Group-Specific Heterogeneity in Short Binary Outcome Panels*

Bjoern Hoeppner†
Bonn Graduate School of Economics, University of Bonn

November 7, 2025

(Most recent version: Google Drive / GitHub)

– Job Market Paper –

Abstract

When the number of time periods is small, identification of binary outcome panel data models that allow for heterogeneity is inherently difficult. This paper presents identification and estimation results for such models in the presence of latent group-specific heterogeneity. We assume that each unit belongs to a time-invariant latent group, so the joint distribution of the binary outcomes conditional on the covariates is a finite mixture model. Under mild conditions, the group-specific conditional outcome models are fully nonparametrically identified, with identification being possible with as few as two time periods. We also provide conditions under which group-specific average marginal effects are identified. In addition, we develop a novel semiparametric estimator for the setup of interest and study its asymptotic properties. Simulations indicate that the estimator performs well in finite samples. We illustrate the estimation procedure and how to interpret the latent groups empirically in the context of homeownership in Germany.

Keywords: Binary Outcome, Group-Specific Heterogeneity, Identification, Mixture Models, Semiparametric Estimation, Short Panels

*I am grateful to my PhD advisors, Joachim Freyberger and Christoph Breunig, for their invaluable guidance and support throughout the preparation of this paper. Additionally, I thank Antonia Antweiler, Christina Brinkmann, Julian Budde, Julius Kappenberg, Chen Lin, Vladislav Morozov, Claudia Noack, Jan Scherer, Nico Thurow, Maximilian Voigt, and the participants of the Bonn Statistics Brown Bag Seminar, the Econometrics Reading Group at the University of Bonn, the 2024 Bonn-Frankfurt-Mannheim PhD Conference, the Bonn–Mannheim–Munich Workshop in Microeconometrics, the IAAE 2025, the Statistische Woche 2025, the Bonn Applied Micro Coffee, and the speakers of the Bonn Statistics Seminar. All errors are my own. *Comments are welcome!*

†Address: Adenauerallee 24-42, 53113 Bonn, Germany. e-Mail: b.hoeppner@uni-bonn.de

1 Introduction

The units we observe are typically heterogeneous in dimensions that are unobserved. To capture this latent heterogeneity in panel data models, researchers often include unit- or time-specific fixed effects in their model specification. However, in short nonlinear panel models with a fixed number of time periods, identification may fail when heterogeneity is introduced. More precisely, in binary outcome linear latent utility models with bounded covariates, identification of the slope coefficient fails in the presence of additive unit- and time-specific fixed effects with two time periods unless the error term is logistically distributed (Chamberlain 2010). Even in the logistic setting, average marginal effects are generally not point identified (Davezies et al. 2024; Dobronyi et al. 2024).

One way to reduce the dimensionality of latent heterogeneity in nonlinear panel models is to impose a group structure. In empirical applications, units can often be meaningfully grouped into a finite number of latent types and are heterogeneous only across these types; for instance, firms with similar production technologies, individuals with similar ability levels, or households with similar consumption behavior. While methods capturing such group-specific heterogeneity have become increasingly popular in econometrics (Bonhomme and Manresa 2015; Su et al. 2016), this literature considers panels with a large number of units and time periods, which can be restrictive in empirical applications.

This paper develops nonparametric identification and semiparametric estimation results that allow for latent group-specific heterogeneity in short binary outcome panel models, that is, settings in which the number of time periods, T, is small and fixed. We assume that each unit can be assigned to one of J disjoint and time-invariant groups, with the group membership being unobserved. We treat J as known for most of our analysis, although we show that it can be identified. In particular, given a sequence of binary outcomes $\{y_{it}\}_{t=1}^{T}$ and a sequence of covariates $\{x_{it}\}_{t=1}^{T}$, we focus on the following class of conditional binary outcome finite mixture models:

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^{T} \mid \{x_{it}\}_{t=1}^{T}) = \sum_{j=1}^{J} \mathbb{P}(g_i = j \mid \{x_{it}\}_{t=1}^{T}) \prod_{t=1}^{T} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j) \qquad (1)$$

with $\{s_t\}_{t=1}^{T} \in \{0,1\}^T$, $i = 1, 2, \ldots, n$, and $g_i \in \{1, \ldots, J\}$ indicating the group of unit i. The

collection of $\mathbb{P}(g_i = j \mid \{x_{it}\}_{t=1}^T)$ for $j = 1, \ldots, J$ is referred to as the component weights, while we refer to $\mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j)$ for $j = 1, \ldots, J$ as the component distributions, which are allowed to vary over time.

Our leading example of model (1) is the linear latent utility model

$$y_{it} = \mathbb{1}(x_{it}^\top \beta_{g_i,t} + \alpha_{g_i,t} - \varepsilon_{it} \geq 0)$$

with time-varying group-specific intercepts as well as slope coefficients. Assuming that $\{\varepsilon_{it}\}_{t=1}^T$ are mutually independent conditional on $\{x_{it}\}_{t=1}^T$ and g_i, and ε_{it} is independent of $\{x_{it}\}_{t=1}^T$ conditional on g_i, this model falls into the class of models in equation (1). For example, when the errors ε_{it} are standard normally distributed conditional on g_i, the component distributions $\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j)$ are probit models with time-varying group-specific intercepts and slope coefficients. Importantly, our identification analysis leaves the dependence between the groups and the covariates almost unrestricted by putting minimal assumptions on the component weights. A special case of this leading example, which is closely related to a setup with additive fixed effects, is a model that only allows the intercept $\alpha_{g_i,t}$ to vary across groups. $\alpha_{g_i,t}$ is referred to as a time-varying group-specific fixed effect (GFE). Unlike unit-specific fixed effects, GFEs are allowed to vary over time but are identical for units within the same group.

Model (1) accommodates broad forms of endogeneity as long as they are fully captured by the latent group structure. For instance, when consumers behave differently across different market segments that are unobserved or only observed up to measurement error, model (1) allows for endogenous selection into these market segments, which correspond to the groups in model (1) (Wedel and Kamakura 2000; Leisch et al. 2018). Alternatively, the group structure can account for an omitted variable bias due to an unobserved discrete regressor. Within this setting, the groups correspond to the support points of this discrete regressor.

We show that the component weights and distributions are nonparametrically identified under mild conditions. To be precise, under the additional simplifying assumption that the component weights are strictly positive, our main identification result implies that the component distributions are identified on the entire covariate support as long as there exists a *single* covariate value at which the component distributions exhibit sufficient variation

across groups. The component weights, on the other hand, are identified at all covariate values at which the component distributions are linearly independent. In the most general setting, we require $T \geq 2\lceil \log_2(J) \rceil + 1$ for identification where $\lceil z \rceil$ is the smallest integer larger than or equal to z. This bound coincides with the bound in Allman et al. (2009) for finite mixtures of binary outcome models in the absence of covariates. Interestingly, we show that this dependence can be relaxed to $T \geq 2\lceil \log_2(J) \rceil$ if the component weights exhibit sufficient variation in the covariates. For two-component binary outcome mixture models, this translates to identification when T is as small as 2; a result that, to the best of our knowledge, is new in the literature.[1]

Once the component distributions are identified, group-specific conditional marginal effects are identified. In empirical applications, it may also be of interest to identify group-specific average marginal effects. We therefore show how to translate the identification of the component weights and distributions into the identification of group-specific average marginal effects.

Our identification argument requires no additional conditions to resolve the so-called *intra-component label switching problem*. Specifically, model (1) has an inherent identification issue: Reassigning the group labels does not change the observed distribution $\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid \{x_{it}\}_{t=1}^T)$. In the presence of covariates, this generic label switching problem cannot be resolved by enforcing an ordering constraint at each covariate value, as doing so does not ensure that the group assignment is consistent across the different covariate values – the intra-component label switching problem (Grün and Leisch 2008a). We use an exclusion restriction due to the panel structure of model (1) to solve this problem. In doing so, our identification argument does not make any assumptions on the smoothness of the component distributions or weights, on the number of covariates in comparison to the number of groups J, does not require the existence of a special regressor or the need of an additional ordering restriction, and allows for discrete as well as continuously distributed regressors. A subset of these assumptions is typically imposed in the existing literature (Grün and Leisch 2008b; Huang et al. 2013; Wang et al. 2014; Gormley and Frühwirth-Schnatter 2019).

To highlight the wide applicability of our identification results, we present various exten-

[1] In this paper, we use the terms *finite mixture model* and *mixture model* interchangeably.

sions of our baseline model. These extensions demonstrate that the number of time periods required for identification depends crucially on the underlying modeling assumptions. *First*, our results generalize to settings with a non-binary outcome variable. More precisely, let S be the cardinality of the support of y_{it} for all t. The component distributions and weights may then be identified when T is as small as $2\lceil \log_S(J) \rceil$. *Second*, when the component weights do not depend on all covariates and the support of the excluded covariates is sufficiently rich, identification in the binary outcome setting may already be possible when $T = 2$ for $J > 2$. Intuitively, when the component weights do not depend on all covariates, the variation in the excluded covariates compensates for the limited variation in the outcome, similar to Kasahara and Shimotsu (2009). *Third*, we extend the binary outcome model to allow for lagged covariates and a dynamic panel model. In the latter case, we require $T \geq 4\lceil \log_2(J) \rceil - 1$ for identification. *Fourth*, we provide conditions similar to Kasahara and Shimotsu (2009), Kasahara and Shimotsu (2014), and Kwon and Mbakop (2021) so that J is identified.

Beyond identification results, we present a novel sieve-based semiparametric estimator for the component weights and distributions in model (1). Motivated by our leading example, this estimator leaves the component weights nonparametric and assumes a parametric structure for the component distributions. This class of estimators nests, for instance, probit and logit models with time-varying group-specific parameters. Additionally, we provide an estimator for group-specific average marginal effects. Using arguments from the literature on sieve estimation (Chen and Shen 1998; Ai and Chen 2003; Newey and Powell 2003; Chen 2007), we show that the proposed estimators are consistent, derive their convergence rates, and present asymptotic normality results for the estimators of the parameters of the component distributions and the group-specific average marginal effects. In a simulation study, we demonstrate that the proposed estimators perform well in finite samples.

To illustrate how to interpret the latent groups empirically and to estimate the number of groups J when it is unknown, we use the proposed estimators to study homeownership in Germany. Among high-income countries, Germany has one of the lowest homeownership rates, at around 47.2% in 2024 (Eurostat 2025). Using data from the German Socio-Economic Panel, we identify three different latent types of households: (i) higher-income and older households that have a high propensity to own; (ii) lower-income households with higher

long-term unemployment that have a high propensity to rent; and (iii) younger, medium- to high-income households that are transitioning into homeownership. In line with economic intuition, the estimates of the group-specific average marginal effects indicate that the homeownership decision of the last group is most responsive to income transfers or other changes in their covariate set. Intuitively, these households are close to their borrowing constraint which may become tighter/looser when the covariates change. Conversely, households in the other latent groups face either very loose or strongly binding borrowing constraints, making their homeownership decisions relatively insensitive to marginal covariate changes. From a policy perspective, if the goal is to effectively increase the homeownership rate, our analysis suggests that a policy should target younger households with medium to high income.

Overall, the empirical application highlights that model (1) allows a researcher to uncover the latent groups that are more receptive to an intervention and illustrates how the component weights can be informative about the group membership of a unit, thus allowing for a more targeted and potentially welfare-improving intervention assignment. In addition, the empirical application illustrates how the group structure can capture important sources of endogeneity, such as unobserved wealth and borrowing constraints.

Literature. While, to the best of our knowledge, model (1) has not been studied in the literature, Hall and Zhou (2003), Hall et al. (2005), Allman et al. (2009), and Bonhomme et al. (2016), among others, study finite mixture models that factorize similarly but do not condition on covariates. Some of our identification arguments are closely related to this literature. However, due to the conditioning on covariates, we have to solve the intra-component label switching problem. At the same time, we can leverage the variation of the component weights in the covariates to strengthen the identification results. In particular, we require fewer time periods for identification, and the standard identification conditions for unconditional mixture models are not required to hold on the entire covariate support.

Kasahara and Shimotsu (2009) study identification of mixtures of dynamic discrete choice models. Similar to the previously mentioned papers, they focus on the mixture structure of the joint distribution of the outcome variable and the covariates across all time periods, whereas we focus on the mixture structure of the conditional distribution of the outcome variable given the covariates. The group structure induced by the joint need not be the same

as the group structure induced by the conditional distribution. Similarly, Hu and Shum (2012) and Higgins and Jochmans (2023) focus on the unconditional distribution in a dynamic discrete choice setting and improve some of the identification results in Kasahara and Shimotsu (2009). Our identification arguments are also related to the tensor decomposition arguments in Leurgans et al. (1993) and similar spectral decompositions in the measurement error literature (Hu 2008; Carroll et al. 2010; Hu and Shum 2012; Freyberger 2018).[2]

Our paper is related and contributes to various strands of the literature. First, we contribute to the growing literature on *latent group-specific heterogeneity* in economics (among many others, Hahn and Moon (2010), Bonhomme and Manresa (2015), Saggio (2016), Su et al. (2016), and Mugnier (2023)). In particular, our leading example that includes only time-varying GFEs is a special case of the models studied in Saggio (2016) and Mugnier (2023). Different from us, these papers focus on large panels with $T \to \infty$, which they leverage to consistently estimate the group membership of each individual. We argue that many objects of interest are already identified in short panels with T as small as 2. This is of empirical relevance, as many panels we observe are not large and the assumption that the group membership is time-invariant is more palatable in short panels. Additionally, we develop a new semiparametric estimator for the short panel setting. Second, apart from the aforementioned literature on *nonparametric identification of mixture models* of the form of equation (1) in the absence of covariates, our identification argument is related to the partial identification arguments of Henry et al. (2014) in that it leverages an exclusion restriction in the component distributions. Our identification results complement Aguiar and Kashaev (2025) who develop identification results for conditional mixture models in the context of dynamic discrete choice models with unobserved choice sets.[3] Compared to Browning and Carro (2014), who discuss local identification in a dynamic binary outcome panel mixture model with and without covariates, we present global identification results. Like Kitamura and Laage (2018), whose nonparametric identification results do not apply to our setting,

[2]We refer to Hu (2025) for an instructive overview on the measurement error literature and the respective spectral decomposition arguments.

[3]Aguiar and Kashaev (2025) do not impose an exclusion restriction on the component distributions. Such an exclusion restriction arises naturally in our setting. Unlike us, however, they impose a stationarity assumption on the component distributions, require at least five time periods for identification, and solve the intra-component label switching problem by leveraging the structural interpretation of the choice sets.

our results highlight the identification power of covariates. Third, the class of models we consider can be interpreted as a *mixture of expert models* – see, for instance, Gormley and Frühwirth-Schnatter (2019) or Yao and Xiang (2024) for textbook discussions of these models, and Grün and Leisch (2008b) for a parametric setting that is more closely related to ours. In the context of panel data models, our identification arguments offer a systematic framework for identifying nonparametric mixtures of expert models and solving the intra-component label switching problem, something that seems to be lacking in the literature (Gormley and Frühwirth-Schnatter 2019). Last, our results complement the recent literature on partial and point identification in the context of *static and dynamic logit* as well as *general binary outcome models in the presence of additive unit-specific fixed effects* (Khan et al. 2023; Aguirregabiria and Carro 2024; Davezies et al. 2024; Dobronyi et al. 2024).

The remainder of this paper is organized as follows. The next section introduces the model and some examples that fit our modeling framework. In Section 3, we discuss our main identification results. Section 4 proposes a semiparametric estimator and analyzes its asymptotic properties. Section 5 studies the proposed estimator in a simulation study, while Section 6 illustrates the estimator in the context of an application. Section 7 presents additional identification results that go beyond the baseline setting. Section 8 concludes. The appendix includes proofs and additional discussions.

2 Model

This section introduces and motivates the class of mixture models considered in this paper. In particular, we discuss a leading example from this class and illustrate how a group structure can arise empirically.

For the remainder of this paper, unless noted otherwise, we assume that a researcher observes a sample $\{(y_{it}, x_{it}^\top)\}_{t=1}^T$ for $i = 1, \dots, n$ where $x_{it} \in \text{Supp}(x_{it}) =: \mathbb{X}_t \subseteq \mathbb{R}^K$ is a vector of covariates and does not include a constant. For all $t = 1, \dots, T$, $y_{it} \in \{0, 1\}$ is a binary outcome variable. We denote by $\{y_{it}\}_{t=1}^T \in \{0, 1\}^T$ the vector of observed outcomes over time. While the sampling scheme is irrelevant for the identification arguments, we use the index i to distinguish between random variables and their realizations, for instance, x_{it} is a random variable and x_t is a realization of x_{it}.

We recall the model of interest

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i) = \sum_{j=1}^J \mathbb{P}(g_i = j \mid X_i) \prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j) \tag{2}$$

for $\{s_t\}_{t=1}^T \in \{0,1\}^T$ and $X_i = (x_{i1}^\top, \ldots, x_{iT}^\top)^\top \in \text{Supp}(X_i) =: \mathbb{X} \subseteq \mathbb{R}^{TK}$. We assume J to be known. In Section 7.4, we discuss that J is identified as the rank of an observed matrix under appropriate conditions. While we do not explicitly treat J as a function of the covariates, our identification results allow for such a setting. In this case, J denotes the maximum number of distinct groups over the support of X_i.

We define $\pi_j(X_i) = \mathbb{P}(g_i = j \mid X_i)$ and refer to it as the *component weight* of group j. $\mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j)$ for $t = 1, \ldots, T$ are the *component distributions* of group j. Crucially, we allow the component distributions to vary over time, which allows us to cover a setting with time-varying GFEs. If unit i belongs to group j, then, at time t, y_{it} given x_{it} is drawn from $\mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j)$.

To fix ideas, we state the main modeling assumption, which we implicitly assume to hold for the remainder of this text if not indicated otherwise.

Assumption M-1 *(Model) Conditional on X_i, the distribution of $\{y_{it}\}_{t=1}^T$ follows the mixture model in equation (2) with a known number of groups J.*

Assumption M-1 requires that the conditional component distribution $\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i, g_i = j)$ factorizes into the conditional marginal distributions given x_{it} and g_i. We will leverage this factorization in our identification argument. As Assumption M-1 may be too strong in some settings, we relax it in our extensions by allowing for a dynamic setting (Section 7.3) or lagged covariates in the conditioning set (Appendix A.1.8).

To motivate the class of models in equation (2), we present our leading example for a data generating process that falls into this class.

Example 2.1 *(Linear latent utility model with time-varying group-specific coefficients and error terms) We consider a linear latent utility model in which both the coefficients and the distributions of the error terms are permitted to be group-specific and time-varying, that is,*

$$y_{it} = \mathbb{1}(x_{it}^\top \beta_{g_i,t} + \alpha_{g_i,t} - \varepsilon_{g_i,it} \geq 0) = \sum_{j=1}^J \mathbb{1}(g_i = j)\mathbb{1}(x_{it}^\top \beta_{j,t} + \alpha_{j,t} - \varepsilon_{j,it} \geq 0)$$

where J is known and g_i may be (arbitrarily) correlated with X_i. We make the following assumptions: (i) $\varepsilon_{j,it} \perp\!\!\!\perp X_i \mid g_i$, and (ii) $\{\varepsilon_{j,it}\}_{t=1}^T$ are mutually independent conditional on $(X_i^\top, g_i)$. These assumptions are generalizations of standard textbook assumptions to the case with latent groups. Importantly, they allow the covariates to be endogenous. However, once we condition on the unobserved group membership, this endogeneity problem vanishes.[4] *Letting F_{jt} denote the distribution of $\varepsilon_{j,it}$ conditional on $g_i = j$, the assumptions imply*

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i) = \sum_{j=1}^J \pi_j(X_i) \prod_{t=1}^T F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t})^{s_t}(1 - F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t}))^{1-s_t}$$

where the component distributions are $\mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j) = F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t})^{s_t}(1 - F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t}))^{1-s_t}$.[5] *When $\varepsilon_{j,it}$ follows a standard normal (logistic) distribution, this model nests a time-varying group-specific probit (logit) model. When the distribution of $\varepsilon_{j,it}$ is unknown, the model reduces to a single index model with time-varying group-specific index parameters and link functions. More broadly, we can interpret this model as a correlated random coefficient binary choice model where the support of the coefficients is discrete.*

Ignoring the latent heterogeneity in model (2) or Example 2.1 may introduce a heterogeneity bias. In the context of estimating average marginal effects (AMEs), this implies that a standard analysis that ignores the latent group structure does not estimate the group-averaged AME; see Appendix A.1.2 for a detailed discussion.

We conclude this section with a few empirical examples that illustrate how a group structure may arise in applications.

Example 2.2 *(Endogenous selection models) In the context of Example 2.1, the groups may capture the purchasing behavior of consumers in different geographic markets (Taylor 2017) or market segments (Wedel and Kamakura 2000; Leisch et al. 2018) where consumers are heterogeneous across markets but homogeneous within a market. Often, it is not observed (or only up to measurement error) which market a unit is a member of. Such a setting falls into the class of models we study in this paper. Importantly, we do not make any assumptions on*

[4] If the endogeneity problem persists even after conditioning on g_i, our results can be extended using a control function approach.

[5] Throughout, we view group-specific parameters as parameters to be estimated. Alternatively, one may interpret these group-specific parameters as random variables. Then, our analysis is conditional on the realizations of these random variables.

how consumers sort into different geographic regions or market segments but remain agnostic about the sorting mechanism. Specifically, we allow the selection to be endogenous as long as the source of endogeneity is fully captured by the group membership.

Example 2.3 *(Omitted variable bias) The group membership may also be motivated through an unobserved discrete covariate. For example, information on race, ethnicity, or religion is sometimes not recorded due to privacy concerns (Fiscella and Fremont 2006) or simply missing in historic data sets (Abramitzky et al. 2024; Cummins and Gráda 2025). In such a setting, the groups in model (2) correspond to the unobserved ethnic groups. Our results imply the identification of the ethnicity-specific outcome models.*

In empirical practice, researchers commonly try to overcome the missingness of ethnicity by imputing the missing covariate using the name of an individual or its recorded address, say R_i, and then use the imputed ethnicity as a covariate (Fiscella and Fremont 2006; McCartan et al. 2024). Since these predictions may be incorrect, this estimation scheme results in inconsistent estimators (McCartan et al. 2024). However, as R_i may indeed be informative about the ethnicity of an individual, one can use it in our analysis by conditioning on it. Under similar assumptions as in Example 2.1 and the additional assumption that $\{\varepsilon_{j,it}\}_{t=1}^{T} \perp\!\!\!\perp R_i \mid X_i, g_i$ for all j,[6] the conditional outcome distribution is a mixture model of the form $\mathbb{P}(\{y_{it} = s_t\}_{t=1}^{T} \mid X_i, R_i) = \sum_{j=1}^{J} \pi_j(X_i, R_i) \prod_{t=1}^{T} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j)$, which we study as a direct extension of model (2) in Appendix A.1.6.

Furthermore, settings in which treatment indicators are not observed or only observed up to some measurement error may fall into the class of mixture models we study. Then, the groups correspond to the different treatment arms. In addition, group membership may represent a latent class that is potentially correlated with covariates; for example, in the context of university dropout, it is well understood that ability is one of the determinants of why students drop out of university (Aina et al. (2022) and references therein). In this case, different ability levels, for instance, low, medium, and high, would correspond to the groups. For a textbook treatment of latent class models and an overview of applications, we

[6]This assumption corresponds to Assumption CI-YS in McCartan et al. (2024), who additionally require full knowledge of the distribution of ethnicity and how ethnicity relates to R_i or X_i (Assumption ACC). Our identification results do not require such knowledge.

refer to Hagenaars and McCutcheon (2002). In the context of migration, the propensity to migrate is a function of an individual's risk attitude, which is typically unobserved (Jaeger et al. 2010). Different levels of risk attitude correspond to the group structure in our model. More mechanically, a special case of model (2) is a setting in which a researcher splits a sample into multiple parts based on some unobserved or mismeasured variable because she believes that there is some heterogeneity in the dimension of this variable.

3 Identification results

We present our main identification results in this section. To fix ideas, we begin with an intuitive discussion and an outline of the key identification issues.

For notational and expositional convenience, we impose the following assumption. Appendix A.1.9 details how this assumption can be relaxed.

Assumption I-1 *(Constant support) For all* $t' = 1, \dots, T$ *and all* $\{x_t\}_{t \neq t'}$, $Supp(x_{it'} \mid \{x_{it}\}_{t \neq t'} = \{x_t\}_{t \neq t'}) = Supp(x_{it'}) = \mathbb{X}_{t'}$.

Assumption I-1 specifies that the conditional support of x_{it} does not depend on the values of the covariates of other time periods and thus excludes time-invariant covariates. It does allow the distributions of the covariates to vary across different conditioning sets. The arguments below can be readily adapted to allow for a varying support, potentially at the cost of a weaker identification result. Often, however, the identification results are not affected by a changing support as long as the covariate value $\{x_t\}_{t=1}^T$ that satisfies the upcoming assumptions does not restrict the conditional supports; see Appendix A.1.9 for a further discussion on this and how to incorporate time-invariant covariates.

3.1 Intuition

We consider a simplified version of the model in equation (2) and set $T = 3$ and $J = 2$. We show that identification may already be achieved with $T = 2$, but a setting with three time periods allows us to keep notation light and relate our identification argument to the

existing literature more easily. For $\{s_t\}_{t=1}^3 \in \{0,1\}^3$, the model of interest is

$$\begin{aligned}\mathbb{P}(\{y_{it} = s_t\}_{t=1}^3 \mid X_i) = \pi(X_i) \prod_{t=1}^3 \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = 1) \\ + (1 - \pi(X_i)) \prod_{t=1}^3 \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = 2)\end{aligned} \tag{3}$$

where $\pi(X_i) = \mathbb{P}(g_i = 1 \mid X_i)$. We treat $\mathbb{P}(\{y_{it} = s_t\}_{t=1}^3 \mid X_i)$ as known for identification purposes. Fixing X_i at $X = (x_1^\top, x_2^\top, x_3^\top)^\top \in \mathbb{X}$, model (3) essentially becomes an unconditional mixture model, so the component weight and distributions are identified at X under the conditions specified in the literature on nonparametric identification of unconditional mixture models, for instance, Hall and Zhou (2003) or Bonhomme et al. (2016). Taking these identification results as given, one may naïvely enforce the required conditions at each $X \in \mathbb{X}$ to identify the component weight and distributions on the entire support of the covariates. However, because the group labels may switch across values of X_i, such a pointwise approach fails to achieve identification. Our results improve upon this pointwise argument in important respects:

1. We link the latent groups across different values of X_i, that is, we solve the intra-component label switching problem.
2. We do not require the respective assumptions to hold at every $X \in \mathbb{X}$.
3. We use the variation in the component weight induced by the covariates to reduce the number of time periods required for identification.

We proceed to give a heuristic idea of how we solve the intra-component label switching problem using the panel structure. To this end, we fix $X_i = X \in \mathbb{X}$ and assume that $\pi(X) \in (0,1)$ and $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = 1) \neq \mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = 2)$ for $t = 1, 2, 3$. Then, the component weight and distributions are identified at X (Hall and Zhou 2003; Bonhomme et al. 2016); we take this identification result as given for now but present it at the end of this section. Similar to Hall and Zhou (2003), Hall et al. (2005), Kasahara and Shimotsu (2009), and Bonhomme et al. (2016), the main source of identification in model (3) stems from the fact that lower-dimensional submodels of the model satisfy the same group structure. Specifically, let $\mathcal{I} \subseteq \{1,2,3\}$, then we refer to $\mathcal{I}$ as a *submodel*. For the submodel

$\mathcal{I} = \{1\}$

$$\begin{aligned}\mathbb{P}(y_{i1} = s_1 \mid X_i = X) = \pi(X)\mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = 1) \\ + (1 - \pi(X))\mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = 2)\end{aligned} \tag{4}$$

where $s_1 \in \{0, 1\}$. Rewriting (4) yields

$$\pi(X) = \frac{\mathbb{P}(y_{i1} = s_1 \mid X_i = X) - \mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = 2)}{\mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = 1) - \mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = 2)}$$

Since $\mathbb{P}(y_{i1} = s_1 \mid X_i = X)$ is observed for all $X \in \mathbb{X}$ and the component distributions depend on x_1 only, we conclude that $\pi(X)$ is identified at x_1 for all values of x_2 and x_3. This observation can be used to identify the component distributions of periods 2 and 3 at $\tilde{x}_t \neq x_t$. To this intent, we focus on the submodel $\mathcal{I}' = \{1, 2\}$. Under the assumption that $\pi(x_1, \tilde{x}_2, x_3) \in (0, 1)$, rewriting the analogue of (4) for this submodel gives

$$\begin{aligned}&\mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = 1)\mathbb{P}_2(y_{i2} = s_2 \mid x_{i2} = \tilde{x}_2, g_i = 1) \\ &= \frac{\mathbb{P}(\{y_{it} = s_t\}_{t=1}^2 \mid X_i = (x_1^\top, \tilde{x}_2^\top, x_3^\top)^\top)}{\pi(x_1, \tilde{x}_2, x_3)} \\ &\quad - \frac{(1 - \pi(x_1, \tilde{x}_2, x_3))\mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = 2)\mathbb{P}_2(y_{i2} = s_2 \mid x_{i2} = \tilde{x}_2, g_i = 2)}{\pi(x_1, \tilde{x}_2, x_3)}\end{aligned}$$

where $\mathbb{P}_1(y_{i1} = s_1 \mid x_{i1} = x_1, g_i = j)$ is identified for $j = 1, 2$ and $s_1 \in \{0, 1\}$, and, from the previous argument, $\pi(x_1, \tilde{x}_2, x_3)$ is identified. Now, by varying $s_1 \in \{0, 1\}$, we get two equations with the two component distributions of period $t = 2$ as unknowns. Since we can solve for the unknowns under the stated assumptions, $\mathbb{P}_2(y_{i2} = s_2 \mid x_{i2} = \tilde{x}_2, g_i = j)$ is identified for all $\tilde{x}_2$ such that $\pi(x_1, \tilde{x}_2, x_3) \in (0, 1)$. By examining the respective submodels, this idea can be applied to other time periods, too. Therefore, if we additionally assume that the component weights are positive on the entire support of the covariates, we can conclude that the component distributions are identified on the entire support of the covariates for all time periods and groups. Specifically, the intra-component label switching problem has been resolved by enforcing some ordering constraint at the initially fixed X only. This ordering was then carried forward when identifying the component weight and distributions at other covariate values.[7] Importantly, the identification argument does not require the assumptions

[7]The generic label switching problem still persists, that is, we can still reassign the group memberships (consistently across all values of X) without affecting the observed distribution. We will therefore write that identification holds up to relabeling, that is, up to enforcing an ordering constraint at some value of X.

of the unconditional mixture model literature to hold at every covariate value, that is, we allow $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = \tilde{x}_t, g_i = 1) = \mathbb{P}_t(y_{it} = 1 \mid x_{it} = \tilde{x}_t, g_i = 2)$ for $\tilde{x}_t \neq x_t$.

Setting the stage. We now formalize the identification argument for the component weight and distributions at the initially fixed X that we took as given above. Additionally, we introduce our notation in the context of the simplified model (3). The notation will play a crucial role in our assumptions and identification arguments below. For every t, we define

$$\mathcal{P}_t(x_{it}) = \begin{pmatrix} \mathbb{P}_t(y_{it} = 0 \mid x_{it}, g_i = 1) & \mathbb{P}_t(y_{it} = 0 \mid x_{it}, g_i = 2) \\ \mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = 1) & \mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = 2) \end{pmatrix}$$

$$\Pi(X_i) = \begin{pmatrix} \pi(X_i) & 0 \\ 0 & 1 - \pi(X_i) \end{pmatrix}$$

Now, the lower-dimensional submodel induced by $\mathcal{I}$ is

$$\mathbb{P}(\{y_{it} = s_t\}_{t \in \mathcal{I}} \mid X_i) = \pi(X_i) \prod_{t \in \mathcal{I}} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = 1) + (1 - \pi(X_i)) \prod_{t \in \mathcal{I}} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = 2)$$

For the submodel $\mathcal{I} = \{1, 2\}$, we have

$$\mathbf{P}_{\mathcal{I}}(X_i) = \begin{pmatrix} \mathbb{P}(y_{i1} = 0, y_{i2} = 0 \mid X_i) & \mathbb{P}(y_{i1} = 0, y_{i2} = 1 \mid X_i) \\ \mathbb{P}(y_{i1} = 1, y_{i2} = 0 \mid X_i) & \mathbb{P}(y_{i1} = 1, y_{i2} = 1 \mid X_i) \end{pmatrix} = \mathcal{P}_1(x_{i1}) \Pi(X_i) \mathcal{P}_2(x_{i2})^\top$$

where $\mathbf{P}_{\mathcal{I}}(X_i)$ is identified. Next, we define $\text{diag}(a)$ for $a \in \mathbb{R}^J$ as the $J \times J$ diagonal matrix with diagonal a, and let $[A]_{k,\cdot}$ denote the k-th row of the matrix A. Then, for the submodel $\tilde{\mathcal{I}} = \{1, 2, 3\}$ and $k = 1, 2$:

$$\begin{aligned} \mathbf{P}_{\tilde{\mathcal{I}},k,3}(X_i) = & \begin{pmatrix} \mathbb{P}(y_{i1} = 0, y_{i2} = 0, y_{i3} = k - 1 \mid X_i) & \mathbb{P}(y_{i1} = 0, y_{i2} = 1, y_{i3} = k - 1 \mid X_i) \\ \mathbb{P}(y_{i1} = 1, y_{i2} = 0, y_{i3} = k - 1 \mid X_i) & \mathbb{P}(y_{i1} = 1, y_{i2} = 1, y_{i3} = k - 1 \mid X_i) \end{pmatrix} \\ & = \mathcal{P}_1(x_{i1}) \text{diag}([\mathcal{P}_3(x_{i3})]_{k,\cdot}) \Pi(X_i) \mathcal{P}_2(x_{i2})^\top \end{aligned}$$

We now argue that the component weight and distributions are identified at some fixed $X_i = X$. To this end, we restate the previous two assumptions equivalently as *(i)* $\mathcal{P}_t(x_t)$ has full rank for all t and *(ii)* $\Pi(X)$ has full rank. In Appendix A.1.3, we show that these assumptions are jointly testable. Under these assumptions, we have for $k = 1, 2$[8]

$$\mathbf{P}_{\tilde{\mathcal{I}},k,3}(X) \mathbf{P}_{\mathcal{I}}(X)^{-1} = \mathcal{P}_1(x_1) \text{diag}([\mathcal{P}_3(x_3)]_{k,\cdot}) \mathcal{P}_1(x_1)^{-1}$$

[8] Alternatively, the identification arguments in Bonhomme et al. (2016) apply in this setup with $J = 2$.

Hence, for $k = 1, 2$, $[\mathcal{P}_3(x_3)]_{k,\cdot}$ is identified up to permutation by the eigenvalues of $\mathbf{P}_{\tilde{\mathcal{I}},k,3}(X)\mathbf{P}_{\mathcal{I}}(X)^{-1}$. Up to the same permutation, the columns of $\mathcal{P}_1(x_1)$ are identified as the eigenvectors of $\mathbf{P}_{\tilde{\mathcal{I}},k,3}(X)\mathbf{P}_{\mathcal{I}}(X)^{-1}$ for $k = 1, 2$ (De Lathauwer et al. 2004). The scaling of the eigenvectors is pinned down since the columns of $\mathcal{P}_1(x_1)$ sum up to 1. Using symmetric arguments, $\mathcal{P}_2(x_2)$ is identified up to the same permutation of the groups. Now, once the component distributions are identified, $\Pi(X) = \mathcal{P}_1(x_1)^{-1}\mathbf{P}_{\mathcal{I}}(X)(\mathcal{P}_2(x_2)^\top)^{-1}$ is identified.

3.2 Main identification results

This section presents our main identification results for model (2). Throughout, we implicitly assume that Assumption M-1 holds. The first part of this section is agnostic as to whether $\underline{\pi}(X_i) = (\pi_1(X_i), \ldots, \pi_J(X_i))^\top$, the vector of component weights, varies in X_i, while the second part discusses how variation of $\underline{\pi}(X_i)$ in X_i can be leveraged to reduce the number of time periods required for identification.

To state our identification results, we introduce additional notation and, for notational convenience, fix $X_i = X \in \mathbb{X}$. The assumptions below specify the conditions that must hold at this fixed value X. First, we extend the definitions of $\mathcal{P}_t(x_t)$ and $\Pi(X)$ to the case with J groups:

$$\mathcal{P}_t(x_t) = \begin{pmatrix} \mathbb{P}_t(y_{it} = 0 \mid x_{it} = x_t, g_i = 1) & \ldots & \mathbb{P}_t(y_{it} = 0 \mid x_{it} = x_t, g_i = J) \\ \mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = 1) & \ldots & \mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = J) \end{pmatrix}$$

$$\Pi(X) = \operatorname{diag}(\underline{\pi}(X))$$

In the following, we sometimes abuse the notation by reordering the arguments of $\Pi(X) = \Pi(\{x_t\}_{t=1}^T)$ for expositional purposes if the context allows us to do so.

When $J > 2$, $\mathcal{P}_t(x_t)$ does not have full rank so that the arguments from Section 3.1 do not apply. However, we may combine multiple time periods to construct a matrix that can act like $\mathcal{P}_t(x_t)$ previously. To this end, for some submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$, we define the $2^{|\mathcal{I}|} \times J$ matrix

$$\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}}) = \overset{\text{col}}{\otimes}_{t\in\mathcal{I}} \mathcal{P}_t(x_t)$$

where $\overset{\text{col}}{\otimes}$ denotes the Khatri-Rao product and $|A|$ denotes the cardinality of the set A. Specifically, for two matrices A_1 and A_2 of dimensions $a_1 \times J$ and $a_2 \times J$, respectively, the

Khatri-Rao product is $A_1 \overset{\text{col}}{\otimes} A_2 = ([A_1]_{\cdot,1} \otimes [A_2]_{\cdot,1}, \ldots, [A_1]_{\cdot,J} \otimes [A_2]_{\cdot,J})$ where $[A]_{\cdot,j}$ denotes the j-th column of some matrix A and $\otimes$ denotes the Kronecker product, that is, $\overset{\text{col}}{\otimes}$ is simply the column-wise Kronecker product of two matrices. Hence, the j-th column of $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ collects all values of the j-th component distribution of $\{y_{it}\}_{t\in\mathcal{I}}$ given $X_i = X$ and $g_i = j$, that is, $\prod_{t\in\mathcal{I}} \mathbb{P}_t(y_{it} = s_t \mid x_{it} = x_t, g_i = j)$ for $\{s_t\}_{t\in\mathcal{I}} \in \{0,1\}^{|\mathcal{I}|}$. If $\mathcal{I} = \{t\}$ is a singleton, we use the convention that $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}}) = \mathcal{P}_t(x_t)$.

We make the following assumptions, which are analogous to the assumptions in Section 3.1. In Appendix A.1.3, we argue that Assumptions I-2 and I-3 are jointly testable.

Assumption I-2 *(Variation in component distributions) There exist $t' \in \{1, \ldots, T\}$ as well as disjoint and non-empty submodels $\mathcal{I}_1$ and $\mathcal{I}_2$ with $t' \notin \mathcal{I}_1 \cup \mathcal{I}_2$ such that $\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\mathcal{I}_\ell})$ has full column rank for $\ell = 1, 2$ and $\mathcal{P}_{t'}(x_{t'})$ has distinct columns.*

Assumption I-3 *(Positive component weights) $\Pi(X)$ has full rank.*

Remark 1. A necessary condition for Assumption I-2 is that $J \leq 2^{\lfloor (T-1)/2 \rfloor}$ where $\lfloor z \rfloor$ denotes the smallest integer less than or equal to z. Rewriting this condition in terms of the dependence of T on J, we have $T \geq 2\lceil \log_2(J) \rceil + 1$. Allman et al. (2009) derive the same condition in the context of generic identification of mixtures of binary outcome models without covariates.

Remark 2. Assumption I-2 mirrors the identifiability conditions in Leurgans et al. (1993) and the rank conditions in Allman et al. (2009). Different from the arguments in Allman et al. (2009) and similar to Bonhomme et al. (2016), our identification argument is constructive.

Remark 3. Lemma 13 in Allman et al. (2009) implies that generically $\text{rank}(\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})) = \min\{2^{|\mathcal{I}|}, J\}$ for any submodel $\mathcal{I}$. In Appendix A.1.10, we present a sufficient and necessary condition when $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank for a given submodel $\mathcal{I}$. Intuitively speaking, $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank if and only if no component distribution can be expressed as the same (possibly non-convex) mixture of the other component distributions for all submodels of the submodel $\mathcal{I}$. Additionally, Appendix A.1.10 presents a sufficient condition for Assumption I-2 based on the Kruskal rank of $\mathcal{P}_t(x_t)$ for all $t \in \mathcal{I}$. In particular, if there exist $2J - 1$ periods and $X \in \mathbb{X}$ such that $\mathcal{P}_t(x_t)$ has distinct columns for all $t = 1, \ldots, T$, then Assumption I-2 is satisfied.

$\mathcal{P}_{t'}(x_{t'})$ in Assumption I-2 has distinct columns if and only if $\mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j) \neq \mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j')$ for all $j \neq j'$, which immediately implies the following.

Lemma 3.1 *The columns of $\mathcal{P}_{t'}(x_{t'})$ are generically distinct.*

Remark 4. Assumption I-3 holds if and only if $\pi_j(X) > 0$ for all $j = 1, \ldots, J$. This puts mild restrictions on the relationship between the covariates and group membership. Assumption I-3 excludes settings in which the covariates perfectly predict the group membership. However, our arguments can be easily adapted to a setting when some groups are mutually exclusive and a researcher knows on which subsets of the covariate support these groups have zero probability.

We are now ready to state our first main identification result.

Theorem 3.2 *There exists $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ such that Assumptions I-2 and I-3 are satisfied with submodels $\mathcal{I}_1$ and $\mathcal{I}_2$, and Assumption I-1 holds. Then,*

1. *for all $t \notin \mathcal{I}_2$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified at $\tilde{x}_t \in \mathbb{X}_t$ for which there exists $\{x^*_{t^*}\}_{t^* \notin \mathcal{I}_2 \cup \{t\}}$ such that $\pi_j(\{x_t\}_{t \in \mathcal{I}_2}, \{x^*_{t^*}\}_{t^* \notin \mathcal{I}_2 \cup \{t\}}, \tilde{x}_t) > 0$,*
2. *for all $t \notin \mathcal{I}_1$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified at $\tilde{x}_t \in \mathbb{X}_t$ for which there exists $\{x^*_{t^*}\}_{t^* \notin \mathcal{I}_1 \cup \{t\}}$ such that $\pi_j(\{x_t\}_{t \in \mathcal{I}_1}, \{x^*_{t^*}\}_{t^* \notin \mathcal{I}_1 \cup \{t\}}, \tilde{x}_t) > 0$, and*
3. *for any submodel $\mathcal{I}$ and $\{\tilde{x}_t\}_{t \in \mathcal{I}}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t \in \mathcal{I}})$ has full column rank and is identified given 1. or 2., $\underline{\pi}(\{\tilde{x}_t\}_{t \in \mathcal{I}}, \{x_t\}_{t \notin \mathcal{I}})$ is identified at $\{\tilde{x}_t\}_{t \in \mathcal{I}}$ for all $\{x_t\}_{t \notin \mathcal{I}}$.*

All objects are identified up to the same relabeling of the groups.

As a direct corollary of Theorem 3.2, we obtain the following result for the case where the component weights are positive on $\mathbb{X}$.

Corollary 3.2.1 *Assumption I-1 holds and there exists $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ such that Assumption I-2 holds and $\pi_j(\tilde{X}) > 0$ for all $\tilde{X} \in \mathbb{X}$ and $j = 1, \ldots, J$. Then*

- *$\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified for all $\tilde{x}_t \in \mathbb{X}_t$, $t = 1, \ldots, T$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, and*

- $\underline{\pi}(\tilde{X})$ *is identified at all* $\tilde{X} \in \mathbb{X}$ *such that there exists a submodel* $\mathcal{I} \subseteq \{1, \ldots, T\}$ *with* $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t \in \mathcal{I}})$ *having full column rank.*

All objects are identified up to the same relabeling of the groups.

Importantly, Theorem 3.2 and Corollary 3.2.1 allow for both discrete and continuous covariates and require that Assumptions I-2 and I-3 hold at a single value of X_i only, while the component distributions and weights are potentially identified at multiple realizations of X_i. In particular, Corollary 3.2.1 states conditions under which the component distributions are identified on the entire support of the covariates, while the component weights are identified at all values $\tilde{X}$ such that there exists some submodel for which the component distributions are not linearly dependent. The intra-component label switching problem has been solved without any assumptions on the smoothness of the component weights and distributions. While Theorem 3.2 highlights that the condition that $\pi_j(\tilde{X}) > 0$ for all $\tilde{X} \in \mathbb{X}$ and $j = 1, \ldots, J$ is not necessary for the assertion of Corollary 3.2.1 to hold, this condition is easily interpretable and common in the literature on semi/nonparametric mixtures of regression models (Huang and Yao 2012; Huang et al. 2013).

A few additional remarks are in order.

Remark 5. The proof of Theorem 3.2 highlights that the set of values of $\tilde{x}_t$ at which $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified is larger than stated in Theorem 3.2. We characterize the iterative definition of the larger set in the proof. Additionally, Theorem 3.2 fixes two submodels $\mathcal{I}_1$ and $\mathcal{I}_2$ such that Assumptions I-2 and I-3 are satisfied. By applying our arguments to more submodels, the identification result may be strengthened further.

