_{p+1}}, Y_{t_{p+1}}, \ldots, Y_{t_T})$

$$\tilde{L}_{21} : \mathcal{L}_{V_1} \to \mathcal{L} \qquad [\tilde{L}_{21} m](v_2) = \int \frac{f_{YD|Z}(y,d;z)}{\prod_{t=2}^{T} f_{D_t|Y^{t-1}D^{t-1}Z}(d_t; y^{t-1}, d^{t-1}, z) f_{D_1|Z}(d_1; z)} m(v_1) dv_1$$

$$\tilde{L}_{2\lambda} : \mathcal{L} \to \mathcal{L} \qquad [\tilde{L}_{2\lambda} m](v_2) = \int \prod_{i=p+1}^{T} f_{Y_{t_i}(d_{t_i})|Z\lambda}(y_{t_i}; z, v) f_{\lambda|Z}(v) m(v) dv$$

$$\tilde{L}_{\lambda 1} : \mathcal{L}_{V_1} \to \mathcal{L} \qquad [\tilde{L}_{\lambda 1} m](v) = \int \prod_{i=1}^{p} f_{Y_{t_i}(d_{t_i})|Z\lambda}(y_{t_i}; z, v) m(v_1) dv_1$$

As before, $\tilde{L}_{21} = \tilde{L}_{2\lambda}\tilde{L}_{\lambda 1}$. The proceeding arguments imply $\tilde{L}_{2\lambda}$ is identified and injective, and under Assumption L4(A) $\tilde{L}_{21}$ is identified. It follows that $\tilde{L}_{\lambda 1}$ is identified as $\tilde{L}_{2\lambda}^{-1}\tilde{L}_{21} = \tilde{L}_{\lambda 1}$ yielding identification of $\alpha_t(d_t), \beta_t(d_t), F_t(d_t), f_{\epsilon(d_t)|Z\lambda}$.

□

B Estimation

B.1 Consistency of sieve MLE

In this section we introduce conditions for the sieve maximum likelihood estimator (4) to be consistent for the true model parameters. We begin by imposing smoothness restrictions on the unknown functions. To do so, given $\gamma > 0$, $\omega \geq 0$ and $\mathcal{X}$ a subset of a Euclidean space, let $\Lambda^\lambda(\mathcal{X})$ denote a Hölder space equipped with the Hölder norm $\|h\|_{\Lambda^\gamma}$ (that is, for k the largest integer smaller than γ, $\Lambda^\lambda(\mathcal{X})$ is a space of functions $h\colon \mathcal{X} \to \mathbf{R}$ having at least k continuous derivatives, the kth of which is Hölder continuous with exponent $\gamma - k$). Then define a weighted Hölder ball with radius $c \in (0, \infty)$ as $\Lambda_c^{\gamma,\omega}(\mathcal{X}) = \{h \in \Lambda^\gamma(\mathcal{X})\colon \|h(\cdot)[1 + \|\cdot\|_E^2]^{-\omega}\|_{\Lambda^\gamma} \leq c\}$, where $\|\cdot\|_E$ is the Euclidean norm.

Without loss of generality, suppose the CCP function $\bar{h}_t(d^t, z^t, y^{t-1}, v_k)$ depends on (d^t, z^t, y^{t-1}) via some vector-valued function $(d^t, z^t, y^{t-1}) \mapsto j_t(d^t, z^t, y^{t-1})$ where j_t is known up to $\left((\alpha_{st}, \beta_{st}, F_{kst}, F_{ust}, \sigma_{st})_{st=1}^T, \Sigma_u\right)$. This is without loss of generality since j_t may be identity. Other examples include rational learning where $j_t(d^t, z^t, y^{t-1}) \in \mathbb{R}^{p(p+3)/2}$ are sufficient statistics for λ_u given realized random variables at time t, and a sort of myopia where $j_t(d^t, z^t, y^{t-1}) \in \mathbb{R}^3$ depends only on (d_t, z_t, y_t). Denote $X_{it} \equiv j_t(D_i^t, Z_i^t, Y_i^{t-1})$ and

$$\mathcal{H}_t = \Lambda_c^{\gamma_1,\omega_1}\left(\text{Supp}(\lambda_k) \times \text{Supp}(X_{it})\right),$$
$$\mathcal{F}_\lambda = \Lambda_c^{\gamma_2,\omega_2}\left(\text{Supp}(\lambda_k)\right).$$

The use of a weighted Holder space allows us to enable the support of $\lambda_{k,i}$ and X_{it} to be unbounded. Though not required for consistency, Assumption E7 places restrictions on (γ_1, γ_2), the parameters that govern the smoothness of the unknown functions.

To simplify notation we make the following assumption which strengthens Assumption KL1:

Assumption E1. For some $X_{it} \in \mathbb{R}_t$, $\bar{h}(d^t, z^t, y^{t-1}, v_k) = \bar{h}(x_{i,t}, z_t, d_t, v_k)$ For any t, $F_{Z_{t+1}|Y_t D_t Z_t} = F_{Z_{t+1}|D_t Z_t}$ and $\text{Supp}(Z_t)$ is finite.

Assumption E1 reduces the dimension of the problem by assuming that the conditional distribution of the state variable Z_t is a parametric object. This is common in applied settings, such as dynamic discrete choice models.