Remark 6. Theorem 3.2 allows for identification of $\mathcal{P}_t(x_t)$ at values of x_t even when $\mathcal{P}_t(x_t)$ has identical columns, that is, the component distributions are identical. Similarly, we may identify $\underline{\pi}(X)$ at values of X even when $\pi_j(X) = 0$ for some j, that is, in this general setting, the number of "active"/non-zero probability groups is allowed to vary across the covariate support; or, put differently, J is allowed to be a function of the covariates.

Remark 7. Theorem 3.2 does not require any variation in the component weights. However, such variation can be used to relax the assumptions. In particular, the conclusion of Theorem 3.2 continues to hold when there does not exist $t' \notin \mathcal{I}_1 \cup \mathcal{I}_2$ such that $\mathcal{P}_{t'}(x_{t'})$ has distinct

columns, but there exist $t' \notin \mathcal{I}_1 \cup \mathcal{I}_2$ and $\tilde{x}_{t'} \neq x_{t'} \in \mathbb{X}_{t'}$ such that $\Pi(\tilde{X})$ with $\tilde{X} = (\{x_t\}_{t \neq t'}^\top, \tilde{x}_{t'}^\top)^\top$ has full rank and $\Pi(\tilde{X})\Pi(X)^{-1}$ has distinct diagonal entries, that is, the component weights exhibit some variation in $x_{t'}$.

While Theorem 3.2 and its corollary require weak assumptions, they do not fully leverage the exclusion restriction that is present in the submodels. Specifically, the variability of $\underline{\pi}(X)$ across different values of X, if present, can be used to decrease the dependence of T on J. In the following, we use the notational convention that if $\{1, \ldots, T\} = \mathcal{I}_1 \cup \mathcal{I}_2$, then $\underline{\pi}(\{x_t\}_{t \in \mathcal{I}_1}, \{x_t\}_{t \in \mathcal{I}_2}, \{x_t\}_{t \notin \mathcal{I}_1 \cup \mathcal{I}_2}) = \underline{\pi}(\{x_t\}_{t \in \mathcal{I}_1}, \{x_t\}_{t \in \mathcal{I}_2})$. We make the following assumption.

Assumption I-4 *(Variation in the component weights and distributions) There exist disjoint and non-empty submodels $\mathcal{I}_1$ and $\mathcal{I}_2$ such that $\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t \in \mathcal{I}_\ell})$ has full column rank for $\ell = 1, 2$ and there exist $\{\tilde{x}_t\}_{t \in \mathcal{I}_1} \neq \{x_t\}_{t \in \mathcal{I}_1}$ and $\{\underline{x}_t\}_{t \in \mathcal{I}_2} \neq \{x_t\}_{t \in \mathcal{I}_2}$ such that $\mathcal{P}_{\mathcal{I}_1}(\{\tilde{x}_t\}_{t \in \mathcal{I}_1})$ and $\mathcal{P}_{\mathcal{I}_2}(\{\underline{x}_t\}_{t \in \mathcal{I}_2})$ have full column rank, and*

$$\begin{aligned} &\Pi(\{x_t\}_{t \in \mathcal{I}_1}, \{x_t\}_{t \in \mathcal{I}_2}, \{x_t\}_{t \notin \mathcal{I}_1 \cup \mathcal{I}_2}) \Pi(\{\tilde{x}_t\}_{t \in \mathcal{I}_1}, \{x_t\}_{t \in \mathcal{I}_2}, \{x_t\}_{t \notin \mathcal{I}_1 \cup \mathcal{I}_2})^{-1} \\ \times\ &\Pi(\{\tilde{x}_t\}_{t \in \mathcal{I}_1}, \{\underline{x}_t\}_{t \in \mathcal{I}_2}, \{x_t\}_{t \notin \mathcal{I}_1 \cup \mathcal{I}_2}) \Pi(\{x_t\}_{t \in \mathcal{I}_1}, \{\underline{x}_t\}_{t \in \mathcal{I}_2}, \{x_t\}_{t \notin \mathcal{I}_1 \cup \mathcal{I}_2})^{-1} \end{aligned}$$

is well-defined and has distinct and non-zero diagonal entries.

In contrast to Assumption I-2, a necessary condition for Assumption I-4 is that $T \geq 2\lceil \log_2(J) \rceil$, that is, when the component weights exhibit sufficient variation in the covariates, the number of time periods required for identification may be reduced by one period. In Appendix A.1.3, we discuss that Assumption I-4 is testable.

If Assumption I-4 holds, an analogue of Theorem 3.2 can be established.

Theorem 3.3 *There exists $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ such that Assumption I-4 holds with submodels $\mathcal{I}_1$ and $\mathcal{I}_2$, and Assumption I-1 holds. Then,*

1. *for all $t \notin \mathcal{I}_2$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \overline{x}_t, g_i = j)$ is identified at $\overline{x}_t \in \mathbb{X}_t$ for which there exists $\{x_{t^*}^*\}_{t^* \notin \mathcal{I}_2 \cup \{t\}}$ such that $\pi_j(\{x_t\}_{t \in \mathcal{I}_2}, \{x_{t^*}^*\}_{t^* \notin \mathcal{I}_2 \cup \{t\}}, \overline{x}_t) > 0$ or $\pi_j(\{\underline{x}_t\}_{t \in \mathcal{I}_2}, \{x_{t^*}^*\}_{t^* \notin \mathcal{I}_2 \cup \{t\}}, \overline{x}_t) > 0$,*
2. *for all $t \notin \mathcal{I}_1$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \overline{x}_t, g_i = j)$ is identified at $\overline{x}_t \in \mathbb{X}_t$ for which there exists $\{x_{t^*}^*\}_{t^* \notin \mathcal{I}_1 \cup \{t\}}$ such that $\pi_j(\{x_t\}_{t \in \mathcal{I}_1}, \{x_{t^*}^*\}_{t^* \notin \mathcal{I}_1 \cup \{t\}}, \overline{x}_t) > 0$ or $\pi_j(\{\tilde{x}_t\}_{t \in \mathcal{I}_1}, \{x_{t^*}^*\}_{t^* \notin \mathcal{I}_1 \cup \{t\}}, \overline{x}_t) > 0$,*

3. *for any submodel* $\mathcal{I}$ *and* $\{\overline{x}_t\}_{t\in\mathcal{I}}$ *such that* $\mathcal{P}_{\mathcal{I}}(\{\overline{x}_t\}_{t\in\mathcal{I}})$ *has full column rank and is identified given 1. or 2.,* $\underline{\pi}(\{\overline{x}_t\}_{t\in\mathcal{I}}, \{x_t\}_{t\notin\mathcal{I}})$ *is identified at* $\{\overline{x}_t\}_{t\in\mathcal{I}}$ *for all* $\{x_t\}_{t\notin\mathcal{I}}$.

All objects are identified up to the same relabeling of the groups.

A direct corollary of Theorem 3.3 is the case where the component weights are positive on $\mathbb{X}$.

Corollary 3.3.1 *There exists* $X \in \mathbb{X}$ *such that Assumption I-4 holds,* $\pi_j(\tilde{X}) > 0$ *for all* $\tilde{X} \in \mathbb{X}$ *and* $j = 1, \ldots, J$, *and Assumption I-1 holds. Then*

- $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ *is identified for all* $\tilde{x}_t \in \mathbb{X}_t$, $t = 1, \ldots, T$, $j = 1, \ldots, J$, *and* $s_t \in \{0, 1\}$, *and*
- $\underline{\pi}(\tilde{X})$ *is identified at all* $\tilde{X} \in \mathbb{X}$ *such that there exists a submodel* $\mathcal{I} \subseteq \{1, \ldots, T\}$ *with* $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ *having full column rank.*

All objects are identified up to the same relabeling of the groups.

Comparing the conditions of Theorem 3.2 and Theorem 3.3 reveals a potential weak identification issue in the context of mixture models. When $\underline{\pi}(X)$ exhibits only little variation in X, the conditions of Assumption I-4 may be satisfied in the population, while in finite samples we may not be able to pick up this little variation. If that is the case, one may want to revert to the setting of Theorem 3.2 as it does not require any variation in the component weights for identification.

3.3 Identification of marginal effects

Once the component distributions are identified, identification of group-specific conditional marginal effects such as $\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j)$ follows directly. Therefore, known functionals of group-specific conditional marginal effects are immediately identified, too; for instance, group-specific weighted average effects, $\int \frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x, g_i = j)v(x)dx$, with known weight function $v(\cdot)$.

To identify group-specific AMEs, we require additional assumptions. To see this, we let $\text{AME}_{j,t,k} = \mathbb{E}\left[\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j) \mid g_i = j\right]$ be the group-specific AME of group j when $x_{it,k}$ is continuously distributed and $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = j)$ is differentiable

in $x_{it,k}$.[9] Letting $x_{it,-k}$ contain all entries of x_{it} but the k-th one, we make the following simplifying assumption.

Assumption I-5 *(AMEs – Regularity) Conditional on $g_i = j$ and $x_{it,-k}$, $x_{it,k}$ is continuously distributed and $\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j)$ is differentiable in $x_{it,k}$. $AME_{j,t,k}$ exists.*

We focus on the case where $x_{it,k}$ is continuously distributed and $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = j)$ is differentiable in $x_{it,k}$ for expositional purposes. The discrete case follows from similar arguments. Letting $\pi_j = \mathbb{E}[\pi_j(X_i)]$, which is assumed to be positive, the law of iterated expectations allows us to rewrite $\text{AME}_{j,t,k}$ as follows

$$\text{AME}_{j,t,k} = \frac{\mathbb{E}\left[\pi_j(X_i)\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j)\right]}{\pi_j}$$

Hence, identification of $\text{AME}_{j,t,k}$ requires identification of the component distribution, $\pi_j(X)$, and π_j.[10] We assume

Assumption I-6 *(AMEs - Component distributions) $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = j)$ is identified for almost all $x_t \in \mathbb{X}_t$.*

Our previous results highlight that Assumption I-6 is satisfied under mild conditions. However, to identify π_j, we require a stronger additional assumption. Specifically, we assume

Assumption I-7 *(AMEs - Weights) $\pi_j(X)$ is identified for almost all $X \in \mathbb{X}$ and $\pi_j > 0$.*

The second part of Assumption I-7 is weak and satisfied when $\pi_j(X_i) > 0$ with positive probability. On the other hand, when the component weights and component distributions are left completely nonparametric, the first part of Assumption I-7 may be too strong in some settings as it requires that for almost all $X \in \mathbb{X}$ there exists a submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$ such that $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank, that is, the component distributions are not allowed

[9] For the discrete case, we could consider average partial effects (APEs) defined as $\text{APE}_{j,t,k}(x_k^{(1)}, x_k^{(2)}) = \mathbb{E}[\mathbb{P}_t(y_{it} = 1 \mid x_{it,k} = x_k^{(1)}, x_{it,-k}, g_i = j) - \mathbb{P}_t(y_{it} = 1 \mid x_{it,k} = x_k^{(2)}, x_{it,-k}, g_i = j) \mid g_i = j]$ where $x_{it,-k}$ contains all entries of x_{it} but the k-th one. Alternatively, one may also consider CAMEs of the following type $\mathbb{E}\left[\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j) \mid g_i = j, x_{it,k} = x_{t,k}\right]$. Identification of APEs and CAMEs follows from analogous arguments as the ones presented in this section. Hence, we do not discuss their identification in detail.

[10] In the discrete case, we have $\text{APE}_{j,t,k}(x_k^{(1)}, x_k^{(2)}) = \frac{1}{\pi_j}\mathbb{E}[\pi_j(X_i)\{\mathbb{P}_t(y_{it} = 1 \mid x_{it,k} = x_k^{(1)}, x_{it,-k}, g_i = j) - \mathbb{P}_t(y_{it} = 1 \mid x_{it,k} = x_k^{(2)}, x_{it,-k}, g_i = j)\}]$.

to be linearly dependent for all submodels on a subset of $\mathbb{X}$ that has positive probability. However, if one is willing to put more structure on the problem by modeling the component distributions or weights (semi-)parametrically, Assumption I-7 is mild, too.[11] If a researcher is worried about Assumption I-7 but knows that it holds on a known subset of the support $\mathbb{B} \subseteq \mathbb{X}$, we sketch a partial identification argument in Appendix A.1.4. Alternatively, one may focus on the group-specific AMEs conditional on $X \in \mathbb{B}$ which are identified. We conclude our discussion with the following lemma.

Lemma 3.4 *(Identification of AMEs) Assume that Assumptions I-5 to I-7 hold, then $AME_{j,t,k}$ is identified.*

Lemma 3.4 follows immediately from the previous discussion so we omit its proof.

3.4 Interpreting the groups

Due to the label switching issue, the groups themselves have no economic meaning without any further analysis. We briefly discuss how such an analysis may proceed and refer to Section 6 for an example. As a first step, the component weights should be examined in detail. Specifically, if some covariates are known to be predictive of the group membership, this can be used to assign the groups their economic meaning. For example, when $J = 2$, income is an entry of X, and $\pi_1(X)$ is strictly increasing in income, while $\pi_2(X)$ is strictly decreasing in income, then the first group may be interpreted as the group of high-income individuals and the second group may be interpreted as the group of low-income individuals. In a second step, group-specific means of the outcome and covariates can be estimated. In particular, for the covariates we have $\mathbb{E}[x_{it} \mid g_i = j] = \mathbb{E}[\pi_j(X_i)x_{it}]/\pi_j$ which can be consistently estimated via a sample analog estimator. If of interest, other group-specific moments can be estimated analogously. These group-specific summary statistics highlight the differences across groups and thereby facilitate their economic interpretation. Additionally, we suggest predicting the group membership of each individual using the posterior probability

[11] For example, our arguments in Section 7.2 suggest that when $\pi_j(X)$ does not depend on all entries of X, we can vary these excluded entries to achieve identification of the component weights. Alternatively, when $\pi_j(X)$ depends on X only through an index, we can vary X while keeping the index fixed to achieve that at X there exists some submodel $\mathcal{I}$ such that $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank.

to be in a group, that is,

$$\mathbb{P}(g_i = j \mid X_i, \{y_{it}\}_{t=1}^T) = \frac{\pi_j(X_i) \prod_{t=1}^T \mathbb{P}_t(y_{it} \mid x_{it}, g_i = j)}{\sum_{j=1}^J \pi_j(X_i) \prod_{t=1}^T \mathbb{P}_t(y_{it} \mid x_{it}, g_i = j)}$$

The predicted group of an individual is $g_i^* = \arg\max_{j \in \{1,\ldots,J\}} \mathbb{P}(g_i = j \mid X_i, \{y_{it}\}_{t=1}^T)$. While these predictions are not consistent, the predicted group gives an idea of the group an individual most likely belongs to and may allow a researcher to "correlate" the group membership with other variables of interest.

4 Estimation

This section presents a semiparametric estimation procedure of model (2) that parameterizes the component distributions and leaves the component weights flexible. Specifically, the model of interest is

$$\begin{aligned} &\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i; \pi^{(0)}, \beta^{(0)}) \\ =&\sum_{j=1}^J \pi_j^{(0)}(X_i) \prod_{t=1}^T F_{jt}\left(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}\right)^{s_t} \left(1 - F_{jt}\left(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}\right)\right)^{1-s_t} \end{aligned} \tag{5}$$

where $F_{jt}(\cdot)$ is assumed to be known, for instance, the CDF of a standard normally or logistically distributed random variable. Model (5) is motivated by Example 2.1 and nests, for instance, time-varying group-specific logit or probit models. We use the superscript (0) to denote the true parameter values $\theta^{(0)} = (\beta^{(0)\top}, \pi^{(0)}) \in \Theta = \mathcal{B} \times \mathcal{H}$ with $\beta^{(0)} = \left\{(\beta_{j,t}^{(0)\top}, \alpha_{j,t}^{(0)})\right\}_{t=1,\ldots,T;j=1,\ldots,J} \in \mathcal{B}$ and $\pi^{(0)} = (\pi_1^{(0)}(\cdot), \ldots, \pi_{J-1}^{(0)}(\cdot)) \in \mathcal{H}$. $\pi^{(0)}$ does not include $\pi_J^{(0)}(\cdot)$ since $\pi_J^{(0)}(\cdot) = 1 - \sum_{j=1}^{J-1} \pi_j^{(0)}(\cdot)$. Model (5) induces the population log-likelihood $\mathcal{L}(\theta) = \mathbb{E}\left[\log(\mathbb{P}(y_i \mid X_i; \theta))\right]$ with sample analog $\hat{\mathcal{L}}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log\left(\mathbb{P}(y_i \mid X_i; \theta)\right)$. For the remainder of the paper, we make the following assumption on the sampling process.

Assumption E-1 *(Random sample)* $\{(\{y_{it}\}_{t=1}^T, X_i^\top, g_i)\}_{i=1}^n$ *is a random sample.*

Although our theoretical results focus on the case where $\pi_j(X)$ is left fully nonparametric and all our results hold under sufficient smoothness assumptions on the component weights, the resulting estimator may perform poorly in finite samples due to the curse of dimensionality. We therefore suggest placing further restrictions on $\pi_j(X)$ and modeling the weights semiparametrically; for instance, the component weights may depend on the time average

of the covariates or only on a subset of the covariates. We impose such restrictions in both the simulation studies and the empirical illustration. However, since the asymptotic results for these semiparametric specifications are special cases of our more general framework, we focus on the more general setup here.

Importantly, we note that the semiparametric specification of model (5) is only one model that is covered by our identification results. For example, our identification results can be readily applied to settings where $F_{jt}(\cdot)$ is unknown, for example, a single index model (Ichimura 1993). An estimator for this kind of model may be implemented by replacing $F_{jt}(\cdot)$ in model (5) with an appropriate series approximation. Studying this estimator formally goes beyond the scope of this paper.

We do not claim that the previously derived dependence of $T \geq 2\lceil \log_2(J) \rceil$ is tight in the semiparametric specification of model (5). Ignoring the intra-component label switching problem, the arguments of Bajari et al. (2011) imply that at least $T \geq \lceil \log_2(J) \rceil$ time periods are required for the semiparametric information bound of $\beta^{(0)}$ to be nonzero. Although these two bounds are close in settings in which J is not too large, we believe it is interesting to study whether the latter bound can be achieved in the context of model (5). We leave this for future research.[12]

4.1 Estimation procedure

Component distributions and weights. To introduce our estimation procedure, we first reparameterize model (5). To this end, we assume that $\pi_j^{(0)}(\cdot) > 0$ for all $j = 1, \ldots, J$, and let $w_j^{(0)}(X) := \log(\pi_j^{(0)}(X)/\pi_J^{(0)}(X))$ denote the log-odds ratio for $j = 1, \ldots, J-1$. For $j = 1, \ldots, J-1$, we can now reparameterize the component weights by introducing a link function, that is,

$$\pi_j^{(0)}(X_i) = \frac{\exp(w_j^{(0)}(X_i))}{1 + \sum_{j'=1}^{J-1} \exp(w_{j'}^{(0)}(X_i))}$$

Letting $\theta_w^{(0)} = (\beta^{(0)\top}, w^{(0)}) \in \Theta^{(w)} = \mathcal{B} \times \mathcal{W}$ with $w^{(0)} = (w_1^{(0)}, \ldots, w_{J-1}^{(0)})$ and $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_{J-1}$, we similarly reparameterize the likelihood by introducing a link function. With a slight abuse of notation, we will use $\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i; \theta_w^{(0)})$, $\hat{\mathcal{L}}_n(\theta_w)$, and $\mathcal{L}(\theta_w)$ to denote

[12] While Bajari et al. (2011) sketch an idea for an estimation procedure, they do not provide any identification results that apply to our setting.

the reparameterized versions of the likelihood, the log-likelihood, and the population analog of the log-likelihood, respectively.

Using a link function has the advantage that the estimated component weights are automatically constrained to map into the unit interval. However, the use of a link function is not without loss of generality because we additionally require that $\pi_j^{(0)}(\cdot) > 0$ for all $j = 1, \ldots, J$. Alternatively, the constraint $0 \leq \pi_j(X) \leq 1$ for all $X \in \mathbb{X}$ and $j = 1, \ldots, J$ may be directly enforced in the optimization routine without the use of a link function. We conjecture that this constrained estimation routine can relax the overlap condition but leave the analysis of the constrained estimator for future research.

Since $\mathcal{W}$ is infinite-dimensional, maximizing $\hat{\mathcal{L}}_n(\theta_w)$ over $\Theta^{(w)}$ directly is infeasible. We therefore replace $\Theta^{(w)}$ by a finite-dimensional sieve space $\Theta^{(w)}_{d(n)} = \mathcal{B} \times \mathcal{W}_{d(n)}$ that is dense in $\Theta^{(w)}$ as $d(n) \to \infty$ when $n \to \infty$; see Chen (2007) for a comprehensive treatment of sieve estimators. Specifically, we approximate $w_j^{(0)}(\cdot)$ for $j = 1, \ldots, J-1$ with $\sum_{d=0}^{d(n)} \rho_{j,d}(\cdot)\gamma_{j,d}$ where $\{\rho_{j,d}\}_{d=0}^{\infty}$ is a collection of known basis functions for $j = 1, \ldots, J-1$, usually a tensor-product basis.[13] We provide a formal definition of the sieve space $\mathcal{W}_{d(n)}$ in Appendix A.4.

We estimate $\theta_w^{(0)}$ with an approximate sieve maximum likelihood estimator $\hat{\theta}_{w,n}$ that satisfies $\hat{\mathcal{L}}_n(\hat{\theta}_{w,n}) \geq \sup_{\theta_w \in \Theta^{(w)}_{d(n)}} \hat{\mathcal{L}}_n(\theta_w) - o_p(n^{-1/2})$.[14] Given the sieve dimension $d(n)$, this maximization routine translates into maximizing

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{j=1}^{J} \mathbf{G}_j(X_i;\gamma_n) \prod_{t=1}^{T} F_{jt}\left(x_{it}^\top \beta_{j,t} + \alpha_{j,t}\right)^{y_{it}} \left(1 - F_{jt}\left(x_{it}^\top \beta_{j,t} + \alpha_{j,t}\right)\right)^{1-y_{it}}\right) \tag{6}$$

with respect to β and $\gamma_n = \{\gamma_{n,j}\}_{j=1}^{J-1}$ where for $j = 1, \ldots, J-1$

$$\mathbf{G}_j(X_i;\gamma_n) = \frac{\exp\left(\gamma_{n,j}^\top \rho_j^{d(n)}(X_i)\right)}{1 + \sum_{j'=1}^{J-1} \exp\left(\gamma_{n,j'}^\top \rho_{j'}^{d(n)}(X_i)\right)}$$

and $\mathbf{G}_J(X_i;\gamma_n) = 1 - \sum_{j=1}^{J-1} \mathbf{G}_j(X_i;\gamma_n)$. Following the literature on mixture models, we maxi-

[13] While $d(n)$, the sieve dimension, may also depend on j, we do not consider such a setting here to keep the notation light. However, an extension to this case is immediate.

[14] Alternatively, one may leverage the constructive nature of the identification arguments of Section 3.2. This has the advantage that it avoids any numerical optimization routines that are required for the MLE. However, this estimator would require us to estimate eigenvectors, which are notoriously difficult to estimate when eigengaps are small; a situation likely to occur. Another interesting avenue for future research would be to adapt the estimator of Gu and Koenker (2022) to the setting at hand.

mize (6) via an Expectation-Maximization (EM) algorithm (Dempster et al. 1977; Frühwirth-Schnatter et al. 2019). We present the EM algorithm we use in the simulations and the application in Appendix A.1.11. Letting $\hat{\beta}_n$ and $\hat{\gamma}_n$ denote an (approximate) maximizer of (6), we estimate $\theta_w^{(0)}$ with $\hat{\theta}_{w,n} = (\hat{\beta}_n^\top, \hat{w}_n(\cdot)) = \left(\{\hat{\beta}_{j,t}^\top, \hat{\alpha}_{j,t}\}_{t=1,\dots,T;j=1,\dots,J}, \{\hat{\gamma}_{n,j}^\top \rho_j^{d(n)}(\cdot)\}_{j=1}^{J-1}\right)$. Using the estimated log-odds ratios, an estimator for the component weights $\pi^{(0)}(\cdot)$ is $\hat{\pi}_n(\cdot) = \{\mathbf{G}_j(\cdot;\hat{\gamma}_n)\}_{j=1}^{J-1}$ and $\hat{\pi}_J(\cdot)$ is estimated with $\mathbf{G}_J(\cdot;\hat{\gamma}_n)$. We denote our estimator for $\theta^{(0)}$ with $\hat{\theta}_n = (\hat{\beta}_n^\top, \hat{\pi}_n(\cdot))$.

AMEs. Focusing on the continuous case, a sample analog estimator of $\mathrm{AME}_{j,t,k}$ is

$$\widehat{\mathrm{AME}}_{j,t,k} = \frac{\hat{\beta}_{j,t,k}}{\hat{\pi}_j} \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_j(X_i) F'_{jt}(x_{it}^\top \hat{\beta}_{j,t} + \hat{\alpha}_{j,t})$$

with $\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^n \hat{\pi}_j(X_i) = \frac{1}{n}\sum_{i=1}^n \mathbf{G}_j(X_i;\hat{\gamma}_n)$. An estimator of the APE for a discrete covariate can be constructed analogously. In our application, we also consider the time average of $\widehat{\mathrm{AME}}_{j,t,k}$, which we denote by $\widehat{\mathrm{AME}}_{j,k}$.

Group-specific conditional marginal effects and known functionals thereof can be estimated using plug-in estimators. In particular, an estimator for the group-specific conditional marginal effect $\frac{\partial}{\partial x_{it,k}} \mathbb{P}_t(y_{it} = 1 \mid x_{it} = x, g_i = j)$ is $\hat{\beta}_{j,t,k} F'_{jt}(x^\top \hat{\beta}_{j,t} + \hat{\alpha}_{j,t})$.

4.2 Consistency

This section presents a consistency result for the proposed estimators. For the remainder of this paper, we make the following assumptions:

Assumption E-2 *(Overlap)* $\pi_j^{(0)}(X) \in (0,1)$ *for all* $X \in \mathbb{X}$ *and* $j = 1, \dots, J$.

Assumption E-3 *(Covariates)* $\mathbb{X}$ *is compact with nonempty interior,* X_i *is jointly continuously distributed, and Assumption I-1 holds.*

Assumption E-2 allows us to reparameterize the log-likelihood via a link function as described in the previous section. Although Assumption E-3 excludes discrete covariates, it is straightforward to allow for them by fully saturating the model to be estimated; if the support of the discrete covariate is too large, the dimension of the fully saturated model is allowed to increase with n. While we leverage the compactness assumption on $\mathbb{X}$ in our proofs, a more careful argument may relax this assumption by using a weighted norm;

see, for instance, Freyberger and Masten (2019). At first sight, Assumptions E-2 and E-3 may appear more restrictive than they are. As noted previously, Assumption I-1 can be relaxed; we provide a detailed discussion in Appendix A.1.9. In general, if one is concerned about Assumptions E-2 and E-3, the estimation procedure can be restricted to a compact subset $\tilde{\mathbb{X}} \subset \mathbb{X}$ of the covariate space that satisfies these assumptions. The finite-dimensional parameter $\beta^{(0)}$ and the component weights $\pi^{(0)}$ can then be estimated on this compact subset. Subsequently, our constructive identification result can be leveraged to extend the estimator of $\pi^{(0)}$ to points in the covariate space outside this compact subset. We leave the exploration of such an estimation scheme for future research.

Given our results in Section 3, the next set of assumptions ensures that the component distributions and weights are identified.

Assumption E-4 *(Identification – Component distributions) (i) There exists $X^* \in \mathbb{X}$ such that Assumption I-2 or Assumption I-4 holds, (ii) $F_{jt} : \mathbb{R} \to [0,1]$ is continuously differentiable and strictly monotone for all j and t, and (iii) $\mathbb{E}[\tilde{x}_{it}\tilde{x}_{it}^\top]$ exists and has full rank with $\tilde{x}_{it} = (1, x_{it}^\top)^\top$ for all t.*

Assumption E-5 *(Identification – Component weights) There exists $\mathcal{X} \subseteq \mathbb{X}$ with $\mathbb{P}(\mathcal{X}) = 1$ such that for all $X \in \mathcal{X}$ there exists some submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$ such that $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank.*

Importantly, Assumption E-4(i) only requires that there exists a single X^* satisfying Assumption I-2 or Assumption I-4. Assumption E-4(ii) is satisfied, for instance, when the group-specific binary outcome models are probit or logit models. We note that the assumptions can be relaxed to allow for link functions that are strictly monotone on a subset of the real line. Doing so formally, however, goes beyond the scope of this paper. Overall, Assumption E-4 ensures that $\beta^{(0)}$ is identified, whereas Assumption E-5 ensures that $\pi_j^{(0)}(X)$ is identified F_X-almost surely for all $j = 1, \ldots, J$. When only $\beta^{(0)}$ is of interest, Assumption E-5 may be relaxed by using a weaker norm to show consistency, for instance, the Fisher norm induced by the objective function.

Following Ai and Chen (2003), we define the Hölder-norm for some $\eta > 0$ as follows:

$$\|f\|_{\Lambda^\eta} = \sup_{x \in \mathbb{X}} |f(x)| + \max_{|\lambda|=\underline{\eta}} \sup_{x \neq y} \left| \frac{\nabla^\lambda f(x) - \nabla^\lambda f(y)}{\|x - y\|_E^{\eta - \underline{\eta}}} \right|$$

where $\|\cdot\|_E$ is the Euclidean norm, $\underline{\eta}$ is the largest integer satisfying $\underline{\eta} < \eta$, $f : \mathbb{X} \to \mathbb{R}$ is $\underline{\eta}$-times continuously differentiable, and for any vector $\lambda = (\lambda_1, \ldots, \lambda_{d_X})$

$$\nabla^\lambda f(x) = \frac{\partial^{|\lambda|}}{\partial x_1^{\lambda_1} \ldots \partial x_{d_X}^{\lambda_{d_X}}} f(x)$$

where $|\lambda| = \sum_{k=1}^{d_X} \lambda_k$. Letting $\mathcal{C}_{\underline{\eta}}(\mathbb{X})$ be the space of $\underline{\eta}$-times continuously differentiable functions $f : \mathbb{X} \to \mathbb{R}$, we define a Hölder ball $\Lambda_M^\eta(\mathbb{X}) = \left\{ f \in \mathcal{C}_{\underline{\eta}}(\mathbb{X}) : \|f\|_{\Lambda^\eta} \leq M \right\}$ for some radius $M < \infty$. Letting $\|f\|_\infty = \sup_{X \in \mathbb{X}} |f(X)|$ be the supremum norm, we assume

Assumption E-6 *(Parameter space)* $\beta^{(0)} \in \mathcal{B} \subset \mathbb{R}^{dim(\beta)}$*, which is compact, and for all* $j = 1, \ldots, J-1$*,* $w_j^{(0)} = \log(\pi_j^{(0)}/\pi_J^{(0)}) \in \mathcal{W}_j = \Lambda_M^\eta(\mathbb{X})$ *for some* $M < \infty$ *and* $\eta > 0$*.*

Assumption E-7 *(Denseness of sieve space) (i)* $d(n) \to \infty$ *as* $n \to \infty$*, and (ii) there exists* $w_{d(n)} \in \mathcal{W}_{d(n)}$ *such that* $\sum_{j=1}^{J-1} \|w_{j,d(n)} - w_j^{(0)}\|_\infty \to 0$ *as* $n \to \infty$*.*

The second part of Assumption E-6 can be readily adapted by replacing $\Lambda_M^\eta(\mathbb{X})$ with another parameter space that is compact under the norm $\|\cdot\|_{c,\infty}$ we define below. Here, we choose $\Lambda_M^\eta(\mathbb{X})$ for concreteness. Given Assumption E-2 and the fact that the (multinomial) logit link function is Lipschitz continuous, it is straightforward, albeit tedious, to show that if $\pi_j^{(0)} \in \Lambda_{\tilde{M}}^\eta(\mathbb{X})$ for all $j = 1, \ldots, J-1$ and some $\tilde{M} < \infty$ and $\eta > 0$, then, under our assumptions, there exists $M < \infty$ such that $w_j^{(0)} \in \Lambda_M^\eta(\mathbb{X})$ for all $j = 1, \ldots, J-1$. Assumption E-7 is standard and simply states that the log-odds ratios can be approximated by the sieve space. Assumption E-7 holds for standard choices of basis functions $\rho^{d(n)}(X)$ (Ai and Chen 2003; Chen 2007).

For $\theta \in \Theta$, we define the norm $\|\theta\|_{c,\infty} = \sum_{j=1}^J \sum_{t=1}^T \|\beta_{j,t}\|_E + \sum_{j=1}^{J-1} \|\pi_j\|_\infty$. We are ready to state our consistency result.

Theorem 4.1 *(Consistency – Mixture parameters) Under Assumptions E-1 to E-7,* $\|\hat{\theta}_{w,n} - \theta_w^{(0)}\|_{c,\infty} = o_p(1)$ *and* $\|\hat{\theta}_n - \theta^{(0)}\|_{c,\infty} = o_p(1)$*.*

Of course, the result holds up to joint relabeling of the groups, which we suppress in

this section.[15] A direct corollary of Theorem 4.1 is that $\sum_{j=1}^{J-1} \|\pi_j^{(0)} - \hat{\pi}_j\|_{L_2(Q)} = o_p(1)$ for any probability measure Q where $\|\pi_j\|_{L_2(Q)} = \sqrt{\int \pi_j(X)^2 dQ(X)}$. We include the proof of Theorem 4.1 in Appendix A.4.1.1. Once consistency of the mixture parameters is established, the following result is immediate.

Corollary 4.1.1 *(Consistency – AMEs) Under Assumptions E-1 to E-7,* $\widehat{AME}_{j,t,k} = AME_{j,t,k} + o_p(1)$ *for all* j, t, k.

The proof is provided in Appendix A.4.1.2. Clearly, Corollary 4.1.1 implies that $\widehat{\text{AME}}_{j,k}$ converges to its population counterpart, $\text{AME}_{j,k}$, in probability, too. We conclude this section with a remark.

Remark 8. When $\mathbb{P}(g_i = j \mid X_i) = \mathbb{P}(g_i = j \mid m(X_i))$ for all $j = 1, \ldots, J$ and some known function $m : \mathbb{X} \to \mathbb{X}^* \subset \mathbb{R}^\ell$ with $\ell < \dim(\mathbb{X})$, Theorem 4.1 and Corollary 4.1.1 continue to hold with the original covariates X_i being replaced with $X_i^* = m(X_i)$ as long as the assumptions hold for X_i^* and $\mathbb{X}^*$. For instance, $\pi_j^{(0)}(X_i)$ may depend on a subset of X_i for all $j = 1, \ldots, J$. Another example is the case where $\pi_j^{(0)}(X_i) = \pi_j(\overline{X}_{i,1}, \ldots, \overline{X}_{i,K})$ where $\overline{X}_{i,k} = \frac{1}{T}\sum_{t=1}^{T} x_{it,k}$ for all $j = 1, \ldots, J$.

Our results can be easily adapted to a case where the component weights are functions of generalized linear models, for instance, when the latent group membership arises from an ordered choice model with a known error term distribution and an additive model as the index function.

4.3 Asymptotic distributions

In this section, we briefly discuss our results on the convergence rate of $\hat{\theta}_{w,n}$ and argue that the low-dimensional parameters $\hat{\beta}_n$ and $\widehat{\text{AME}}_{j,t,k}$ are $\sqrt{n}$-consistent. Using the results and arguments of, among others, Shen (1997), Chen and Shen (1998), Chen (2007), and Hu and Schennach (2008), we provide an in-depth discussion of the results in Appendix A.4.

Following Chen and Shen (1998) and Ai and Chen (2003), we first show that $\|\hat{\theta}_{w,n} - \theta^{(0)}\| = o_p(n^{-1/4})$ under standard conditions where $\|\cdot\|$ denotes the Fisher norm that is induced by the objective function and formally defined in Appendix A.4. Combining this

[15] From an estimation standpoint, the optimization routine simply picks one of the permutations of the groups. Subsequently, the econometrician assigns the groups their economic meaning.

rate with the arguments of Hu and Schennach (2008) allows us to show that $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimator for $\beta^{(0)}$. Additionally, under a combination of the conditions in Ai and Chen (2003) and Hu and Schennach (2008), the convergence rate result on $\hat{\theta}_{w,n}$ can be strengthened to hold under the stronger L_2-norm (Corollary A.4.4.1 in Appendix A.4.3.1). This result is helpful in deriving the limiting distribution of $\widehat{\text{AME}}_{j,t,k}$, which we show to be $\sqrt{n}$-consistent. The main results of Appendix A.4 are collected in the following proposition.

Proposition 4.1.1 *Under the assumptions of Theorem A.4.3,* $\sqrt{n}(\hat{\beta}_n - \beta^{(0)}) \xrightarrow{d} N(0, (V^*)^{-1})$. *Under the assumptions of Theorem A.4.5,* $\sqrt{n}(\widehat{AME}_{j,t} - AME_{j,t}) \xrightarrow{d} N(0, A_{j,t})$ *for all* j, t *and* $\sqrt{n}(\widehat{AME}_j - AME_j) \xrightarrow{d} N(0, A_j)$ *for all* j *with* $\widehat{AME}_{j,t} = (\widehat{AME}_{j,t,1}, \ldots, \widehat{AME}_{j,t,K})^\top$ *and* $AME_{j,t}$, $\widehat{AME}_j$, *and* AME_j *defined similarly.* V^*, $A_{j,t}$, *and* A_j *are defined in Appendix A.4.3.*

As the asymptotic covariance expressions are lengthy, we omit them here for brevity. We prove Proposition 4.1.1 using a combination of theorems in Appendix A.4.3. While we do not provide an estimator for asymptotic variances, Proposition 4.1.1 and the respective proofs motivate and justify the use of a nonparametric bootstrap for inference under standard regularity conditions. In particular, one may mimic our proofs on the bootstrap sample to derive the same asymptotic linearizations on the bootstrap sample. Then, Theorem 1 in Mammen (1992) justifies the use of the nonparametric bootstrap. Alternatively, the arguments in Ackerberg et al. (2012) and Chen and Liao (2015) suggest that once the sieve dimension is chosen we can derive an estimator for the asymptotic variance pretending that the model is parametric. We implement and study the latter estimator in our simulation study.

5 Simulation study

In this section, we study the finite-sample properties of the proposed estimators in a simulation study. The baseline setup is motivated by Example 2.1 but the slope coefficients do not vary over time, that is,

$$y_{it} = \sum_{j=1}^{J} \mathbb{1}(g_i = j) \mathbb{1}\left(\sum_{k=1}^{3} x_{it,k}^\top \beta_{j,k}^{(0)} + \alpha_{j,t}^{(0)} - \varepsilon_{j,it} \geq 0 \right) \text{ for } i = 1, \ldots, n \text{ and } t = 1, \ldots, T$$

In our baseline settings, $J = 2$, $n \in \{1000, 2000\}$, and $T = 3$. We replicate each simulation 1000 times and draw $\varepsilon_{j,it} \sim N(0,1)$ for all i, t, and j. The covariates X_i are generated as follows: We first draw $\tilde{X}_i \sim N(\mathbf{0}, \Sigma)$ where $\mathbf{0}$ is the $3T$-dimensional zero vector and $Var(\tilde{x}_{it,j}) = 1$, $Cov(\tilde{x}_{it^*,k}, \tilde{x}_{it,j}) = 0.5^{|k-j|}$ if $t^* = t$, and $Cov(\tilde{x}_{it^*,k}, \tilde{x}_{it,j}) = \mathbb{1}(j = k)0.2^{|t^*-t|}$ if $t^* \neq t$. Then, $X_i = 2\Phi(\tilde{X}_i) - 1$ where $\Phi(\cdot)$ is the CDF of a standard normally distributed random variable and the transformation is applied element-wise. Hence, $x_{it,k}$ is marginally uniformly distributed on $[-1, 1]$ for all k and t.

In the setting with two groups and three time periods, the parameters are

$$\beta_1^{(0)} = (0.8, -0.4, -0.3)^\top \; ; \quad \beta_2^{(0)} = (-1.3, -0.2, -1.0)^\top$$
$$\alpha_1^{(0)} = (1, -1.5, -0.6)^\top \; ; \quad \alpha_2^{(0)} = (-0.7, 1.0, 0.7)^\top$$

In each simulation run, we initialize the parameter estimates for $\beta^{(0)}$ around the true parameters plus uniform noise on $[-0.5, 0.5]$, whereas the initial γ_n is drawn from a uniform distribution with support $[-1, 1]$.[16] To approximate the weights, we use orthogonal Legendre polynomials of order $d(n) = \lceil n^{1/7} \rceil$. This choice of $d(n)$ is of similar order as in Ai and Chen (2003) and chosen to satisfy the rate requirements imposed on $d(n)$ in Appendix A.4.

The group membership is generated by an ordered probit model, which we outline below. We focus on two different group membership generating processes: first, the component weights depend on X_i only through the time averages of the covariates; second, the component weights depend solely on $\{x_{it,1}\}_{t=1}^T$. The first modeling approach may be empirically justified by interpreting temporal averages as sufficient statistics for the group membership, which may be reasonable in short panels. The second modeling approach introduces an additional exclusion restriction in the weights and assumes that conditional on $\{x_{it,1}\}_{t=1}^T$ $\{x_{it,k}\}_{t=1,\ldots,T;k\neq 1}$ are independent of g_i. Such exclusion restrictions may be motivated by additional domain knowledge. Both settings allow us to put more structure on the component weights, which we enforce in the estimation procedure. We proceed to discuss the two settings separately.[17] Appendix A.11 includes additional simulation results.

[16] In empirical practice, we recommend to use multiple starting values and choose the estimate that maximizes the log-likelihood. We do so in our empirical application.

[17] Another semiparametric restriction, which may be reasonable in some settings, is to assume that the component weights are symmetric in their arguments across time.

5.1 Weights – Time averages

We define $\overline{X}_{i,k} = \frac{1}{T}\sum_{t=1}^{T} x_{it,k}$ for $k = 1, 2, 3$. For $J = 2$, the groups are generated as $g_i = \mathbb{1}\left(\zeta_0 + \zeta_1 \overline{X}_{i,1}^2 + \zeta_2 \log(\overline{X}_{i,2}^2 + 1) + \zeta_3 \overline{X}_{i,3} - u_i \geq 0\right) + 2\mathbb{1}\left(\zeta_0 + \zeta_1 \overline{X}_{i,1}^2 + \zeta_2 \log(\overline{X}_{i,2}^2 + 1) + \zeta_3 \overline{X}_{i,3} - u_i < 0\right)$ with $u_i \sim N(0,1)$ independent of X_i and $\varepsilon_{j,it}$ for all j and t, and $\zeta = (0, 0.8, -1, 1.2)^\top$. Under this group assignment mechanism, the two groups are of similar size in expectation.

Table 1 reports the simulation results for the estimator of $\beta^{(0)}$ and for the percentage of individuals correctly assigned to their true group based on their posterior group probabilities for sample sizes $n \in \{1000, 2000\}$. Approximately 91% of all units are assigned to the correct group. The bias of the estimator is small and the standard deviations scale approximately with $1/\sqrt{2}$ as we double the sample size from 1000 to 2000, which is to be expected since the estimator is $\sqrt{n}$-consistent.

Table 1: Simulation results for the coefficients. Weights depend on time averages and $J = 2$.

	$n = 1000$ and $T = 3$						$n = 2000$ and $T = 3$					
	Bias		**St. dev.**		**RMSE**		**Bias**		**St. dev.**		**RMSE**	
	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2
α_1	0.012	-0.004	0.102	0.102	0.103	0.102	0.003	-0.006	0.067	0.067	0.067	0.068
α_2	-0.011	0.010	0.144	0.136	0.145	0.136	-0.005	0.012	0.101	0.093	0.101	0.094
α_3	-0.009	0.013	0.073	0.097	0.074	0.098	0.000	0.007	0.051	0.070	0.051	0.070
β_1	0.011	-0.018	0.122	0.127	0.123	0.129	0.001	-0.006	0.085	0.091	0.085	0.092
β_2	-0.002	0.003	0.116	0.105	0.116	0.105	-0.003	-0.005	0.078	0.072	0.078	0.073
β_3	0.001	-0.012	0.106	0.117	0.106	0.118	0.000	-0.003	0.072	0.077	0.072	0.077
Group assign.	91.34%						91.76%					

Table 2 presents the simulation results for the group-specific AMEs for $n \in \{1000, 2000\}$. All estimators have a small bias and, in line with our theoretical results, the RMSEs decrease by a factor of approximately $\sqrt{2}$ as the sample size increases from 1000 to 2000. Using an estimator for the asymptotic variance as suggested by Ackerberg et al. (2012), the actual coverage rates of the confidence intervals are close to the nominal rate of 95% when $n = 1000$ and mostly match this nominal rate when $n = 2000$.

Table 3 summarizes the simulation results for the group-specific AMEs in a setting with two time periods. In this setup, the GFEs are $\alpha_1^{(0)} = (1, -1.5)^\top$ and $\alpha_2^{(0)} = (-0.7, 1.0)^\top$, with the slope coefficients unchanged. As discussed in Section 3.2, compared to a setting with three time periods, identification with two time periods requires additional variation in

Table 2: Simulation results for the AMEs. Weights depend on time averages and $J = 2$.

	$n = 1000$ and $T = 3$						$n = 2000$ and $T = 3$					
	Gr. 1			Gr. 2			Gr. 1			Gr. 2		
	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$
Bias												
x_1	-0.0006	0.0009	0.0004	-0.0008	-0.0002	0.0000	-0.0009	-0.0002	-0.0005	0.0004	0.0016	0.0009
x_2	0.0010	-0.0002	0.0008	0.0012	0.0011	0.0013	-0.0002	-0.0003	-0.0006	-0.0010	-0.0006	-0.0008
x_3	0.0011	-0.0001	0.0011	-0.0004	-0.0002	0.0001	0.0002	-0.0002	0.0001	0.0005	0.0014	0.0009
RMSE												
x_1	0.0265	0.0215	0.0325	0.0216	0.0208	0.0214	0.0180	0.0147	0.0234	0.0157	0.0156	0.0152
x_2	0.0278	0.0179	0.0344	0.0253	0.0202	0.0226	0.0188	0.0117	0.0236	0.0176	0.0139	0.0158
x_3	0.0262	0.0167	0.0322	0.0237	0.0222	0.0226	0.0179	0.0112	0.0219	0.0158	0.0145	0.0145
Coverage												
x_1	0.9410	0.9360	0.9350	0.9570	0.9550	0.9600	0.9470	0.9450	0.9130	0.9480	0.9450	0.9520
x_2	0.9570	0.9330	0.9490	0.9610	0.9560	0.9580	0.9570	0.9470	0.9490	0.9560	0.9580	0.9530
x_3	0.9410	0.9250	0.9370	0.9380	0.9370	0.9360	0.9550	0.9450	0.9470	0.9480	0.9530	0.9540

Note: The nominal coverage rate is 95%.

the component weights, which may be difficult to detect in finite samples. Combining this observation with the fact that the panel is one time period shorter, we expect the group-specific AME estimators to have larger RMSEs in a setting with two time periods. This is what we observe in Table 3. Importantly, however, the RMSEs still scale with approximately $1/\sqrt{n}$ as we increase the sample size. The additional statistical difficulty of the problem is also reflected in the actual coverage rates. When $n = 1000$, the confidence intervals slightly undercover. However, when the sample size increases to 2000, the actual coverage rates increase and are closer to or around the nominal coverage rate of 95%.[18]

We extend the baseline setting to a case with $J = 3$ groups; we outline the exact data-generating process in Appendix A.11.1. Although similar arguments to those in Section 7.2 indicate that identification may be achieved with fewer periods, we restrict our attention to a setting with $T = 4$ that is in line with our main identification results of Section 3. Table 4 reports the simulation results for the group-specific AMEs when $n = 2000$. As expected, the percentage of correctly assigned units decreases as the number of groups increases. Similar to the baseline specification, the bias of the estimators for the AMEs is small, while the

[18]In a future version of this paper, we plan to study whether the coverage can be improved by using an estimator for the asymptotic variance that does not impose the Fisher information equality. Such an estimator should approximate the curvature of the finite sample objective function better and may therefore be a better finite sample approximation of the Fisher norm we use to derive the variance estimator; we refer to Appendix A.4.2.1 for the definition of the Fisher norm and in what sense the Fisher information identity holds.

Table 3: Simulation results for the AMEs. Weights depend on time averages, $J = 2$, and $T = 2$.

	$n = 1000$ and $T = 2$				$n = 2000$ and $T = 2$			
	Gr. 1		Gr. 2		Gr. 1		Gr. 2	
	$t = 1$	$t = 2$	$t = 1$	$t = 2$	$t = 1$	$t = 2$	$t = 1$	$t = 2$
Bias								
x_1	-0.0023	-0.0001	0.0011	0.0022	0.0000	-0.0003	0.0009	0.0015
x_2	0.0009	0.0000	-0.0010	-0.0003	0.0013	0.0009	-0.0003	-0.0001
x_3	0.0032	0.0003	0.0001	0.0008	-0.0001	-0.0004	-0.0009	-0.0002
RMSE								
x_1	0.0414	0.0277	0.0334	0.0297	0.0274	0.0193	0.0217	0.0190
x_2	0.0441	0.0256	0.0368	0.0274	0.0313	0.0176	0.0242	0.0183
x_3	0.0401	0.0244	0.0348	0.0314	0.0273	0.0165	0.0223	0.0197
Coverage								
x_1	0.9080	0.9130	0.9060	0.9160	0.9320	0.9240	0.9430	0.9470
x_2	0.9270	0.9160	0.9320	0.9350	0.9280	0.9290	0.9510	0.9490
x_3	0.9310	0.9180	0.9280	0.9010	0.9430	0.9360	0.9510	0.9440

Note: For $T = 2$, the GFEs are $\alpha_1^{(0)} = (1, -1.5)^\top$ and $\alpha_2^{(0)} = (-0.7, 1.0)^\top$. The nominal coverage rate is 95%.

actual coverage rates are mostly close to their nominal counterpart of 95%. In Appendix A.11.1, we also include the results for $n = 1000$. Consistent with our theoretical results, the RMSEs of the estimators scale approximately with $1/\sqrt{n}$.

Table 4: Simulation results for the AMEs. Weights depend on time averages and $J = 3$.

	Gr. 1				Gr. 2				Gr. 3			
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 1$	$t = 2$	$t = 3$	$t = 4$
Bias												
x_1	-0.0023	-0.0018	0.0001	0.0020	0.0008	-0.0012	0.0000	0.0003	-0.0007	-0.0013	0.0006	-0.0004
x_2	0.0004	0.0004	-0.0011	-0.0023	0.0018	0.0013	0.0017	0.0018	0.0017	0.0013	0.0015	0.0015
x_3	0.0011	0.0008	0.0004	-0.0003	0.0009	-0.0006	0.0004	0.0005	0.0009	0.0009	0.0005	0.0007
RMSE												
x_1	0.0192	0.0188	0.0211	0.0228	0.0202	0.0199	0.0203	0.0204	0.0178	0.0186	0.0200	0.0186
x_2	0.0207	0.0152	0.0272	0.0312	0.0224	0.0203	0.0223	0.0236	0.0207	0.0202	0.0152	0.0182
x_3	0.0177	0.0124	0.0230	0.0262	0.0186	0.0180	0.0185	0.0191	0.0212	0.0205	0.0144	0.0182
Coverage												
x_1	0.9400	0.9180	0.9420	0.9370	0.9400	0.9480	0.9500	0.9400	0.9400	0.9310	0.9360	0.9390
x_2	0.9280	0.8990	0.9310	0.9330	0.9490	0.9520	0.9520	0.9530	0.9320	0.9270	0.9310	0.9300
x_3	0.9360	0.9230	0.9330	0.9370	0.9560	0.9570	0.9550	0.9520	0.9350	0.9390	0.9400	0.9350
Group assign.	85.12%											
n	2000											
T	4											

5.2 Weights – Single covariate

Now, X_i enters the component weights solely through $\{x_{it,1}\}_{t=1}^3$. In particular, for $J = 2$ and $T = 3$, the groups are generated with $g_i = \mathbb{1}\left(\zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + \zeta_3 x_{i3,1}^2 - u_i \geq 0\right) + 2\mathbb{1}\left(\zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + \zeta_3 x_{i3,1}^2 - u_i < 0\right)$ where $u_i \sim N(0,1)$ is independent of X_i as well as $\varepsilon_{j,it}$ for all j and t, and $\zeta = (0.25, 1.2, -1.4, -0.2)^\top$. In expectation, the two groups are of similar size. We include additional results for this setting in Appendix A.11.2.

Table 5 displays the simulation results for the estimator of $\beta^{(0)}$ and for the percentage of correctly assigned individuals based on their posterior group probabilities for $n \in \{1000, 2000\}$. Approximately 92% of all units are assigned to the correct group. As in the previous simulation setting, the bias of the estimator is small and the standard deviations scale approximately with $1/\sqrt{2}$ as we double the sample size from 1000 to 2000. The same holds true for the estimators of the group-specific AMEs (Table 6).

Table 5: Simulation results for the coefficients. Weights depend on the first covariate only and $J = 2$.

	$n = 1000$ and $T = 3$						$n = 2000$ and $T = 3$					
	Bias		**St. dev.**		**RMSE**		**Bias**		**St. dev.**		**RMSE**	
	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2
α_1	0.012	-0.007	0.107	0.103	0.108	0.103	0.005	-0.005	0.074	0.067	0.074	0.067
α_2	-0.012	0.020	0.143	0.137	0.144	0.138	-0.007	0.010	0.093	0.083	0.093	0.084
α_3	-0.006	0.012	0.081	0.091	0.081	0.092	0.000	0.004	0.054	0.062	0.054	0.062
β_1	0.002	-0.015	0.132	0.123	0.132	0.124	0.003	-0.003	0.088	0.082	0.088	0.082
β_2	-0.003	0.003	0.122	0.094	0.122	0.094	-0.004	-0.002	0.085	0.068	0.085	0.068
β_3	0.001	-0.018	0.110	0.105	0.110	0.106	-0.002	-0.007	0.077	0.070	0.077	0.071
Group assign.	91.9%						92.43%					

6 Empirical illustration

We illustrate the proposed estimators in the context of homeownership in Germany. Focusing on two-adult households,[19] we model the homeownership decision of a household and explore how this decision varies across different latent household types.

[19] Two-adult households are households with a household head and a partner of the household head. This contrasts, for instance, with single-adult households which include only a household head.

Table 6: Simulation results for the AMEs. Weights depend on the first covariate only and $J = 2$.

	$n = 1000$ and $T = 3$						$n = 2000$ and $T = 3$					
	Gr. 1			Gr. 2			Gr. 1			Gr. 2		
	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$
Bias												
x_1	-0.0020	-0.0009	-0.0020	-0.0003	0.0019	0.0009	-0.0004	-0.0002	-0.0002	0.0008	0.0016	0.0009
x_2	0.0006	0.0002	0.0004	0.0012	0.0012	0.0012	-0.0004	-0.0003	-0.0009	-0.0003	0.0000	-0.0002
x_3	0.0008	0.0002	0.0011	-0.0019	0.0000	-0.0010	-0.0002	-0.0002	-0.0003	-0.0008	0.0001	-0.0005
RMSE												
x_1	0.0275	0.0235	0.0365	0.0237	0.0184	0.0196	0.0191	0.0161	0.0250	0.0163	0.0138	0.0143
x_2	0.0252	0.0194	0.0371	0.0244	0.0186	0.0211	0.0179	0.0134	0.0260	0.0177	0.0136	0.0154
x_3	0.0235	0.0181	0.0342	0.0212	0.0190	0.0202	0.0166	0.0123	0.0242	0.0149	0.0129	0.0135

6.1 Background

Compared to other high-income countries, the homeownership rate in Germany is low with around 47.2% in 2024 (Eurostat 2025). This low homeownership rate can be partly explained by a large social housing sector with lenient eligibility requirements, high transaction costs when purchasing a dwelling, and the absence of a mortgage interest tax reduction scheme for owner-occupiers (Kaas et al. 2021). Although homeownership reduces mobility and may increase the investment risk of homeowners due to a large, potentially highly levered initial investment (Flavin and Yamashita 2002; Dietz and Haurin 2003), it has been associated with positive and socially desirable outcomes. For example, homeownership allows households to hedge against rent risk and inflation (Sinai and Souleles 2005; Malmendier and Wellsjo 2024), may boost small business employment (Adelino et al. 2015), allows households to accumulate wealth over time (Haurin et al. (2002) and Dietz and Haurin (2003) and references therein), results in stronger social and political engagement (DiPasquale and Glaeser 1999), and is associated with better health as well as improved child outcomes in terms of years of schooling and lifetime income (Green and White 1997; Dietz and Haurin 2003).

Given these previous points, increasing homeownership rates appears desirable from a welfare perspective. Kaas et al. (2019) find, among other results, that a policy that abolishes social housing and provides a subsidy for low-income households increases the homeownership rate and has positive welfare effects across the entire income distribution. Interpreting a subsidy as an income transfer, we shall more generally study which types of households

would most strongly react to a subsidy by increasing their likelihood to own. Our results suggest that income effects are heterogeneous across three latent household types.

6.2 Empirical strategy

Our empirical analysis uses a subset of the German Socio-Economic Panel (GSOEP) (Goebel et al. 2019). In particular, we use the years 2014 to 2019 of version 39 of the GSOEP (SOEP 2024). The GSOEP is an annual longitudinal survey of private households in Germany that has been conducted since 1984.

The empirical analysis is based on the following outcome equation:

$$y_{it} = \mathbb{1}\left(x_{1,it}\beta_{1,g_i} + x_{2,it}^{\top}\beta_2 + \sum_{f=1}^{15} \mathbb{1}(f_{it} = f)\lambda_f + \alpha_{g_i,t} - \varepsilon_{it} \geq 0 \right) \tag{7}$$

where i indexes a household and t a time period. The dependent variable y_{it} is an indicator for homeownership equal to 1 if household i is a homeowner at time t and equal to 0 otherwise. We distinguish between two classes of covariates based on whether the associated parameter is permitted to vary across groups. $x_{1,it}$ is equal to the signed fourth root of the yearly net household income deflated to 2015 euros using the CPI published by the German Federal Statistical Office (2025).[20] $x_{2,it}$ contains a dummy for the presence of children younger than 18 years old, the age and age squared of the household head,[21] a dummy for long-term unemployment of the household head, a dummy for the marital status of the household head, a gender dummy for the household head with female coded as 1, and two educational dummies for the household head. Additionally, we include 15 (out of 16) state dummies to allow for state fixed effects. f_{it} denotes the federal state household i resides in at time t. Model (7) is a similar, albeit simplified version of the model in Diaz-Serrano (2005) who studies the effect of labor income uncertainty on homeownership. We define an individual to be long-term unemployed if she has been unemployed in at least two consecutive survey periods. We include one educational dummy for general elementary and middle vocational education (*general-elementary educ.*) and one educational dummy for Abitur

[20] Similar to Honoré and Weidner (2025), though in a different setting, we use the signed fourth root to capture decreasing marginal effects of income on the homeownership decision.

[21] Even though the covariate age does not satisfy the support condition of Assumption I-1, the discussions in Appendix A.1.9 imply that all model parameters are still identified.

plus vocational training and higher vocational training (*Abi-vocational educ.*).[22] This leaves tertiary education as the reference group because we drop all observations from our analysis for which the household head or partner is still in school or classified as having inadequate schooling.[23] An individual has inadequate schooling if she has attended at most four years of some kind of primary education.

Model (7) allows for various sources of endogeneity. In particular, the GFEs $\alpha_{g_i,t}$ capture different baseline probabilities to own a home across the different latent groups. For example, when the latent groups reflect different wealth levels, the GFEs capture wealth effects in that wealthier households are more likely to be homeowners all else equal. Since wealth is correlated with income and age, ignoring the group structure would introduce an endogeneity bias (Di 2007; Turner and Luea 2009). At the same time, homeownership itself affects the wealth accumulation of households (Di et al. 2007). This simultaneity problem may be captured by the GFEs, too. Similarly, the GFEs may capture borrowing constraints that a household is facing but are unobserved to the econometrician. Since borrowing constraints are highly correlated with household income and wealth, ignoring the group structure introduces an endogeneity problem (Acolin et al. 2016). Because wealth accumulation and borrowing constraints evolve over time, it is essential that the GFEs are allowed to be time-varying. This is an advantage of our method over an approach that only includes additive individual-specific fixed effects. Of course, in doing so, we assume that the source of endogeneity and heterogeneity can be captured by a finite number of groups.

Another advantage of model (7) over typical binary outcome models is that it allows the slope coefficient on net household income to vary across groups. This feature accounts for heterogeneity in the homeownership decision to income changes: For example, households that are close to the borrowing constraint may strongly benefit from an income increase as it becomes cheaper to finance a home, while low-wealth and high-wealth households may react less to income changes because the borrowing constraint is either tightly or far from binding. Further sources of endogeneity may arise from heterogeneity in risk preferences or

[22] These dummies are based on the variable *pgisced97* and its categories. *general-elementary educ.* includes individuals that have some kind of secondary education and *Abi-vocational educ.* corresponds approximately to post-secondary but non-tertiary education.

[23] The results are robust to including the households with inadequately schooled household head or partner in the analysis, see Appendix A.2.3.

income uncertainty across households (Diaz-Serrano 2005).

Appendix A.2.4 provides a detailed description of the construction of the data set. Most importantly, we drop all observations from our analysis for which the dependent variable or any of the covariates is missing for either the household head or her partner. Also, while our theoretical results can be adapted to unbalanced panels under standard missing-at-random assumptions, we balance the panel. This leaves us with a panel of $T = 6$ time periods and $n = 2089$ household observations. Approximately 11.4% of these households change their ownership situation over the survey years we consider and 60.5% of the household-time observations are homeowners.[24]

Following Diaz-Serrano (2005), we assume that $\varepsilon_{it} \mid g_i \sim N(0,1)$ for all i and t so that, under the assumptions of Example 2.1, model (7) falls into the class of mixture models studied in this paper and is a special case of the model we propose an estimator for. Similar to the simulation setup of Section 5.1, we assume that the component weights $\pi_j(X)$ are a function of the time averages of the covariates. Specifically, the component weights depend on the time average of the deflated yearly net household income, $\overline{x^4}_{1,i} = \frac{1}{6}\sum_{t=1}^{6} x^4_{1,it}$, and the time-averaged age of the household head, $\overline{Age}_i = \frac{1}{6}\sum_{t=1}^{6} Age_{it}$. This allows the homeownership decision to systematically differ for households with different (time-averaged) income or age of the household head and captures the sources of endogeneity discussed previously. While we exclude some covariates from the component weights, these covariates are still allowed to be correlated with the group indicator g_i as long as this correlation is through the time-averaged household income or the age of the household head.

As in the simulation studies, we let $d(n) = \lceil 2089^{1/7} \rceil = 3$. Following the literature on mixture models, we choose the number of groups J via the Bayesian Information Criterion (BIC) defined as $-2n\hat{\mathcal{L}}_n^{d(n)}(\beta, \gamma_n) + \log(n)(\dim(\beta) + (J-1)(d(n)+1))$ where $\hat{\mathcal{L}}_n^{d(n)}(\beta, \gamma_n)$ is the empirical log-likelihood once the sieve dimension is fixed (equation (6)) and $\dim(\beta) + (J-1)(d(n)+1)$ are the degrees of freedom of the model (Celeux et al. 2019; Hansen 2022). While we do not show that the BIC is consistent for the true number of groups J in our

[24]The larger share of homeowners is due to the focus on two-adult households and balancing the panel. If we include other types of households as well while keeping the remaining sample selection restrictions identical, the homeownership rate drops to 43.7%. In the unbalanced version of our panel, the homeownership rate is about 54%.

setting, the BIC has been shown to have desirable consistency properties when the mixture model of interest is identified (Celeux et al. 2019). We leave it for future research to study the properties of the BIC for model selection in our setting. To perform tests and construct confidence intervals, we use the nonparametric bootstrap.

6.3 Results

The BIC selects $J = 3$ groups in our setting. Table 7 collects the estimated coefficients. We find that the different groups have different baseline probabilities to be a homeowner as expressed through the different GFEs. While the GFEs are similar over time for Group 1 and 3, they increase for Group 2, indicating that households in this group become increasingly likely to be homeowners over time. A mechanism at play may be an increase in wealth of Group 2 households that is not captured by income or age. Additionally, we see that the coefficient on the fourth root of net household income varies across the groups. With 0.246 and 0.269, respectively, it is of similar magnitude and statistically significantly different from zero for Group 1 and 2. The estimate is not significantly different from zero for the third group. Because the coefficient itself carries little meaning beyond its sign, we study the group-specific effects of the household income on homeownership in more detail below by considering the group-specific AMEs. The remaining estimates that are constant across the groups are of expected sign: In previous studies, a similar negative relationship of long-term unemployment with homeownership and a similar positive relationship of marriage with homeownership have been reported (Diaz-Serrano 2005; Thomas and Mulder 2016).

Before turning our attention to the group-specific AMEs, we illustrate how to make the latent groups interpretable. To this end, Figure 1 presents three contour plots of the component weights.[25] Since only six observations fall into the regions of low income/low age and high income/low age, we truncate the lower-left and lower-right corners of the contour plots to prevent misinterpretation arising from extrapolation of the nonparametric estimates beyond the data support. Figure A.3 in Appendix A.2.2 depicts which subsets of the support

[25]When the component weights are a function of more than two variables, it is impossible to plot the component weights as in Figure 1. Then, one may plot "marginalized component weights", that is, one may consider, for instance, $\mathbb{P}(g_i = j \mid \overline{Age}_i) = \mathbb{E}[\pi_j(\overline{x^4}_{1,i}, \overline{Age}_i) \mid \overline{Age}_i]$, which can be estimated via a plug-in series estimator $\widehat{\mathbb{E}}[\hat{\pi}_j(\overline{x^4}_{1,i}, \overline{Age}_i) \mid \overline{Age}_i]$. To illustrate this, we include the plots of the marginalized component weights for the setup at hand in Appendix A.2.1.

Table 7: Parameter estimates

	Group 1	**Group 2**	**Group 3**
Time-varying group-specific fixed effects			
$t = 1$	-4.610^{**}	-8.264^{***}	-7.449^{**}
$t = 2$	-3.537^{*}	-11.295^{***}	-7.899^{***}
$t = 3$	-4.038^{*}	-7.047^{***}	-7.734^{**}
$t = 4$	-4.247^{**}	-6.358^{**}	-8.328^{**}
$t = 5$	-4.275^{**}	-5.422^{**}	-8.276^{***}
$t = 6$	-4.343^{**}	-4.739^{*}	-7.444^{**}
Group-specific slope coefficient			
Net household income	0.246^{***}	0.269^{**}	0.162
Group-invariant slope coefficients			
Age		0.072	
Age-squared		0.000	
Children		0.108	
Female		0.101	
Long-term unemployed		-0.411^{**}	
Married		0.636^{***}	
Abi-vocational educ.		0.461^{**}	
General-elementary educ.		0.329^{**}	

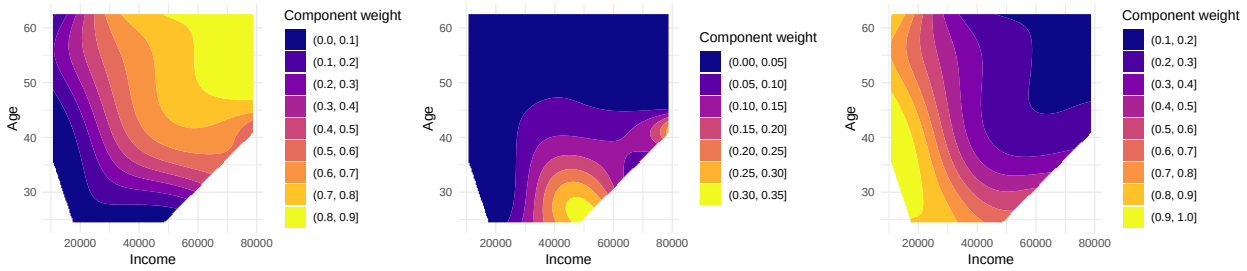
Note: Significantly different from 0 at 10%(*), 5%(**), 1%(***) using a two-sided t-test. The critical values are the respective quantiles of the nonparametric bootstrap distribution. We do not include the state-specific dummies for readability purposes.

are affected by this extrapolation issue. The brighter a region in Figure 1, the larger the component weight, that is, the more likely a household is to be a member of the respective group.

Focusing on Figure 1a first, households with a higher income and higher age of the household head are more likely to be in this group. Taken together with the evidence in Table 7, which shows comparatively large GFEs for Group 1 and thus indicates a higher baseline probability of homeownership, this suggests that Group 1 may be interpreted as the group of (always-)homeowners, high-wealth households, or households with loose borrowing constraints. Turning to Group 3 next, we observe the opposite relationship in Figure 1c: Households with lower income and younger household heads are more likely to be a member of this group. When combined with the comparatively small GFEs and therefore a lower baseline probability to own, this observation suggests that Group 3 may be interpreted as the group of (always-)renters, low-wealth households, or households with tight borrowing constraints. Lastly, according to Figure 1b, younger households with medium to high in-

come are comparatively more likely to be members of Group 2. Combining this observation with the fact that the GFEs of Group 2 are increasing over time suggests that Group 2 is comprised of younger, medium- to high-income households that transition into homeownership. Intuitively, this group should react on the margin most strongly to an increase in household income through, for instance, a subsidy.

Figure 1: Contour plot of the component weights as a function of the time averages of income and age



(a) Group 1 (b) Group 2 (c) Group 3

This interpretation of the groups is further supported by Table 8 which reports the group-specific time-averaged means of the outcome variable and covariates, along with the unconditional group membership probabilities. Group 1 is the group of on average higher income and older homeowners, while Group 3 is the group of on average lower income renters who, on average, exhibit a substantially larger long-term unemployment rate compared to the other two groups. Households in Group 2 are on average younger than households in the other groups but have a higher average net household income than the group of (always-)renters. Also, Group 2 households are more likely to have children younger than 18 years old who live at home. To further highlight the interpretation of Group 2 as the households that transition into homeownership over time, we note that the unconditional homeownership probability increases from 5.3% in 2014 to 96.8% in 2019 for this group. In terms of prevalence, Group 1 is the largest, accounting for approximately 58.5% of the population, followed by Group 3 with 36%, while Group 2 is the smallest, representing only 5.5% of the population.

To assess the effects of the covariates on the decision to buy a home across the different groups, Table 9 presents group-specific AMEs for non-binary and group-specific APEs for

Table 8: Group-specific time-averaged means of the variables

	Group 1	**Group 2**	**Group 3**
Homeownership rate			
	0.9807	0.4776	0.0102
π_j			
	0.5853	0.0549	0.3598
Covariates			
Net household income	46,502	43,822	38,543
Age	48.85	39.69	45.42
Children	0.5729	0.7886	0.6087
Female	0.3080	0.4000	0.3499
Long-term unemployed	0.0760	0.0622	0.1221
Married	0.9172	0.8846	0.8951
Abi-vocational educ.	0.1811	0.1880	0.1736
General-elementary educ.	0.5141	0.5247	0.6050
Tertiary educ.	0.3048	0.2873	0.2214

binary covariates. As expected, Group 2 exhibits the strongest income effect. For example, an increase of 13,500 euros from the group-average of 43,822 euros raises the probability to be a homeowner by about 4.5% on average.[26] Although income also has a statistically significant effect on homeownership for Groups 1 and 3, the AMEs are considerably smaller. Overall, Table 9 highlights that marginal changes in the covariates have larger effects on the homeownership decision of Group 2 households compared to households of the other groups. Intuitively, while Group 1 and Group 3 households are likely to continue to own or rent, respectively, Group 2 households are deciding on transitioning into homeownership over the sample period. Consequently, on the margin, these households are expected to react most strongly to changes in their environment that affect their homeownership decision. In particular, it seems as if these households are close to their borrowing constraint that may become (un)binding as the covariates vary minimally. On the other hand, Group 1 and 3 households' borrowing constraints are not much affected by a change in the covariates as these constraints are very loose or tight, respectively, to begin with. To conclude, if the goal of a policy is to increase the homeownership rate, our analysis suggests that such a policy should target households with medium to high income and younger household heads.

[26]The increase of 13,500 euros corresponds approximately to an increase of $x_{1,it}^{1/4}$ by one.

Table 9: Group-specific time-averaged AMEs and APEs

	Group 1	**Group 2**	**Group 3**
Net household income	0.0095^{***}	0.0450^{**}	0.0039^{*}
Age	0.0028	0.0121	0.0017
Age-squared	0.0000	-0.0001	0.0000
Children	0.0043	0.0181	0.0026
Female	0.0038	0.0169	0.0025
Long-term unemployed	-0.0212^{*}	-0.0702^{**}	-0.0070^{**}
Married	0.0382^{*}	0.1104^{**}	0.0089^{***}
Abi-vocational educ.	0.0140^{***}	0.0754^{**}	0.0152^{*}
General-elementary educ.	0.0135^{**}	0.0547^{**}	0.0079^{**}

Note: Significantly different from 0 at 10%(*), 5%(**), 1%(***) using a two-sided t-test. The critical values are the respective quantiles of the nonparametric bootstrap distribution. We do not include the state-specific dummies for readability purposes.

7 Extensions

7.1 Multinomial outcomes

We now discuss the case when the support of y_{it} is $\{1, \dots, S\}$ for $S > 2$. The arguments of Section 3.2 continue to apply with only minor modifications. Specifically, we define the matrix

$$\tilde{\mathcal{P}}_t(x_t) = \begin{pmatrix} \mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = 1) & \dots & \mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, g_i = J) \\ \vdots & \ddots & \vdots \\ \mathbb{P}_t(y_{it} = S \mid x_{it} = x_t, g_i = 1) & \dots & \mathbb{P}_t(y_{it} = S \mid x_{it} = x_t, g_i = J) \end{pmatrix}$$

and for some submodel $\mathcal{I} \subseteq \{1, \dots, T\}$, we have $\tilde{\mathcal{P}}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}}) = \overset{\text{col}}{\otimes}_{t\in\mathcal{I}} \tilde{\mathcal{P}}_t(x_t)$. While we maintain Assumption I-3, we adapt Assumption I-2 as follows

Assumption Ĩ-2 *(Variation in component distributions) There exist $t' \in \{1, \dots, T\}$ as well as disjoint and non-empty submodels $\mathcal{I}_1$ and $\mathcal{I}_2$ with $t' \notin \mathcal{I}_1 \cup \mathcal{I}_2$ such that $\tilde{\mathcal{P}}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\mathcal{I}_\ell})$ has full column rank for $\ell = 1, 2$ and $\tilde{\mathcal{P}}_{t'}(x_{t'})$ has distinct columns.*

A necessary condition for the full column rank conditions in Assumption Ĩ-2 is $T \geq 2\lceil \log_S(J) \rceil + 1$. As $2\lceil \log_S(J) \rceil$ is decreasing in S, identification becomes easier when the support of y_{it} is larger. Alternatively, we may assume

Assumption Ĩ-4 *(Variation in the component weights and distributions) There exist disjoint*

and non-empty submodels $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *such that* $\tilde{\mathcal{P}}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\mathcal{I}_\ell})$ *has full column rank for* $\ell = 1, 2$ *and there exist* $\{\tilde{x}_t\}_{t\in\mathcal{I}_1} \neq \{x_t\}_{t\in\mathcal{I}_1}$ *and* $\{\underline{x}_t\}_{t\in\mathcal{I}_2} \neq \{x_t\}_{t\in\mathcal{I}_2}$ *such that* $\tilde{\mathcal{P}}_{\mathcal{I}_1}(\{\tilde{x}_t\}_{t\in\mathcal{I}_1})$ *and* $\tilde{\mathcal{P}}_{\mathcal{I}_2}(\{\underline{x}_t\}_{t\in\mathcal{I}_2})$ *have full column rank, and*

$$\begin{aligned}&\Pi(\{x_t\}_{t\in\mathcal{I}_1}, \{x_t\}_{t\in\mathcal{I}_2}, \{x_t\}_{t\notin\mathcal{I}_1\cup\mathcal{I}_2})\Pi(\{\tilde{x}_t\}_{t\in\mathcal{I}_1}, \{x_t\}_{t\in\mathcal{I}_2}, \{x_t\}_{t\notin\mathcal{I}_1\cup\mathcal{I}_2})^{-1}\\ \times\ &\Pi(\{\tilde{x}_t\}_{t\in\mathcal{I}_1}, \{\underline{x}_t\}_{t\in\mathcal{I}_2}, \{x_t\}_{t\notin\mathcal{I}_1\cup\mathcal{I}_2})\Pi(\{x_t\}_{t\in\mathcal{I}_1}, \{\underline{x}_t\}_{t\in\mathcal{I}_2}, \{x_t\}_{t\notin\mathcal{I}_1\cup\mathcal{I}_2})^{-1}\end{aligned}$$

is well-defined and has distinct and non-zero diagonal entries.

A necessary condition for Assumption $\tilde{\text{I}}$-4 is that $T \geq 2\lceil \log_S(J)\rceil$, which is weaker than before. On the other hand, Assumption $\tilde{\text{I}}$-4 requires more variation in the component weights.

After having replaced the respective matrices with the ones that contain all realizations of the outcome variable, Assumption I-2 with Assumption $\tilde{\text{I}}$-2 and Assumption I-4 with Assumption $\tilde{\text{I}}$-4, Theorems 3.2 and 3.3 and their corollaries directly apply to the setting of this section. Similarly, the proofs have to be only minimally adapted. We therefore do not include any formal results in this section.

7.2 Additional exclusion restriction in the component weights

In this section, we consider a modification of model (2).[27] In particular, we assume that the component weights $\pi_j(X_i)$ do not depend on all entries of X_i for all $j = 1, \ldots, J$. We focus on the case where the component weights depend only on a single covariate $x_{it,k}$ for all t and some $k \in \{1, \ldots, K\}$,[28] that is,

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i) = \sum_{j=1}^J \pi_j(\{x_{it,k}\}_{t=1}^T) \prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j) \tag{8}$$

Assumption I-2 requires $T \geq 2\lceil \log_2(J)\rceil + 1$. Interestingly, by compensating for the lack of variation in the outcome variable with variation in the covariates $x_{it,k'}$ with $k' \neq k$, the component distributions and weights in model (8) may already be identified if $T = 3$ for an arbitrary number of groups J.

[27] In Appendix A.1.6, we discuss how a different exclusion restriction allows us to relax the necessary condition $T \geq 2\lceil \log_2(J)\rceil + 1$ from Section 3.2 to $T \geq 2\lceil \log_2(J)\rceil$ without enforcing Assumption I-4.

[28] Other cases follow directly from the arguments in this and previous sections. More so, similar arguments as presented in this section also apply to the case when the component weights depend on some index of the covariates, for instance, on time averages of the covariates as in the simulation setup of Section 5.1. Then, one can construct a similar matrix like the $\mathcal{Q}_{\mathcal{I}}(\cdot)$ matrix below by varying the X_i's in such a way that the index remains the same and thus $\underline{\pi}(\cdot)$ does not change.

To keep the notation light, the remainder of this section presents a simplified identification result in the context of model (8). Appendix A.1.5 presents a detailed discussion of the model. To state the simplified result, we require some additional notation: For some time period t and covariate values $x_t^{(r)} \in \mathbb{X}_t$ for $r = 1, \ldots, R$ with $x_{t,k}^{(r)} = x_{t,k}^{(s)}$ for all r, s and $R \geq 1$, define

$$\mathcal{Q}_t(\{x_t^{(r)}\}_{r=1}^R) = \left(\mathcal{P}_t(x_t^{(1)})^\top, \ldots, \mathcal{P}_t(x_t^{(R)})^\top\right)^\top \in \mathbb{R}^{2R \times J}$$

The construction of $\mathcal{Q}_t(\{x_t^{(r)}\}_{r=1}^R)$ captures the idea that variation in the excluded covariates makes up for the lack of variation in the outcome. Similarly, for some submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$ and covariate values $\{x_t^{(r)}\}_{t \in \mathcal{I}}$ for $r = 1, \ldots, R^*$ with $x_{t,k}^{(r)} = x_{t,k}^{(s)}$ for all $t \in \mathcal{I}$ and r, s, and $R^* \geq 1$, we define

$$\mathcal{Q}_{\mathcal{I}}(\{\{x_t^{(r)}\}_{t \in \mathcal{I}}\}_{r=1}^{R^*}) = \left(\mathcal{P}_{\mathcal{I}}(\{x_t^{(1)}\}_{t \in \mathcal{I}})^\top, \ldots, \mathcal{P}_{\mathcal{I}}(\{x_t^{(R^*)}\}_{t \in \mathcal{I}})^\top\right)^\top \in \mathbb{R}^{2^{|\mathcal{I}|} R^* \times J}$$

Additionally, we let $\Pi(\{x_{t,k}\}_{t=1}^T) = \text{diag}((\pi_1(\{x_{t,k}\}_{t=1}^T), \ldots, \pi_J(\{x_{t,k}\}_{t=1}^T))^\top)$.

Since the structure of the problem is similar to that in Section 3.2, we make the following analogous assumption.

Assumption I-2′ *(Variation in the excluded covariates) Given some fixed $X_i = X \in \mathbb{X}$, there exist distinct $\tilde{t}, t^*, t' \in \{1, \ldots, T\}$ such that $\mathcal{P}_{t'}(x_{t'})$ has distinct columns and for which there exist $x_{\tilde{t}}^{(r)} \in \mathbb{X}_{\tilde{t}}$ for $r = 1, \ldots, R$ with $x_{\tilde{t},k}^{(r)} = x_{\tilde{t},k}$ for all r, and, similarly, $x_{t^*}^{(r')} \in \mathbb{X}_{t^*}$ for $r' = 1, \ldots, R'$ with $x_{t^*,k}^{(r')} = x_{t^*,k}$ for all r' such that $\mathcal{Q}_{\tilde{t}}(\{x_{\tilde{t}}^{(r)}\}_{r=1}^R)$ and $\mathcal{Q}_{t^*}(\{x_{t^*}^{(r')}\}_{r'=1}^{R'})$ have full column rank.*

A necessary condition for Assumption I-2′ is that there exist three time periods such that in two of them the covariates that do not enter $\Pi(\cdot)$ have at least $\lceil J/2 \rceil$ points of support. Assumptions I-2′ and I-2 are not mutually exclusive. When additional time periods are available, one may still combine them to larger submodels to ensure that the respective matrices have full rank and, in doing so, allow for more groups. However, to isolate the effect of the additional exclusion restriction in model (8), we do not discuss this extension formally; it follows from a combination of the arguments in this section with those in Section 3.2. We note that, instead of working with $\mathcal{Q}_t(\{x_t^{(r)}\}_{r=1}^R)$, one would then work with $\mathcal{Q}_{\mathcal{I}}(\{\{x_t^{(r)}\}_{t \in \mathcal{I}}\}_{r=1}^{R^*})$ in Assumption I-2′.

Importantly, similar to Theorem 3.3, the number of required time periods can be reduced

to $T = 2$ if there is sufficient variation in the component weights. To this end, one may adapt Assumption I-4 analogously to how Assumption I-2′ adapts Assumption I-2. We do so formally in Appendix A.1.5.

We conclude this section with the following lemma, which is a direct corollary of Theorem A.1.3 in Appendix A.1.5.

Lemma 7.1 *There exists $X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ such that Assumptions I-2′ is satisfied, $\pi_j(\tilde{X}) > 0$ for all $j = 1, \ldots, J$ and $\tilde{X} \in \mathbb{X}$, and Assumption I-1 holds. Then,*

1. *$\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified for all $\tilde{x}_t \in \mathbb{X}_t$, $t = 1, \ldots, T$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, and*
2. *$\underline{\pi}(\{\tilde{x}_{t,k}\}_{t=1}^T)$ is identified at all $\{\tilde{x}_{t,k}\}_{t=1}^T$ for which there exist a submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$ and $\{\underline{x}_t^{(r)}\}_{t \in \mathcal{I}}$ for $r = 1, \ldots, R^*$ with $\underline{x}_{t,k}^{(r)} = \tilde{x}_{t,k}$ for all $t \in \mathcal{I}$ and $r = 1, \ldots, R^*$ such that $\mathcal{Q}_{\mathcal{I}}(\{\{\underline{x}_t^{(r)}\}_{t \in \mathcal{I}}\}_{r=1}^{R^*})$ has full column rank.*

All objects are identified up to the same relabeling of the groups.

7.3 A dynamic model

We extend the baseline model of equation (2) to allow for dynamic panel models. To this end, we make the following assumption on the component distributions.

Assumption M-2 *(Dynamic panel model) For all $s_t \in \{0, 1\}$, $t \in \{1, \ldots, T\}$ and $j \in \{1, \ldots, J\}$, $\mathbb{P}_t(y_{it} = s_t \mid X_i, \{y_{it^*} = s_{t^*}\}_{t^*=0}^{t-1}, g_i = j) = \mathbb{P}_t(y_{it} = s_t \mid x_{it}, y_{it-1} = s_{t-1}, g_i = j)$ with $X_i = (x_{i1}^\top, \ldots, x_{iT}^\top)^\top$.*

Assumption M-2 imposes that the component distributions depend on the lag of the dependent variable and the contemporaneous covariates; see for instance Example A.1.2. Under Assumption M-2, the model of interest reads

$$\begin{aligned}\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i, y_{i0}) = &\sum_{j=1}^J \mathbb{P}(g_i = j \mid X_i, y_{i0}) \mathbb{P}_1(y_{i1} = s_1 \mid x_{i1}, y_{i0}, g_i = j) \\ &\times \prod_{t=2}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, y_{it-1} = s_{t-1}, g_i = j)\end{aligned}$$

where y_{i0} is assumed to be observed. Allowing for y_{it-1} in the component distribution introduces dependence between the component distributions of neighboring time periods.

Specifically, if we condition on $y_{it-1} = s_{t-1}$ in period t, then the dependent variable is fixed in period $t-1$ so that our previous analysis no longer applies. Similar to Kasahara and Shimotsu (2009), we solve this dependency problem by considering every other time period in our analysis. To this end, we focus on submodels $\mathcal{O} \subseteq \{t \in \{1,\dots,T\} : t \text{ is odd}\} =: \mathcal{T}^{(\text{odd})}$ and combine every odd time period with the subsequent even time period. We define $\mathbb{P}_t^{\otimes}(y_t, x_t, y_{t-1}, x_{t+1}, y_{t+1}, j) = \mathbb{P}_{t+1}(y_{it+1} = y_{t+1} \mid x_{it+1} = x_{t+1}, y_{it} = y_t, g_i = j)\mathbb{P}_t(y_{it} = y_t \mid x_{it} = x_t, y_{it-1} = y_{t-1}, g_i = j)$ and for a time period $t \in \mathcal{T}^{(\text{odd})}$

$$\mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1}) = \begin{pmatrix} \mathbb{P}_t^{\otimes}(0, x_t, y_{t-1}, x_{t+1}, y_{t+1}, 1) & \dots & \mathbb{P}_t^{\otimes}(0, x_t, y_{t-1}, x_{t+1}, y_{t+1}, J) \\ \mathbb{P}_t^{\otimes}(1, x_t, y_{t-1}, x_{t+1}, y_{t+1}, 1) & \dots & \mathbb{P}_t^{\otimes}(1, x_t, y_{t-1}, x_{t+1}, y_{t+1}, J) \end{pmatrix}$$

$\mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1})$ allows us to vary y_t, while y_{t+1} is fixed at some value. For every t, we define $\mathcal{P}_t(x_t, y_{t-1})$ analogously to $\mathcal{P}_t(x_t)$ previously. In the upcoming analysis, $\mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1})$ plays the role of $\mathcal{P}_t(x_t)$ in the baseline model of equation (2). As before, we combine time periods together to create larger submodels that satisfy the required regularity conditions. For $\mathcal{O} \subseteq \mathcal{T}^{(\text{odd})}$, we define

$$\mathcal{P}_{\mathcal{O}}^{\otimes}(\{x_t, y_{t-1}, x_{t+1}, y_{t+1}\}_{t\in\mathcal{O}}) = \overset{\text{col}}{\otimes}_{t\in\mathcal{O}} \mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1})$$

where we abuse the notation in that the mapping is only well-defined if the "double-counted" values of y_t for even t are identical.

For the remainder of this section, we assume for simplicity that T is odd.

Assumption I-2″ *(Variation in the component distributions) Given some fixed* $(X_i^\top, y_{i0}) = (X^\top, s)$ *for* $s \in \{0,1\}$ *and* $X \in \mathbb{X}$*, there exists a partition* $\mathcal{T}^{(odd)} = \{\mathcal{O}_1, \mathcal{O}_2, T\}$ *such that (i)* $\mathcal{O}_1$ *and* $\mathcal{O}_2$ *are non-empty and* $\max(\mathcal{O}_1) < \min(\mathcal{O}_2)$*, (ii)* $\mathcal{P}_{\mathcal{O}_j}^{\otimes}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_j})$ *has full column rank for* $j = 1, 2$ *and* $\tilde{s} \in \{0,1\}$*, (iii)* $\mathcal{P}_{\min(\mathcal{O}_2)}^{\otimes}(x_{\min(\mathcal{O}_2)}, \tilde{s}, x_{\min(\mathcal{O}_2)+1}, s) \overset{col}{\otimes} \mathcal{P}_{\mathcal{O}_2\setminus\{\min(\mathcal{O}_2)\}}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2\setminus\{\min(\mathcal{O}_2)\}})$ *has full column rank for* $\tilde{s} \neq s$*,*[29] *(iv)* $\mathcal{P}_T(x_T, s)$ *has distinct columns, and (v)* $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, y_{it-1} = s_{t-1}, g_i = j) \in (0,1)$ *for all* $j \in \{1,\dots,J\}$*,* $t = 1,\dots,T$*, and* $s_{t-1} \in \{0,1\}$*.*

We define $\Pi(X, \{s_t\}_{t=0}^k) = \text{diag}((\mathbb{P}(g_i = 1 \mid X_i = X, \{y_{it} = s_t\}_{t=0}^k), \dots, \mathbb{P}(g_i = J \mid X_i = X, \{y_{it} = s_t\}_{t=0}^k))^\top) = \text{diag}(\underline{\pi}(X, \{s_t\}_{t=0}^k))$.

[29] When $\mathcal{O}_2$ is a singleton, then $\mathcal{P}_{\mathcal{O}_2\setminus\{\min(\mathcal{O}_2)\}}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2\setminus\{\min(\mathcal{O}_2)\}})^\top := \mathbf{1}_J$, the J-dimensional vector containing only ones.

Assumption I-3″ *(Positive weights)* $\Pi(X, \{s_t\}_{t=0}^k)$ *has full rank for* $\{s_t\}_{t=0}^k \in \{0,1\}^{k+1}$ *and* $k = 0, \ldots, \min(\mathcal{O}_2) - 1$ *with* $\mathcal{O}_2$ *of Assumption I-2″, and Supp*$((X_i^\top, \{y_{it}\}_{t=0}^{\min(\mathcal{O}_2)-1})) = \mathbb{X}_1 \times \cdots \times \mathbb{X}_T \times \{0,1\}^{\min(\mathcal{O}_2)}$.