With this assumption define $\mathcal{G}_t$ to be the set of conditional distribution functions $(d_t, z_t) \mapsto z_{t+1}$. Define k_0 to be the cardinality of $\text{Supp}(Z_{i1})$ and $k_{1,t}$ to be the cardinality of $\text{Supp}(D_{it}) \times \text{Supp}(Z_{it})$. Notice that $\Theta = \Theta_1 \times \mathcal{G}_1 \times \cdots \times \mathcal{G}_{T-1} \times \mathcal{H}_1^{k_{1,1}} \times \cdots \times \mathcal{H}_T^{k_{1,T}} \times \mathcal{F}_\lambda^{k_0}$ and we denote an element of Θ as $\theta = (\theta_1, g_1, \ldots, g_{T-1}, \bar{h}_1, \ldots, \bar{h}_T, f_\lambda)$. Let d_L indicate the Levy metric. Define the following norms on $\mathcal{H}_t^{k_{1,t}}$ and $\mathcal{F}_\lambda^{k_0}$ as follows:

$$\|\bar{h}_t\|_{\infty,\omega_1} = \sup_{\substack{z \in \text{Supp}(Z_{it}) \\ d \in \text{Supp}(D_{it})}} \|\bar{h}_t(d, z, \cdot, \cdot)[1 + \|\cdot\|_E^2]^{-\omega_1}\|_\infty,$$

$$\|f_\lambda\|_{\infty,\omega_2} = \sup_{z \in \text{Supp}(Z_{i1})} \|f_\lambda(z, \cdot)[1 + \|\cdot\|_E^2]^{-\omega_2}\|_\infty,$$

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

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

for scalars $\tilde{\omega}_1, \tilde{\omega}_2$.

Let $\mathcal{H}_{n,t}$ and $\mathcal{F}_{n,\lambda}$ be sieve spaces for $\mathcal{H}_t$ and $\mathcal{F}_\lambda$ respectively. Then $\Theta_n = \Theta_1 \times \mathcal{G}_1 \times \cdots \times \mathcal{G}_{T-1} \times \mathcal{H}_{n,1}^{k_{1,1}} \times \ldots \mathcal{H}_{n,T}^{k_{1,T}} \times \mathcal{F}_{n,\lambda}^{k_0}$ 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.

Let $W_{it} = (D_{it}, Z_{it}, Y_{it})$, so that W_i^t is the period-t history of assignment, state variables and outcomes. Denote $W_i = (W_{it})_{t=1}^T$.

Assumption E3. $W_i = (D_{it}, Z_{it}, Y_{it})_{t=1}^T$ are iid.

Assumption E4. $\Theta_n \subseteq \Theta_{n+1} \subseteq \Theta$ for each $n \geq 1$, Θ_n are compact under d, and $\lim_{n\to\infty} \min_{\theta\in\Theta_n} d(\theta, \theta_0) = 0$.

Assumption E5. $\mathrm{E}[\ell(\theta, W_i)]$ is continuous at $\theta = \theta^*$

Assumption E6.

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

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

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

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

B.2 Plug-in sieve estimator

We assume a linear sieve space and limit its complexity.

Assumption E7. Let q be the length of the vector X_{it}. (1) For each t, $\mathcal{H}_{n,t}$ is a linear sieve of length J_{Hn} and $F_{n,\lambda}$ is a linear sieve of length $J_{\lambda n}$. Furthermore, $J_{Hn} = O(n^{\frac{1}{2\gamma_1/q+1}})$ and $J_{\lambda n} = O(n^{\frac{1}{2\gamma_2+1}})$. (2) $\min\{\gamma_1/q, \gamma_2\} > 1/2$.

Assumption E7 ensures the sieve spaces grow do not grow too quickly. To achieve this rate of growth, the functions are assumed to have particular smoothness. Recall that

identification requires $q \geq p$, where p is the dimension of λ_k, the initially-unknown latent factor. Part (2) of Assumption E7 requires that the CCP functions have adequate smoothness to compensate for the dimensionality of λ_k. In applications, it is common to assume a parametric model for $\bar{h}_t$, in which case the above curse-of-dimensionality is avoided.

The following strengthens Assumption E4.

Assumption E8. (1) $\min_{\theta \in \Theta_n} d(\theta, \theta^*) = o(n^{-1/4})$. (2)

This assumption ensures the sieve space grows sufficiently fast.

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

$$\langle \theta_1 - \theta^*, \theta_2 - \theta^* \rangle = -\frac{\partial^2}{\partial \tau_1 \partial \tau_2} E\left[\ell\left(\theta^* + \tau_1\left(\theta_1 - \theta^*\right) + \tau_2\left(\theta_2 - \theta^*\right), W\right)\right]\Big|_{\tau_1=0, \tau_2=0}, \quad (9)$$

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

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

Assumption E9. There is $C_1 > 0$ such that for all small $\varepsilon > 0$

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

Assumption E10. For any $\delta > 0$, there exists a constant $s \in (0, 2)$ such that

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

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

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

Theorem 5. Suppose the distribution of $(Y_t, D_t, Z_t)_{t=1}^T$ is observed for $T \geq 2p+1$ and that Assumptions KL1-KL5 and Assumptions E1-E10 hold.Then $\|\hat{\theta} - \theta^*\| = o_p(n^{-1/4})$

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

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

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

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

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

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

Under the following regularity conditions (Assumption E10), $\mathcal{V}$ and $\mathcal{V}_n$ are Hilbert spaces under the inner product defined in equation 9.

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

Assumption E10 provides sufficient conditions for the set $\mathcal{V}$ to be a Hilbert space under $\langle \cdot, \cdot \rangle$. 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(1).

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

(3) $$\sup_{\theta \in \mathcal{N}_0} \frac{\left|f(\theta) - f(\theta^*) - \frac{\partial f(\theta^*)}{\partial \theta}[\theta - \theta^*]\right|}{\|v_n^*\|} = o(n^{-1/2}).$$

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

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

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

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

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

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

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