Remark 9. Assumption I-3″ corresponds to Assumption I-3, while Assumption I-2″ is an analog of Assumption I-2. The additional assumptions allow us to handle the dynamic structure of the problem and to consistently assign the group memberships across different conditioning sets of $\{y_{it}\}_{t=0}^{T-1}$. The second part of Assumption I-3″ corresponds to Assumption I-1. Similar to Assumption I-1, it can be relaxed. A more careful argument may also eliminate the need to partition $\mathcal{T}^{(\text{odd})}$ in Assumption I-2″.

Remark 10. Different from the baseline model (2), a necessary condition for Assumption I-2″ is that $T \geq 4\lceil \log_2(J) \rceil + 1$. However, under an additional assumption on the variability of the component weights, analogous to Assumption I-4, this dependence can be relaxed to $T \geq 4\lceil \log_2(J) \rceil - 1$. We show this formally in Appendix A.1.7. Interestingly, under additional exclusion restrictions on the component weights, identification is possible with $T = 2$ for arbitrary J. To show this, one would fix the outcome variable at some value and, using the arguments of Section 7.2, only exploit the variation in the excluded covariates.[30]

In the following, we say that $\mathcal{O} \subseteq \mathcal{T}^{(\text{odd})}$ is a submodel of adjacent periods if $(\max(\mathcal{O}) - \min(\mathcal{O}))/2 + 1 = |\mathcal{O}|$. For instance, $\{1,3,5\}$ is a submodel of adjacent periods, while $\{1,3,7,9\}$ is not. We define

$$\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{x_t, y_{t-1}, x_{t+1}, y_{t+1}\}_{t\in\mathcal{O}}) = \overset{\text{col}}{\otimes}_{t\in\mathcal{O}\setminus\max(\mathcal{O})} \mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1}) \overset{\text{col}}{\otimes} \mathcal{P}_{\max(\mathcal{O})}(x_{\max(\mathcal{O})}, y_{\max(\mathcal{O})-1})$$

where, for notational convenience, we keep $x_{\max(\mathcal{O})+1}$ and $y_{\max(\mathcal{O})+1}$ as arguments of $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\cdot)$.[31]

We conclude this section with a simplified version of our main identification result, which we present in Theorem A.1.5 in Appendix A.1.7. For this purpose, given some submodel $\mathcal{O} \subseteq \mathcal{T}^{(\text{odd})}$, we define $\mathcal{O}^{\oplus} = \{t+1 : t \in \mathcal{O}\}$ and $\underline{\mathcal{O}} = \mathcal{O} \cup \mathcal{O}^{\oplus}$.

Lemma 7.2 *There exist* $X_i = X \in \mathbb{X}$ *and* $s \in \{0,1\}$ *such that Assumption I-2″ holds. Additionally, assume that* $\mathbb{P}(g_i = j \mid X_i = X, \{y_{it} = s_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})}) > 0$ *for all* $j = 1, \ldots, J$,

[30] As this argument follows almost directly from our discussions, we do not formally present it. Details are available upon request.

[31] When $\mathcal{O}$ is a singleton, $\overset{\text{col}}{\otimes}_{t\in\mathcal{O}\setminus\max(\mathcal{O})} \mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1})^\top := \mathbf{1}_J$, the vector of length J containing only ones.

$X \in \mathbb{X}$, *and* $\{s_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)} \in \{0,1\}^{\max(\underline{\mathcal{O}}_1)+1}$, *and that the support condition of Assumption I-3″ holds. Then*

- $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, y_{it-1} = \tilde{s}, g_i = j)$ *is identified for all* $(\tilde{x}_t^\top, \tilde{s})^\top \in \mathbb{X}_t \times \{0,1\}$, $t = 1, \ldots, T$, $j = 1, \ldots, J$, *and* $s_t \in \{0,1\}$, *and*
- $\underline{\pi}(\tilde{X}, y_{i0} = \tilde{s})$ *is identified at all* $(\tilde{X}^\top, \tilde{s})^\top \in \mathbb{X} \times \{0,1\}$ *such that there exists a submodel* $\mathcal{O} \subseteq \mathcal{T}^{(odd)}$ *of adjacent periods with* $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}})$ *having full column rank for all* $\tilde{s} \in \{0,1\}$.

All objects are identified up to the same relabeling of the groups.

We discuss the identification of group-specific AMEs in Appendix A.1.7.

7.4 Selecting the number of groups

Similar to the discussions in Kasahara and Shimotsu (2009), Kasahara and Shimotsu (2014), and Kwon and Mbakop (2021), we have the following identification result for the number of groups J in model (2):

Lemma 7.3 *Fix* $X_i = X \in \mathbb{X}$ *and two disjoint and non-empty submodels* $\mathcal{I}_1 \subset \{1, \ldots, T\}$ *and* $\mathcal{I}_2 \subset \{1, \ldots, T\}$, *then rank*$(\mathbf{P}_{\mathcal{I}_1 \cup \mathcal{I}_2}(X)) \leq J$ *where* $\mathbf{P}_{\mathcal{I}_1 \cup \mathcal{I}_2}(X)$ *is an observed matrix we define formally in Appendix A.1.3. If* $\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t \in \mathcal{I}_\ell})$ *has full column rank for* $\ell = 1, 2$ *and* $\Pi(X)$ *has full rank, then rank*$(\mathbf{P}_{\mathcal{I}_1 \cup \mathcal{I}_2}(X)) = J$.

Under Assumptions I-2 and I-3 or under Assumption I-4, Lemma 7.3 implies $\text{rank}(\mathbf{P}_{\mathcal{I}_1 \cup \mathcal{I}_2}(X)) = J$ so that J is identified. Motivated by Lemma 7.3, an interesting avenue for future research is to construct an estimator that considers multiple covariate values jointly to estimate (a lower bound for) the number of groups.

8 Concluding remarks

This paper establishes identification and estimation results for short binary outcome panel data models in the presence of latent group-specific heterogeneity. In particular, we develop new nonparametric identification results for a broad class of conditional binary outcome mixture models. This class includes, for instance, probit and logit models with

time-varying group-specific parameters and allows for rich forms of endogeneity, provided they are captured by the latent group structure. Notably, when the probabilities to belong to a group are positive on the entire covariate support and the group-specific conditional outcome distributions exhibit sufficient variation at a *single* point in the covariate space, these group-specific distributions are nonparametrically identified on the entire support of the covariates. Unlike much of the econometric literature on group-specific heterogeneity, our analysis does not require the time dimension of the panel to grow to infinity. We show that many objects of interest, including the group-specific distributions of the outcomes given the covariates or the group-specific average marginal effects, may already be identified with as few as two time periods. We highlight the power of our identification arguments in various extensions such as a dynamic model or a model that satisfies additional exclusion restrictions.

We propose novel sieve-based semiparametric estimators for the mixture model and group-specific average marginal effects and study their asymptotic properties. Using data from the German Socio-Economic Panel on homeownership, we illustrate how to empirically interpret the latent groups that are discovered by the estimator. Specifically, we identify three latent groups of households characterized by different wealth levels and borrowing constraints.

Multiple avenues for future research are of interest. First, we use the BIC to select the number of groups in our application without studying the properties of this model selection step. In future research, it is important to extend our analysis to this or similar estimation procedures that jointly estimate J along with the model parameters. Second, we focused exclusively on one semiparametric model that is covered by our identification results. Studying other semiparametric models and estimators that enforce different semiparametric restrictions seems to be a promising avenue for future research. Importantly, our identification results cover most of these models. Third, in order to enhance the numerical stability and computational efficiency of the proposed semiparametric estimator, it is valuable to explore estimation methods beyond the EM algorithm. We intend to address at least some of these questions in our future work.

References

Abramitzky, R., L. Boustan, and D.S. Connor (2024). "Leaving the enclave: Historical evidence on immigrant mobility from the industrial removal office". In: *The Journal of Economic History* 84.2, pp. 352–394.

Ackerberg, D., X. Chen, and J. Hahn (2012). "A practical asymptotic variance estimator for two-step semiparametric estimators". In: *Review of Economics and Statistics* 94.2, pp. 481–498.

Acolin, A., J. Bricker, P. Calem, and S. Wachter (2016). "Borrowing constraints and homeownership". In: *American Economic Review: Papers & Proceedings* 106.5, pp. 625–629.

Adelino, M., A. Schoar, and F. Severino (2015). "House prices, collateral, and self-employment". In: *Journal of Financial Economics* 117.2, pp. 288–306.

Aguiar, V.H. and N. Kashaev (2025). "Identification and estimation of discrete choice models with unobserved choice sets". In: *Journal of Business & Economic Statistics* 43.1, pp. 204–215.

Aguirregabiria, V. and J.M. Carro (2024). *Identification of Average Marginal Effects in Fixed Effects Dynamic Discrete Choice Models*. arXiv: 2107.06141 [econ.EM]. URL: https://arxiv.org/abs/2107.06141.

Ai, C. and X. Chen (2003). "Efficient estimation of models with conditional moment restrictions containing unknown functions". In: *Econometrica* 71.6, pp. 1795–1843.

Aina, C., E. Baici, G. Casalone, and F. Pastore (2022). "The determinants of university dropout: A review of the socio-economic literature". In: *Socio-Economic Planning Sciences* 79, p. 101102.

Allman, E.S., C. Matias, and J.A. Rhodes (2009). "Identifiability of parameters in latent structure models with many observed variables". In: *The Annals of Statistics* 37.6A, pp. 3099–3132.

Bajari, P., J. Hahn, H. Hong, and G. Ridder (2011). "A note on semiparametric estimation of finite mixtures of discrete choice models with application to game theoretic models". In: *International Economic Review* 52.3, pp. 807–824.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016). "Estimating multivariate latent-structure models". In: *The Annals of Statistics* 44.2, pp. 540–563.

Bonhomme, S. and E. Manresa (2015). "Grouped patterns of heterogeneity in panel data". In: *Econometrica* 83.3, pp. 1147–1184.

Browning, M. and J.M. Carro (2014). "Dynamic binary outcome models with maximal heterogeneity". In: *Journal of Econometrics* 178.2, pp. 805–823.

Carroll, R.J., X. Chen, and Y. Hu (2010). "Identification and estimation of nonlinear models using two samples with nonclassical measurement errors". In: *Journal of Nonparametric Statistics* 22.4, pp. 379–399.

Celeux, G., S. Frühwirth-Schnatter, and C.P. Robert (2019). "Model selection for mixture models – Perspectives and strategies". In: *Handbook of Mixture Analysis*. Chapman and Hall/CRC, pp. 117–154.

Chamberlain, G. (2010). "Binary response models for panel data: Identification and information". In: *Econometrica* 78.1, pp. 159–168.

Chen, X. (2007). "Large sample sieve estimation of semi-nonparametric models". In: *Handbook of Econometrics* 6, pp. 5549–5632.

Chen, X. and Z. Liao (2015). "Sieve semiparametric two-step GMM under weak dependence". In: *Journal of Econometrics* 189.1, pp. 163–186.

Chen, X. and X. Shen (1998). "Sieve extremum estimates for weakly dependent data". In: *Econometrica* 66.2, pp. 289–314.

Cummins, N.J. and C.Ó. Gráda (2025). "The Irish in England". In: *The Journal of Economic History* 85.1, pp. 180–214.

Davezies, L., X. D'Haultfœuille, and L. Laage (2024). *Identification and Estimation of Average Causal Effects in Fixed Effects Logit Models*. arXiv: 2105.00879 [econ.EM]. URL: https://arxiv.org/abs/2105.00879.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2004). "Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition". In: *SIAM Journal on Matrix Analysis and Applications* 26.2, pp. 295–327.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Di, Z.X. (2007). "Do homeowners have higher future household income?" In: *Housing Studies* 22.4, pp. 459–472.

Di, Z.X., E. Belsky, and X. Liu (2007). "Do homeowners achieve more household wealth in the long run?" In: *Journal of Housing Economics* 16.3-4, pp. 274–290.

Diaz-Serrano, L. (2005). "Labor income uncertainty, skewness and homeownership: A panel data study for Germany and Spain". In: *Journal of Urban Economics* 58.1, pp. 156–176.

Dietz, R.D. and D.R. Haurin (2003). "The social and private micro-level consequences of homeownership". In: *Journal of Urban Economics* 54.3, pp. 401–450.

DiPasquale, D. and E.L. Glaeser (1999). "Incentives and social capital: Are homeowners better citizens?" In: *Journal of Urban Economics* 45.2, pp. 354–384.

Dobronyi, C., J. Gu, K.i. Kim, and T.M. Russell (2024). *Identification of Dynamic Panel Logit Models with Fixed Effects*. arXiv: 2104.04590 [econ.EM]. URL: https://arxiv.org/abs/2104.04590.

Eurostat (2025). *Distribution of population by tenure status, type of household and income group*. DOI: https://doi.org/10.2908/ILC_LVHO02.

Fiscella, K. and A.M. Fremont (2006). "Use of geocoding and surname analysis to estimate race and ethnicity". In: *Health Services Research* 41.4p1, pp. 1482–1500.

Flavin, M. and T. Yamashita (2002). "Owner-occupied housing and the composition of the household portfolio". In: *American Economic Review* 92.1, pp. 345–362.

Freyberger, J. (2018). "Non-parametric panel data models with interactive fixed effects". In: *The Review of Economic Studies* 85.3, pp. 1824–1851.

Freyberger, J. and M.A. Masten (2019). "A practical guide to compact infinite dimensional parameter spaces". In: *Econometric Reviews* 38.9, pp. 979–1006.

Frühwirth-Schnatter, S., G. Celeux, and C.P. Robert (2019). *Handbook of Mixture Analysis*. Chapman and Hall/CRC.

German Federal Statistical Office (2025). *Consumer price index: Germany, years. Code: 61111-0001*. URL: https://www-genesis.destatis.de/datenbank/online/statistic/61111/table/61111-0001.

Goebel, J., M.M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp (2019). "The German socio-economic panel (SOEP)". In: *Jahrbücher für Nationalökonomie und Statistik* 239.2, pp. 345–360.

Gormley, I.C. and S. Frühwirth-Schnatter (2019). "Mixture of experts models". In: *Handbook of Mixture Analysis*. Ed. by S. Frühwirth-Schnatter, G. Celeux, and C.P. Robert. Chapman and Hall/CRC, pp. 271–307.

Green, R.K. and M.J. White (1997). "Measuring the benefits of homeowning: Effects on children". In: *Journal of Urban Economics* 41.3, pp. 441–461.

Grün, B. and F. Leisch (2008a). "Finite Mixtures of Generalized Linear Regression Models". In: *Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg*. Physica-Verlag HD, pp. 205–230.

— (2008b). "Identifiability of finite mixtures of multinomial logit models with varying and fixed effects". In: *Journal of Classification* 25.2, pp. 225–247.

Gu, J. and R. Koenker (2022). "Nonparametric maximum likelihood methods for binary response models with random coefficients". In: *Journal of the American Statistical Association* 117.538, pp. 732–751.

Hagenaars, J.A. and A.L. McCutcheon (2002). *Applied latent class analysis*. Cambridge University Press.

Hahn, J. and H.R. Moon (2010). "Panel data models with finite number of multiple equilibria". In: *Econometric Theory* 26.3, pp. 863–881.

Hall, P., A. Neeman, R. Pakyari, and R. Elmore (2005). "Nonparametric inference in multivariate mixtures". In: *Biometrika* 92.3, pp. 667–678.

Hall, P. and X.-H. Zhou (2003). "Nonparametric estimation of component distributions in a multivariate mixture". In: *The Annals of Statistics* 31.1, pp. 201–224.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Haurin, D.R., R.D. Dietz, and B.A. Weinberg (2002). "The impact of neighborhood homeownership rates: A review of the theoretical and empirical literature". In: *Journal of Housing Research*, pp. 119–151.

Henry, M., Y. Kitamura, and B. Salanié (2014). "Partial identification of finite mixtures in econometric models". In: *Quantitative Economics* 5.1, pp. 123–144.

Higgins, A. and K. Jochmans (2023). "Identification of mixtures of dynamic discrete choices". In: *Journal of Econometrics* 237.1, p. 105462.

Honoré, B.E. and M. Weidner (2025). "Moment conditions for dynamic panel logit models with fixed effects". In: *Review of Economic Studies* 92.5, pp. 3112–3137.

Hu, Y. (2008). "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution". In: *Journal of Econometrics* 144.1, pp. 27–61.

— (2025). *The Econometrics of Unobservables*. Tech. rep. Manuscript.

Hu, Y. and S.M. Schennach (2008). "Instrumental variable treatment of nonclassical measurement error models". In: *Econometrica* 76.1, pp. 195–216.

Hu, Y. and M. Shum (2012). "Nonparametric identification of dynamic models with unobserved state variables". In: *Journal of Econometrics* 171.1, pp. 32–44.

Huang, M., R. Li, and S. Wang (2013). "Nonparametric mixture of regression models". In: *Journal of the American Statistical Association* 108.503, pp. 929–941.

Huang, M. and W. Yao (2012). "Mixture of regression models with varying mixing proportions: a semiparametric approach". In: *Journal of the American Statistical Association* 107.498, pp. 711–724.

Ichimura, H. (1993). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models". In: *Journal of Econometrics* 58.1-2, pp. 71–120.

Jaeger, David A, Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, and Holger Bonin (2010). "Direct evidence on risk attitudes and migration". In: *The Review of Economics and Statistics* 92.3, pp. 684–689.

Kaas, L., G. Kocharkov, and E. Preugschat (2019). "Wealth inequality and homeownership in Europe". In: *Annals of Economics and Statistics* 136, pp. 27–54.

Kaas, L., G. Kocharkov, E. Preugschat, and N. Siassi (2021). "Low homeownership in Germany—A quantitative exploration". In: *Journal of the European Economic Association* 19.1, pp. 128–164.

Kasahara, H. and K. Shimotsu (2009). "Nonparametric identification of finite mixture models of dynamic discrete choices". In: *Econometrica* 77.1, pp. 135–175.

— (2014). "Non-parametric identification and estimation of the number of components in multivariate mixtures". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76.1, pp. 97–111.

Khan, S., M. Ponomareva, and E. Tamer (2023). "Identification of dynamic binary response models". In: *Journal of Econometrics* 237.1, p. 105515.

Kitamura, Y. and L. Laage (2018). *Nonparametric Analysis of Finite Mixtures*. arXiv: `1811.02727 [econ.EM]`. URL: `https://arxiv.org/abs/1811.02727`.

Kwon, C. and E. Mbakop (2021). "Estimation of the number of components of nonparametric multivariate finite mixture models". In: *The Annals of Statistics* 49.4, pp. 2178–2205.

Leisch, F., S. Dolnicar, and B. Grün (2018). *Market segmentation analysis: Understanding it, doing it, and making it useful*. Springer.

Leurgans, S.E., R.T. Ross, and R.B. Abel (1993). "A decomposition for three-way arrays". In: *SIAM Journal on Matrix Analysis and Applications* 14.4, pp. 1064–1083.

Malmendier, U. and A.S. Wellsjo (2024). "Rent or buy? inflation experiences and homeownership within and across countries". In: *The Journal of Finance* 79.3, pp. 1977–2023.

Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. 1st ed. Springer.

McCartan, C., R. Fisher, J. Goldin, D.E. Ho, and K. Imai (2024). *Estimating racial disparities when race is not observed*. Tech. rep. National Bureau of Economic Research.

Mugnier, M. (2023). "Unobserved Clusters of Time-Varying Heterogeneity in Nonlinear Panel Data Models". Working paper.

Newey, W.K. and J.L. Powell (2003). "Instrumental variable estimation of nonparametric models". In: *Econometrica* 71.5, pp. 1565–1578.

Saggio, R. (2016). "Discrete Unobserved Heterogeneity in Discrete Choice Panel Data Models". Working paper.

Shen, X. (1997). "On methods of sieves and penalization". In: *The Annals of Statistics* 25.6, pp. 2555–2591.

Sinai, T. and N.S. Souleles (2005). "Owner-occupied housing as a hedge against rent risk". In: *The Quarterly Journal of Economics* 120.2, pp. 763–789.

Socio-Economic Panel (SOEP) (2024). *Data for years 1984 – 2022 (SOEP-Core v39, EU Edition)*. DOI: 10.5684/soep.core.v39eu.

Su, L., Z. Shi, and P.C.B. Phillips (2016). "Identifying latent structures in panel data". In: *Econometrica* 84.6, pp. 2215–2264.

Taylor, L.O. (2017). "Hedonics". In: *A Primer on Nonmarket Valuation*. Ed. by P.A. Champ, K.J. Boyle, and T.C. Brown. Springer, pp. 235–292.

Thomas, M.J. and C.H. Mulder (2016). "Partnership patterns and homeownership: a cross-country comparison of Germany, the Netherlands and the United Kingdom". In: *Housing Studies* 31.8, pp. 935–963.

Turner, T.M. and H. Luea (2009). "Homeownership, wealth accumulation and income status". In: *Journal of Housing Economics* 18.2, pp. 104–114.

Wang, S., W. Yao, and M. Huang (2014). "A note on the identifiability of nonparametric and semiparametric mixtures of GLMs". In: *Statistics & Probability Letters* 93, pp. 41–45.

Wedel, M. and W.A. Kamakura (2000). *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media.

Yao, W. and S. Xiang (2024). *Mixture Models: Parametric, Semiparametric, and New Directions*. CRC Press.

Appendix: Group-Specific Heterogeneity in Short Binary Outcome Panels

A.1 Additional results and discussions

A.1.1 Extending Example 2.1

While our identification analysis in Section 3.2 focuses on the mixture model in equation (2), the extensions of Section 7.3 and Appendix A.1.8 allow the component distributions to depend on the lagged outcome variable or lagged covariates. We motivate these extensions in the context of a linear latent utility model.

Example A.1.1 *(Lagged covariates) We extend Example 2.1 to include x_{it-1} as an additional covariate. We assume that x_{i0} is observed. The model of interest becomes*

$$y_{it} = \sum_{j=1}^{J} \mathbb{1}(g_i = j)\mathbb{1}(x_{it}^\top \beta_{j,t} + x_{it-1}\gamma_{j,t} + \alpha_{j,t} - \varepsilon_{j,it} \geq 0)$$

Letting $X_i = (x_{i0}^\top, \ldots, x_{iT}^\top)^\top$ and adopting the same assumptions as in Example 2.1, we have

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i) = \sum_{j=1}^{J} \pi_j(X_i) \prod_{t=1}^{T} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, x_{it-1}, g_i = j)$$

Example A.1.2 *(Dynamic binary choice model) We modify Example 2.1 to include y_{it-1} as an additional covariate. We assume that y_{i0} is observed. The model of interest is a dynamic binary choice model*

$$y_{it} = \sum_{j=1}^{J} \mathbb{1}(g_i = j)\mathbb{1}(x_{it}^\top \beta_{j,t} + y_{it-1}\gamma_{j,t} + \alpha_{j,t} - \varepsilon_{j,it} \geq 0)$$

Due to the presence of the lagged outcome variable as a covariate, the previous factorization argument does not hold. However, one can iteratively factorize the joint distribution of $\{y_{it}\}_{t=1}^T$. To avoid modeling the initial period, we condition on y_{i0} and assume (i) $\varepsilon_{j,it} \perp\!\!\!\perp (X_i^\top, \{y_{ik}\}_{k=0}^{t-1}) \mid g_i$ for $t = 1, \ldots, T$ and (ii) $\varepsilon_{j,it} \mid g_i = j \sim F_{jt}$. Then

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i, y_{i0} = s_0) = \sum_{j=1}^{J} \pi_j(X_i, s_0) \prod_{t=1}^{T} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, y_{it-1} = s_{t-1}, g_i = j)$$

where $\mathbb{P}_t(y_{it} = s_t \mid x_{it}, y_{it-1}, g_i = j) = F_{jt}(x_{it}^\top \beta_{j,t} + y_{it-1}\gamma_{j,t} + \alpha_{j,t})^{s_t}(1 - F_{jt}(x_{it}^\top \beta_{j,t} + y_{it-1}\gamma_{j,t} + \alpha_{j,t}))^{1-s_t}$ are the component distributions.

A.1.2 Heterogeneity bias

Ignoring the latent heterogeneity in model (2) may introduce a heterogeneity bias. To see this, we consider a researcher who is interested in estimating the average marginal effect (AME) of $x_{it,1}$ on y_{it}. Instead of considering the model (2), the researcher may simply ignore the latent group structure and perform a standard analysis to estimate the AME in the hope that doing so produces a weighted combination of the group-specific AMEs. Unfortunately, this is not the case. Assuming that the required derivatives exist, we have

$$E\left[\frac{\partial}{\partial x_{it,1}}\mathbb{P}(y_{it}=1 \mid x_{it})\right] = \sum_{j=1}^{J} E\left[\pi_{jt}(x_{it})\frac{\partial}{\partial x_{it,1}}\mathbb{P}(y_{it}=1 \mid x_{it}, g_i=j)\right] + \sum_{j=1}^{J} E\left[\left(\frac{\partial}{\partial x_{it,1}}\pi_{jt}(x_{it})\right)\mathbb{P}(y_{it}=1 \mid x_{it}, g_i=j)\right]$$

where we use that model (2) implies that $\mathbb{P}(y_{it}=1 \mid x_{it}) = \sum_{j=1}^{J}\pi_{jt}(x_{it})\,\mathbb{P}(y_{it}=1 \mid x_{it}, g_i = j)$ and $\pi_{jt}(x_{it}) = E[\pi_j(X_i) \mid x_{it}]$. The first term is exactly the weighted combination of the AMEs the researcher hoped to estimate, while the second term is the heterogeneity bias, which may be non-negligible. The heterogeneity bias is equal to 0 if the group structure is independent from $x_{it,1}$. However, even in the case of independence, a group-level analysis may be more informative than the standard analysis. This is why we provide conditions to identify and estimate group-specific AMEs.

A.1.3 Testability

We argue that Assumptions I-2 and I-3 are jointly testable. To this end, we introduce some additional notation. We let $\mathcal{I}_1$ and $\mathcal{I}_2$ be two disjoint and non-empty submodels and define the $2^{|\mathcal{I}_1|} \times 2^{|\mathcal{I}_2|}$ matrix

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X) = \left\{\left\{\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{I}_1}, \{y_{it}=s_t^*\}_{t\in\mathcal{I}_2} \mid X)\right\}_{\{s_t\}_{t\in\mathcal{I}_1}\in\{0,1\}^{|\mathcal{I}_1|}}\right\}_{\{s_t^*\}_{t\in\mathcal{I}_2}\in\{0,1\}^{|\mathcal{I}_2|}}$$

which is known. A row of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)$ fixes $\{s_t\}_{t\in\mathcal{I}_1} \in \{0,1\}^{|\mathcal{I}_1|}$ and iterates through all $2^{|\mathcal{I}_2|}$ combinations of $\{s_t^*\}_{t\in\mathcal{I}_2} \in \{0,1\}^{|\mathcal{I}_2|}$, whereas a column fixes $\{s_t^*\}_{t\in\mathcal{I}_2}$ and iterates through all $2^{|\mathcal{I}_1|}$ combinations of $\{s_t\}_{t\in\mathcal{I}_1}$. Importantly, these combinations of outcomes align with the combinations in the rows of $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$ and $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$. For $k=1,2$ and $t' \notin \mathcal{I}_1\cup\mathcal{I}_2$, we similarly define

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k} = \left\{\left\{\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{I}_1}, \{y_{it}=s_t^*\}_{t\in\mathcal{I}_2}, y_{it'}=k-1 \mid X)\right\}_{\{s_t\}_{t\in\mathcal{I}_1}\in\{0,1\}^{|\mathcal{I}_1|}}\right\}_{\{s_t^*\}_{t\in\mathcal{I}_2}\in\{0,1\}^{|\mathcal{I}_2|}}$$

Given these definitions, simple algebra yields

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X) \;= \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top$$

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X) \;= \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{D}_{t',k}(x_{t'})\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top$$

for $k = 1, 2$ and where $\mathcal{D}_{t',k}(x_{t'}) = \text{diag}([\mathcal{P}_{t'}(x_{t'})]_{k,\cdot})$ with $\text{diag}(a)$ for $a \in \mathbb{R}^J$ being a $J \times J$ diagonal matrix with diagonal a and $[A]_{k,\cdot}$ denoting the k-th row of a matrix A.

Now, letting $A^\dagger$ denote the Moore-Penrose inverse of some matrix A, we have the following lemma.

Lemma A.1.1 *Fix $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$, $t' \in \{1, \ldots, T\}$ and disjoint and non-empty submodels $\mathcal{I}_1$ and $\mathcal{I}_2$ with $t' \notin \mathcal{I}_1 \cup \mathcal{I}_2$, then rank$(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) = J$ and $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ has at least $J-1$ non-zero eigenvalues of algebraic multiplicity 1 if and only if Assumptions I-2 and I-3 hold.*

When X is jointly continuously distributed and all component distributions are continuous in X, Lemma A.1.1 readily implies that when Assumptions I-2 and I-3 are satisfied at some $X \in \mathbb{X}$, then there exists $\delta > 0$ such that these assumptions are satisfied for all $X \in B_\delta(X)$, a δ-ball around X. This follows from the specific structure of the Moore-Penrose inverse and the facts that the rank of a matrix is lower semi-continuous and that eigenvalues are continuous.

Similarly, I-4 is testable.

Lemma A.1.2 *Fix $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$. Assumption I-4 is satisfied at X if and only if rank$(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) =$ rank$(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})) =$ rank$(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})) =$ rank$(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})) = J$ and $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger$ has J distinct non-zero eigenvalues where $\tilde{X} = (\{\tilde{x}_t\}_{t\in\mathcal{I}_1}^\top, \{x_t\}_{t\notin\mathcal{I}_1}^\top)^\top$, $\underline{X} = (\{\underline{x}_t\}_{t\in\mathcal{I}_2}^\top, \{x_t\}_{t\notin\mathcal{I}_2}^\top)^\top$, and $\underline{\tilde{X}} = (\{\tilde{x}_t\}_{t\in\mathcal{I}_1}^\top, \{\underline{x}_t\}_{t\in\mathcal{I}_2}^\top, \{x_t\}_{t\notin\mathcal{I}_1\cup\mathcal{I}_2}^\top)^\top$.*

We prove Lemmas A.1.1 and A.1.2 in Appendix A.3.

A.1.4 Partial identification of group-specific AMEs

We consider a setting where a researcher believes that Assumption I-7 holds only on a known subset of the support $\mathbb{B} \subseteq \mathbb{X}$. Then, $\text{AME}_{j,t,k}$ is partially identified. To see this, we note

$$\begin{aligned}\text{AME}_{j,t,k} = \; & \underbrace{\mathbb{E}\left[\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it}=1 \mid x_{it}, g_i=j) \mid g_i=j, X_i\in\mathbb{B}\right]}_{=:\text{AME}_{j,t,k}^{(\mathbb{B})}}\mathbb{P}(X_i\in\mathbb{B}\mid g_i=j) \\ & + \underbrace{\mathbb{E}\left[\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it}=1 \mid x_{it}, g_i=j) \mid g_i=j, X_i\notin\mathbb{B}\right]}_{=:\text{AME}_{j,t,k}^{(\mathbb{B}^c)}}(1-\mathbb{P}(X_i\in\mathbb{B}\mid g_i=j))\end{aligned}$$

where $\text{AME}_{j,t,k}^{(\mathbb{B})}$ is identified because

$$\text{AME}_{j,t,k}^{(\mathbb{B})} = \mathbb{E}\left[\pi_j(X_i)\mathbb{1}(X_i \in \mathbb{B})\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it}=1 \mid x_{it}, g_i = j)\right] / \mathbb{P}(g_i = j, X_i \in \mathbb{B})$$

where $\pi_j(X_i)\mathbb{1}(X_i \in \mathbb{B})\frac{\partial}{\partial x_{it,k}}\mathbb{P}_t(y_{it}=1 \mid x_{it}, g_i = j)$ is identified on $\mathbb{X}$ and $\mathbb{P}(g_i = j, X_i \in \mathbb{B})$ is identified with $E[\pi_j(X_i)\mathbb{1}(X_i \in \mathbb{B})]$. At the same time, since $\mathbb{P}_t(y_{it}=1 \mid x_{it}, g_i = j)$ is identified for almost all $x_{it} \in \mathbb{X}_t$ by Assumption I-6, we can bound $\text{AME}_{j,t,k}^{(\mathbb{B}^c)}$ with identified quantities as follows

$$L := \inf_{X \notin \mathbb{B}} \frac{\partial}{\partial x_{t,k}}\mathbb{P}_t(y_{it}=1 \mid x_t, g_i = j) \leq \text{AME}_{j,t,k}^{(B^c)} \leq \sup_{X \notin \mathbb{B}} \frac{\partial}{\partial x_{t,k}}\mathbb{P}_t(y_{it}=1 \mid x_t, g_i = j) =: U$$

$\mathbb{P}(X_i \in \mathbb{B} \mid g_i = j) = \mathbb{P}(X_i \in \mathbb{B}, g_i = j)/\pi_j$ is not identified because π_j is not identified. However, $\pi_j \in [\mathbb{P}(g_i = j, X_i \in \mathbb{B}), \mathbb{P}(g_i = j, X_i \in \mathbb{B}) + \mathbb{P}(X_i \notin \mathbb{B})]$. Now, these bounds can be combined to construct an identified set for $\text{AME}_{j,t,k}$. Such an identified set may be sharpened under additional assumptions. For instance, in some settings, it may be reasonable to assume that $\mathbb{P}(g_i = j \mid X_i \in \mathbb{B}) = \pi_j$, then the identified set is of the following form

$$\begin{aligned}\text{AME}_{j,t,k} \in \Big[&\text{AME}_{j,t,k}^{(\mathbb{B})}\,\mathbb{P}(X_i \in \mathbb{B} \mid g_i = j) + L(1 - \mathbb{P}(X_i \in \mathbb{B} \mid g_i = j)),\\ &\text{AME}_{j,t,k}^{(\mathbb{B})}\,\mathbb{P}(X_i \in \mathbb{B} \mid g_i = j) + U(1 - \mathbb{P}(X_i \in \mathbb{B} \mid g_i = j))\Big]\end{aligned}$$

We leave it for future work to study this partial identification approach in detail.

A.1.5 Additional exclusion restrictions in the component weights

In this section, we discuss the model of Section 7.2 in more detail. To this end, we recall the model and some notation first:

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i) = \sum_{j=1}^J \pi_j(\{x_{it,k}\}_{t=1}^T)\prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j) \tag{A.1}$$

For some time period t and covariate values $x_t^{(r)} \in \mathbb{X}_t$ for $r = 1, \ldots, R$ with $x_{t,k}^{(r)} = x_{t,k}^{(s)}$ for all r, s and $R \geq 1$, define

$$\mathcal{Q}_t(\{x_t^{(r)}\}_{r=1}^R) = \left(\mathcal{P}_t(x_t^{(1)})^\top, \ldots, \mathcal{P}_t(x_t^{(R)})^\top\right)^\top \in \mathbb{R}^{2R \times J}$$

Similarly, for some submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$ and covariate values $\{x_t^{(r)}\}_{t \in \mathcal{I}}$ for $r = 1, \ldots, R^*$ with $x_{t,k}^{(r)} = x_{t,k}^{(s)}$ for all $t \in \mathcal{I}$ and r, s, and $R^* \geq 1$, we define

$$\mathcal{Q}_{\mathcal{I}}(\{\{x_t^{(r)}\}_{t \in \mathcal{I}}\}_{r=1}^{R^*}) = \left(\mathcal{P}_{\mathcal{I}}(\{x_t^{(1)}\}_{t \in \mathcal{I}})^\top, \ldots, \mathcal{P}_{\mathcal{I}}(\{x_t^{(R^*)}\}_{t \in \mathcal{I}})^\top\right)^\top \in \mathbb{R}^{2^{|\mathcal{I}|}R^* \times J}$$

We note that $\overset{\text{col}}{\otimes}_{t \in \mathcal{I}} \mathcal{Q}_t(\{x_t^{(r_t)}\}_{r_t=1}^{R_t})$ is a special case of $\mathcal{Q}_{\mathcal{I}}(\{\{x_t^{(r)}\}_{t \in \mathcal{I}}\}_{r=1}^{R^*})$ by choosing $\{\{x_t^{(r)}\}_{t \in \mathcal{I}}\}_{r=1}^{R^*}$ accordingly. Additionally, we let $\Pi(\{x_{t,k}\}_{t=1}^T) = \text{diag}((\pi_1(\{x_{t,k}\}_{t=1}^T), \ldots, (\pi_J(\{x_{t,k}\}_{t=1}^T))^\top$ and use $\mathcal{D}_{t,k}(x_t)$ for $k = 1, 2$ as in Section

3.1. Next, we fix $X_i = X$ and time periods $\tilde{t} \neq t^* \neq t'$. We consider the two sequences $x_{\tilde{t}}^{(r)}$ for $r = 1, \ldots, R$, and $x_{t^*}^{(r')}$ for $r' = 1, \ldots, R'$ where $x_{\tilde{t},k}^{(r)} = x_{\tilde{t},k}^{(s)}$ for all r, s and, similarly, $x_{t^*,k}^{(r')} = x_{t^*,k}^{(s')}$ for all r', s'. Additionally, we define the vectors $X^{(r,r')} = (\{x_t\}_{t \notin \{\tilde{t}, t^*\}}, x_{\tilde{t}}^{(r)\top}, x_{t^*}^{(r')\top})^\top$ for $r = 1, \ldots, R$ and $r' = 1, \ldots, R'$. Last, we define the two observed matrices

$$
\begin{aligned}
&\mathbf{P}_{\{\tilde{t},t^*\}}(\{x_t\}_{t\notin\{\tilde{t},t^*\}}, \{x_{\tilde{t}}^{(r)}\}_{r=1}^{R}, \{x_{t^*}^{(r')}\}_{r'=1}^{R'}) \\
&= \begin{pmatrix}
\left\{\begin{pmatrix} \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=0 \mid X_i = X^{(1,r')}) & \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=1 \mid X_i = X^{(1,r')}) \\ \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=0 \mid X_i = X^{(1,r')}) & \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=1 \mid X_i = X^{(1,r')}) \end{pmatrix}\right\}_{r'=1}^{R'} \\
\vdots \\
\left\{\begin{pmatrix} \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=0 \mid X_i = X^{(R,r')}) & \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=1 \mid X_i = X^{(R,r')}) \\ \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=0 \mid X_i = X^{(R,r')}) & \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=1 \mid X_i = X^{(R,r')}) \end{pmatrix}\right\}_{r'=1}^{R'}
\end{pmatrix} \\
&\mathbf{P}_{\{\tilde{t},t^*,t'\},k}(\{x_t\}_{t\notin\{\tilde{t},t^*\}}, \{x_{\tilde{t}}^{(r)}\}_{r=1}^{R}, \{x_{t^*}^{(r')}\}_{r'=1}^{R'}) \\
&= \begin{pmatrix}
\left\{\begin{pmatrix} \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=0, y_{it'}=k-1 \mid X_i = X^{(1,r')}) & \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=1, y_{it'}=k-1 \mid X_i = X^{(1,r')}) \\ \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=0, y_{it'}=k-1 \mid X_i = X^{(1,r')}) & \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=1, y_{it'}=k-1 \mid X_i = X^{(1,r')}) \end{pmatrix}\right\}_{r'=1}^{R'} \\
\vdots \\
\left\{\begin{pmatrix} \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=0, y_{it'}=k-1 \mid X_i = X^{(R,r')}) & \mathbb{P}(y_{i\tilde{t}}=0, y_{it^*}=1, y_{it'}=k-1 \mid X_i = X^{(R,r')}) \\ \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=0, y_{it'}=k-1 \mid X_i = X^{(R,r')}) & \mathbb{P}(y_{i\tilde{t}}=1, y_{it^*}=1, y_{it'}=k-1 \mid X_i = X^{(R,r')}) \end{pmatrix}\right\}_{r'=1}^{R'}
\end{pmatrix}
\end{aligned}
$$

where we use the convention that for a sequence of matrices $\{A_r\}_{r=1}^R$ is equal to $(A_1, \ldots, A_R)$ so that both matrices are of dimension $2R \times 2R'$. Similar to Appendix A.1.3, the following relationships hold

$$
\begin{aligned}
\mathbf{P}_{\{\tilde{t},t^*\}}(\{x_t\}_{t\notin\{\tilde{t},t^*\}}, \{x_{\tilde{t}}^{(r)}\}_{r=1}^{R}, \{x_{t^*}^{(r')}\}_{r'=1}^{R'}) &= \mathcal{Q}_{\tilde{t}}(\{x_{\tilde{t}}^{(r)}\}_{r=1}^{R}) \Pi(\{x_{t,k}\}_{t=1}^{T}) \mathcal{Q}_{t^*}(\{x_{t^*}^{(r')}\}_{r'=1}^{R'})^\top \\
\mathbf{P}_{\{\tilde{t},t^*,t'\},k}(\{x_t\}_{t\notin\{\tilde{t},t^*\}}, \{x_{\tilde{t}}^{(r)}\}_{r=1}^{R}, \{x_{t^*}^{(r')}\}_{r'=1}^{R'}) &= \mathcal{Q}_{\tilde{t}}(\{x_{\tilde{t}}^{(r)}\}_{r=1}^{R}) \Pi(\{x_{t,k}\}_{t=1}^{T}) \mathcal{D}_{t',k}(x_{t'}) \mathcal{Q}_{t^*}(\{x_{t^*}^{(r')}\}_{r'=1}^{R'})^\top
\end{aligned}
$$

We state an additional assumption that is an analogue of Assumption I-3.

Assumption I-3′ *(Positive weights)* $\Pi(\{x_{t,k}\}_{t=1}^T)$ *has full column rank.*

Under Assumptions I-2′ and I-3′, we establish the following theorem which is to be interpreted as an analogue of Theorem 3.2 in Section 3.2.

Theorem A.1.3 *There exists* $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ *such that Assumptions I-2′ and I-3′ are satisfied with periods* $\tilde{t}$*,* t^* *and* t'*, and Assumption I-1 holds. Then,*

1. *for all* $t \neq \tilde{t}$*,* $j = 1, \ldots, J$*, and* $s_t \in \{0, 1\}$*,* $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ *is identified at* $\tilde{x}_t \in \mathbb{X}_t$ *for which there exists* $\{x_{\underline{t},k}^*\}_{\underline{t} \notin \{t,\tilde{t}\}}$ *such that* $\pi_j(x_{\tilde{t},k}, \{x_{\underline{t},k}^*\}_{\underline{t}\notin\{t,\tilde{t}\}}, \tilde{x}_{t,k}) > 0$
2. *for all* $t \neq t^*$*,* $j = 1, \ldots, J$*, and* $s_t \in \{0, 1\}$*,* $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ *is identified at* $\tilde{x}_t \in \mathbb{X}_t$ *for which there exists* $\{x_{\underline{t},k}^*\}_{\underline{t}\notin\{t,t^*\}}$ *such that* $\pi_j(x_{t^*,k}, \{x_{\underline{t},k}^*\}_{\underline{t}\notin\{t,t^*\}}, \tilde{x}_{t,k}) > 0$*, and*

3. *for any submodel* $\mathcal{I} \subseteq \{1, \ldots, T\}$ *and* $\{\tilde{x}_t^{(r)}\}_{t\in\mathcal{I}}$ *with* $r = 1, \ldots, R$ *and* $\tilde{x}_{t,k}^{(r)} = \tilde{x}_{t,k}^{(s)}$ *for all* $t \in \mathcal{I}$ *and* r, s *such that* $\mathcal{Q}_{\mathcal{I}}(\{\{\tilde{x}_t^{(r)}\}_{t\in\mathcal{I}}\}_{r=1}^{R})$ *has full column rank and is identified given 1. or 2.,* $\underline{\pi}(\{\tilde{x}_{t,k}\}_{t\in\mathcal{I}}, \{x_{t,k}\}_{t\notin\mathcal{I}})$ *is identified at* $\{\tilde{x}_{t,k}\}_{t\in\mathcal{I}}$ *for all* $\{x_{t,k}\}_{t\notin\mathcal{I}}$.

All objects are identified up to the same relabeling of the groups.

An important implication of Theorem A.1.3 is that whenever $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified at $\tilde{x}_t$, it is also identified for all other values of $x_{t,k'}$ with $k' \neq k$, holding $\tilde{x}_{t,k}$ fixed, since $\pi_j(\cdot)$ is not a function of $x_{it,k'}$ for $k' \neq k$. Similar to Remark 5 for Theorem 3.2, Theorem A.1.3 does not state the full identification result due to the iterative structure of the argument; we elaborate on this in the proof. Analogous to Theorem 3.3 in Section 3.2, one can reduce the number of required time periods to $T = 2$ if there is sufficient variation in the component weights. We discuss this in more detail at the end of this section.

A direct and important corollary of Theorem A.1.3 is the following:

Corollary A.1.3.1 *Assume that* $\pi_j(X) = \pi_j > 0$ *for all* $j = 1, \ldots, J$ *and* $X \in \mathbb{X}$*, Assumption I-1 holds, and there exists* $X_i = X$ *at which Assumption I-2′ is satisfied. Then, (i)* $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ *is identified for all* $t = 1, \ldots, T$*,* $s_t \in \{0, 1\}$*,* $j = 1, \ldots, J$*, and* $\tilde{x}_t \in \mathbb{X}_t$*, and (ii)* π *is identified. All objects are identified up to the identical relabeling of the groups.*

Importantly, Corollary A.1.3.1 allows for an arbitrary number of groups J as long as Assumption I-2′ is satisfied. Corollary A.1.3.1 is closely related to, albeit not a special case of Proposition 4 in Kasahara and Shimotsu (2009). It highlights the power of the presence of covariates if they only enter the component distributions.

Leveraging variability in the component weights. We end this section with a brief discussion on how the variability of the component weights in the covariates, if present, may be used to decrease the number of time periods required for identification. To this intent, we make the following assumption.

Assumption I-4′ *(Variation in the component weights and distributions) Given some fixed* $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$*, there exist* $\tilde{t}, t^* \in \{1, \ldots, T\}$ *with* $\tilde{t} \neq t^*$ *such that*

1. *there exist* $x_{\tilde{t}}^{(r)} \in \mathbb{X}_{\tilde{t}}$ *for* $r = 1, \ldots, R$ *with* $x_{\tilde{t},k}^{(r)} = x_{\tilde{t},k}$ *for all* r*, and, similarly,* $x_{t^*}^{(r')} \in \mathbb{X}_{t^*}$ *for* $r' = 1, \ldots, R'$ *with* $x_{t^*,k}^{(r')} = x_{t^*,k}$ *for all* r' *such that* $\mathcal{Q}_{\tilde{t}}(\{x_{\tilde{t}}^{(r)}\}_{r=1}^{R})$ *and* $\mathcal{Q}_{t^*}(\{x_{t^*}^{(r')}\}_{r'=1}^{R'})$ *have full column rank,*
2. *there exist* $\tilde{x}_{\tilde{t}} \in \mathbb{X}_{\tilde{t}}$ *and* $\tilde{x}_{\tilde{t}}^{(\tilde{r})} \in \mathbb{X}_{\tilde{t}}$ *for* $\tilde{r} = 1, \ldots, \tilde{R}$ *with* $\tilde{x}_{\tilde{t},k}^{(\tilde{r})} = \tilde{x}_{\tilde{t},k}$ *for all* $\tilde{r}$*, and, similarly,* $\underline{x}_{t^*} \in \mathbb{X}_{t^*}$ *and* $\underline{x}_{t^*}^{(\underline{r})} \in \mathbb{X}_{t^*}$ *for* $\underline{r} = 1, \ldots, \underline{R}$ *with* $\underline{x}_{t^*,k}^{(\underline{r})} = \underline{x}_{t^*,k}$ *for all* $\underline{r}$ *such that* $\mathcal{Q}_{\tilde{t}}(\{\tilde{x}_{\tilde{t}}^{(\tilde{r})}\}_{\tilde{r}=1}^{\tilde{R}})$ *and* $\mathcal{Q}_{t^*}(\{\underline{x}_{t^*}^{(\underline{r})}\}_{\underline{r}=1}^{\underline{R}})$ *have full column rank and*

$$\Pi(x_{\tilde{t},k}, x_{t^*,k}, \{x_{t,k}\}_{t\notin\{\tilde{t},t^*\}})\Pi(\tilde{x}_{\tilde{t},k}, x_{t^*,k}, \{x_{t,k}\}_{t\notin\{\tilde{t},t^*\}})^{-1}$$

$$\times\ \Pi(\tilde{x}_{\tilde{t},k}, \underline{x}_{t^*,k}, \{x_{t,k}\}_{t\notin\{\tilde{t},t^*\}})\Pi(x_{\tilde{t},k}, \underline{x}_{t^*,k}, \{x_{t,k}\}_{t\notin\{\tilde{t},t^*\}})^{-1}$$

is well-defined and has distinct and non-zero diagonal entries.[1]

Assumption I-4′ requires as few as two time periods, that is, *in the presence of an additional exclusion restriction, identification in the binary outcome setting may be possible with two time periods even for* $J > 2$ as long as the covariates induce sufficient variation in the component weights and distributions in the sense of Assumption I-4′. We conclude this section with the following theorem.

Theorem A.1.4 *There exists* $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ *such that Assumption I-4′ is satisfied with periods* $\tilde{t}$ *and* t^**, and Assumption I-1 holds. Then,*

1. *for all* $t \neq \tilde{t}$*,* $j = 1, \ldots, J$*, and* $s_t \in \{0, 1\}$*,* $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \overline{x}_t, g_i = j)$ *is identified at* $\overline{x}_t \in \mathbb{X}_t$ *for which there exists* $\{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}$ *such that* $\pi_j(x_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \overline{x}_{t,k}) > 0$ *or* $\pi_j(\tilde{x}_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \overline{x}_{t,k}) > 0$,[2]
2. *for all* $t \neq t^*$*,* $j = 1, \ldots, J$*, and* $s_t \in \{0, 1\}$*,* $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \overline{x}_t, g_i = j)$ *is identified at* $\overline{x}_t \in \mathbb{X}_t$ *for which there exists* $\{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,t^*\}}$ *such that* $\pi_j(x_{t^*,k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,t^*\}}, \overline{x}_{t,k}) > 0$ *or* $\pi_j(\underline{x}_{t^*,k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,t^*\}}, \overline{x}_{t,k}) > 0$*, and*
3. *for any submodel* $\mathcal{I} \subseteq \{1, \ldots, T\}$ *and* $\{\tilde{x}_t^{(r)}\}_{t\in\mathcal{I}}$ *with* $r = 1, \ldots, R$ *and* $\tilde{x}_{t,k}^{(r)} = \tilde{x}_{t,k}^{(s)}$ *for all* $t \in \mathcal{I}$ *and* r, s *such that* $\mathcal{Q}_\mathcal{I}(\{\{\tilde{x}_t^{(r)}\}_{t\in\mathcal{I}}\}_{r=1}^R)$ *has full column rank and is identified given 1. or 2.,* $\underline{\pi}(\{\tilde{x}_{t,k}\}_{t\in\mathcal{I}}, \{x_{t,k}\}_{t\notin\mathcal{I}})$ *is identified at* $\{\tilde{x}_{t,k}\}_{t\in\mathcal{I}}$ *for all* $\{x_{t,k}\}_{t\notin\mathcal{I}}$*.*

All objects are identified up to the same relabeling of the groups.

Importantly, the typical corollary for the case of $\pi_j(X) > 0$ for all $X \in \mathbb{X}$ and $j = 1, \ldots, J$ applies here, too, that is, an analogue of Corollary 3.3.1 holds.

A.1.6 An additional exclusion restriction

We consider a modification of the main model in equation (2). Specifically,

$$\begin{aligned}\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i, w_i) &= \sum_{j=1}^J \mathbb{P}(g_i = j \mid X_i, w_i) \prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j) \\ &= \sum_{j=1}^J \pi_j(X_i, w_i) \prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = j)\end{aligned}$$

that is, the component weights additionally depend on some covariate w_i that does not enter any component distribution. Such a model arises, for instance, in Example 2.3 or when

[1] When $T = 2$, $\Pi(x_{\tilde{t},k}, x_{t^*,k}, \{x_{t,k}\}_{t\notin\{\tilde{t},t^*\}})$ is to be understood as $\Pi(x_{\tilde{t},k}, x_{t^*,k})$.

[2] When $T = 2$, then, for instance, $\pi_j(x_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \overline{x}_{t,k})$ is to be understood as $\pi_j(x_{\tilde{t},k}, \overline{x}_{t,k})$.

a researcher wants to split the sample on a variable that does not enter the component distributions and is measured with measurement error. We shall argue that an additional exclusion restriction of this kind may act like an additionally observed time period.[3] Letting $\mathcal{I}_1$ and $\mathcal{I}_2$ be two non-empty disjoint submodels, we have

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X_i, w_i) = \mathcal{P}_{\mathcal{I}_1}(\{x_{it}\}_{t\in\mathcal{I}_1})\Pi(X_i, w_i)\mathcal{P}_{\mathcal{I}_2}(\{x_{it}\}_{t\in\mathcal{I}_2})^\top \tag{A.2}$$

where the notation is identical to Section 3.2 and Appendix A.1.3, and $\Pi(X_i, w_i) = \text{diag}(\pi_1(X_i, w_i), \ldots, \pi_J(X_i, w_i))$. Since equation (A.2) holds for all w_i, we may vary w_i in (A.2) to get a system of equations which we can use in our identification argument. In the following, we briefly sketch the idea of the identification argument. To fully formalize this sketch, similar arguments as in the proof of Theorem 3.2 may be employed.

We fix $X_i = X$ and w_i at w as well as $\tilde{w}$ so that

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X, w)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top$$
$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, \tilde{w}) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X, \tilde{w})\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top$$

We make the following assumptions: *(i)* $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}_\ell}$ has full column rank for $\ell = 1, 2$, *(ii)* $\Pi(X, w)$ and $\Pi(X, \tilde{w})$ are invertible, and *(iii)* $\Pi(X, \tilde{w})\Pi(X, w)^{-1}$ has distinct diagonal entries. Crucially, we do not require an additional period $t' \notin \mathcal{I}_1 \cup \mathcal{I}_2$ anymore. Thus, we only need $T \geq 2\lceil\log_2(J)\rceil$. Now,

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, \tilde{w})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w)^\dagger = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X, \tilde{w})\Pi(X, w)^{-1}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger$$

It follows that up to identical permutations of the groups, $\Pi(X, \tilde{w})\Pi(X, w)^{-1}$ is identified as the non-zero eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, \tilde{w})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w)^\dagger$ and the columns of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ are identified as the corresponding eigenvectors of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, \tilde{w})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w)^\dagger$ where we use that the non-zero eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, \tilde{w})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w)^\dagger$ are distinct. The sign and scaling of the eigenvectors are pinned down by the facts that all entries of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ are non-negative and all its columns sum up to 1. At the same time

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, \tilde{w})^\top(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w)^\top)^\dagger = \mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})\Pi(X, \tilde{w})\Pi(X, w)^{-1}\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\dagger$$

so that the columns of $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$ are up to the same permutation identified as the eigenvectors of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, \tilde{w})^\top(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w)^\top)^\dagger$ associated with the non-zero eigenvalues. Hence, $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$, $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$, and $\Pi(X, w)\Pi(X, \tilde{w})^{-1}$ are identified up to the same relabeling of the groups. Additionally,

$$\Pi(X, w) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X, w)(\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top)^\dagger$$

and similarly for $\Pi(X, \tilde{w})$. Hence, these objects are also identified up to the same relabeling

[3] I thank Koen Jochmans for mentioning this idea to me in a private conversation.

of the groups.

To solve the intra-component label switching problem, similar arguments as in the proof of Theorem 3.2 apply. In these arguments, we can additionally use the variation due to w to ensure that $\Pi(\cdot)$ has full rank, if required.

A.1.7 More on the dynamic panel setting

We begin with the main identification result that summarizes the discussion on the dynamic panel setting of Section 7.3 and implies Lemma 7.2.

Theorem A.1.5 *There exist $X_i = X \in \mathbb{X}$ and $y_{it} = s$ for even t and $t = 0$ such that Assumptions I-2″ and I-3″ are satisfied with submodels $\mathcal{O}_1$ and $\mathcal{O}_2$. For some submodel $\mathcal{O} \subseteq \mathcal{T}^{(odd)}$, we define $\mathcal{O}^{\oplus} = \{t+1 : t \in \mathcal{O}\}$ and $\underline{\mathcal{O}} = \mathcal{O} \cup \mathcal{O}^{\oplus}$. Then,*

1. *for all $t \in \underline{\mathcal{O}_1}$, $j = 1, \ldots, J$, and $s_t \in \{0,1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, y_{it-1} = \tilde{s}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{s})^\top \in \mathbb{X}_t \times \{0,1\}$ for which there exist $\{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \notin \{t,t+1,\ldots,T-1\}}$ and $\{\underline{s}_t\}_{t=0}^{t-2}$ such that $\pi_j(\{x_t\}_{t \in \{t+1,\ldots,T-1\}}, \tilde{x}_t, \{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \notin \{t,t+1,\ldots,T-1\}}, y_{it-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{t-2}) > 0$,*
2. *for all $t \in \underline{\mathcal{O}_2}$, $j = 1, \ldots, J$, and $s_t \in \{0,1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, y_{it-1} = \tilde{s}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{s})^\top \in \mathbb{X}_t \times \{0,1\}$ for which there exists $\{\tilde{x}_{\tilde{t}}\}_{\tilde{t} > t}$ so that $\pi_j(\{x_{t^*}\}_{t^* \in \{1,\ldots,t-1\}}, \tilde{x}_t, \{\tilde{x}_{\tilde{t}}\}_{\tilde{t} > t}, y_{i0} = \tilde{s}) > 0$,*
3. *for all $j = 1, \ldots, J$ and $s_T \in \{0,1\}$, $\mathbb{P}_T(y_{iT} = s_T \mid x_{iT} = \tilde{x}_T, y_{iT-1} = \tilde{s}, g_i = j)$ is identified at $(\tilde{x}_T^\top, \tilde{s})^\top \in \mathbb{X}_T \times \{0,1\}$ for which there exist $\{\tilde{x}_t\}_{t \notin \underline{\mathcal{O}_2} \cup \{T\}}$ and $\{\underline{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1}$ such that $\pi_j(\{x_t\}_{t \in \underline{\mathcal{O}_2}}, \tilde{x}_T, \{\tilde{x}_t\}_{t \notin \underline{\mathcal{O}_2} \cup \{T\}}, y_{i \max(\underline{\mathcal{O}_1})} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1}) > 0$, and*
4. *for any submodel $\mathcal{O} \subseteq \mathcal{T}^{(odd)}$ of adjacent periods and $\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}}$ such that $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}})$ has full column rank for all $\tilde{s} \in \{0,1\}$ and is identified given 1., 2., or 3., $\underline{\pi}(\tilde{x}_{t \in \underline{\mathcal{O}}}, \{x_t\}_{t \notin \underline{\mathcal{O}}}, y_{i0} = s)$ is identified at $\{\tilde{x}_t\}_{t \in (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ for all $\{x_t\}_{t \notin (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ and $s \in \{0,1\}$.*

All objects are identified up to the same relabeling of the groups.

We proof Theorem A.1.5 in Appendix A.7.1.

Leveraging the variability in the component weights. Next, we discuss how the dependence of T on J can be relaxed. For simplicity, we assume that T is odd for the remainder of this section. We recall

$$\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{x_t, y_{t-1}, x_{t+1}, y_{t+1}\}_{t \in \mathcal{O}}) = \overset{\text{col}}{\otimes}_{t \in \mathcal{O} \setminus \max(\mathcal{O})} \mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1}) \overset{\text{col}}{\otimes} \mathcal{P}_{\max(\mathcal{O})}(x_{\max(\mathcal{O})}, y_{\max(\mathcal{O})-1})$$

where, for notational convenience, we keep $x_{\max(\mathcal{O})+1}$ and $y_{\max(\mathcal{O})+1}$ as arguments of $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\cdot)$. When $\mathcal{O}$ is a singleton, $\overset{\text{col}}{\otimes}_{t \in \mathcal{O} \setminus \max(\mathcal{O})} \mathcal{P}_t^{\otimes}(x_t, y_{t-1}, x_{t+1}, y_{t+1})^\top := \mathbf{1}_J$, the vector of lengths J containing only ones. Crucially, $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{x_t, y_{t-1}, x_{t+1}, y_{t+1}\}_{t \in \mathcal{O}})$ contains one time period fewer

than $\mathcal{P}^{\otimes}_{\mathcal{O}}(\{x_t, y_{t-1}, x_{t+1}, y_{t+1}\}_{t\in\mathcal{O}})$. Specifically, it does not include period $\max(\mathcal{O})+1$. We are ready to present our assumptions.

Assumption $\tilde{\text{I}}$-4′ *(Variation in the component distributions) Given some fixed $(X_i^\top, y_{i0}) = (X^\top, s)$ for $s \in \{0,1\}$ and $X \in \mathbb{X}$, there exists a partition $\mathcal{T}^{(odd)} = \{\mathcal{O}_1, \mathcal{O}_2\}$ such that (i) $\mathcal{O}_1$ and $\mathcal{O}_2$ are non-empty and disjoint and $\max(\mathcal{O}_1) < \min(\mathcal{O}_2)$, (ii) $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})$ and $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_2})$ have full column rank for $\tilde{s} \in \{0,1\}$, (iii) $\mathcal{P}^{\otimes}_{\mathcal{O}_1\setminus\{\max(\mathcal{O}_1)\}}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_1\setminus\{\max(\mathcal{O}_1)\}}) \overset{col}{\otimes} \mathcal{P}^{\otimes}_{\max(\mathcal{O}_1)}(x_{\max(\mathcal{O}_1)}, s, x_{\max(\mathcal{O}_1)+1}, \tilde{s})$ has full column rank for $\tilde{s} \neq s$, (iv) $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, y_{it-1} = s_{t-1}, g_i = j) \in (0,1)$ for all $j \in \{1,\dots,J\}$, $t = 1,\dots,T$, and $s_{t-1} \in \{0,1\}$ and (v) there exist $\{\tilde{x}_t, \tilde{x}_{t+1}\}_{t\in\mathcal{O}_1} \neq \{x_t, x_{t+1}\}_{t\in\mathcal{O}_1}$ and $(\{\underline{x}_t, \underline{x}_{t+1}\}_{t\in\mathcal{O}_2\setminus\{T\}}, \underline{x}_T^\top) \neq (\{x_t, x_{t+1}\}_{t\in\mathcal{O}_2\setminus\{T\}}, x_T^\top)$[4] such that*

$$\begin{aligned}&\Pi(\{x_t, x_{t+1}\}_{t\in\mathcal{O}_1}, \{x_t, x_{t+1}\}_{t\in\mathcal{O}_2\setminus\{T\}}, x_T, \tilde{s})\Pi(\{\tilde{x}_t, \tilde{x}_{t+1}\}_{t\in\mathcal{O}_1}, \{x_t, x_{t+1}\}_{t\in\mathcal{O}_2\setminus\{T\}}, x_T, \tilde{s})^{-1}\\ \times\ &\Pi(\{\tilde{x}_t, \tilde{x}_{t+1}\}_{t\in\mathcal{O}_1}, \{\underline{x}_t, \underline{x}_{t+1}\}_{t\in\mathcal{O}_2\setminus\{T\}}, \underline{x}_T, \tilde{s})\Pi(\{x_t, x_{t+1}\}_{t\in\mathcal{O}_1}, \{\underline{x}_t, \underline{x}_{t+1}\}_{t\in\mathcal{O}_2\setminus\{T\}}, \underline{x}_T, \tilde{s})^{-1}\end{aligned}$$

is well-defined and has distinct and non-zero diagonal entries for $\tilde{s} \in \{0,1\}$, and that $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})$ and $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{\underline{x}_t, \tilde{s}, \underline{x}_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_2})$ have full column rank for $\tilde{s} \in \{0,1\}$ as well as that $\mathcal{P}^{\otimes}_{\mathcal{O}_1\setminus\{\max(\mathcal{O}_1)\}}(\{\tilde{x}_t, s, \tilde{x}_{t+1}, s\}_{t\in\mathcal{O}_1\setminus\{\max(\mathcal{O}_1)\}}) \overset{col}{\otimes} \mathcal{P}^{\otimes}_{\max(\mathcal{O}_1)}(\tilde{x}_{\max(\mathcal{O}_1)}, s, \tilde{x}_{\max(\mathcal{O}_1)+1}, \tilde{s})$[5] *has full column rank for $\tilde{s} \neq s$.*

Assumption $\tilde{\text{I}}$-3′ *(Positive weights) $\Pi(X, \{s_t\}_{t=0}^k) = diag(\underline{\pi}(X, \{s_t\}_{t=0}^k))$ has full rank for $\{s_t\}_{t=0}^k \in \{0,1\}^{k+1}$ and $k = 0,\dots,\min(\mathcal{O}_2)-1$ with $\mathcal{O}_2$ and X of Assumption $\tilde{I}$-4′. Additionally, $Supp((X_i^\top, \{y_{it}\}_{t=0}^{\min(\mathcal{O}_2)-1})) = \mathbb{X}_1 \times \cdots \times \mathbb{X}_T \times \{0,1\}^{\min(\mathcal{O}_2)}$.*

A necessary condition for Assumption $\tilde{\text{I}}$-4′ to hold is $T \geq 4\lceil \log_2(J) \rceil - 1$, that is, with $J = 2$ identification is possible with only four time periods. To keep the assumptions interpretable they are substantially stronger than required, as revealed by the proof. For instance, we do not require that $\mathcal{O}_1$ and $\mathcal{O}_2$ partition $\mathcal{T}^{(\text{odd})}$ but only that these submodels are adjacent submodels and of adjacent time periods, that is, $\max(\mathcal{O}_1) + 2 = \min(\mathcal{O}_2)$. The conditions of the assumptions appear involved because we have to ensure that varying y_{i0} is possible while keeping the group assignment identical. Previously, we used the final period T to do so. We arrive at the following theorem.

Theorem A.1.6 *There exist $X_i = X = (x_1^\top, \dots, x_T^\top)^\top \in \mathbb{X}$ and $y_{it} = s$ for even t and $t = 0$ such that Assumptions $\tilde{I}$-4′ and $\tilde{I}$-3′ are satisfied with submodels $\mathcal{O}_1$ and $\mathcal{O}_2$. For some submodel $\mathcal{O} \subseteq \mathcal{T}^{(odd)}$, we define $\mathcal{O}^{\oplus} = \{t+1 : t \in \mathcal{O}\}$ and $\underline{\mathcal{O}} = \mathcal{O} \cup \mathcal{O}^{\oplus}$. Then,*

[4] When $|\mathcal{O}_2| = 1$, then $\mathcal{O}_2 = \{T\}$ and $(\{x_t, x_{t+1}\}_{t\in\mathcal{O}_2\setminus\{T\}}, x_T^\top)$ is to be understood as $x_T^\top$.

[5] $\mathcal{P}^{\otimes}_{\mathcal{O}_1\setminus\max(\mathcal{O}_1)}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_1\setminus\max(\mathcal{O}_1)})^\top := \mathbf{1}_J \in \mathbb{R}^J$ if $\mathcal{O}_1 \setminus \max(\mathcal{O}_1) = \emptyset$.

1. *for all* $t \in \underline{\mathcal{O}_1}$, $j = 1, \ldots, J$, *and* $s_t \in \{0,1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, y_{it-1} = \tilde{s}, g_i = j)$ *is identified at* $(\tilde{x}_t^\top, \tilde{s})^\top \in \mathbb{X}_t \times \{0,1\}$ *for which there exist* $\{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \notin \{t,t+1,\ldots,T-1\}}$ *and* $\{\underline{s}_t\}_{t=0}^{t-2}$ *such that* $\pi_j(\{x_t\}_{t \in \{t+1,\ldots,T\}}, \tilde{x}_t, \{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \notin \{t,t+1,\ldots,T\}}, y_{it-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{t-2}) > 0$,
2. *for all* $t \in \underline{\mathcal{O}_2} \setminus \{T+1\}$, $j = 1, \ldots, J$, *and* $s_t \in \{0,1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, y_{it-1} = \tilde{s}, g_i = j)$ *is identified at* $(\tilde{x}_t^\top, \tilde{s})^\top \in \mathbb{X}_t \times \{0,1\}$ *for which there exists* $\{\tilde{x}_{\tilde{t}}\}_{\tilde{t} > t}$ *so that* $\pi_j(\{x_{t^*}\}_{t^* \in \{1,\ldots,t-1\}}, \tilde{x}_t, \{\tilde{x}_{\tilde{t}}\}_{\tilde{t} > t}, y_{i0} = \tilde{s}) > 0$,
3. *for any submodel* $\mathcal{O} \subseteq \mathcal{T}^{(odd)}$ *of adjacent periods and* $\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}}$ *such that* $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}})$ *has full column rank for all* $\tilde{s} \in \{0,1\}$ *and is identified given 1. or 2.,* $\underline{\pi}(\tilde{x}_{t \in \underline{\mathcal{O}}}, \{x_t\}_{t \notin \underline{\mathcal{O}}}, y_{i0} = s)$ *is identified at* $\{\tilde{x}_t\}_{t \in (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ *for all* $\{x_t\}_{t \notin (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ *and* $s \in \{0,1\}$.

All objects are identified up to the same relabeling of the groups.

A simple corollary is the following.

Corollary A.1.6.1 *There exist* $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ *and* $s \in \{0,1\}$ *such that Assumption* $\tilde{I}$*-4′ holds. Additionally, assume that* $\mathbb{P}(g_i = j \mid X_i = X, \{y_{it} = s_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})}) > 0$ *for all* $j = 1, \ldots, J$, $X \in \mathbb{X}$, *and* $\{s_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})} \in \{0,1\}^{\max(\underline{\mathcal{O}_1})+1}$ *and that the support condition of Assumption* $\tilde{I}$*-3′ holds. Then*

- $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, y_{it-1} = \tilde{s}, g_i = j)$ *is identified for all* $(\tilde{x}_t^\top, \tilde{s})^\top \in \mathbb{X}_t \times \{0,1\}$, $t = 1, \ldots, T$, $j = 1, \ldots, J$, *and* $s_t \in \{0,1\}$, *and*
- $\underline{\pi}(\tilde{X}, y_{i0} = \tilde{s})$ *is identified at all* $(\tilde{X}^\top, \tilde{s})^\top \in \mathbb{X} \times \{0,1\}$ *such that there exists a submodel* $\mathcal{O} \subseteq \mathcal{T}^{(odd)}$ *of adjacent periods with* $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}})$ *having full column rank for all* $\tilde{s} \in \{0,1\}$.

All objects are identified up to the same relabeling of the groups.

We include the proofs of Theorem A.1.6 and Corollary A.1.6.1 in Appendix A.7.

Identification of AMEs. Identification of a group-specific AME with respect to the lagged dependent variable, that is, $\mathbb{E}\left[\mathbb{P}_t(y_{it} = 1 \mid x_{it}, y_{it-1} = 1, g_i = j) - \mathbb{P}_t(y_{it} = 1 \mid x_{it}, y_{it-1} = 0, g_i = j) \mid g_i = j\right]$, follows under the same assumptions as the identification of the AMEs in Section 3.2. Identification of the group-specific AME of the following kind $\text{AME}_{j,t,k} = \mathbb{E}\left[\frac{\partial}{\partial x_{it,k}} \mathbb{P}_t(y_{it} = 1 \mid x_{it}, y_{it-1}, g_i = j) \mid g_i = j\right] = \mathbb{E}\left[\mathbb{P}(g_i = j \mid X_i, y_{it-1}) \frac{\partial}{\partial x_{it,k}} \mathbb{P}_t(y_{it} = 1 \mid x_{it}, y_{it-1}, g_i = j)\right] / \pi_j$ is a bit more delicate as it requires us to identify $\mathbb{P}(g_i = j \mid X_i, y_{it-1})$. Under an analogous assumption to Assumption I-7 in Section 3.2, our previous results only imply that $\mathbb{P}(g_i = j \mid X_i, y_{i0})$ is identified. We proceed to argue that $\mathbb{P}(g_i = j \mid X_i = X, y_{it-1} = s_{t-1})$ is identified for all t

and $(X^\top, s_{t-1}) \in \mathbb{X} \times \{0,1\}$ under analogous assumptions as in Section 3.2. To see this, we note

$$\mathbb{P}(g_i = j \mid X_i = X, y_{it-1} = s_{t-1}) = \frac{\mathbb{P}(g_i = j, y_{it-1} = s_{t-1} \mid X_i = X)}{\mathbb{P}(y_{it-1} = s_{t-1} \mid X_i = X)}$$

where denominator is identified. We rewrite the numerator further

$$\begin{aligned}\mathbb{P}(g_i = j, y_{it-1} = s_{t-1} \mid X_i = X) &= \sum_{s_{t-2} \in \{0,1\}} \mathbb{P}(g_i = j, y_{it-1} = s_{t-1}, y_{it-2} = s_{t-2} \mid X_i = X) \\ &= \sum_{s_{t-2} \in \{0,1\}} \Big\{ \mathbb{P}_{t-1}(y_{it-1} = s_{t-1} \mid x_{it-1} = x_{t-1}, y_{it-2} = s_{t-2}, g_i = j) \\ &\quad \times \mathbb{P}(g_i = j \mid X_i = X, y_{it-2} = s_{t-2})\, \mathbb{P}(y_{it-2} = s_{t-2} \mid X_i = X) \Big\}\end{aligned}$$

where we used Assumption M-2. For $t = 2$, all three terms are identified so that $\mathbb{P}(g_i = j \mid X_i = X, y_{i1} = s_1)$ is identified, which in turn implies that the three final terms are identified for $t = 3$. Hence, $\mathbb{P}(g_i = j \mid X_i = X, y_{it-1} = s_{t-1})$ is recursively identified for all t. Now, the identification of $\text{AME}_{j,t,k}$ follows from the arguments from Section 3.2 under analogous assumptions.

A.1.8 Lagged covariates

In this section, we allow lagged covariates to enter the component distributions. All proofs are included in Appendix A.8. Specifically, we make the following assumption:

Assumption M-3 *(Lagged covariates) For all $s_t \in \{0,1\}$ and $j \in \{1, \ldots, J\}$, $\mathbb{P}_t(\{y_{it} = s_t\}_{t=1}^T \mid X_i, g_i = j) = \prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, x_{it-1}, g_i = j)$ with $X_i = (x_{i0}^\top, x_{i1}^\top, \ldots, x_{iT}^\top)^\top$.*

Under Assumption M-3, the mixture model of interest becomes

$$\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T \mid X_i) = \sum_{j=1}^J \mathbb{P}(g_i = j \mid X_i) \prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, x_{it-1}, g_i = j)$$

where we assume that x_{i0} is observed. An interesting special case of this model is when $J = 2$ and $\prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, x_{it-1}, g_i = 1) = \prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, g_i = 1)$ and $\prod_{t=1}^T \mathbb{P}_t(y_{it} = s_t \mid x_{it}, x_{it-1}, g_i = 2)$ depends on x_{it-1}, that is, the groups are motivated by whether the component distributions are Markovian of order 1 or not.[6] For any submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$, we still have

$$\mathbb{P}(\{y_{it} = s_t\}_{t \in \mathcal{I}} \mid X_i) = \sum_{j=1}^J \mathbb{P}(g_i = j \mid X_i) \prod_{t \in \mathcal{I}} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, x_{it-1}, g_i = j)$$

While we focus on lags of order 1, our arguments can be generalized to lags of higher order. Allowing for lagged covariates in the component distributions does not affect the

[6] I am grateful to Matias Cattaneo for drawing my attention to this motivational perspective.

identification argument at a fixed $X_i = X$. It, however, affects the exclusion restriction we have previously used to solve the intra-component label switching problem. To state our assumptions, we let $\overline{\mathcal{I}} = \mathcal{I} \cup \{\mathcal{I} \ominus 1\}$ for some submodel $\mathcal{I}$ where $\mathcal{I} \ominus 1 = \{t - 1 : t \in \mathcal{I}\}$. Next, we define $\mathcal{P}_t(x_t, x_{t-1})$ and $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\overline{\mathcal{I}}})$ for some submodel $\mathcal{I}$ analogous to $\mathcal{P}_t(x_t)$ and $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ in Section 3.2. We assume

Assumption I-2‴ *(Variation in the component distributions) Let $t' = \lceil T/2 \rceil$. There exist two non-empty submodels $\mathcal{I}_1 \subseteq \{1, \ldots, t'-1\}$ and $\mathcal{I}_2 \subseteq \{t'+1, \ldots, T\}$ such that $\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\overline{\mathcal{I}}_\ell})$ has full column rank for $\ell = 1, 2$ and $\mathcal{P}_{t'}(x_{t'}, x_{t'-1})$ has distinct columns.*

Assumption I-2‴ ensures sufficient spacing between the two submodels that we use for identification, that is, we can use one submodel to identify the component distributions for time periods in the other submodel at different values of the covariates. The assumption may be weakened as long as the time periods in $\mathcal{I}_1$ and $\mathcal{I}_2$ are well-separated. We arrive at the following theorem.

Theorem A.1.7 *There exists $X_i = X = (x_0^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ such that Assumptions I-2‴ and I-3 are satisfied with submodels $\mathcal{I}_1$ and $\mathcal{I}_2$, and Assumption I-1 holds. Then,*[7]

1. *for all $t \in \{1, \ldots, t'-1\}$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}$ such that $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_2}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1}) > 0$,*
2. *for all $t \in \{t'+1, \ldots, T\}$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_1\cup\{t,t-1\}}$ such that $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_1}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_1\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1}) > 0$,*
3. *for all $j = 1, \ldots, J$ and $s_{t'} \in \{0, 1\}$, $\mathbb{P}_{t'}(y_{it'} = s_{t'} \mid x_{it'} = \tilde{x}_{t'}, x_{it'-1} = \tilde{x}_{t'-1}, g_i = j)$ is identified at $(\tilde{x}_{t'}^\top, \tilde{x}_{t'-1}^\top)^\top \in \mathbb{X}_{t'} \times \mathbb{X}_{t'-1}$ for which there exists $\{\tilde{x}_{t^*}\}_{t^*\notin\{t',t'-1\}}$ and a submodel $\mathcal{I} \subseteq \{1, \ldots, T\} \setminus \{t'\}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})$ has full column rank and is identified via 1. or 2., and $\pi_j(\tilde{X}) > 0$ where $\tilde{X} = (\tilde{x}_0^\top, \ldots, \tilde{x}_T^\top)^\top$, and*
4. *for any submodel $\mathcal{I}$ and $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})$ has full column rank and is identified given 1., 2., or 3., $\underline{\pi}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}}, \{x_t\}_{t\notin\overline{\mathcal{I}}})$ is identified at $\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}}$ for all $\{x_t\}_{t\notin\overline{\mathcal{I}}}$.*

All objects are identified up to the same relabeling of the groups.

A few remarks are in order.

Remark 11. As in the previous sections, the identification result may be strengthened at the cost of additional notation. We discuss how to do so in the proof of Theorem A.1.7.

[7] $t \notin \mathcal{I}$ for $\mathcal{I} \subseteq \{0, 1, \ldots, T\}$ is defined with respect to $\{0, 1, \ldots, T\}$ as the reference set.

Remark 12. We abuse the notation heavily in part 1. and 2. of Theorem A.1.7. Specifically, when $J = 2$, it may be the case that some of the sets in part 1. and 2. of Theorem A.1.7 are empty. For instance, if $J = 2$ and $T = 3$, then $t' = 2$, $\mathcal{I}_1 = \{1\}$ with $\overline{\mathcal{I}}_1 = \{0, 1\}$, and $\mathcal{I}_2 = \{3\}$ with $\overline{\mathcal{I}}_2 = \{2, 3\}$. We focus on part 1. of Theorem A.1.7 so that $t = 1$. Then $\{0, \ldots, T\} \setminus (\overline{\mathcal{I}}_2 \cup \{1, 0\}) = \emptyset$ and we understand $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_2}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1})$ to be equal to $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_2}, \tilde{x}_t, \tilde{x}_{t-1})$. Hence, for all $t \in \{1, \ldots, t' - 1\}$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_2}, \tilde{x}_t, \tilde{x}_{t-1}) > 0$. This is for instance satisfied when $\pi_j(X) > 0$ for all $j = 1, \ldots, J$ and $X \in \mathbb{X}$.

Remark 13. Period t' separates the two submodels $\mathcal{I}_1$ and $\mathcal{I}_2$ by one period so that $\overline{\mathcal{I}}_1 \cap \overline{\mathcal{I}}_2 = \emptyset$, which allows for the easy to interpret identification results in part 1. and 2. of Theorem A.1.7. On the other hand, while the component distribution $\mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = \tilde{x}_{t'}, x_{it'-1} = \tilde{x}_{t'-1}, g_i = j)$ may be identified at a large set of covariate values, the set cannot be characterized as easily as before. However, in the presence of an additional time period that separates $\mathcal{I}_1$ and $\mathcal{I}_2$ further, we can restore a result that is similar to before and Corollary 3.2.1. We replace Assumption I-2‴ with the following assumption.

Assumption I-2⁗ *(Additional time period) Let $t' = \lceil T/2 \rceil$. There exist two non-empty submodels $\mathcal{I}_1 \subseteq \{1, \ldots, t' - 2\}$ and $\mathcal{I}_2 \subseteq \{t' + 1, \ldots, T\}$ or $\mathcal{I}_1 \subseteq \{1, \ldots, t' - 1\}$ and $\mathcal{I}_2 \subseteq \{t' + 2, \ldots, T\}$ such that $\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\overline{\mathcal{I}}_\ell})$ has full column rank for $\ell = 1, 2$ and $\mathcal{P}_{t'}(x_{t'}, x_{t'-1})$ has distinct columns.*

A necessary condition for Assumption I-2⁗ is $T \geq 2\lceil \log_2(J) \rceil + 2$ in contrast to $T \geq 2\lceil \log_2(J) \rceil + 1$ previously. We present an analogue to Theorem A.1.7 under Assumption I-2⁗.

Theorem A.1.8 *There exists $X_i = X = (x_0^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ such that Assumptions I-2⁗ and I-3 are satisfied with submodels $\mathcal{I}_1 \subseteq \{1, \ldots, t' - 2\}$ [$\mathcal{I}_1 \subseteq \{1, \ldots, t' - 1\}$] and $\mathcal{I}_2 \subseteq \{t' + 1, \ldots, T\}$ [$\mathcal{I}_2 \subseteq \{t' + 2, \ldots, T\}$], and Assumption I-1 holds. We define $\overline{\mathcal{I}}_\ell = \mathcal{I}_\ell \cup \{\mathcal{I}_\ell \ominus 1\}$ for $\ell = 1, 2$ where for some submodel $\mathcal{I}$, $\mathcal{I} \ominus 1 = \{t - 1 : t \in \mathcal{I}\}$. Then,*

1. *for all $t \in \{1, \ldots, t' - 1\}$ [$t \in \{1, \ldots, t'\}$], $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}$ such that $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_2}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1}) > 0$,*
2. *for all $t \in \{t', \ldots, T\}$ [$t \in \{t' + 1, \ldots, T\}$], $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_1\cup\{t,t-1\}}$ such that $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_1}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_1\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1}) > 0$, and*

3. *for any submodel* $\mathcal{I}$ *and* $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ *such that* $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})$ *has full column rank and is identified given 1. or 2.,* $\underline{\pi}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}}, \{x_t\}_{t\notin\overline{\mathcal{I}}})$ *is identified at* $\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}}$ *for all* $\{x_t\}_{t\notin\overline{\mathcal{I}}}$.

All objects are identified up to the same relabeling of the groups.

We conclude this remark with the following corollary.

Corollary A.1.8.1 *There exists* $X_i = X = (x_0^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ *such that Assumptions I-2‴, I-2⁗,* $\pi_j(X) > 0$ *for all* $j = 1, \ldots, J$ *and* $X \in \mathbb{X}$*, and Assumption I-1 holds. Then*

- $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ *is identified for all* $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$*,* $t = 1, \ldots, T$*,* $j = 1, \ldots, J$*, and* $s_t \in \{0, 1\}$*, and*
- $\underline{\pi}(\tilde{X})$ *is identified at all* $\tilde{X} \in \mathbb{X}$ *such that there exists a submodel* $\mathcal{I} \subseteq \{1, \ldots, T\}$ *with* $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})$ *having full column rank.*

All objects are identified up to the same relabeling of the groups.

A.1.9 Violation of Assumption I-1

It is easy to see that our results continue to hold when the when the submodels $\mathcal{I}_1$ and $\mathcal{I}_2$ as well as the point X such that Assumptions I-2 and I-3 hold are chosen cleverly, that is, such that conditioning on $\{x_t\}_{t\neq t'}$ does not restrict the support of $x_{it'}$ for all $t' \in \{1, \ldots, T\}$. In this section, we consider three more complicated but still simple settings to discuss how to relax Assumption I-1 in the context of the baseline model of Section 3.2. A leading example in economics for a violation of Assumption I-1 is the assignment of treatment at a single point in time $t^* \in \{1, \ldots, T\}$. Let $D_i \in \{0, 1\}$ denote the treatment status of unit i and define the treatment indicator $D_{it} = \mathbb{1}(t \geq t^*)D_i$. In the first part of this section, we focus on including D_{it} as a covariate, while the second part focuses on including covariates Z_i that do not vary over time, for instance, pre-treateatment covariates or fixed covariates Z_{it} that vary of time but are pre-determined such as the age of an individual. The main difficulty that arises in these examples is to ensure the consistent assignment of group labels across the different values of D_{it} and Z_i, respectively. Last, we provide a general discussion of how to achieve identification when Assumption I-1 does not hold.

When D_{it} is included as covariate in the model of equation (2), Assumption I-1 is violated because conditioning on $D_{it} = d$ for some $d \in \{0, 1\}$ and $t \geq t^*$ implies $D_{i\tilde{t}} = d$ for all $\tilde{t} \geq t^*$. Hence, changing the value of the conditioning variable D_{it} changes the support of $D_{i\tilde{t}}$. In the following, we let $\tilde{x}_{it} = (x_{it}^\top, D_{it})^\top$ denote the vector of covariates. For simplicity, we assume

Assumption I-1′ *(Varying support – D_{it}) For all* $t' \in \{1, \ldots, T\}$ *and* $\{x_t, d_t\}_{t\neq t'}$*, Supp*$(x_{it'} \mid \{x_t, d_t\}_{t\neq t'}) = \mathbb{X}_{t'}$*, Supp*$(d_{t'} \mid \{x_t, d_t\}_{t\neq t'}) = \{d_{t^*}\}$ *for* $t' > t^*$*, and Supp*$(d_{t'} \mid \{x_t, d_t\}_{t\neq t'}) =$

$\{0\}$ *for* $t' < t^*$. *Also,* $Supp(d_{t^*} \mid \{x_t, d_t\}_{t \neq t^*}) = \{d_{t^*+1}\}$ *if* $t^* < T$ *and* $Supp(d_{t^*} \mid \{x_t, d_t\}_{t \neq t^*}) = \{0, 1\}$, *otherwise.*

To keep the notation light, Assumption I-1′ requires that $\mathbb{X}_t$ is not a function of the conditioning set, while the support of D_{it} is a function of the conditioning set for $t \geq t^*$.We proceed to present an analogue of Theorem 3.2 under Assumption I-1′. The key observation in establishing this theorem is that the support of D_{it} for $t < t^*$ is not a function of the conditioning set $\{D_{i\tilde{t}}\}_{\tilde{t} \geq t^*}$, while the support of $D_{i\tilde{t}}$ for $\tilde{t} \geq t^*$ is not a function of the conditioning set $\{D_{it}\}_{t<t^*}$. To formalize our argument, we let $\mathbb{D}_t$ denote the support of D_{it}, that is, $\mathbb{D}_t = \{0\}$ for $t \in \{1, \ldots, t^* - 1\}$ and $\mathbb{D}_t = \{0, 1\}$ for $t \in \{t^*, \ldots, T\}$. The support of $\tilde{x}_{it}$ is denoted by $\tilde{\mathbb{X}}_t = \mathbb{X}_t \times \mathbb{D}_t$, while the support of $\tilde{X}_i = (\tilde{x}_{i1}^\top, \ldots, \tilde{x}_{iT}^\top)^\top$ is $\tilde{\mathbb{X}} = \prod_{t=1}^T \tilde{\mathbb{X}}_t$. Lastly, the support of $\tilde{x}_{it'}$ conditional on $\{\tilde{x}_{it} = \tilde{x}_t\}_{t \in \mathcal{I}}$ for some submodel $\mathcal{I}$ with $t' \notin \mathcal{I}$ is denoted by $\tilde{\mathbb{X}}_{t'}(\{d_t\}_{t \in \mathcal{I}})$ where we use Assumption I-1′ to simplify the notation.

Theorem A.1.9 *There exists* $\tilde{X}_i = \tilde{X} = (x_1^\top, d_1, \ldots, x_T^\top, d_T)^\top \in \tilde{\mathbb{X}}$ *such that Assumptions I-2 and I-3 are satisfied with submodels* $\mathcal{I}_1$ *and* $\mathcal{I}_2$.[8] *We assume that* $\mathcal{I}_1 \subseteq \{1, \ldots, t^* - 1\}$ *and that Assumption I-1′ holds. Then,*

1. *for all* $t \notin \mathcal{I}_2$, $j = 1, \ldots, J$, *and* $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid \tilde{x}_{it} = x_t^*, g_i = j)$ *is identified at* $x_t^* = \tilde{x}_t$ *and at* $x_t^* \in \tilde{\mathbb{X}}_t(\{d_t\}_{t \in \mathcal{I}_2})$ *for which there exists* $\{x_{t^*}^*\}_{t^* \notin \mathcal{I}_2 \cup \{t\}} \in \prod_{t^* \notin \mathcal{I}_2 \cup \{t\}} \tilde{\mathbb{X}}_{t^*}(\{d_t\}_{t \in \mathcal{I}_2 \cup \{t\}})$ *such that* $\pi_j(\{\tilde{x}_t\}_{t \in \mathcal{I}_2}, \{x_{t^*}^*\}_{t^* \notin \mathcal{I}_2 \cup \{t\}}, x_t^*) > 0$,
2. *for all* $t \notin \mathcal{I}_1$, $j = 1, \ldots, J$, *and* $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid \tilde{x}_{it} = x_t^*, g_i = j)$ *is identified at* $x_t^* = \tilde{x}_t$ *and at* $x_t^* \in \tilde{\mathbb{X}}_t$ *for which there exists* $\{x_{t^*}^*\}_{t^* \notin \mathcal{I}_1 \cup \{t\}} \in \prod_{t^* \notin \mathcal{I}_1 \cup \{t\}} \tilde{\mathbb{X}}_{t^*}(d_t)$ *such that* $\pi_j(\{\tilde{x}_t\}_{t \in \mathcal{I}_1}, \{x_{t^*}^*\}_{t^* \notin \mathcal{I}_1 \cup \{t\}}, x_t^*) > 0$, *and*
3. *for any submodel* $\mathcal{I}$ *and* $\{\tilde{x}_t\}_{t \in \mathcal{I}}$ *such that* $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t \in \mathcal{I}})$ *has full column rank and is identified given 1. or 2.,* $\underline{\pi}(\{\tilde{x}_t\}_{t \in \mathcal{I}}, \{\tilde{x}_t\}_{t \notin \mathcal{I}})$ *is identified at* $\{\tilde{x}_t\}_{t \in \mathcal{I}}$ *for all* $\{\tilde{x}_t\}_{t \notin \mathcal{I}} \in \prod_{t^* \notin \mathcal{I}} \tilde{\mathbb{X}}_{t^*}(\{d_t\}_{t \in \mathcal{I}})$.

All objects are identified up to the same relabeling of the groups.

Remark 14. Importantly, in part 1 $\tilde{\mathbb{X}}_t(\{d_t\}_{t \in \mathcal{I}_2}) = \tilde{\mathbb{X}}_t$ for $t \in \mathcal{I}_1$. In part 2, we have $\tilde{\mathbb{X}}_t(\{d_t\}_{t \in \mathcal{I}_1}) = \tilde{\mathbb{X}}_t$ for all $t \notin \mathcal{I}_1$ as conditioning on $\{D_{it}\}_{t \in \mathcal{I}_1}$ does not put any restrictions on the support of the covariates of the periods $t \notin \mathcal{I}_1$.

Remark 15. In the context of D_{it} being a treatment indicator, Theorem A.1.9 differs from Theorem 3.2 in its assumptions by requiring that there exist sufficiently many pretreatment periods that can be used to construct the submodel $\mathcal{I}_1$ of Assumption I-2. To this end, a necessary condition is the availability of $\lceil \log_2(J) \rceil$ pretreatment periods.

[8] Specifically, replace X with $\tilde{X}$ in Assumptions I-2 and I-3.

The proof of Theorem A.1.9 follows from the same arguments as the proof of Theorem 3.2, albeit with a bit more care given to the support of the covariates. We therefore omit the proof.

While in the presence of D_{it} as a covariate, we are able to essentially restore the conclusion of Theorem 3.2, this is not the case if a time-invariant Z_i or fixed covariate Z_{it} is included as a covariate because by conditioning on some value of Z_i/Z_{it}, we fix either the value of Z_i over time in our conditional analysis for all time periods or the trajectory of Z_{it} is fixed over time. Our previous identification results crucially leverage the fact that we can fix the covariates for some (potentially cleverly chosen) submodel $\mathcal{I}$, while varying the covariates for the periods that are not in $\mathcal{I}$. To keep notation light, we will focus on the case of a time-invariant Z_i here but note that all our arguments apply to the setting with fixed covariates, too. To fix ideas, we impose the following simplifying assumption.

Assumption I-1″ *(Varying support – Z_i) For all $t' \in \{1, \dots, T\}$ and $(\{x_t\}_{t \neq t'}, z)$, Supp$(x_{it'} \mid (\{x_t\}_{t \neq t'}, z)) = \mathbb{X}_{t'}$.*

Similar to the setting with D_{it}, Assumption I-1″ requires that the support of $x_{it'}$, the other covariates excluding Z_i, is not a function of the conditioning set $(\{x_t\}_{t \neq t'}, z)$.

If we condition on $Z_i = z$ for $z \in \mathbb{Z}$, the support of Z_i, the conclusions of Theorem 3.2 apply for this conditional model given $Z_i = z$, that is, we identify the component distributions and weights for all values of x_t and $X = (x_1^\top, \dots, x_T^\top)^\top$ given $Z_i = z$ as asserted in Theorem 3.2 – of course, this assumes that conditional on $Z_i = z$ there exists an $X \in \mathbb{X}$ such that Assumptions I-2 and I-3 hold. However, we cannot link the conditional models across different values of z, which makes it impossible to ensure that the group labels are consistently assigned across these models without further assumptions. If Z_i was continuously distributed, such an additional assumption may be the transversality condition imposed by Huang et al. (2013) and Wang et al. (2014). The following three assumptions and final heuristic argument ensure the consistent assignment of group labels across all values of Z_i even when Z_i is discrete:

1. The component weights do not depend on Z_i, that is, $\underline{\pi}(X_i, Z_i) = \underline{\pi}(X_i)$ for all $X_i \in \mathbb{X}$ and $Z_i \in \mathbb{Z}$, or equivalently $g_i \perp\!\!\!\perp Z_i \mid X_i$ – this assumption may be weakened to hold on a subset of $\mathbb{X}$. Then the arguments of Section 7.2 apply. Specifically, the conclusions of Theorem A.1.3 apply. This argument highlights that the intra-component label switching problem vanishes when $\underline{\pi}(X) = \underline{\pi}$ for all $X \in \mathbb{X}$.
2. There exists some time period t and $x_t \in \mathbb{X}_t$ such that $\mathcal{P}_t(x_t, Z_i = z) = \mathcal{P}_t(x_t)$ for all $z \in \mathbb{Z}$ and $\mathcal{P}_t(x_t)$ has distinct columns and is identified given Theorem 3.2. As the group labels are consistently assigned given $Z_i = z$, one can then uniquely match the

columns of $\mathcal{P}_t(x_t, Z = z)$ with the columns of $\mathcal{P}_t(x_t, Z = z')$ with $z \neq z'$ to align the group labels across the different values of Z_i.

3. In Section 4, we put more structure on the component distributions. In the notation of this section, we assume that $\mathbb{P}_t(y_{it} = 1 \mid Z_i, x_{it}, g_i = j) = F_{jt}(Z_i\beta_{j,t}^{(Z)} + x_{it}^\top\beta_{j,t}^{(x)} + \alpha_{j,t})$ where $F_{jt}(\cdot)$ is known.[9] Here, Z is a scalar random variable for the ease of notation. Assume that the analysis of Sections 3.2 and 4.2 applies conditional on $Z_i = z$ for all $z \in \mathbb{Z}$ at a value of $X \in \mathbb{X}$ so that we achieve identification of the component weights and distributions for all values of Z_i. Specifically, we are only left with aligning the group labels across different values of Z_i. Now, we assume that $F_{jt}(\cdot) = F_t(\cdot)$, $F_t(\cdot)$ is strictly increasing, and $\beta_{j,t}^{(Z)} = \beta_t^{(Z)}$. Under these assumptions, it is straightforward to realign the groups across the different values of Z_i. To see this, we fix X and vary Z_i from z to z' with $z' > z$ without loss of generality. By assumption, $F_t(z\beta_t^{(Z)} + x_t^\top\beta_{j,t}^{(x)} + \alpha_{j,t})$ and $F_t(z'\beta_t^{(Z)} + x_t^\top\beta_{j,t}^{(x)} + \alpha_{j,t})$ are identified for all $j = 1, \ldots, J$. We fix the group labels at z via the order statistics of $\{\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, Z_i = z, g_i = j)\}_{j=1}^J = \{F_t(z\beta_t^{(Z)} + x_t^\top\beta_{j,t}^{(x)} + \alpha_{j,t})\}_{j=1}^J$ or, equivalently, since $F_t(\cdot)$ is strictly increasing, via the order statistics of $\{z\beta_t^{(Z)} + x_t^\top\beta_{j,t}^{(x)} + \alpha_{j,t}\}_{j=1}^J$. Without loss of generality, we may assume that $F_t(z\beta_t^{(Z)} + x_t^\top\beta_{j,t}^{(x)} + \alpha_{j,t}) \neq F_t(z\beta_t^{(Z)} + x_t^\top\beta_{j',t}^{(x)} + \alpha_{j',t})$ for all $j' \neq j$ under Assumption I-2 or I-4.

   Considering the q-th order statistic with groups labels fixed at the initial z via the order statistics, we have

$$x_t^\top\beta_{q,t} + z\beta_t^{(Z)} + \alpha_{q,t} \begin{cases} < x_t^\top\beta_{j,t} + z\beta_t^{(Z)} + \alpha_{j,t} & q < j \\ > x_t^\top\beta_{j,t} + z\beta_t^{(Z)} + \alpha_{j,t} & q > j \end{cases}$$

$$\Leftrightarrow x_t^\top\beta_{q,t} + \alpha_{q,t} \begin{cases} < x_t^\top\beta_{j,t} + \alpha_{j,t} & q < j \\ > x_t^\top\beta_{j,t} + \alpha_{j,t} & q > j \end{cases}$$

   Hence, for all $z \in \mathbb{Z}$ the proposed group assignment mechanism is equivalent to an assignment of the groups via the order statistics of $\{x_t^\top\beta_{j,t}^{(x)} + \alpha_{j,t}\}_{j=1}^J$. Since this group assignment mechanism does not depend on z, the ordering of the groups (as fixed at the initial z) does not vary across the different values of $z \in \mathbb{Z}$ when based on the order statistics of $\{\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, Z_i = z, g_i = j)\}_{j=1}^J$. Put differently, the order statistics of $\{\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, Z_i = z, g_i = j)\}_{j=1}^J$ correspond to the same true groups for all $z \in \mathbb{Z}$. This allows us to align the group labels across different values of Z_i, thus solving the intra-component label switching problem.

   To provide some intuition we note that a direct implication of this discussion is that the

[9] The assumption that $F_{jt}(\cdot)$ is known is not necessary. The component distributions may also be a single index model with unknown link function (Ichimura 1993).

intra-component label switching problem can be easily resolved when only $\alpha_{j,t}$ varies across groups and all slope coefficients $\beta_{j,t} = \beta_t$ for all $j = 1, \ldots, J$. Specifically, we consider the model

$$\mathbb{P}(y_{it} = 1 \mid x_{it}, z_{it}) = \sum_{j=1}^{J} \pi_j(X_i) F(\alpha_{j,t} + x_{it}^\top \beta_t)$$

where $F(\cdot)$ is some known strictly increasing function and $\alpha_{j,t} \neq \alpha_{j',t}$ for $j \neq j'$. Now, assume that the component weights and distributions are identified at the two different pairs of values X and X'. We fix the labels at X and order the component distributions increasingly. Since only the intercept varies across groups and $F(\cdot)$ is strictly increasing, the order of the groups is the same at X', which allows us to align the group labels across different values of X.

4. When the groups are economically interpretable for each $z \in \mathbb{Z}$, then this may allow to match the groups across different value of $z \in \mathbb{Z}$ through their economic meaning.

Lastly, we consider the case when the support of a covariate conditional on all other time periods is a subset of the unconditional support. Given some submodel $\mathcal{I}$, we let $\mathbb{X}_t(\{x_t\}_{t\in\mathcal{I}})$ denote support of x_{it} conditional on the covariates of the chosen submodel $\mathcal{I}$. Under this notation, we may easily restate the assertion of Theorem 3.2.

Theorem A.1.10 *There exists $X_i = X = (x_1^\top, \ldots, x_T^\top)^\top \in \mathbb{X}$ such that Assumptions I-2 and I-3 are satisfied with submodels $\mathcal{I}_1$ and $\mathcal{I}_2$. Then,*

1. *for all $t \notin \mathcal{I}_2$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified at $\tilde{x}_t \in \mathbb{X}_t(\{x_t\}_{t\in\mathcal{I}_2})$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}} \in \prod_{t^*\notin\mathcal{I}_2\cup\{t\}} \mathbb{X}_{t^*}(\{x_t\}_{t\in\mathcal{I}_2}, \tilde{x}_t)$ such that $\pi_j(\{x_t\}_{t\in\mathcal{I}_2}, \{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}}, \tilde{x}_t) > 0$,*
2. *for all $t \notin \mathcal{I}_1$, $j = 1, \ldots, J$, and $s_t \in \{0, 1\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified at $\tilde{x}_t \in \mathbb{X}_t(\{x_t\}_{t\in\mathcal{I}_1})$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_1\cup\{t\}} \in \prod_{t^*\notin\mathcal{I}_1\cup\{t\}} \mathbb{X}_{t^*}(\{x_t\}_{t\in\mathcal{I}_1}, \tilde{x}_t)$ such that $\pi_j(\{x_t\}_{t\in\mathcal{I}_1}, \{x^*_{t^*}\}_{t^*\notin\mathcal{I}_1\cup\{t\}}, \tilde{x}_t) > 0$,*
3. *for any submodel $\mathcal{I}$ and $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ such that $\mathcal{P}_\mathcal{I}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ has full column rank and is identified given 1. or 2., $\underline{\pi}(\{\tilde{x}_t\}_{t\in\mathcal{I}}, \{x_t\}_{t\notin\mathcal{I}})$ is identified at $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ for all $\{x_t\}_{t\notin\mathcal{I}} \in \prod_{t\notin\mathcal{I}} \mathbb{X}_t(\{\tilde{x}_t\}_{t\in\mathcal{I}})$.*

All objects are identified up to the same relabeling of the groups.

A similar adaptation of Theorem 3.3 is straightforward.

Remark 16. (**Parametric component distributions**) When the component distributions are modeled parametrically as in Section 4, the nonparametric identification result of Theorem A.1.10 may already suffice to identify the parameters of the component distributions.

In particular, borrowing the notation from Section 4 and letting $\mathbb{X}_{j,t}^{(iden)}$ be the collection of values of x_t at which $\mathbb{P}(y_{it} = 1 \mid x_{it} = x_t, g_i = j)$ is identified, then $\beta_{j,t}^{(0)}$ is identified for group j at time t, for instance, when $F_{jt}(\cdot)$ is strictly increasing and $E\left[\mathbb{1}(x_{it} \in \mathbb{X}_{j,t}^{(iden)}) x_{it} x_{it}^\top\right]$ has full rank, that is, the covariates must be linearly independent on $\mathbb{X}_{j,t}^{(iden)}$. When $\pi_j(X) > 0$ for all $j = 1, \ldots, J$ and $X \in \mathbb{X}$, we have $\mathbb{X}_t(\{x_t\}_{t \in \mathcal{I}_2}) \subseteq \mathbb{X}_{j,t}^{(iden)}$ for $t \notin \mathcal{I}_1$ and analogously when $t \notin \mathcal{I}_2$.[10] Hence, the parameters of the component distribution are identified when the initial X does not restrict the support of the covariates to a linear subspace of $\mathbb{X}$.

Clearly, once $\beta_{j,t}^{(0)}$ is identified for all $j = 1, \ldots, J$ and $t = 1, \ldots, T$, the component distributions are identified at all $x_t \in \mathbb{X}_t$ and therefore $\underline{\pi}(\tilde{X})$ is identified at all $\tilde{X} \in \mathbb{X}$ such that there exists a submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$ with $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t \in \mathcal{I}})$ having full column rank.

Remark 17. Similar to Remark 5 in the main text, the identified set is potentially larger than the one described in Theorem A.1.10. We present an algorithm that highlights in which sense the identified set may be larger and how to enlarge the identified set without having problems with the intra-component label switching problem:

Step 0 Given Theorem A.1.10, let $\mathbb{X}_{j,t}^{(iden,0)}$ be the values of x_t at which $\mathbb{P}(y_{it} = 1 \mid x_{it} = x_t, g_i = j)$ is identified and, similarly, let $\mathbb{X}_{\pi}^{(iden,0)}$ be the collection of covariate values $X \in \mathbb{X}$ at which $\underline{\pi}(X)$ is identified.

Step 1 Let X_0 be the covariate value fixed in Theorem A.1.10 and set $s = 1$.

1. **While** there exists $X_s \in \prod_t \cap_j \mathbb{X}_{j,t}^{(iden,s-1)}$ with $X_s \notin \{X_0, \ldots, X_{s-1}\}$ and some submodel $\mathcal{I}$ such that $\mathcal{P}_{\mathcal{I}}(\{x_{s,t}\}_{t \in \mathcal{I}})$ has full column rank, do the following:
   (a) Apply the conclusion of Theorem A.1.10 to X_s and let $\mathbb{X}_{j,t}^{(iden,s)} = \mathbb{X}_{j,t}^{(iden,s-1)} \cup \mathbb{X}_{j,t}^{(iden)}$ as well as $\mathbb{X}_{\pi}^{(iden,s)} = \mathbb{X}_{\pi}^{(iden,s-1)} \cup \mathbb{X}_{\pi}^{(iden)}$ where $\mathbb{X}_{\pi}^{(iden)}$ and $\mathbb{X}_{j,t}^{(iden)}$ are the identified sets using X_s.
   (b) If $\mathbb{X}_{j,t}^{(iden,s)} = \mathbb{X}_{j,t}^{(iden,s-1)}$, $\mathbb{X}_{\pi}^{(iden,s)} = \mathbb{X}_{\pi}^{(iden,s-1)}$ and there does not exist $X_s^* \in \prod_t \cap_j \mathbb{X}_{j,t}^{(iden,s)}$ with $X_s \notin \{X_0, \ldots, X_{s-1}, X_s\}$ such that $\mathcal{P}_{\mathcal{I}}(\{x_{s,t}\}_{t \in \mathcal{I}})$ has full column rank, stop the algorithm. Otherwise, set $s = s + 1$.

Importantly, since $X_s \in \prod_t \cap_j \mathbb{X}_{j,t}^{(iden,s-1)}$ for all s, the component distributions are identified at X_s already in step $s - 1$, which allows us to identify all objects up to the same relabeling across all iterations.

Heuristically speaking, the above iterative procedure may allow us to recover the conclusion of Theorem 3.2 when the covariate values X for which there exists a submodel $\mathcal{I}$ with $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t \in \mathcal{I}})$ having full column are "closely linked", in the sense that exploring all the conditional covariate spaces leads to exploring $\mathbb{X}$.

[10] The set inclusion holds as the assertion of Theorem A.1.10 may be strengthened as discussed in the next remark.

A.1.10 Sufficient condition for $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ to have full rank

To better understand the exact settings when $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank for a given submodel $\mathcal{I}$, we first present a necessary and sufficient condition. Subsequently, we discuss an additional sufficient condition based on the Kruskal ranks of $\mathcal{P}_t(x_t)$ for all $t \in \mathcal{I}$.

We start with the following lemma.

Lemma A.1.11 *$\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank if and only if there does not exist $\gamma \neq 0 \in \mathbb{R}^J$ and $j^* \in \{1,\ldots,J\}$ such that $[\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})]_{\cdot,j^*} = \sum_{j\neq j^*} \frac{\gamma_j}{-\gamma_{j^*}} [\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})]_{\cdot,j}$ with $\sum_{j\neq j^*} \frac{\gamma_j}{-\gamma_{j^*}} = 1$ for all $\mathcal{I}^* \subseteq \mathcal{I}$ if and only if the following $2^{|\mathcal{I}|} \times J$ matrix has full rank*

$$V(\{x_t\}_{t\in\mathcal{I}}) = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \prod_{t\in\mathcal{I}_1^*}[\mathcal{P}_t(x_t)]_{1,1} & \prod_{t\in\mathcal{I}_1^*}[\mathcal{P}_t(x_t)]_{1,2} & \ldots & \prod_{t\in\mathcal{I}_1^*}[\mathcal{P}_t(x_t)]_{1,J} \\ \vdots & \vdots & \ddots & \vdots \\ \prod_{t\in\mathcal{I}_{2^{|\mathcal{I}|}-1}^*}[\mathcal{P}_t(x_t)]_{1,1} & \prod_{t\in\mathcal{I}_{2^{|\mathcal{I}|}-1}^*}[\mathcal{P}_t(x_t)]_{1,2} & \ldots & \prod_{t\in\mathcal{I}_{2^{|\mathcal{I}|}-1}^*}[\mathcal{P}_t(x_t)]_{1,J} \end{pmatrix}$$

where $\{\mathcal{I}_1^*,\ldots,\mathcal{I}_{2^{|\mathcal{I}|}-1}^*\}$ is the power set of $\mathcal{I}$ excluding the empty set, that is the collection of all distinct submodels of $\mathcal{I}$.

Intuitively speaking, $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full rank column rank if no component distribution can be expressed as the same (possibly non-convex) mixture of the other component distributions for all submodels of the submodel $\mathcal{I}$. The proof of Lemma A.1.11 is included in Appendix A.9.1.

Let $\text{krk}(A)$ denote the Kruskal rank of a matrix A. Specifically, the Kruskal rank of A is the largest number k such that any subset of k columns of A are linearly independent. Clearly, $\text{krk}(A) \leq \text{rank}(A)$. A repeated application of Lemma 1 in Sidiropoulos and Bro (2000) gives the following result:

Lemma A.1.12 *Assume that $krk(P_t(x_t)) \geq 1$ for all $t \in \mathcal{I}$, then $rank(\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})) \geq krk(\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})) \geq \min\{\sum_{t\in\mathcal{I}} krk(\mathcal{P}_t(x_t)) - |\mathcal{I}| + 1, J\}$.*

A trivial implication of Lemma A.1.12 is the following:

Corollary A.1.12.1 *Assume there exist $2J-1$ periods such that $\mathcal{P}_t(x_t)$ has distinct columns for all these periods at some $X \in \mathbb{X}$, then Assumption I-2 is satisfied at X.*

The condition in Corollary A.1.12.1 corresponds to the identification condition in Blischke (1964) in the context of mixtures of binomial distributions.

Proof. Due to the fact that all columns are distinct, $\text{krk}(\mathcal{P}_t(x_t)) = 2$ for all $2J-1$ periods. Now, take one period as t' and equally split the remaining periods into the submodels $\mathcal{I}_1$ and $\mathcal{I}_2$. From Lemma A.1.12, we conclude that $\mathcal{P}_{\mathcal{I}_j}(\{x_t\}_{t\in\mathcal{I}_j})$ has full column rank for $j = 1, 2$. As $\mathcal{P}_{t'}(x_{t'})$ has distinct columns by assumption, Assumption I-2 is therefore satisfied at X. □

We conclude this discussion with the following simple corollary that highlights that we do not require $\mathcal{P}_t(x_t)$ to vary over time for $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ to have full column rank.

Corollary A.1.12.2 *Let $\mathcal{I}$ be a submodel such that $|\mathcal{I}| \geq J-1$ and for all $t \in \mathcal{I}$*

$$\mathcal{P}_t(x_t) = \begin{pmatrix} p_1 & p_2 & \dots & p_J \\ 1-p_1 & 1-p_2 & \dots & 1-p_J \end{pmatrix}$$

with $p_j \neq p_{j'}$ for $j \neq j'$. Then, $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank.

The proof follows directly from Lemma A.1.12.

A.1.11 EM algorithm

We recall the empirical log-likelihood over the sieve space

$$\hat{\mathcal{L}}_n^{d(n)}(\beta, \gamma_n) := \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{j=1}^{J} \mathbf{G}_j(X_i;\gamma_n) \prod_{t=1}^{T} F_{jt}\left(x_{it}^\top\beta_{j,t} + \alpha_{j,t}\right)^{s_t}\left(1 - F_{jt}\left(x_{it}^\top\beta_{j,t} + \alpha_{j,t}\right)\right)^{1-s_t}\right) \tag{A.3}$$

For the purpose of the EM algorithm, we fix the sieve dimension $d(n)$ and treat the maximization problem for fixed $d(n)$ as parametric.

To begin with, the complete-data log-likelihood associated with $\hat{\mathcal{L}}_n^{d(n)}(\cdot)$ is

$$\begin{aligned}\hat{\mathcal{L}}_{n,CD}^{d(n)}(\beta,\gamma) = &\sum_{i=1}^{n}\sum_{j=1}^{J} \mathbb{1}(g_i = j)\log(\mathbf{G}_j(X_i;\gamma_n)) \\ &+ \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{J} \mathbb{1}(g_i = j) y_{it}\log(F_{jt}(x_{it}^\top\beta_{j,t} + \alpha_{j,t})) \\ &+ \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{J} \mathbb{1}(g_i = j)(1-y_{it})\log(1 - F_{jt}(x_{it}^\top\beta_{j,t} + \alpha_{j,t}))\end{aligned}$$

which follows from the fact that the complete-data likelihood for unit i is approximated in the sieve space as follows

$$\begin{aligned}&\mathbb{P}(\{y_{it} = s_t\}_{t=1}^T, g_i = j \mid X_i) \\ \approx &\prod_{j=1}^{J}\left(\mathbf{G}_j(X_i;\gamma_n)\prod_{t=1}^{T} F_{jt}\left(x_{it}^\top\beta_{j,t} + \alpha_{j,t}\right)^{s_t}\left(1 - F_{jt}\left(x_{it}^\top\beta_{j,t} + \alpha_{j,t}\right)\right)^{1-s_t}\right)^{\mathbb{1}(g_i=j)}\end{aligned}$$

Given some initial estimates of $\gamma_n^{(0)}$[11] and $\beta^{(0)}$ with $\tilde{\gamma}$ and $\tilde{\beta}$, the conditional complete-data log-likelihood is

$$\begin{aligned} Q(\beta, \gamma_n \mid \tilde{\gamma}, \tilde{\beta}) = & \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{P}(g_i = j \mid X_i, y_i; \tilde{\gamma}, \tilde{\beta}) \log\left(\mathbf{G}_j(X_i; \gamma_n)\right) \\ & + \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbb{P}(g_i = j \mid X_i, y_i; \tilde{\gamma}, \tilde{\beta}) y_{it} \log(F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t})) \\ & + \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbb{P}(g_i = j \mid X_i, y_i; \tilde{\gamma}, \tilde{\beta})(1 - y_{it}) \log(1 - F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t})) \end{aligned}$$

where $y_i = (y_{i1}, \ldots, y_{iT})^\top$ and

$$\mathbb{P}(g_i = j \mid X_i, y_i; \tilde{\gamma}, \tilde{\beta}) = \frac{\mathbf{G}_j(X_i; \tilde{\gamma}) \prod_{t=1}^{T} \left\{ F_{jt}(x_{it}^\top \tilde{\beta}_{j,t} + \tilde{\alpha}_{j,t})^{y_{it}} (1 - F_{jt}(x_{it}^\top \tilde{\beta}_{j,t} + \tilde{\alpha}_{j,t}))^{1-y_{it}} \right\}}{\sum_{j=1}^{J} \mathbf{G}_j(X_i; \tilde{\gamma}) \prod_{t=1}^{T} \left\{ F_{jt}(x_{it}^\top \tilde{\beta}_{j,t} + \tilde{\alpha}_{j,t})^{y_{it}} (1 - F_{jt}(x_{it}^\top \tilde{\beta}_{j,t} + \tilde{\alpha}_{j,t}))^{1-y_{it}} \right\}}$$

The EM algorithm, which is similar to the one discussed in Chapter 11 in Frühwirth-Schnatter et al. (2019), is then Algorithm 1.

Algorithm 1 EM algorithm

0. Initialize $\gamma = \gamma^0$ and $\beta = \beta^0$ and set $s \leftarrow 1$, stop $\leftarrow$ `false`, $s_{\max} \in \mathbb{N}$, and `tol` $\leftarrow \varepsilon$ for $\varepsilon > 0$.

while stop = `false` and $s < s_{\max}$ **do**

1. **E-step**: Calculate $\kappa_{ij}^s = \mathbb{P}(g_i = j \mid X_i, y_i; \gamma^s, \beta^s)$ for $j = 1, \ldots, J$ and $i = 1, \ldots, n$ with

$$\kappa_{ij}^s = \frac{\mathbf{G}_j(X_i; \gamma^s) \prod_{t=1}^{T} \left\{ F_{jt}(x_{it}^\top \beta_{j,t}^s + \alpha_{j,t}^s)^{y_{it}} (1 - F_{jt}(x_{it}^\top \beta_{j,t}^s + \alpha_{j,t}^s))^{1-y_{it}} \right\}}{\sum_{j=1}^{J} \mathbf{G}_j(X_i; \gamma^s) \prod_{t=1}^{T} \left\{ F_{jt}(x_{it}^\top \beta_{j,t}^s + \alpha_{j,t}^s)^{y_{it}} (1 - F_{jt}(x_{it}^\top \beta_{j,t}^s + \alpha_{j,t}^s))^{1-y_{it}} \right\}}$$

2. **M-step**: Maximize $Q(\gamma, \beta \mid \gamma^s, \beta^s)$ with respect to γ and β.
   1. Maximize $Q(\gamma, \beta \mid \gamma^s, \beta^s)$ with respect to γ: Find γ^{s+1} via some numerical optimization scheme.
   2. Maximize $Q(\gamma, \beta \mid \gamma^s, \beta^s)$ with respect to $(\beta_{j,t}^\top, \alpha_{j,t})^\top$ for all $j = 1, \ldots, J$ and $t = 1, \ldots, T$: Find $\beta_{j,t}^{s+1}$ and $\alpha_{j,t}^{s+1}$ via some numerical optimization scheme.
3. **Stopping rule**: Calculate stop_criterion.
   if stop_criterion $<$ `tol` **then**
      stop $\leftarrow$ `true`
   else
      Set $s \leftarrow s + 1$
   end if

end while

4. Set $\hat{\gamma}_n = \gamma^s$ and $\hat{\beta} = \beta^s$.

Various choices of stopping criteria exist. We use an observed-data log-likelihood-based stopping criterion. Specifically, stop_criterion $= \hat{\mathcal{L}}_n^{d(n)}(\gamma^s, \beta^s) - \hat{\mathcal{L}}_n^{d(n)}(\gamma^{s-1}, \beta^{s-1})$. For a

[11] $\gamma_n^{(0)}$ is the pseudo-true parameter that maximizes the population version of (A.3).

detailed discussion of other stopping criteria, we refer to Chapter 2 of McLachlan and Peel (2000).

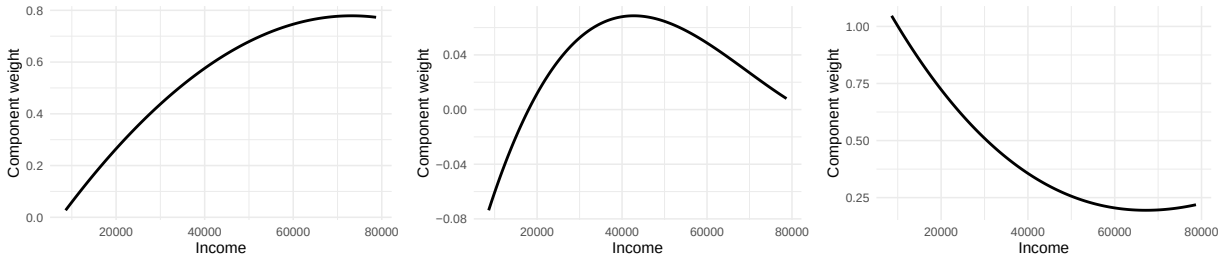
A good starting value is crucial for the EM algorithm to perform well because starting close to the true value "ensures" that the EM algorithm converges to the global maximum and speeds up to the EM algorithm. In empirical practice, it makes therefore sense to try out multiple starting values and choose the final parameter estimate that maximizes the sieve version of the observed-data log-likelihood, $\hat{\mathcal{L}}_n^{d(n)}(\cdot)$.

A.2 Additional empirical results

A.2.1 Marginalized component weights

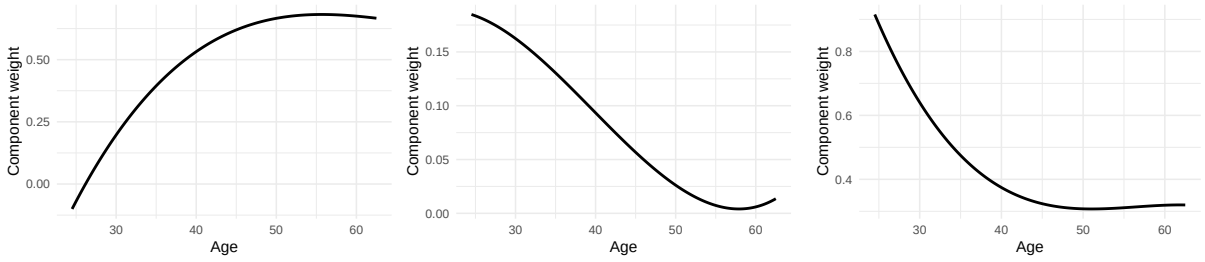
This section presents the marginalized component weights as discussed in Section 6.3. As a basis for the series estimator, we use Legendre polynomials up to order 3. Since Figures A.1 and A.2 are for illustrative purposes, we did not constrain the estimator to map into the unit interval so that the estimated probabilities may fall outside the unit interval in some plots.

Figure A.1: Marginalized component weights for Income



(a) Group 1 (b) Group 2 (c) Group 3

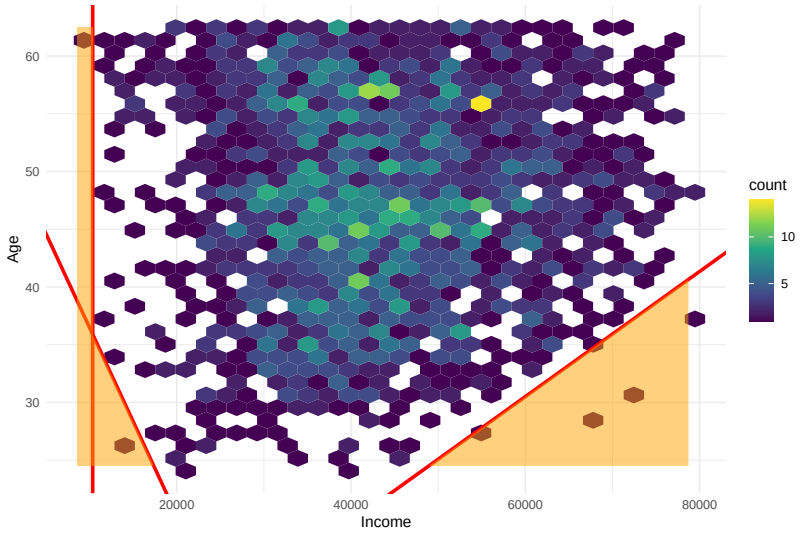
Figure A.2: Marginalized component weights for Age



(a) Group 1 (b) Group 2 (c) Group 3

A.2.2 Joint support of age and income

Figure A.3: Hexagonal binned plot of the joint distribution of the time-average of the net household income and age. The regions in orange are the ones we exclude when plotting the component weights in Figure 1.



A.2.3 Results including all education levels

Here we present the results of the empirical application when we also include households in our analysis in which the household head or partner has an education level that is deemed inadequate according to the *pgisced97* variable in the GSOEP. Table A.1 presents the estimated coefficients, Table A.2 contains the estimated AMEs and APEs, respectively, and Table A.3 contains the time-averaged summary statistics for the different groups. Overall, the results are similar to the ones presented in Section 6.3 in the main text.

Table A.1: Parameter estimates

	Group 1	**Group 2**	**Group 3**
Time-varying group-specific fixed effects			
$t = 1$	-4.180^{*}	-7.585^{***}	-6.538^{**}
$t = 2$	-3.103	-9.974^{***}	-6.977^{**}
$t = 3$	-3.606^{*}	-6.432^{**}	-6.781^{**}
$t = 4$	-3.812^{*}	-5.754^{**}	-7.236^{**}
$t = 5$	-3.839^{*}	-4.863^{*}	-7.537^{**}
$t = 6$	-3.903^{*}	-4.323^{*}	-6.515^{**}
Group-specific slope coefficient			
Net household income	0.244^{***}	0.253^{**}	0.124
Group-invariant slope coefficients			
Age		0.077	
Age-squared		-0.001	
Children		0.071	
Female		0.135	
Long-term unemployed		-0.398^{**}	
Married		0.655^{**}	
Inadequate educ.		0.044	
General-elementary educ.		-0.168	
Tertiary educ.		-0.508^{**}	

Note: Significantly different from 0 at 10%(*), 5%(**), 1%(***) using a two-sided t-test. The critical values are the respective quantiles of the nonparametric bootstrap distribution. We do not include the state-specific dummies for readability purposes.

A.2.4 Cleaning the data – Overview

We use Version 39 of the GSOEP (SOEP 2024). For information how to get access to the GSOEP, we refer to the data access page of the GSOEP. The GSOEP is comprised of multiple data sets. We use the following data sets and variables from the respective data set – for exact details on the variables, we refer to the documentation of the GSOEP, the SOEPcompanion or the GSOEP documentation of paneldata.org.

- **pgen**: preprocessed and generated data on individuals; checked for consistency
  - *pgpartz*: indicator whether cohabitating with partner

Table A.2: Group-specific time-averaged AMEs/APEs

	Group 1	**Group 2**	**Group 3**
Net household income	0.0094^{***}	0.0444^{**}	0.0029
Age	0.0030	0.0136	0.0018
Age-squared	0.0000	−0.0001	0.0000
Children	0.0028	0.0126	0.0016
Female	0.0050	0.0235	0.0033
Long-term unemployed	-0.0202^{*}	-0.0713^{**}	-0.0066^{**}
Married	0.0397^{*}	0.1194^{**}	0.0086^{***}
Inadequate educ.	0.0016	0.0076	0.0011
General-elementary educ.	−0.0065	−0.0294	−0.0041
Tertiary educ.	-0.0243^{*}	-0.0898^{**}	-0.0091^{**}

Note: Significantly different from 0 at 10%(*), 5%(**), 1%(***) using a two-sided t-test. The critical values are the respective quantiles of the nonparametric bootstrap distribution. We do not include the state-specific dummies for readability purposes.

Table A.3: Group-specific time-averaged variables

	Group 1	**Group 2**	**Group 3**
Outcome			
	0.9808	0.4735	0.0097
Uncond. group prob.			
	0.5754	0.0559	0.3686
Covariates			
Net household income	46,439	44,087	37,987
Age	48.85	39.83	45.31
Children	0.5746	0.7892	0.6179
Female	0.3069	0.3976	0.3482
Long-term unemployed	0.0778	0.0632	0.1286
Married	0.9177	0.8847	0.8959
Inadequate educ.	0.0056	0.0066	0.0167
General-elementary educ.	0.5128	0.5184	0.6033
Abi-vocational educ.	0.1791	0.1867	0.1685
Tertiary educ.	0.3024	0.2883	0.2115

  - *pgfamstd*: marital status
  - *pgisced97*: education level
  - *pgemplst*: employment status

- **ppathl**: information on all persons that ever lived in a GSOEP household at the time of survey
  - *gebjahr*: year of birth
  - *sex*: sex/gender
- **pbrutto**: contains general information on individuals
  - *stell_h*: relationship to household head

- **hgen**: contains generated data on the households
  - *hgowner*: tenant or owner of dwelling
  - *hghinc*: current monthly household net income
- **hbrutto**: contains general information on households
  - *bula_h*: federal state household is located in
- **kidlong**: contains data on children based on the information collected in annual waves
  - *year_of_birth*: year of birth of the respective child

All data sets are keyed by *hid* (household ID), *syear* (survey year), and, if available, *pid* (personal ID).

To deflate relevant incomes, we use the CPI as published by the German Federal Statistical Office (accessible here).

Next, we describe how we generated the covariates that are relevant for our analysis.

- *marital_status*: Based on the variable on *pgfamstd*, we construct a binary marital status variable. *marital_status* equals 1 when $pgfamstd$ equals married (but separated), husband/wife abroad, or registered same-sex partnership (living apart), and *marital_status* equals 0 when an individual is single, divorced, or widowed.
- **Education dummies**: Based on *pgisced97*, we construct the following education dummies:
  - *general_elementary*: general elementary, and middle vocational
  - *vocational*: Abi + vocational, and higher vocational
  - *tertiary*: higher education
- **Emplyoment status**: Based on *pgemplst*, we construct two dummies for employment and unemployment (the dummmies take value `NA` whenever *pgemplst* takes negative values, that is, when *pgemplst* is missing).
  - *ft_employed*: full-time employment
  - *unemployed*: not employed
- *long_term_unemployed*: Based on the previously constructed *unemployed* variable, we construct a binary variable for long-term unemployment. *long_term_unemployed* takes value 1 if an individual has been unemployed at least in the current and previous survey year.
- *age*: We calculate the age of an individual by taking the difference of *syear* and *gebjahr*.
- **Relationship to household head**: Using *stell_h*, we construct a binary variable *hhead_partner_ind* indicating whether an individual is the household head (1) or the

partner of the household head (0) – here a partner is a spouse, same-sex partner or life partner as indicated by *stell_h*. For all other values of *stell_h*, *hhead_partner_ind* takes the value `NA`. For information on *stell_h*, we refer to the description of stell_h on paneldata.org.

- **Homeownership**: Based on the variable *hgowner*, we construct variable *howner* that is equal to 1 whenever an individual is an owner as indicated by *hgowner*, equal to 0 whenever an individual is a (main or sub-) tenant or living in a shared accomodation as indicated by *hgowner*. *howner* takes negative values when *hgowner* takes negative values/is missing.
- **Number of children**: Based on the *year_of_birth* variable in **kidlong**, we determine the age of a child and save it in the *age_child* variable. This is done after some data cleaning. Specifically, we impute missing birth years in some years whenever they are available in other years for the same individual. In case different years of birth are reported, we take the smallest birth year. Based on *age_child*, we construct a dummy *children_ind* that is equal 1 if a child under or equal to the age of 17 is recorded for the household in **kidlong** and 0 otherwise.[12]
- *yearly_hh_labor_income*: We deflate *hghinc* to real 2015 Euros using the CPI by the German Federal Statistical Office and calculate *yearly_hh_labor_income* by multiplying the resulting value with 12.

Next, we discuss how we subset the data set. In particular, we only include individuals in the age bracket 21 to 65 and focus on households with one household head and one partner of the household head as indicated by *hhead_partner_ind*. We drop all observations (household head or partner) for which at least one of the following holds:

- *yearly_hh_labor_income* is missing.
- *pgisced97* is missing or individuals are still in school or have an inadequate education level according to *pgisced97*.
- *sex* is missing.
- *marital_status* is missing.
- *unemployed* or *long_term_unemployed* is missing.
- *bula_h* is missing.
- *age_child* takes on negative values or is larger than 100.

That is, we drop all observations from our analysis for which the dependent variable or any of the covariates is missing for either the household head or her partner. Also, to ensure

[12]The actual construction of the dummy in the code is a bit more complicated but this complexity is not relevant to the analysis at hand.

that our theoretical support conditions on the covariates hold, we exclude households with a deflated yearly net household income of more than 85,000 euros from our analysis. This value approximately corresponds to the 96th percentile of the deflated annual net household income distribution from 1991 to 2022 of households for which we observe the covariates and outcome variable. In a final step, we balance the panel.

A.3 Proofs for Section 3.2 and Appendix A.1.3

Throughout this section, we use the additional notation introduced in Appendix A.1.3.

A.3.1 A helpful lemma

The first lemma is Proposition 2.1 in Leurgans et al. (1993). The proof directly follows from the definition of the Moore-Penrose inverse.

Lemma A.3.1 *Let A and B be two matrices of full column rank and D a non-singular square matrix, then $(ADB^\top)^\dagger = (B^\dagger)^\top D^{-1} A^\dagger$ with $A^\dagger = (A^\top A)^{-1} A^\top$ and $B^\dagger = (B^\top B)^{-1} B^\top$.*

A.3.2 Proof of Lemma 3.1

Proof. It is easy to see that $\mathcal{P}_{t'}(x_{t'})$ has distinct columns if and only if $\mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j) \neq \mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j')$ for all $j \neq j'$. We therefore need to show that the set $\mathcal{E} := \{\exists j \neq j' : \mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j) = \mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j')\}$ has Lebesgue measure 0. To this end, we rewrite $\mathcal{E} = \cup_{j \neq j'} \{\mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j) = \mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j')\}$ and note that $(\mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = 1), \ldots, \mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = J))^\top \in \mathbb{R}^J$. Thus, for all $j \neq j'$, the set $\{\mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j) = \mathbb{P}_{t'}(y_{it'} = 1 \mid x_{it'} = x_{t'}, g_i = j')\}$ is $(J-1)$-dimensional and therefore has Lebesgue measure 0. Combing this with Boole's inequality gives $\lambda(\mathcal{E}) = 0$ where $\lambda(A)$ denotes the Lebesgue measure of some measurable set A. This concludes the proof. □

A.3.3 Proof of Lemma A.1.1

Proof. Sufficiency:

1. If $\text{rank}(\mathbf{P}_{\mathcal{I}_1 \cup \mathcal{I}_2}(X)) = J$, then, by definition, $\text{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t \in \mathcal{I}_1}) \Pi(X) \mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t \in \mathcal{I}_2})^\top) = J$. From standard arguments, we know $J = \text{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t \in \mathcal{I}_1}) \Pi(X) \mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t \in \mathcal{I}_2})^\top) \leq \min\{\text{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t \in \mathcal{I}_1})), \text{rank}(\Pi(X)), \text{rank}(\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t \in \mathcal{I}_2}))\} \leq J$ where the final inequality follows from the dimensions of the matrices. Since all matrices are at most of rank J, it follows that $\text{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t \in \mathcal{I}_1}) = \text{rank}(\Pi(X)) = \text{rank}(\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t \in \mathcal{I}_2})) = J$, that is, all matrices have full column rank.

2. Given the previous argument, all required matrices have full column rank so that

$$\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\mathcal{I}_\ell})^\dagger = (\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\mathcal{I}_\ell})^\top \mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\mathcal{I}_\ell}))^{-1}\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\mathcal{I}_\ell})^\top \text{ for } \ell = 1,2$$

and by Lemma A.3.1

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger = (\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\dagger)^\top \Pi(X)^{-1}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger$$

It follows

$$\begin{aligned}&\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\mathcal{D}_{t',1}(x_{t'})\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger\\ \Rightarrow\ &\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1}) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\mathcal{D}_{t',1}(x_{t'})\end{aligned}$$

Thus, $[\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$ is up to permutation equal to the first J eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ sorted in a decreasing order.[13] If the algebraic multiplicity of the first $J-1$ non-zero eigenvalues is 1, $[\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$ has distinct entries so that the columns of $\mathcal{P}_{t'}(x_{t'})$ are distinct.[14]

Necessity:

1. From the same arguments as in the sufficiency part, we have $\text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) \leq J$. At the same time, since $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ and $\Pi(X)$ have full rank, $\text{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)) = J$. Thus, Sylvester's rank inequality gives $\text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) \geq J$ where we use that $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$ has full column rank. It follows that $\text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) = J$.
2. Because $\mathcal{P}_{t'}(x_{t'})$ has distinct columns, no two entries in $[\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$ are the same. Otherwise, as $[\mathcal{P}_{t'}(x_{t'})]_{2,\cdot} = 1 - [\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$, it would be the case that two columns are identical. We note that this condition allows one entry in $[\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$ to be equal to 0. Now, following the identical arguments as in the sufficiency part and recalling that $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ has at most J non-zero eigenvalues, we know that $[\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$ is up to permutation equal to the first J eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ sorted in a decreasing order. Hence, it follows that $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ has at least $J-1$ non-zero eigenvalues of algebraic multiplicity 1.

□

[13] We note that $\text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger) \leq J$ so that $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ has at most J non-zero eigenvalues. By requiring that at least $J-1$ non-zero eigenvalues have algebraic multiplicity 1, we allow $\mathbb{P}_{t'}(y_{it'} = 0 \mid x_{it} = x_{t'}, g_i = j) = 0$ for some j.

[14] We can focus on $[\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$ here without loss of generality since $[\mathcal{P}_{t'}(x_{t'})]_{2,\cdot} = 1 - [\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$.

A.3.4 Proof of Theorem 3.2

Proof. We recall from Appendix A.1.3

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top \tag{A.4}$$

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{D}_{t',k}(x_{t'})\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top \tag{A.5}$$

The proof proceeds in multiple steps. For readability, we use $\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{I}} \mid X_i = X)$ and $\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{I}} \mid X)$ interchangeably in the following.

Step 1: Identification of $\mathcal{P}_{t'}(x_{t'})$ and $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$. From Lemma A.3.1, $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger = (\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\dagger)^\top\Pi(X)^{-1}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger$ so that for $k=1,2$

$$\begin{aligned}&\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\mathcal{D}_{t',k}(x_{t'})\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger\\ \Rightarrow\ &\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1}) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\mathcal{D}_{t',k}(x_{t'})\end{aligned}$$

Hence, the diagonal entries of $\mathcal{D}_{t',1}(x_{t'})$ are identified up to permutation as the first J eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ sorted in a decreasing order.[15] Since $\mathcal{D}_{t',2}(x_{t'}) = I_{J\times J} - \mathcal{D}_{t',1}(x_{t'})$ with $I_{J\times J}$ being the $J\times J$ identity matrix, $\mathcal{D}_{t',2}(x_{t'})$ is identified up to the same permutation and thus the columns of $\mathcal{P}_{t'}(x_{t'})$ are identified up to the same permutation. For the identification of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$, we perform a case analysis:

1. *All diagonal entries of $\mathcal{D}_{t',1}(x_{t'})$ are positive*: Since all diagonal entries of $\mathcal{D}_{t',1}(x_{t'})$ are distinct, the columns of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ are identified as the eigenvectors of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ associated with the eigenvalues $[\mathcal{P}_{t'}(x_{t'})]_{1,\cdot}$ up to the same permutation as $\mathcal{D}_{t',1}(x_{t'})$ is identified. Since $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ contains non-negative entries only and all columns sum up to 1, the scaling of the respective eigenvector is uniquely pinned down.
2. *One diagonal element of $\mathcal{D}_{t',1}(x_{t'})$ is equal to* 0: As in the previous case, all columns of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ associated with positive diagonal entries of $\mathcal{D}_{t',1}(x_{t'})$ are identified up to the same permutation as the eigenvectors associated with the positive eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},1}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ where the scaling is pinned down by the properties of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$. The group for which the diagonal element in $\mathcal{D}_{t',1}(x_{t'})$ equals 0 has value 1 in $\mathcal{D}_{t',2}(x_{t'})$. Thus, the remaining column of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ still to be identified is the eigenvector associated with the eigenvalue 1 of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$.[16] Hence, all columns of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ are again identified up to the same permutation as $\mathcal{D}_{t',k}(x_{t'})$ for $k=1,2$. Clearly, a symmetric argument applies if one diagonal element

[15] We sort the eigenvalues in a decreasing order since $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)^\dagger$ has at least $2^{|\mathcal{I}_1|}-J$ and at most $2^{|\mathcal{I}_1|}-J+1$ zero-valued eigenvalues.

[16] Since the diagonal entries of $\mathcal{D}_{t',2}(x_{t'})$ are distinct, the eigenvalue 1 has algebraic multiplicity 1.

of $\mathcal{D}_{t',2}(x_{t'})$ is equal to 0.

It follows that the columns of $\mathcal{P}_{t'}(x_{t'})$ and $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ are identified up to the same relabeling of the groups.

Step 2: Identification of $\underline{\pi}(X)$ (1). We define $\mathbf{P}_{\mathcal{I}_1}(X) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\underline{\pi}(X) = \{\sum_{j=1}^{J}\pi_j(X)\prod_{t\in\mathcal{I}_1}\mathbb{P}_t(y_{it}=s_t \mid x_{it}=x_t, g_i=j)\}_{\{s_t\}_{t\in\mathcal{I}_1}\in\{0,1\}^{|\mathcal{I}_1|}} = \{\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{I}_1} \mid X)\}_{\{s_t\}_{t\in\mathcal{I}_1}\in\{0,1\}^{|\mathcal{I}_1|}} \in \mathbb{R}^{2^{|\mathcal{I}_1|}}$, which is identified. Then

$$\underline{\pi}(X) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^{\dagger}\mathbf{P}_{\mathcal{I}_1}(X)$$

Assuming the relabeling issue away for a second, $\underline{\pi}(X)$ is identified because the RHS is identified. More so, as $\mathbf{P}_{\mathcal{I}_1}(X)$ is identified for all X, $\underline{\pi}(X)$ is identified at $\{x_t\}_{t\in\mathcal{I}_1}$ for all $\{x_t\}_{t\notin\mathcal{I}_1}$. It remains to be argued that the relabeling of $\underline{\pi}(X)$ agrees with the relabeling of the groups in $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$. To see this, let Δ denote any permutation matrix of the columns of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$. Recalling that $\Delta^{-1}=\Delta^{\top}$, it follows that $(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Delta)^{\dagger} = \Delta^{\top}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^{\dagger}$ so that $(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Delta)^{\dagger}\mathbf{P}_{\mathcal{I}_1}(X) = \Delta^{\top}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^{\dagger}\mathbf{P}_{\mathcal{I}_1}(X)$. Thus, $\underline{\pi}(X)$ is identified up to the same group relabeling as $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$.

Step 3: Identification of $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$. For equation (A.4), we have

$$\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^{\top} = \Pi(X)^{-1}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^{\dagger}\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)$$

where the RHS is identified. To argue that the group relabeling is consistent, we note that for any permutation matrix Δ of the columns of $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$, $\Delta^{\top}\Pi(X)\Delta$ is the relabelled/permuted version of $\Pi(X)$. Thus, $(\Delta^{\top}\Pi(X)\Delta)^{-1}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Delta)^{\dagger}\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X) = \Delta^{\top}\Pi(X)^{-1}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^{\dagger}\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)$ so that $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$ is identified up to the same group relabeling as $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$.

Step 4: Identification of $\underline{\pi}(X)$ (2). Combining *Step 2* and *Step 3*, it follows $\underline{\pi}(X)$ is identified at $\{x_t\}_{t\in\mathcal{I}_2}$ for all $\{x_t\}_{t\notin\mathcal{I}_2}$ up to the same group relabeling as $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$.

Step 5: Identification of $\mathcal{P}_t(x_t)$ for all t. We consider two cases

1. $t\notin\mathcal{I}_2$: We consider the submodel $\mathcal{I}_2\cup\{t\}$, that is,

$$\Big\{\mathbb{P}(\{y_{it'}=s_{t'}\}_{t'\in\mathcal{I}_2\cup\{t\}} \mid X)\Big\}_{\{s_{t'}\}_{t'\in\mathcal{I}_2\cup\{t\}}\in\{0,1\}^{|\mathcal{I}_2|+1}}$$
$$= \mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X) = \mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})\Pi(X)\mathcal{P}_t(x_t)^{\top}$$

where the LHS is identified for all $X\in\mathbb{X}$. Now, Assumptions I-2 and I-3 imply

$$\mathcal{P}_t(x_t)^{\top} = \Pi(X)^{-1}\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^{\dagger}\mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X) \tag{A.6}$$

Hence, given the same arguments as in *Step 3*, $\mathcal{P}_t(x_t)$ is identified up to the same group relabeling as $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$.

To strengthen the identification result, we first discuss a simplified version to avoid convoluting the argument through heavy notation. To this end, we fix $\{x_t\}_{t\in\mathcal{I}_2}$, then $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\dagger$ is identified and well-defined, while $\Pi(X)$ and $\mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X)$ are identified at $\{x_t\}_{t\in\mathcal{I}_2}$ for all $\{x_t\}_{t\notin\mathcal{I}_2}$ – the former follows from *Step 4*. Given equation (A.6) and that $t\notin\mathcal{I}_2$, $\mathcal{P}_t(\tilde{x}_t)$ is therefore identified at $\tilde{x}_t\in\mathbb{X}_t$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}}$ such that $\Pi(\{x_t\}_{t\in\mathcal{I}_2},\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}},\tilde{x}_t)$ has full rank, which is equivalent to $\pi_j(\{x_t\}_{t\in\mathcal{I}_2},\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}},\tilde{x}_t)>0$ for all $j=1,\dots,J$.

Now, we strengthen the identification result. Focusing on identification of $\mathbb{P}_t(y_{it}=s_t\mid x_{it}=\tilde{x}_t,g_i=j^*)$ for $s_t\in\{0,1\}$, we assume that there exists $\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}}$ such that $\pi_{j^*}(\{x_t\}_{t\in\mathcal{I}_2},\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}},\tilde{x}_t)>0$. We define $\mathbf{J}=\{j\in\{1,\dots,J\}:\pi_j(\{x_t\}_{t\in\mathcal{I}_2},\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}},\tilde{x}_t)>0\}$, which is the set of groups with positive component weight. Without loss, we assume that the elements of $\mathbf{J}$ are ordered increasingly. $\mathbf{J}$ is identified from *Step 4*. Next, we let $[\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})]_{\cdot,j\in\mathbf{J}}$ and $[\mathcal{P}_t(x_t)]_{\cdot,j\in\mathbf{J}}$ denote the collection of the columns of the respective matrices with column index $j\in\mathbf{J}$ and use $[\Pi(X)]_{j\in\mathbf{J},j\in\mathbf{J}}$ to denote the $|\mathbf{J}|\times|\mathbf{J}|$ submatrix of $\Pi(X)$ that only contains the component weights of the groups in $\mathbf{J}$ on the diagonal.[17] Since $\mathbf{J}$ is identified, $[\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})]_{\cdot,j\in\mathbf{J}}$ and $[\Pi(X)]_{j\in\mathbf{J},j\in\mathbf{J}}$ are identified. Furthermore, $[\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})]_{\cdot,j\in\mathbf{J}}$ has full column rank and $[\Pi(X)]_{j\in\mathbf{J},j\in\mathbf{J}}$ is invertible. Last, we notice

$$\begin{aligned}\mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X)&=\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})\Pi(X)\mathcal{P}_t(x_t)^\top\\&=[\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})]_{\cdot,j\in\mathbf{J}}[\Pi(X)]_{j\in\mathbf{J},j\in\mathbf{J}}[\mathcal{P}_t(x_t)]^\top_{\cdot,j\in\mathbf{J}}\end{aligned}\tag{A.7}$$

Now, our previous arguments imply that $[\mathcal{P}_t(\tilde{x}_t)]_{\cdot,j\in\mathbf{J}}$ is identified at $\tilde{x}_t$. To conclude, since $j^*\in\mathbf{J}$, this argument implies that $\mathbb{P}_t(y_{it}=s_t\mid x_{it}=\tilde{x}_t,g_i=j^*)$ is identified for all $s_t\in\{0,1\}$ at $\tilde{x}_t\in\mathbb{X}_t$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}}$ such that $\pi_{j^*}(\{x_t\}_{t\in\mathcal{I}_2},\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_2\cup\{t\}},\tilde{x}_t)>0$.

2. $t\notin\mathcal{I}_1$: A symmetric argument using the submodel $\mathcal{I}_1\cup\{t\}$ with $\mathbf{P}_{\mathcal{I}_1\cup\{t\}}(X)=\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{P}_t(x_t)^\top$ implies that for all $j\in\{1,\dots,J\}$ and $s_t\in\{0,1\}$, $\mathbb{P}_t(y_{it}=s_t\mid x_{it}=\tilde{x}_t,g_i=j)$ is identified at $\tilde{x}_t\in\mathbb{X}_t$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_1\cup\{t\}}$ such that $\pi_j(\{x_t\}_{t\in\mathcal{I}_1},\{x^*_{t^*}\}_{t^*\notin\mathcal{I}_1\cup\{t\}},\tilde{x}_t)>0$.

[17] More precisely, let $\mathbf{J}=\{j_1,\dots,j_\ell\}$ for $\ell\le J$ and $j_\ell<j_{\ell'}$ for $\ell<\ell'$. Then,

$$\begin{aligned}[\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})]_{\cdot,j\in\mathbf{J}}&=\begin{pmatrix}[\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})]_{\cdot,j_1}&\dots&[\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})]_{\cdot,j_\ell}\end{pmatrix}\\[\mathcal{P}_t(x_t)]_{\cdot,j\in\mathbf{J}}&=\begin{pmatrix}[\mathcal{P}_t(x_t)]_{\cdot,j_1}&\dots&[\mathcal{P}_t(x_t)]_{\cdot,j_\ell}\end{pmatrix}\\[\Pi(X)]_{j\in\mathbf{J},j\in\mathbf{J}}&=\operatorname{diag}((\pi_{j_1}(X),\dots,\pi_{j_\ell}(X))).\end{aligned}$$

Step 6: Identification of $\underline{\pi}(X)$ (3). For any $\mathcal{I}$ and $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ has full column rank and is identified, we have for all $\{x_t\}_{t\notin\mathcal{I}}$

$$\underline{\pi}(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{x_t\}_{t\notin\mathcal{I}}) = \mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})^{\dagger}\mathbf{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{x_t\}_{t\notin\mathcal{I}})$$

Since $\mathbf{P}_{\mathcal{I}}(X)$ is identified for all $X\in\mathbb{X}$, we conclude that for any $\mathcal{I}$ and $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ has full column rank and is identified, $\underline{\pi}(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{x_t\}_{t\notin\mathcal{I}})$ is identified at $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ for all $\{x_t\}_{t\notin\mathcal{I}}$.

This concludes the proof of the theorem. However, as mentioned in Remark 5, we may strengthen the identification result. We now discuss how to do so:

Step 7: Iterative identification argument. We fix some time period t^* and consider any submodel $\mathcal{I}$ and $\{\tilde{x}\}_{t\in\mathcal{I}}$ such that $t^*\notin\mathcal{I}$ and $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ has full column rank and is identified; clearly, this is the case if and only if $\mathcal{P}_t(\tilde{x}_t)$ is identified for all $t\in\mathcal{I}$.[18] Hence, the previous arguments give a set of the values of $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ is identified. Once this set is established, it is immediate to check whether $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ has full column rank. Let $\mathcal{X}_{\mathcal{I}}^{(1)} = \{\{\tilde{x}_t\}_{t\in\mathcal{I}} : \mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ has full column rank and is identified based on *Step 5*$\}$. Now, for all $\{\tilde{x}_t\}_{t\in\mathcal{I}}\in\mathcal{X}_{\mathcal{I}}^{(1)}$, $\underline{\pi}(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{x_t\}_{t\notin\mathcal{I}})$ is identified for all $\{x_t\}_{t\notin\mathcal{I}}$ by *Step 6*. From the arguments in *Step 5*, we know that for all $\{\tilde{x}_t\}_{t\in\mathcal{I}}\in\mathcal{X}_{\mathcal{I}}^{(1)}$ and all $\{\tilde{x}_t\}_{t\notin\mathcal{I}}$, we have

$$\mathcal{P}_{t^*}(\tilde{x}_t)^{\top} = \Pi(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{\tilde{x}_t\}_{t\notin\mathcal{I}})^{-1}\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})^{\dagger}\mathbf{P}_{\mathcal{I}\cup\{t^*\}}(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{\tilde{x}_t\}_{t\notin\mathcal{I}}) \tag{A.8}$$

as long as $\Pi(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{\tilde{x}_t\}_{t\notin\mathcal{I}})$ has full rank – alternatively, one can follow the arguments for the subcollection $\mathbf{J}$ of groups with positive component weights as in *Step 5*. Then, $\mathcal{P}_{t^*}(\tilde{x}_{t^*})$ is identified as the RHS of equation (A.8) is identified. Thus, we can (weakly) strengthen the previous identification result as follows:

- For all $t^*=1,\ldots,T$, $\mathbb{P}_{t^*}(y_{it^*}=s_{t^*}\mid x_{it^*}=\tilde{x}_{t^*},g_i=j)$ is identified for all $s_{t^*}\in\{0,1\}$ at all $\tilde{x}_{t^*}\in\mathbb{X}_{t^*}$ for which there exists a submodel $\mathcal{I}$ with $t^*\notin\mathcal{I}$ and $\{\tilde{x}_t\}_{t\in\mathcal{I}}\in\mathcal{X}_{\mathcal{I}}^{(1)}$ such that there exists $\{\tilde{x}_t\}_{t\notin\mathcal{I}\cup\{t^*\}}$ with $\pi_j(\{\tilde{x}_t\}_{t\in\mathcal{I}},\{\tilde{x}_t\}_{t\notin\mathcal{I}\cup\{t^*\}},\tilde{x}_{t^*})>0$.

Different from the statement in Theorem 3.2, we allow for an arbitrary submodel $\mathcal{I}$ as long as it satisfies the desired assumptions. More importantly, however, we do not fix the value of $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ at the initially fixed $\{x_t\}_{t\in\mathcal{I}}$ of Theorem 3.2 that satisfies Assumptions I-2 and I-3 but allow for values of $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ such that equation (A.8) holds and the RHS of this equation is identified and well-defined.

Now, the argument from *Step 6* applies again to the additionally identified values/submodels, which can then be used to work through this *Step 7* again based on $\mathcal{X}_{\mathcal{I}}^{(2)} = \{\{\tilde{x}_t\}_{t\in\mathcal{I}}$:

[18] This follows from $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}}) =^{\text{col}}\otimes \mathcal{P}_t(\tilde{x}_t)$ and the fact that summing over the respective rows of $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ gives $\mathcal{P}_t(\tilde{x}_t)$.

$\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\mathcal{I}})$ has full column rank and is identified based on *Step 7*} for some submodel $\mathcal{I}$. This analysis may now be performed iteratively. As of now, we have do not have a characterization of the final set of the values $\tilde{x}_t$ at which $\mathcal{P}_t(\tilde{x}_t)$ is identified for $t = 1, \ldots, T$.

This concludes the proof. □

A.3.5 Proof of Corollary 3.2.1

Proof. The corollary follows immediately from Theorem 3.2.

□

A.3.6 Proof of Lemma A.1.2

Proof. Using the notation from Appendix A.1.3, we note

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top;\ \mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X}) = \mathcal{P}_{\mathcal{I}_1}(\{\tilde{x}_t\}_{t\in\mathcal{I}_1})\Pi(\tilde{X})\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top$$
$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X}) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(\underline{X})\mathcal{P}_{\mathcal{I}_2}(\{\underline{x}_t\}_{t\in\mathcal{I}_2})^\top;\ \mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}}) = \mathcal{P}_{\mathcal{I}_1}(\{\tilde{x}_t\}_{t\in\mathcal{I}_1})\Pi(\underline{\tilde{X}})\mathcal{P}_{\mathcal{I}_2}(\{\underline{x}_t\}_{t\in\mathcal{I}_2})^\top$$

1. $\Rightarrow$: Under Assumption I-4, Lemma A.3.1 implies

$$\begin{aligned}
&\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger \\
&= \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger \\
&\Rightarrow \mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1}) \\
&= \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}
\end{aligned}$$

   The above is an eigendecomposition of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger$. As the diagonal of $\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}$ contains J distinct and non-zero values, $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger$ has J distinct non-zero eigenvalues – all remaining eigenvalues are clearly equal to 0.
   Next, the fact that $\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}$ is well-defined and has non-zero diagonal entries implies that $\Pi(X)$, $\Pi(\tilde{X})$, $\Pi(\underline{X})$, and $\Pi(\underline{\tilde{X}})$ have full rank. Combining this observation with the full column rank conditions on the component distributions in Assumption I-4 and the arguments in the necessity part of the proof of Lemma A.1.1, we conclude that $\text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) = \text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})) = \text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})) = \text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})) = J$.
2. $\Leftarrow$: If $\text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) = \text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})) = \text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})) = \text{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})) = J$, then similar arguments as in the sufficiency proof of Lemma A.1.1 imply that $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$, $\mathcal{P}_{\mathcal{I}_1}(\{\tilde{x}_t\}_{t\in\mathcal{I}_1})$, $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$, and $\mathcal{P}_{\mathcal{I}_2}(\{\underline{x}_t\}_{t\in\mathcal{I}_2})$ have full column rank and that $\Pi(X)$, $\Pi(\tilde{X})$, $\Pi(\underline{X})$, and $\Pi(\underline{\tilde{X}})$ have full rank. Thus, Lemma A.3.1 implies

that $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger$ is defined as above. Therefore, the diagonal of the $J\times J$ matrix $\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}$ contains (up to permutation) the J non-zero eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger$. As these eigenvalues are assumed to be distinct, the diagonal entries of $\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}$ are distinct and non-zero.

This concludes the proof. □

A.3.7 Proof of Theorem 3.3

Proof. Following the arguments in the proof of Lemma A.1.2

$$\begin{aligned}
&\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger\\
&= \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})^\dagger\\
&\Rightarrow \mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\\
&= \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}
\end{aligned}$$

where $\Pi(X)\Pi(\tilde{X})^{-1}\Pi(\underline{\tilde{X}})\Pi(\underline{X})^{-1}$ is a diagonal matrix with distinct and positive entries, that is, the J non-zero eigenvalues of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger$ are all distinct. Hence, $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ is identified up to permutation as the eigenvectors of $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\tilde{X})^\dagger\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{\tilde{X}})\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(\underline{X})^\dagger$. The scaling and sign of the respective eigenvectors are pinned down by the fact that $\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})$ contains only non-negative entries and its columns must add to 1. *Step 2* from the proof of Theorem 3.2 implies that up to identical permutation of the groups $\underline{\pi}(X)$ is identified at $\{x_t\}_{t\in\mathcal{I}_1}$ for all $\{x_t\}_{t\notin\mathcal{I}_1}$. Upon noticing that under Assumption I-4 $\Pi(\tilde{X})$, $\Pi(X)$, and $\Pi(\underline{X})$ have full rank, the arguments of *Step 3* in the proof of Theorem 3.2 imply that $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})$ and $\mathcal{P}_{\mathcal{I}_2}(\{\underline{x}_t\}_{t\in\mathcal{I}_2})$ are identified up to identical group permutation. This result itself can be leveraged to show that $\mathcal{P}_{\mathcal{I}_1}(\{\tilde{x}_t\}_{t\in\mathcal{I}_1})$ is identified up to identical permutation of the groups. Now, the remainder of the proof follows from *Step 2* to *Step 7* of the proof of Theorem 3.2. □

A.4 Additional discussions and proofs for Section 4

We start this appendix with a few additional discussions. We begin with a formal definition of the sieve space: For $j = 1, \ldots, J-1$

$$\mathcal{W}_{j,d(n)} = \left\{w \in \mathcal{W}_j : w(x) = \gamma_{j,n}^\top \rho_j^{d(n)}(X) \text{ with } \gamma_{j,n} \in \mathbb{R}^{d(n)+1}\right\}$$

where $\mathcal{W}_j$ is as defined in Assumption E-6. Then, $\mathcal{W}_{d(n)} = \mathcal{W}_{1,d(n)} \times \cdots \times \mathcal{W}_{J-1,d(n)}$.

Throughout the proofs, we use the following immediate results:

Fact 1 As $\mathcal{B}$ and $\mathbb{X}$ are assumed to be compact, there exists some $C < \infty$ such that $\sup_{x\in\mathbb{X}_t,(\beta^\top,\alpha)^\top\in\mathcal{B}} |x^\top\beta + \alpha| \leq C$.

Fact 2 As $F_{jt}(\cdot)$ is strictly monotone on $\mathbb{R}$, Fact 1 implies that there exist $\underline{f} \in (0,1)$ and $\overline{f} \in (0,1)$ such that $\underline{f} \leq F_{jt}(\cdot) \leq \overline{f}$ for all j and t.

Fact 3 Since $F_{jt}(\cdot)$ is assumed to be continuously differentiable and strictly monotone on $\mathbb{R}$, the compactness of $\mathbb{X}$ and $\mathcal{B}$ implies that there exits $C_F < \infty$ such that $|F'_{jt}(x^\top\beta+\alpha)| \leq C_F$ uniformly over $\mathcal{B}$ and $\mathbb{X}$ for all j and t where $F'_{jt}(x) = \frac{\partial}{\partial x}F_{jt}(x)$.

Fact 4 Since $\pi_j^{(0)}(X)$ is assumed to be continuous, the compactness of $\mathbb{X}$ in combination with Assumption E-2 implies that there exist $\overline{p} \in (0,1)$ and $\underline{p} \in (0,1)$ such that $\underline{p} \leq \pi_j(X) \leq \overline{p}$ for all $X \in \mathbb{X}$ and $j = 1,\ldots,J$.

Fact 5 : By compactness of $\mathbb{X}$ there exists $C_X < \infty$ such that $\|x_{it}\|_E < C_X$ for all $t = 1,\ldots,T$.

Additionally, we use the following notation

$$L_j(X;w) = \begin{cases} \frac{\exp(w_j(X))}{1+\sum_{j=1}^{J-1}\exp(w_j(X))} & \text{for } j = 1,\ldots,J-1 \\ \frac{1}{1+\sum_{j=1}^{J-1}\exp(w_j(X))} & \text{for } j = J \end{cases}$$

$$\ell(y,X;w,\beta) = \log\left(\sum_{j=1}^{J} L_j(X;w)\prod_{t=1}^{T} F_{jt}(x_t^\top\beta_{j,t}+\alpha_{j,t})^{y_t}(1-F_{jt}(x_t^\top\beta_{j,t}+\alpha_{j,t}))^{1-y_t}\right)$$

with $y = \{y_t\}_{t=1}^T \in \{0,1\}^T$. We abbreviate *with probability approaching 1* with wpa 1. For notational simplicity, we assume that $\rho_j^{d(n)}(\cdot) = \rho_{j'}^{d(n)}(\cdot)$ for all j, j' in this section. The proofs and assumptions can be straightforwardly adapted to the case where different component weights are approximated using different sets of basis functions.

All results in this Appendix hold up to the joint relabeling of the groups. To keep notation light, we suppress this in our arguments. For estimation purposes, this is also practically irrelevant because the optimization routine will simply pick one of the permutations of the groups. In a subsequent step, the groups have to be interpreted economically, which pins them down from an economic standpoint. This latter step is not affected by the generic label switching problem.

A.4.1 Consistency

A.4.1.1 Proof of Theorem 4.1

Proof. First, we prove that $\|\hat{\theta}_{w,n} - \theta_w^{(0)}\|_{c,\infty} = o_p(1)$ by checking the conditions of Lemma A1 in Newey and Powell (2003) similar to Hu and Schennach (2008). In a subsequent step, we argue that this implies $\|\hat{\theta}_n - \theta^{(0)}\|_{c,\infty} = o_p(1)$.

1. $\mathcal{L}(\theta_w^{(0)}) > -\infty$: By Jensen's inequality

$$\begin{aligned} T^{-1}\mathcal{L}(\theta_w^{(0)}) \geq T^{-1}\,\mathbb{E}\Bigg[& \sum_{j=1}^{J}\sum_{t=1}^{T} \pi_j^{(0)}(X_i)\Big\{ y_{it}\log(F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})) \\ & + (1-y_{it})\log(1 - F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))\Big\}\Bigg] \\ \geq\ \mathbb{E}\Bigg[& \min_{j,t}\Big\{ y_{it}\log(F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})) \\ & + (1-y_{it})\log(1 - F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))\Big\}\Bigg] \\ >\ & -\infty \end{aligned}$$

   where the last inequality follows from Fact 2 above.

2. $\mathcal{L}(\theta_w)$ has a unique maximum on $\Theta^{(w)}$ at $\theta_w^{(0)}$: From standard maximum likelihood arguments, $\mathcal{L}(\theta_w)$ has a unique maximum on $\Theta^{(w)}$ at $\theta_w^{(0)}$ when for each $\theta_w' \in \Theta^{(w)}$, $\mathbb{P}[\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X_i; w', \beta') = \mathbb{P}(\{y_{it}\}_{t=1}^T \mid X_i; w^{(0)}, \beta^{(0)})] = 1$ implies that $\theta_w' = \theta_w^{(0)}$ or, more precisely, $\|\theta_w' - \theta_w^{(0)}\|_{c,\infty} = 0$. Since $\mathbb{P}(\{y_t\}_{t=1}^T \mid X; w^{(0)}, \beta^{(0)})$ and $\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X_i; w', \beta')$ are continuous in X for all $\{y_t\}_{t=1}^T \in \{0,1\}^T$, and $\mathbb{P}(\{y_t\}_{t=1}^T \mid X; w^{(0)}, \beta^{(0)}) > 0$ for all $\{y_t\}_{t=1}^T$ and X, $\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X; w', \beta') = \mathbb{P}(\{y_{it}\}_{t=1}^T \mid X; w^{(0)}, \beta^{(0)})$ almost surely implies that $\mathbb{P}(\{y_t\}_{t=1}^T \mid X; w', \beta') = \mathbb{P}(\{y_t\}_{t=1}^T \mid X; w^{(0)}, \beta^{(0)})$ for all $X \in \mathbb{X}$ and $\{y_t\}_{t=1}^T \in \{0,1\}^T$.[19] Then, under Assumptions E-2, E-4, and E-5, the assertion of Corollary 3.2.1 or Corollary 3.3.1 applies for both $\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X; w', \beta')$ and $\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X; w^{(0)}, \beta^{(0)})$, that is, the component distributions, say $\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j; \beta^{(0)})$ and $\mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j; \beta')$, are identified for all $j = 1, \ldots, J$ and $x_t \in \mathbb{X}_t$, and the component weights, say $\pi_j^{(0)}(X)$ and $\pi_j'(X)$, are identified for all $X \in \mathcal{X}$ and $j = 1, \ldots, J$. From Assumption E-6, we know that $w_j^{(0)}(X)$ and $w_j'(X)$ are continuous in X, which under Assumption E-2 implies that $\pi_j^{(0)}(X)$ and $\pi_j'(X)$ are continuous so we conclude that $\pi_j^{(0)}(X)$ and $\pi_j'(X)$ are identified for all $X \in \mathbb{X}$ and $j = 1, \ldots, J$. As the component weights and distributions are derived from $\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X; w^{(0)}, \beta^{(0)}) = \mathbb{P}(\{y_t\}_{t=1}^T \mid X; w', \beta')$, they are identical for the two parameter specifications $(w^{(0)}, \beta^{(0)})$ and (w', β'). Hence, we can conclude that $\pi_j^{(0)}(X) = \pi_j'(X)$ for all $j = 1, \ldots, J$ and $X \in \mathbb{X}$. Therefore, $w_j'(X) = \log(\pi_j'(X)/\pi_J'(X)) = \log(\pi_j^{(0)}(X)/\pi_J^{(0)}(X)) = w_j^{(0)}(X)$ for all $X \in \mathbb{X}$ and $j = 1, \ldots, J$. Similarly, we have for the component distributions $\mathbb{P}(y_{it} = 1 \mid x_{it}, g_i = j; \beta^{(0)}) = F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}) = F_{jt}(x_t^\top \beta_{j,t}' + \alpha_{j,t}') = \mathbb{P}(y_{it} = 1 \mid x_{it}, g_i = j; \beta')$ for all

[19] Notice that a similar argument continues to hold even in the presence of discrete regressors.

$x_t \in \mathbb{X}_t$. Then, under Assumption E-4(iii)

$$(\alpha'_{j,t}, \beta'^{\top}_{j,t})^{\top} = (\alpha^{(0)}_{j,t}, \beta^{(0)\top}_{j,t})^{\top} = \mathbb{E}[\tilde{x}_{it}\tilde{x}_{it}^{\top}]^{-1} E\left[\tilde{x}_{it} F_{jt}^{-1}\big(\mathbb{P}(y_{it} = 1 \mid x_{it}, g_i = j)\big)\right]$$

We note that this argument can be extended to a case with discrete covariates. We conclude that $w' = w^{(0)}$ and $\beta' = \beta^{(0)}$, that is, $\|\theta'_w - \theta^{(0)}_w\|_{c,\infty} = 0$. Thus, $\mathcal{L}(\theta_w)$ is uniquely maximized at $\theta^{(0)}_w$.

3. $\Theta^{(w)}$ is compact: $\Theta^{(w)}$ is indeed compact under $\|\cdot\|_{c,\infty}$; see, for instance, Theorems 1 and 2 in Freyberger and Masten (2019) as well as discussions therein.
4. $\hat{\mathcal{L}}_n(\theta_w)$ and $\mathcal{L}(\theta_w)$ are continuous under $\|\cdot\|_{c,\infty}$: Recall

$$\ell(y, X; w, \beta) = \log\left(\sum_{j=1}^{J} L_j(X; w) \prod_{t=1}^{T} F_{jt}(x_t^{\top}\beta_{j,t} + \alpha_{j,t})^{y_t}(1 - F_{jt}(x_t^{\top}\beta_{j,t} + \alpha_{j,t}))^{1-y_t}\right)$$

For any $\theta_{w,1} = (\beta_1^{\top}, w_1) \in \Theta^{(w)}$, $\theta_{w,2} = (\beta_2^{\top}, w_2) \in \Theta^{(w)}$ and arbitrary $X \in \mathbb{X}$ as well as $y = \{y_t\}_{t=1}^{T} \in \{0,1\}^T$ a mean value expansion gives

$$\ell(y, X; w_1, \beta_1) = \ell(y, X; w_2, \beta_2) + \frac{\partial}{\partial \tau}\ell(y, X; \overline{w} + \tau(w_1 - w_2), \overline{\beta} + \tau(\beta_1 - \beta_2))\Big|_{\tau=0}$$

where $\overline{w} = w_2 + \overline{\tau}(w_1 - w_2) \in \mathcal{W}$ and $\overline{\beta} = \beta_2 + \overline{\tau}(\beta_1 - \beta_2) \in \mathcal{B}$ are intermediate values for some $\overline{\tau} \in (0,1)$.[20] Particularly, we have $\sum_{j=1}^{J} L_j(X; \overline{w}) = 1$ and $0 < L_j(X; \overline{w}) < 1$. Letting $\tilde{\beta}_{jt} = (\alpha_{j,t}, \beta_{j,t}^{\top})^{\top}$ and $\tilde{x}_{it} = (1, \tilde{x}_{it}^{\top})^{\top}$, some calculus yields

$$\begin{aligned}
&\ell(y, X; w_1, \beta_1) - \ell(y, X; w_2, \beta_2) \\
&= \frac{1}{\sum_{j=1}^{J} L_j(X; \overline{w}) \prod_{t=1}^{T} F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t})^{y_t}(1 - F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t}))^{1-y_t}} \\
&\quad \Bigg\{ \sum_{j=1}^{J-1} L_j(X; \overline{w})(w_{1,j}(X) - w_{2,j}(X)) \Bigg[\prod_{t=1}^{T} F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t})^{y_t}(1 - F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t}))^{1-y_t} \\
&\quad - \sum_{j=1}^{J} L_j(X; \overline{w}) \prod_{t=1}^{T} F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t})^{y_t}(1 - F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t}))^{1-y_t} \Bigg] \\
&\quad + \sum_{j=1}^{J} L_j(X; \overline{w}) \sum_{t^*=1}^{T} \prod_{t \neq t^*} F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t})^{y_t}(1 - F_{jt}(x_t^{\top}\overline{\beta}_{j,t} + \overline{\alpha}_{j,t}))^{1-y_t} \\
&\quad \times (\mathbb{1}(y_{t^*} = 1) - \mathbb{1}(y_{t^*} = 0)) F'_{jt^*}(x_{t^*}^{\top}\overline{\beta}_{j,t^*} + \overline{\alpha}_{j,t^*}) \tilde{x}_{t^*}^{\top}(\tilde{\beta}_{1,j,t^*} - \tilde{\beta}_{2,j,t^*}) \Bigg\}
\end{aligned}$$

where $F'_{jt}(x) = \frac{\partial}{\partial x} F_{jt}(x)$. Thus

$$|\ell(y, X; w_1, \beta_1) - \ell(y, X; w_2, \beta_2)|$$

[20] $\mathcal{W}$ is convex and it is without loss to assume that $\mathcal{B}$ is convex. Assume $\mathcal{B}$ is not convex, then we may redefine the parameter space to be the closed convex hull of $\mathcal{B}$, which is compact – see Theorem 5.35 in Aliprantis and Border (2006).

$$
\begin{aligned}
&\leq \frac{\sum_{j=1}^{J-1} \|w_{1,j} - w_{2,j}\|_\infty + C_F C_X \sum_{t^*=1}^{T} \sum_{j=1}^{J} \|\tilde{\beta}_{1,j,t^*} - \tilde{\beta}_{2,j,t^*}\|_E}{\min_j \left\{ \prod_{t=1}^{T} F_{jt}(x_t^\top \overline{\beta}_{j,t} + \overline{\alpha}_{j,t})^{y_t} (1 - F_{jt}(x_t^\top \overline{\beta}_{j,t} + \overline{\alpha}_{j,t}))^{1-y_t} \right\}} \\
&\leq \frac{(1 + C_F C_X) \|\theta_{w,1} - \theta_{w,2}\|_{c,\infty}}{\inf_{\beta \in \mathcal{B}} \min_j \left\{ \prod_{t=1}^{T} F_{jt}(x_t^\top \beta_{j,t} + \alpha_{j,t})^{y_t} (1 - F_{jt}(x_t^\top \beta_{j,t} + \alpha_{j,t}))^{1-y_t} \right\}} \\
&\leq \frac{1}{\min\{\underline{f}, 1 - \overline{f}\}^T} (1 + C_F C_X) \|\theta_{w,1} - \theta_{w,2}\|_{c,\infty}
\end{aligned}
$$

where the last two inequalities leverage Fact 2 from above. We now conclude that $\mathcal{L}(\theta_w)$ and $\hat{\mathcal{L}}_n(\theta_w)$ are Lipschitz-continuous in θ_w.

5. $\sup_{\theta_w \in \Theta^{(w)}} |\hat{\mathcal{L}}_n(\theta_w) - \mathcal{L}(\theta_w)| = o_p(1)$: We check the conditions of Lemma A2 in Newey and Powell (2003). $\Theta^{(w)}$ is compact under $\|\cdot\|_{c,\infty}$ and by the previous argument condition (iii) is trivially satisfied.[21] It remains to be shown that $\hat{\mathcal{L}}_n(\theta_w) \xrightarrow{p} \mathcal{L}(\theta_w)$ for all $\theta_w \in \Theta^{(w)}$. From a similar mean-value argument as in the previous part,

$$
\begin{aligned}
&\sup_{\theta_w \in \Theta^{(w)}} |\ell(y_i, X_i; w, \beta)| \\
&\leq |\ell(y_i, X_i; w^{(0)}, \beta^{(0)})| \\
&\quad + \left| \frac{1 + C_F C_X}{\inf_{\beta \in \mathcal{B}} \min_j \left\{ \prod_{t=1}^{T} F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t})^{y_t} (1 - F_{jt}(x_{it}^\top \beta_{j,t} + \alpha_{j,t}))^{1-y_t} \right\}} \right| 2 \sup_{\theta_w \in \Theta^{(w)}} \|\theta_w\|_{c,\infty} \\
&\leq |\ell(y_i, X_i; w^{(0)}, \beta^{(0)})| + \frac{1 + C_F C_X}{\min\{\underline{f}, 1 - \overline{f}\}^T} 2C
\end{aligned}
$$

where the second inequality uses that $\Theta^{(w)}$ is compact under $\|\cdot\|_{c,\infty}$ so there exists $C < \infty$ such that $\sup_{\theta_w \in \Theta^{(w)}} \|\theta_w\|_{c,\infty} \leq C$. We conclude that for all $\theta_w \in \Theta^{(w)}$, $\mathbb{E}[|\ell(y_i, X_i; w, \beta)|] \leq \mathbb{E}[\sup_{\theta_w \in \Theta^{(w)}} |\ell(y_i, X_i; w, \beta)|] < \infty$.[22] Thus, a standard WLLN readily implies that $\hat{\mathcal{L}}_n(\theta_w) \xrightarrow{p} \mathcal{L}(\theta_w)$ for all $\theta_w \in \Theta^{(w)}$ and Lemma A2 of Newey and Powell (2003) applies and implies $\sup_{\theta_w \in \Theta^{(w)}} |\hat{\mathcal{L}}_n(\theta_w) - \mathcal{L}(\theta_w)| = o_p(1)$.

6. For $d \geq 1$, $\Theta_d^{(w)} \subseteq \Theta_{d+1}^{(w)} \cdots \subseteq \Theta^{(w)}$ are compact: The inclusion follows by definition of the sieve spaces. Since $\mathcal{B}$ does not change and is compact, it remains to argue that $\mathcal{W}_d$ is compact under $\|\cdot\|_{c,\infty}$ for all $d \geq 1$. Since $\mathcal{W}_d$ is a subset of a $\|\cdot\|_{c,\infty}$-compact set, it is totally bounded under $\|\cdot\|_{c,\infty}$. Additionally, $\mathcal{W}_d$ is $\|\cdot\|_{c,\infty}$-closed so that $\mathcal{W}_d$ is $\|\cdot\|_{c,\infty}$-compact.
7. For $\theta_w^{(0)} \in \Theta^{(w)}$, there exists $\theta_{w,d(n)} \in \Theta_{d(n)}^{(w)}$ such that $\|\theta_{w,d(n)} - \theta_w^{(0)}\|_{c,\infty} \to 0$:[23] This

[21] Condition (iii) requires that for all $\theta_w, \tilde{\theta}_w \in \Theta^{(w)}$ $|\hat{\mathcal{L}}_n(\theta_w) - \hat{\mathcal{L}}_n(\tilde{\theta}_w)| \leq B_n \|\theta_w - \tilde{\theta}_w\|_{c,\infty}^v$ for some $v > 0$ and $B_n = O_p(1)$. Choose $v = 1$ and B_n as in the previous step, which is finite and thus $B_n = O_p(1)$.

[22] $\mathbb{E}[|\ell(y_i, X_i; w^{(0)}, \beta^{(0)})|] < \infty$ by the first point in this proof and the observation that $\ell(y, X; w^{(0)}, \beta^{(0)}) \leq 0$.

[23] The approximation condition in Newey and Powell (2003) is only required to hold for $\theta^{(0)}$; see, for

follows directly from Assumption E-7.[24]

All conditions of Lemma A1 in Newey and Powell (2003) are satisfied. We conclude $\|\hat{\theta}_{w,n} - \theta_w^{(0)}\|_{c,\infty} = o_p(1)$.

We proceed to argue that $\|\hat{\theta}_n - \theta^{(0)}\|_{c,\infty} = o_p(1)$. To this end, we observe that for $j = 1, \dots, J-1$

$$\hat{\pi}_j(X) = \frac{\exp(\hat{w}_j(X))}{1 + \sum_{j=1}^{J-1} \exp(\hat{w}_j(X))} \; ; \quad \pi_j^{(0)}(X) = \frac{\exp(w_j^{(0)}(X))}{1 + \sum_{j=1}^{J-1} \exp(w_j^{(0)}(X))}$$

At all $X \in \mathbb{X}$, a mean-value expansion yields

$$\begin{aligned}\hat{\pi}_j(X) = \pi_j^{(0)}(X) &+ L_j(X; \tilde{w}(X))(\hat{w}_j(X) - w_j^{(0)}(X)) \\ &- L_j(X; \tilde{w}(X)) \sum_{j^*=1}^{J-1} L_{j^*}(X; \tilde{w}(X))(\hat{w}_{j^*}(X) - w_{j^*}^{(0)}(X))\end{aligned}$$

$\tilde{w}(X) = w^{(0)}(X) + \tilde{\tau}(\hat{w}_n(X) - w^{(0)}(X))$ for some $\tilde{\tau} \in (0,1)$. It follows that uniformly over $\mathbb{X}$, $|\hat{\pi}_j(X) - \pi_j^{(0)}(X)| \leq \sum_{j=1}^{J-1} |\hat{w}_j(X) - w_j^{(0)}(X)|$ where we use that $L_j(X; \tilde{w}) \in (0,1)$. Since the supremum over a sum is smaller than or equal to the sum of the suprema, it follows $\|\hat{\pi}_j - \pi_j^{(0)}\|_\infty \leq \sum_{j=1}^{J-1} \|\hat{w}_j - w_j^{(0)}\|_\infty = o_p(1)$. Since the chosen j was arbitrary and $J-1$ is finite, we conclude that $\sum_{j=1}^{J-1} \|\hat{\pi}_j - \pi_j^{(0)}\|_\infty = o_p(1)$. Noting that $\|\hat{\theta}_{w,n} - \theta_w^{(0)}\|_{c,\infty} = o_p(1)$ only if $\|\hat{\beta}_n - \beta^{(0)}\|_E = o_p(1)$, we conclude $\|\hat{\theta}_n - \theta^{(0)}\|_{c,\infty} = o_p(1)$. We note that $\sum_{j=1}^{J-1} \|\hat{\pi}_j - \pi_j^{(0)}\|_\infty = o_p(1)$ implies that $\|\hat{\pi}_J - \pi_J^{(0)}\|_\infty = o_p(1)$, too. This concludes the proof.

□

A.4.1.2 Proof of Corollary 4.1.1

Proof. Under the assumptions, $\text{AME}_{j,t,k}$ is identified and exists for all j, t, k. For an arbitrary j, t, k-tuple, we recall

$$\widehat{\text{AME}}_{j,t,k} = \frac{\hat{\beta}_{j,t,k}}{\hat{\pi}_j} \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_j(X_i) F'(x_{it}^\top \hat{\beta}_{j,t} + \hat{\alpha}_{j,t})$$

where $\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^n \hat{\pi}_j(X_i)$ with population analog $\pi_j = \mathbb{P}(g_i = j) = \mathbb{E}[\pi_j^{(0)}(X_i)]$.

First, we argue that $\hat{\pi}_j \xrightarrow{p} \pi_j^{(0)}$. To this end, rewrite

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \pi_j^{(0)}(X_i) + \frac{1}{n} \sum_{i=1}^{n} (\hat{\pi}_j(X_i) - \pi_j^{(0)}(X_i))$$

instance, Condition 3.2 for Theorem 3.1 in Chen (2007) or Proposition 10 in Freyberger and Masten (2019).

[24] Since $\beta^{(0)} \in \mathcal{B}$, it remains to argue that there exists $w_{d(n)} \in \mathcal{W}_{d(n)}$ such that $\sum_{j=1}^{J-1} \|w_{j,d(n)} - w_j^{(0)}\|_\infty \to 0$, which is Assumption E-7.

Under Assumption E-1, a standard WLLN implies that $\frac{1}{n}\sum_{i=1}^n \pi_j^{(0)}(X_i) \xrightarrow{p} \mathbb{E}[\pi_j^{(0)}(X_i)] = \pi_j^{(0)}$. At the same time, the conclusion of Theorem 4.1 gives

$$\left|\frac{1}{n}\sum_{i=1}^n(\hat{\pi}_j(X_i) - \pi_j^{(0)}(X_i))\right| \leq \|\hat{\pi}_j - \pi_j^{(0)}\|_\infty = o_p(1)$$

We therefore conclude that $\hat{\pi}_j \xrightarrow{p} \pi_j^{(0)} \in [\underline{p}, \overline{p}]$.

Next, we rewrite

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{\pi}_j(X_i)F'(x_{it}^\top\hat{\beta}_{j,t} + \hat{\alpha}_{j,t}) = &\underbrace{\frac{1}{n}\sum_{i=1}^n \pi_j^{(0)}(X_i)F'(x_{it}^\top\beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})}_{(I)} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n(\hat{\pi}_j(X_i) - \pi_j^{(0)}(X_i))F'(x_{it}^\top\hat{\beta}_{j,t} + \hat{\alpha}_{j,t})}_{(II)} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n \pi_j^{(0)}(X_i)(F'(x_{it}^\top\hat{\beta}_{j,t} + \hat{\alpha}_{j,t}) - F'(x_{it}^\top\beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))}_{(III)}\end{aligned}$$

From Fact 3, $\mathbb{E}[|\pi_j^{(0)}(X_i)F'(x_{it}^\top\beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})|] \leq C_F$ so that a standard WLLN implies

$$(I) \xrightarrow{p} E\left[\pi_j^{(0)}(X_i)F'(x_{it}^\top\beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})\right]$$

Next, we turn to (III). As (III) is continuous in $\hat{\beta}_{j,t}$ and $\hat{\alpha}_{j,t}$ and $\mathbb{E}[\sup_{\beta\in\mathcal{B}}|\pi_j^{(0)}(X_i)(F'(x_{it}^\top\beta_{j,t} + \alpha_{j,t}) - F'(x_{it}^\top\beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))|] \leq 2C_F$ from Fact 3, Lemma 4.3 in Newey and McFadden (1994) applies and implies that $(III) = o_p(1)$.

It remains to be shown that $(II) = o_p(1)$. To this end, we observe $|(II)| \leq C_F\|\hat{\pi} - \pi^{(0)}\|_\infty = o_p(1)$ where the final equality follows from Theorem 4.1.

Combining all results with the continuous mapping theorem and the fact that $\hat{\beta}_{j,t,k} = \beta_{j,t,k}^{(0)} + o_p(1)$ from Theorem 4.1, we conclude $\widehat{\text{AME}}_{j,t,k} = \text{AME}_{j,t,k} + o_p(1)$. Noting that the j,t,k-tuple was arbitrary concludes the proof. □

A.4.2 Convergence rates

A.4.2.1 Fisher norm

We observe that $\mathcal{W}$ is convex and therefore connected. Without loss of generality we may also assume that $\mathcal{B}$ is convex and thus connected, too.[25] Hence, $\Theta^{(w)}$ is convex (and connected). Following Ai and Chen (2003) and Hu and Schennach (2008), we define the

[25] Assume $\mathcal{B}$ is not convex, then we may redefine the parameter space to be the closed convex hull of $\mathcal{B}$, which is compact – see Theorem 5.35 in Aliprantis and Border (2006)

pathwise derivative of $\ell(\cdot)$ in the direction $[\theta_w - \theta_w^{(0)}]$ at $\theta_w^{(0)}$,

$$\frac{d\ell(y, X; \theta_w^{(0)})}{d\theta_w}[\theta_{w,1} - \theta_w^{(0)}] = \left.\frac{\partial \ell(y, X; \theta_w^{(0)} + \tau(\theta_{w,1} - \theta_w^{(0)}))}{\partial \tau}\right|_{\tau=0}$$

which is linear. In our setting

$$\begin{aligned}
&\frac{d\ell(y, X; \theta_w^{(0)})}{d\theta_w}[\theta_{w,1} - \theta_w^{(0)}] \\
&= \frac{1}{\sum_{j=1}^J \pi_j^{(0)}(X) \prod_{t=1}^T F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})^{y_t}(1 - F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))^{1-y_t}} \\
&\left\{ \sum_{j=1}^{J-1} \pi_j^{(0)}(X)(w_{1,j}(X) - w_j^{(0)}(X)) \left[\prod_{t=1}^T F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})^{y_t}(1 - F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))^{1-y_t} \right.\right. \\
&\left. - \sum_{j=1}^J \pi_j^{(0)}(X) \prod_{t=1}^T F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})^{y_t}(1 - F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))^{1-y_t} \right] \\
&+ \sum_{j=1}^J \pi_j^{(0)}(X) \sum_{t^*=1}^T \prod_{t \neq t^*} F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})^{y_t}(1 - F_{jt}(x_t^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))^{1-y_t} \\
&\left. \times (\mathbb{1}(y_{t^*} = 1) - \mathbb{1}(y_{t^*} = 0)) F'_{jt^*}(x_{t^*}^\top \beta_{j,t^*}^{(0)} + \alpha_{j,t^*}^{(0)}) \tilde{x}_{t^*}^\top (\tilde{\beta}_{1,j,t^*} - \tilde{\beta}_{j,t^*}^{(0)}) \right\}
\end{aligned}$$

Following Ai and Chen (2003), we define the following local Fisher norm that is induced by the objective function :[26]

$$\|\theta_{w,1} - \theta_w^{(0)}\|^2 = -\mathbb{E}\left[\left. \frac{\partial^2 \ell(y_i, X_i; \theta_w^{(0)} + \tau(\theta_{w,1} - \theta_w^{(0)}))}{\partial \tau^2} \right|_{\tau=0} \right] = \mathbb{E}\left[\left(\frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[\theta_{w,1} - \theta_w^{(0)}] \right)^2 \right]$$

where, under the assumptions to follow, the second equality follows from the Fisher information identity. Given our previous derivation of the pathwise derivative, we conclude

$$\begin{aligned}
\|\theta_{w,1} - \theta_w^{(0)}\|^2 = \mathbb{E}\Bigg[&\frac{1}{\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X_i; \theta_w^{(0)})^2} \left\{ \sum_{j=1}^{J-1} \pi_j^{(0)}(X_i)(w_{1,j}(X_i) - w_j^{(0)}(X_i)) \right. \\
&\times \left[\prod_{t=1}^T F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})^{y_{it}}(1 - F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))^{1-y_{it}} - \mathbb{P}(\{y_{it}\}_{t=1}^T \mid X_i; \theta_w^{(0)}) \right] \\
&+ \sum_{j=1}^J \pi_j^{(0)}(X_i) \sum_{t^*=1}^T \prod_{t \neq t^*} F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})^{y_{it}}(1 - F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))^{1-y_{it}} \\
&\left. \left. \times (\mathbb{1}(y_{it^*} = 1) - \mathbb{1}(y_{it^*} = 0)) F'_{jt^*}(x_{it^*}^\top \beta_{j,t^*}^{(0)} + \alpha_{j,t^*}^{(0)}) \tilde{x}_{it^*}^\top (\tilde{\beta}_{1,j,t^*} - \tilde{\beta}_{j,t^*}^{(0)}) \right\}^2 \right]
\end{aligned}$$

[26] To be correct, the Fisher norm is a pseudo-norm. Nevertheless, we shall write Fisher norm in the following.

where

$$\mathbb{P}(\{y_{it}\}_{t=1}^T \mid X_i; \theta_w^{(0)}) = \sum_{j=1}^J \pi_j^{(0)}(X_i) \prod_{t=1}^T F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)})^{y_{it}} (1 - F_{jt}(x_{it}^\top \beta_{j,t}^{(0)} + \alpha_{j,t}^{(0)}))^{1-y_{it}}$$

A.4.2.2 Fisher norm is weaker than L_2-norm and $\|\cdot\|_{c,\infty}$

For $\theta_w = (\beta^\top, w) \in \mathcal{B} \times \mathcal{W}$, we let $\|\theta_w\|_{c,L_2(P_0)} = \sum_{j=1}^J \sum_{t=1}^T \|\tilde{\beta}_{j,t}\|_E + \sum_{j=1}^{J-1} \|w_j\|_{L_2(P_0)}$ where P_0 is the distribution of $(y_i^\top, X_i^\top)^\top$ and $\|w_j\|_{L_2(P_0)} = \sqrt{\mathbb{E}[w_j(X_i)^2]}$. Using similar arguments as in the proof of Theorem 4.1, we find

$$\|\theta_{w,1} - \theta_w^{(0)}\| \leq \frac{\sqrt{2}\max\{\sqrt{J-1}, C_X C_F \sqrt{T}\}}{\min\{\underline{f}, 1-\overline{f}\}^T} \|\theta_{w,1} - \theta_w^{(0)}\|_{c,L_2(P_0)}$$

At the same time, $\|\theta_{w,1} - \theta_w^{(0)}\|_{L_2(P_0)} \leq \|\theta_{w,1} - \theta_w^{(0)}\|_{c,\infty}$. We therefore conclude that the Fisher norm is weaker than both the L_2-norm $\|\cdot\|_{c,L_2(P_0)}$ and our consistency norm $\|\cdot\|_{c,\infty}$.

A.4.2.3 Actual rate derivation

Let $C_0 = \sqrt{TJ}\sup_{\beta\in\mathcal{B}}\|\beta\|_E + (J-1)\max\{|\log(\underline{p}/\overline{p}) - \chi|, \log(\overline{p}/\underline{p}) + \chi\}$ for some $\chi > 0$ and define the two local parameter spaces

$$\Theta_0^{(w)} = \left\{\theta_w \in \Theta^{(w)} : \|\theta_w - \theta_w^{(0)}\|_{c,\infty} = o(1), \|\theta_w\|_{c,\infty} \leq C_0\right\}$$

and $\Theta_{0n}^{(w)} = \left\{\theta_w \in \Theta_{d(n)}^{(w)} : \|\theta_w - \theta_w^{(0)}\|_{c,\infty} = o(1), \|\theta_w\|_{c,\infty} \leq C_0\right\}$. We note that $\sup_{\beta\in\mathcal{B}}\|\beta\|_E < \infty$ by compactness of $\mathcal{B}$ and $\hat{\theta}_{w,n} \in \Theta_{0n}^{(w)}$ wpa 1. To see in particular that $\|\hat{\theta}_{w,n}\|_{c,\infty} \leq C_0$ wpa 1, we first observe that $\sum_{j=1}^J \sum_{t=1}^T \|\tilde{\beta}_{j,t}\|_E \leq \sqrt{TJ}\|\beta\|_E$ so that $\sum_{j=1}^J \sum_{t=1}^T \|(\hat{\beta}_{j,t}^\top, \hat{\alpha}_{j,t})^\top\|_E \leq \sqrt{TJ}\sup_{\beta\in\mathcal{B}}\|\beta\|_E$. From Fact 4, $\sup_{X\in\mathbb{X}}|w_j^{(0)}(X)| \leq \max\{|\log(\underline{p}/\overline{p})|, \log(\overline{p}/\underline{p})\}$, which combined with $\|\hat{w}_j(X) - w_j^{(0)}(X)\|_{c,\infty}$ implies that $\sup_{X\in\mathbb{X}}|\hat{w}_j(X)| \leq \max\{|\log(\underline{p}/\overline{p}) - \chi|, \log(\overline{p}/\underline{p}) + \chi\}$ for all $\chi > 0$ wpa 1. Hence, $\|\hat{\theta}_{w,n}\|_{c,\infty} \leq C_0$ wpa 1.

For $\kappa > 0$, $N(\kappa, \Theta, \|\cdot\|)$ is the κ-covering number of Θ under $\|\cdot\|$. We make the following additional assumptions.

Assumption A-1 *(Fisher norm)* $F_{jt} : \mathbb{R} \subseteq [0,1]$ *is twice continuously differentiable for all* j, t.

Assumption A-2 *(Approximation rates) There is* $\theta_{w,d(n)} = (\beta_{d(n)}^\top, w_{d(n)}) \in \Theta_{d(n)}^{(w)}$ *such that* $\|\theta_{w,d(n)} - \theta^{(0)}\| = o(n^{-1/4})$.

Assumption A-3 *(Norm equivalence) There exist* $c_1, c_2 > 0$ *such that*

$$c_1\left[\mathcal{L}(\theta_w^{(0)}) - \mathcal{L}(\theta_w)\right] \leq \frac{1}{2}\left\|\theta_w - \theta_w^{(0)}\right\|^2 \leq c_2\left[\mathcal{L}(\theta_w^{(0)}) - \mathcal{L}(\theta_w)\right]$$

for all $\theta_w \in \Theta_{0n}^{(w)}$.

Assumption A-4 *(Bounded smallest eigenvalue) The smallest eigenvalue of* $\mathbb{E}[\rho^{d(n)}(X_i)\rho^{d(n)}(X_i)^\top]$ *is bounded away from zero uniformly over* $d(n)$.

Assumption A-5 *(Variance control)* $d(n) \times \log\left(\max\left\{\sqrt{\mathbb{E}[\|\rho^{d(n)}(X_i)\|_E^2]}, n\right\}\right) \times n^{-1/2} = o(1)$.

Assumption A-1 allows us to define the local Fisher norm induced by the objective functions and is satisfied in a probit or logit setting.

Assumption A-2 is standard in the literature (Ai and Chen 2003; Ai and Chen 2007). Upon choosing $\beta_{d(n)} = \beta^{(0)}$ in Assumption A-2, it is clear that the rate requirement is an assumption on the log-odds ratios. Since the Fisher norm is weaker than $\|\cdot\|_{c,\infty}$, approximation results for functions in Hölder classes with smoothness $\eta > 0$ of the kind $\|w_{j,d(n)} - w_j^{(0)}\|_\infty = O(d(n)^{-\eta/d_X})$ for all $j = 1, \ldots, J-1$ (see, for instance, Section 2.3.1 in Chen (2007) and references therein) may be used to conclude that $\|w_{j,d(n)} - w_j^{(0)}\| = O(d(n)^{-\eta/d_X})$ for all $j = 1, \ldots, J-1$. We, however, stress that Assumption A-2 requires the approximation rate only in the weaker Fisher norm, not under the sup-norm. Hence, the rate under $\|\cdot\|$ may indeed be much faster than the rate under the sup-norm.

Assumption A-3 can be verified by arguing that the remainder term of a second-order approximation of the objective function is negligible under the Fisher norm. A sufficient condition for Assumption A-3 is that there exist $\nu > 0$ and $C(y, X)$ with $\mathbb{E}[|C(y_i, X_i)|] < \infty$ such that for all $\tilde{\tau} \in (0, 1)$

$$\begin{aligned} &\left|\frac{\partial^2}{\partial\tau^2}\ell(y_i, X_i; \theta_w^{(0)} + \tau(\theta_w - \theta_w^{(0)}))\Big|_{\tau=\tilde{\tau}} - \frac{\partial^2}{\partial\tau^2}\ell(y_i, X_i; \theta_w^{(0)} + \tau(\theta_w - \theta_w^{(0)}))\Big|_{\tau=0}\right| \\ &\leq C(y_i, X_i)\|\theta_w - \theta_w^{(0)}\|^{2+\nu} \quad y_i, X_i\text{-as} \end{aligned} \tag{A.9}$$

which is similar to Assumption A.5(ii) in Carroll et al. (2010). Under the stated conditions, equation (A.9) implies that for all $\theta_w \in \Theta_{0n}^{(w)}$

$$\mathcal{L}(\theta_w^{(0)}) - \mathcal{L}(\theta_w) = \frac{1}{2}\|\theta_w - \theta_w^{(0)}\|^2(1 + o(1))$$

where $o(\cdot)$ is under $\|\cdot\|_{c,\infty}$. This implies Assumption A-3. Hu and Schennach (2008) provide conditions for equation (A.9) via a fourth-order Taylor expansion of the objective function; see their Assumption 23.

Assumption A-4 is standard in the literature (Ai and Chen 2003; Ai and Chen 2007). A sufficient condition for Assumption is when $\rho^{d(n)}(\cdot)$ is orthonormal on $\mathbb{X}$ with respect to Lebesgue measure and the density of X_i is bounded away from zero.

Assumption A-5 allows us to control the convergence rate of the variance of the estimator. For standard choices of series bases, we have under typical assumptions $\sqrt{\mathbb{E}[\|\rho^{d(n)}(X_i)\|_E^2]} \leq$

$\sup_{X\in\mathbb{X}} \|\rho^{d(n)}(X)\|_E \leq C(d(n)+1) \leq Cn$ (Newey 1997). The last inequality follows from the fact that Assumption A-5 implies that $d(n)\log(n)n^{-1/2} = o(1)$ which implies that $d(n)/n = o(1)$. Then, Assumption A-5 reduces to $d(n)\times\log(n)\times n^{-1/2} = o(1)$, which is standard in the literature (Ai and Chen 2003; Ai and Chen 2007; Hu and Schennach 2008). Similarly, under the common assumption that the maximal eigenvalue of $\mathbb{E}[\rho^{d(n)}(X)\rho^{d(n)}(X)^\top]$ is bounded above uniformly over $d(n)$, we have $\sqrt{\mathbb{E}[\|\rho^{d(n)}(X_i)\|_E^2]} \leq C\sqrt{d(n)+1}$ for some positive and finite C.

Before stating the rate result, we state a helpful lemma that allows us to apply Corollary 1 in Chen and Shen (1998) locally on $\Theta_{0n}^{(w)}$.

Lemma A.4.1 *(Localization of rates) Let B_n be some measurable event such that $\mathbb{P}(B_n) \to 1$ as $n \to \infty$ and assume that $\lim_{x\to\infty}\limsup_{n\to\infty}\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq x\varepsilon_n, B_n) = 0$. Then, $\lim_{x\to\infty}\limsup_{n\to\infty}\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq x\varepsilon_n) = 0$ or equivalently $\|\hat{\theta}_n - \theta\| = O_p(\varepsilon_n)$.*

Proof. We observe

$$\begin{aligned}\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq x\varepsilon_n) =&\ \mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq x\varepsilon_n, B_n) + \mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq x\varepsilon_n, B_n^c)\\ \leq&\ \mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq x\varepsilon_n, B_n) + \mathbb{P}(B_n^c)\end{aligned}$$

where B_n^c denotes the complement of B_n. Taking the respective limits gives the desired result since $\mathbb{P}(B_n^c) \to 0$ as $n \to \infty$. □

Letting $B_n = \{\hat{\theta}_{w,n} \in \Theta_{0n}^{(w)}\}$, which is measurable as $\hat{\theta}_{w,n}$ is measurable under our assumptions, Lemma A.4.1 allows us to apply Corollary 1 in Chen and Shen (1998) on the local parameter space, that is, one may replace $\Theta_{d(n)}^{(w)}$ in their proof with $\Theta_{0n}^{(w)}$.

Theorem A.4.2 *(Convergence rates) Under Assumptions E-1 to E-7 and A-1 to A-5, $\|\hat{\theta}_{w,n} - \theta_w^{(0)}\| = o_p(n^{-1/4})$.*

Proof. Throughout the proof, we use the convention that $\dim(\tilde{x}_{it}) = K$ and, if allowed, write $d(n)$ instead of $d(n)+1$ absorbing the one into some constant. All inequalities, if not noted otherwise, are to be understood to hold almost surely. We check the conditions of Corollary 1 in Chen and Shen (1998) on the local parameter space $\Theta_{0n}^{(w)}$.

1. *Norm equivalence* holds by Assumption A-3.
2. *Objective function is uniformly bounded*: Since $\{F_{jt}(\cdot)\}_{j=1,\dots,J;t=1,\dots,T}$ are strictly monotone on $\mathbb{R}$, and $\mathbb{X}$ as well as $\mathcal{B}$ are compact, the objective function is uniformly bounded – a detailed discussion of this is provided when verifying *Condition A.2*.
3. *Stationarity and mixing condition*: By Assumption E-1, $\{(y_i^\top, X_i^\top)^\top\}$ is clearly stationarity and satisfies the required mixing condition.

4. *Condition A.2*: We need to show that for all small $\varepsilon > 0$

$$\sup_{\{\theta_w \in \Theta_{0n}^{(w)} : \|\theta_w^{(0)} - \theta_w\| \leq \varepsilon\}} Var(\ell(y_i, X_i; \theta_w) - \ell(y_i, X_i; \theta_w^{(0)})) \leq C_1 \varepsilon^2$$

Let $h(\mathbb{P}_{\theta_w}, \mathbb{P}_{\theta_w^{(0)}}; X_i) = \sqrt{\frac{1}{2}\sum_{s \in \{0,1\}^T} \left(\sqrt{\mathbb{P}(y_i = s \mid X_i; \theta_w)} - \sqrt{\mathbb{P}(y_i = s \mid X_i; \theta_w^{(0)})} \right)^2}$ be the conditional Hellinger distance where $y_i = (y_{i1}, \ldots, y_{iT})^\top$. We observe

$$-\frac{1}{2}\left[\mathcal{L}(\theta_w^{(0)}) - \mathcal{L}(\theta_w)\right] = -\frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\log(\mathbb{P}(y_i \mid X_i; \theta_w^{(0)})) - \log(\mathbb{P}(y_i \mid X_i; \theta_w)) \mid X_i\right]\right]$$

Now, following identical arguments as in Lemma 1.3 in van de Geer (2000), for $\theta_w \in \{\theta_w \in \Theta_{0n}^{(w)} : \|\theta_w - \theta_w^{(0)}\| \leq \varepsilon\}$

$$-\frac{1}{2}\mathbb{E}\left[\log(\mathbb{P}(y_i \mid X_i; \theta_w^{(0)})) - \log(\mathbb{P}(y_i \mid X_i; \theta_w)) \mid X_i\right] \leq -h^2(\mathbb{P}_{\theta_w}, \mathbb{P}_{\theta_w^{(0)}}; X_i)$$

where we use that, uniformly over $\Theta_{0n}^{(w)}$, $\mathbb{P}(y \mid X; \theta_w) > 0$ for all $(y^\top, X^\top) \in \{0,1\}^T \times \mathbb{X}$. We conclude

$$\mathbb{E}[h^2(\mathbb{P}_{\theta_w}, \mathbb{P}_{\theta_w^{(0)}}; X)] \leq \frac{1}{2}\left[\mathcal{L}(\theta_w^{(0)}) - \mathcal{L}(\theta_w)\right] \leq \frac{1}{4c_1}\|\theta_w - \theta_w^{(0)}\|^2$$

where the final inequality follows from Assumption A-3.

Following arguments similar to Example 2 in Shen and Wong (1994), we define $r(y_i, X_i; \theta_w)^2 = \frac{\mathbb{P}(y_i|X_i;\theta_w)}{\mathbb{P}(y_i|X_i;\theta_w^{(0)})}$. From Fact 2 at the beginning of this section, $\underline{f} \leq F_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{jt}) \leq \overline{f}$ for all j, t with $\overline{f}, \underline{f} \in (0,1)$. Letting $\tilde{x}_{it} = (1, x_{it}^\top)^\top$ and $\tilde{\beta}_{j,t} = (\alpha_{j,t}, \beta_{j,t}^\top)^\top$, we therefore have for all $\theta_w \in \Theta^{(w)}$

$$\begin{aligned}
\min\{\underline{f}, 1-\overline{f}\}^T &\leq \min_j \prod_{t=1}^T F_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t})^{y_{it}}(1 - F_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}))^{1-y_{it}} \\
&\leq \mathbb{P}(y_i \mid X_i; \theta_w) = \sum_{j=1}^J L_j(X_i; w) \prod_{t=1}^T F_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t})^{y_{it}}(1 - F_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}))^{1-y_{it}} \\
&\leq \max_j \prod_{t=1}^T F_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t})^{y_{it}}(1 - F_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}))^{1-y_{it}} \leq \max\{\overline{f}, 1-\underline{f}\}^T
\end{aligned}$$

so that

$$L := \sqrt{\frac{\min\{\underline{f}, 1-\overline{f}\}^T}{\max\{\overline{f}, 1-\underline{f}\}^T}} \leq r(y_i, X_i; \theta_w) \leq \sqrt{\frac{\max\{\overline{f}, 1-\underline{f}\}^T}{\min\{\underline{f}, 1-\overline{f}\}^T}} =: M$$

where $0 < L, M < \infty$ and $L < 1$. Hence, for all $\theta_w \in \{\theta_w \in \Theta_{0n}^{(w)} : \|\theta_w^{(0)} - \theta_w\| \leq \varepsilon\}$

$$Var(\ell(y_i, X_i; \theta_w) - \ell(y_i, X_i; \theta_w^{(0)})) \leq \mathbb{E}\left[(\ell(y_i, X_i; \theta_w) - \ell(y_i, X_i; \theta_w^{(0)}))^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\log\left(\frac{\mathbb{P}(y_i \mid X_i;\theta_w)}{\mathbb{P}(y_i \mid X_i;\theta_w^{(0)})}\right)\right)^2\right] \\
&\leq 4C\,\mathbb{E}\left[(r(y_i,X_i;\theta_w)-1)^2\right] \\
&= 8C\,\mathbb{E}[h^2(\mathbb{P}_{\theta_w},\mathbb{P}_{\theta_w^{(0)}};X_i)] \\
&\leq \frac{2C}{c_1}\varepsilon^2
\end{aligned}
$$

where the second inequality follows from a mean value expansion and the fact that $0 < L \leq r(y,X;\theta) \leq M < \infty$. C is some generic positive and finite constant. We conclude that Condition A.2 is satisfied.

5. *Condition A.3*: Let $\mathcal{F}_n = \{\ell(y,X;\theta_w) - \ell(y,X;\theta_w^{(0)}) : \|\theta_w - \theta_w^{(0)}\| \leq \delta, \theta_w \in \Theta_{0n}^{(w)}\}$. There exists $\delta_n \in (0,1)$ such that $\delta_n = \inf\{\delta > 0 : \delta^{-2}\int_{b\delta^2}^{a\delta} H_{[]}^{1/2}(\kappa,\mathcal{F}_n)d\kappa \leq C_2 n^{1/2}\}$ where $H_{[]}(\kappa,\mathcal{F}_n)$ is the bracketing L_2-metric entropy of the space $\mathcal{F}_n$.
For $\theta_{1,w},\theta_{2,w} \in \Theta_{0n}^{(w)}$

$$
\begin{aligned}
&|\ell(y_i,X_i;w_1,\beta_1) - \ell(y_i,X_i;w^{(0)},\beta^{(0)}) - (\ell(y_i,X_i;w_2,\beta_2) - \ell(y_i,X_i;w^{(0)},\beta^{(0)}))| \\
&= |\ell(y_i,X_i;w_1,\beta_1) - \ell(y_i,X_i;w_2,\beta_2)| \\
&\leq \frac{1}{\min\{\underline{f},1-\overline{f}\}^T}\left(\sum_{j=1}^{J-1}|w_{1,j}(X_i) - w_{2,j}(X_i)| + C_F C_X \sum_{t=1}^{T}\sum_{j=1}^{J}\|\tilde{\beta}_{1,j,t} - \tilde{\beta}_{2,j,t}\|_E\right) \\
&\leq \frac{1}{\min\{\underline{f},1-\overline{f}\}^T}\left(\sum_{j=1}^{J-1}\|\gamma_{1,j,n} - \gamma_{2,j,n}\|_E\|\rho^{d(n)}(X_i)\|_E + C_X C_F \sum_{t=1}^{T}\sum_{j=1}^{J}\|\tilde{\beta}_{1,j,t} - \tilde{\beta}_{2,j,t}\|_E\right) \\
&\leq \frac{\|\rho^{d(n)}(X_i)\|_E + C_F C_X}{\min\{\underline{f},1-\overline{f}\}^T}\left(\sum_{j=1}^{J-1}\|\gamma_{1,j,n} - \gamma_{2,j,n}\|_E + \sum_{t=1}^{T}\sum_{j=1}^{J}\|\tilde{\beta}_{1,j,t} - \tilde{\beta}_{2,j,t}\|_E\right)
\end{aligned}
$$

Letting $\xi_n = (\beta^\top, \gamma_{1,n}^\top, \ldots, \gamma_{J-1,n}^\top)$, we therefore have by Jensen's inequality that there exists some finite constant C^* that does not depend on $d(n)$ such that

$$
|\ell(y_i,X_i;w_1,\beta_1) - \ell(y_i,X_i;w_2,\beta_2)| \leq C^* \frac{\|\rho^{d(n)}(X_i)\|_E + C_F C_X}{\min\{\underline{f},1-\overline{f}\}^T}\|\xi_{1,n} - \xi_{2,n}\|_E
$$

Letting $\Gamma_{j,n}$ be the parameter space of $\gamma_{j,n}$ and $\left\|\|\rho^{d(n)}(X_i)\|_E\right\|_{L_2(P_0)} = \sqrt{\mathbb{E}[\|\rho^{d(n)}(X_i)\|_E^2]}$, Theorem 2.7.17 in van der Vaart and Wellner (2023) implies

$$
\begin{aligned}
H_{[]}(\kappa,\mathcal{F}_n) \leq \sum_{j=1}^{J-1}&\log\left(N\left(\frac{\kappa\min\{\underline{f},1-\overline{f}\}^T}{2C^*(\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)} + C_F C_X)}, \Gamma_{j,n}, \|\cdot\|_E\right)\right) \\
&+ \log\left(N\left(\frac{\kappa\min\{\underline{f},1-\overline{f}\}^T}{2C^*(\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)} + C_F C_X)}, \mathcal{B}, \|\cdot\|_E\right)\right)
\end{aligned}
$$

Let $\lambda_{\min} > 0$ denote the minimal eigenvalue of $\mathbb{E}[\rho^{d(n)}(X)\rho^{d(n)}(X)^\top]$. Under Assump-

tion A-4, we have from the definition of the local parameter space $\Theta_{0n}^{(w)}$ that

$$\lambda_{\min}\|\gamma_j\|_E^2 \leq \gamma_j^\top \mathbb{E}[\rho^{d(n)}(X)\rho^{d(n)}(X)^\top]\gamma_j = \|\gamma_j^\top \rho^{d(n)}(X)\|_{L_2(P_0)}^2 \leq \|\gamma_j^\top \rho^{d(n)}(X)\|_\infty^2 \leq C_0^2$$

Hence $\|\gamma_j\|_E^2 \leq \frac{C_0^2}{\lambda_{\min}} =: \tilde{C}^2$. It follows from Corollary 2.6 in van de Geer (2000) or Exercise 7 on page 143 in van der Vaart and Wellner (2023) that

$$\begin{aligned}
&\log\left(N\left(\frac{\kappa \min\{\underline{f}, 1-\overline{f}\}^T}{2C^*(\|\|\rho^{d(n)}(X)\|_E\|_{L_2(P_0)} + C_F C_X)}, \Gamma_{j,n}, \|\cdot\|_E\right)\right) \\
&\leq d(n)\log\left(\frac{6C^*(\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)} + C_F C_X)\tilde{C}}{\kappa \min\{\underline{f}, 1-\overline{f}\}^T}\right) \text{ and} \\
&\log\left(N\left(\frac{\kappa \min\{\underline{f}, 1-\overline{f}\}^T}{2C^*(\|\|\rho^{d(n)}(X)\|_E\|_{L_2(P_0)} + C_F C_X)}, \mathcal{B}, \|\cdot\|_E\right)\right) \\
&\leq JTK\log\left(\frac{6C^*(\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)} + C_F C_X)C_{\mathcal{B}}}{\kappa \min\{\underline{f}, 1-\overline{f}\}^T}\right)
\end{aligned}$$

where $C_{\mathcal{B}} = \sup_{\beta\in\mathcal{B}}\|\beta\|_E < \infty$ by compactness of $\mathcal{B}$ and we suppress that the first bound should depend $d(n)+1$; the additional plus 1 can be absorbed into a constant. Combining this with the previous bound gives

$$H_{[]}(\kappa, \mathcal{F}_n) \leq ((J-1)d(n) + JTK)\log\left(\frac{6C^*(\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)} + C_F C_X)\overline{C}}{\kappa \min\{\underline{f}, 1-\overline{f}\}^T}\right)$$

where $\overline{C} = \max\{\tilde{C}, C_{\mathcal{B}}\}$. Bounding further yields

$$H_{[]}(\kappa, \mathcal{F}_n) \leq C((J-1)d(n) + JTK)\log\left(\frac{\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)}}{\kappa}\right)$$

for some generic constant C that does not depend on κ, $d(n)$ or $\|\|\rho^{d(n)}(X)\|_E\|_{L_2(P_0)}$. Plugging this into Condition A.3 and setting $a = b = 1$ gives

$$\begin{aligned}
&\frac{1}{\sqrt{n}\delta^2}\int_{\delta^2}^{\delta}\sqrt{H_{[]}(\kappa, \mathcal{F}_n)}d\kappa \\
&\leq \frac{\sqrt{C((J-1)d(n)+JTK)}}{\sqrt{n}\delta}\left[\left[\log\left(\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)}\right)\right]^{1/2} + \left[2\log\left(\frac{1}{\delta}\right)\right]^{1/2}\right]
\end{aligned}$$

Following similar arguments as in Chen et al. (2023) (a working paper version of Chen et al. (2024)), the above is smaller than some constant C^* if

$$\begin{aligned}
&\frac{C((J-1)d(n)+JTK)}{n\delta^2}\log\left(\left\|\|\rho^{d(n)}(X)\|_E\right\|_{L_2(P_0)}\right) \leq C^*/2 \\
&\frac{C((J-1)d(n)+JTK)}{n\delta^2}2\log\left(\frac{1}{\delta}\right) \leq C^*/2
\end{aligned}$$

which is satisfied for all n sufficiently large when

$$\delta \geq C\sqrt{\frac{((J-1)d(n)+JTK)\log\left(\left\|\left\|\rho^{d(n)}(X)\right\|_E\right\|_{L_2(P_0)}\right)}{n}}$$

$$\delta \geq C\sqrt{\frac{((J-1)d(n)+JTK)\log\left(\frac{n}{(J-1)d(n)+JTK}\right)}{n}}$$

where C is some generic constant. We may therefore choose

$$\begin{aligned}\delta_n &= \sqrt{C\frac{((J-1)d(n)+JTK)\left[\log\left(\left\|\left\|\rho^{d(n)}(X)\right\|_E\right\|_{L_2(P_0)}\right)+\log\left(\frac{n}{(J-1)d(n)+JTK}\right)\right]}{n}} \\ &\leq C\sqrt{\frac{((J-1)d(n)+JTK)\log\left(\max\left\{\left\|\left\|\rho^{d(n)}(X)\right\|_E\right\|_{L_2(P_0)},n\right\}\right)}{n}} \\ \implies \delta_n &= O_p\left(\sqrt{\frac{d(n)\log\left(\max\left\{\left\|\left\|\rho^{d(n)}(X)\right\|_E\right\|_{L_2(P_0)},n\right\}\right)}{n}}\right) = o(n^{-1/4})\end{aligned}$$

where C may vary from line to line and the last line follows from Assumption A-5 and the fact that J, T, and K are fixed as $n \to \infty$. Hence, *Condition A.3* is satisfied.

Now, the conclusion of Corollary 1 in Chen and Shen (1998) applies and implies

$$\|\hat{\theta}_{w,n} - \theta_w^{(0)}\| = O_p\left(\max\left\{\delta_n, \max_j \|w_{j,d(n)} - w_j^{(0)}\|\right\}\right) = o_p(n^{-1/4})$$

This concludes the proof. □

A.4.3 Asymptotic normality

To present our asymptotic normality results, we require some additional notation. We borrow most of our notation from Ai and Chen (2003) and the Online Appendices of Hu and Schennach (2008) and Freyberger (2018). Specifically, we let $\overline{\boldsymbol{V}}$ denote the closed linear span of $\Theta_0^{(w)} - \{\theta_w^{(0)}\}$ under the Fisher norm $\|\cdot\|$. Next, we define the following inner product

$$\langle v_1, v_2\rangle = \mathbb{E}\left[\frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[v_1]\frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[v_2]\right]$$

for $v_1, v_2 \in \overline{\boldsymbol{V}}$ and the pathwise derivative is $[y, X\text{-a.s.}]$ defined as in the previous section. The inner product induces $\|\cdot\|$ as defined previously so that $(\overline{\boldsymbol{V}}, \|\cdot\|)$ constitutes a Hilbert space. Specifically, we have $\overline{\boldsymbol{V}} = \mathbb{R}^{d_\beta} \times \overline{\mathcal{U}}$ with $\overline{\mathcal{U}} = \overline{\mathcal{W}_0 - \{w^{(0)}\}}$ where $\mathcal{W}_0 = \{w \in \mathcal{W} : \sum_{j=1}^{J-1} \|w_j - w_j^{(0)}\|_\infty = o(1)\}$.

The remainder of this section proceeds as follows: First, Appendix A.4.3.1 presents an asymptotic normality result for $\hat{\beta}_n$. Next, Appendix A.4.3.2 presents an asymptotic normality result for $\hat{\pi}_j$ which is required to derive the asymptotic distribution of $\widehat{\text{AME}}_{j,t,k}$. Appendix A.4.3.3 then derives the asymptotic distribution of $\widehat{\text{AME}}_{j,t,k}$. Lastly, Appendix A.4.3.4 provides a discussion on necessary condition for $\sqrt{n}$-estimability of $\hat{\beta}_n$. For purposes of readability, we shall often drop the subscript n when talking about estimators in the following.

A.4.3.1 Asymptotic normality of $\hat{\beta}_n$

From the Cramér-Wold device, $\sqrt{n}(\hat{\beta} - \beta^{(0)})$ converges to a multivariate Gaussian distribution if and only if for all $\lambda \neq 0$ $\sqrt{n}\lambda^\top(\hat{\beta} - \beta^{(0)})$ converges to a univariate normal distribution. We therefore study the asymptotic behavior of the linear functional $f(\theta_w^{(0)}) = \lambda^\top \beta^{(0)}$ in the following. The operator norm of $f(\cdot)$ on $(\overline{V}, \|\cdot\|)$ is

$$\sup_{\theta_w \in \overline{V}: \|\theta\| > 0} \frac{|f(\theta_w)|}{\|\theta\|} = \sup_{\|\theta_w - \theta_w^{(0)}\| > 0} \frac{|\lambda^\top(\beta - \beta^{(0)})|}{\|\theta_w - \theta_w^{(0)}\|} \tag{A.10}$$

It is well known that $\hat{\beta}$ is $\sqrt{n}$-estimable if and only if the above operator norm is finite. We will see that this condition is closely linked to the positive-definiteness of the asymptotic covariance matrix of $\hat{\beta}$, that is, the operator norm is finite if and only if the asymptotic variance of $\hat{\beta}$ exists and is finite and positive definite. To see this, we observe that for any $\beta - \beta^{(0)} \neq 0$

$$\begin{aligned}
\frac{d\ell(y, X; \theta_w^{(0)})}{d\theta}[\theta_w - \theta_w^{(0)}] &= \left(\frac{d\ell(y, X; \theta_w^{(0)})}{d\beta'} - \frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\mu] \right) (\beta - \beta^{(0)}) \\
&= \left(D_{\mu_1}(y, X), \ldots, D_{\mu_{d_\beta}}(y, X) \right) (\beta - \beta^{(0)})
\end{aligned}$$

with

$$\frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\mu] = \left(\frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\mu_1], \ldots, \frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\mu_{d_\beta}] \right)$$

and $w - w^{(0)} = -\mu(\beta - \beta^{(0)})$ with $\mu \in \prod_{k=1}^{d_\beta} \overline{\mathcal{U}}$. Thus,

$$\|\theta_w - \theta_w^{(0)}\| = \sqrt{(\beta - \beta^{(0)})^\top \mathbb{E}\left[D_\mu(y_i, X_i) \right)^\top D_\mu(y_i, X_i)] (\beta - \beta^{(0)})}$$

Next, define

$$\mu_k^* \in \arg\min_{\mu_k \in \overline{\mathcal{U}}} \mathbb{E}\left[D_{\mu_k}(y, X)^2 \right]$$

where the solution need not be unique. Define the vector $D_{\mu^*}(y,X) = (D_{\mu_1^*}(y,X),\ldots,D_{\mu_{d_\beta}^*}(y,X))$ and the matrix

$$V^* = \mathbb{E}[D_{\mu^*}(y_i,X_i)^\top D_{\mu^*}(y_i,X_i)]$$

One can show that $\sup_{\|\theta_w-\theta_w^{(0)}\|>0} \frac{(\beta-\beta^{(0)})^\top\lambda\lambda^\top(\beta-\beta^{(0)})}{\|\theta_w-\theta_w^{(0)}\|^2} < \infty$ if and only if V^* is positive definite. Specifically, we have

$$\sup_{\|\theta_w-\theta_w^{(0)}\|>0} \frac{(\beta-\beta^{(0)})^\top\lambda\lambda^\top(\beta-\beta^{(0)})}{\|\theta_w-\theta_w^{(0)}\|^2} = \lambda^\top(V^*)^{-1}\lambda$$

μ^* is referred to as the least favorable direction (or least favorable submodel in the literature on semiparametric efficiency).

For the remainder of this section, we assume

Assumption A-6 *(Asymptotic Covariance) (i) V^* exists and is finite and positive definite. (ii) $\beta^{(0)} \in int(\mathcal{B})$.*

In Section A.4.3.4, we provide typical sufficient conditions that ensure that V^* is positive definite. Under Assumption A-6, the functional of interest is bounded so by the Riesz representation theorem there exists $v^*(\lambda) \in \overline{\boldsymbol{V}}$ such that for all $\theta_w - \theta_w^{(0)} \in \overline{\boldsymbol{V}}$

$$f(\theta_w-\theta_w^{(0)}) = \lambda^\top(\beta-\beta^{(0)}) = \langle \theta_w-\theta_w^{(0)}, v^*(\lambda)\rangle$$
$$\sup_{\|\theta_w-\theta_w^{(0)}\|>0} \frac{(\beta-\beta^{(0)})^\top\lambda\lambda^\top(\beta-\beta^{(0)})}{\|\theta_w-\theta_w^{(0)}\|^2} = \|v^*(\lambda)\|^2 = \lambda^\top(V^*)^{-1}\lambda$$

From Wong and Severini (1991) and Ai and Chen (2003), we know $v^*(\lambda) = (v_\beta^*(\lambda), v_\pi^*(\lambda))$ with $v_\beta^*(\lambda) = V^{*-1}\lambda$ and $v_\pi^*(\lambda) = -\mu^* v_\beta^*$.

We define the following local parameter spaces

$$\mathcal{N}_0 = \{\theta_w \in \Theta_0^{(w)} : \|\theta_w-\theta_w^{(0)}\| = o(n^{-1/4})\} \text{ and } \mathcal{N}_{0n} = \{\theta_w \in \Theta_{0n}^{(w)} : \|\theta_w-\theta_w^{(0)}\| = o(n^{-1/4})\}$$

Next, we define for all $j = 1,\ldots,J-1$ and $d = 0,\ldots,d(n)$

$$\frac{d\ell(y,X;\theta_w^{(0)})}{dw_j}[\rho_d^{d(n)}] = \left.\frac{\partial \ell(y,X;w_j^{(0)}+\tau\rho_d^{d(n)}, \{w_{j^*}^{(0)}\}_{j^*\neq j}, \beta^{(0)})}{\partial\tau}\right|_{\tau=0}$$
$$= \frac{\pi_j^{(0)}(X)\rho_d^{d(n)}(X)\Big[\prod_{t=1}^T F_{jt}(\tilde{x}_t^\top\tilde{\beta}_{j,t}^{(0)})^{y_t}(1-F_{jt}(\tilde{x}_t^\top\tilde{\beta}_{j,t}^{(0)}))^{1-y_t} - \mathbb{P}(\{y_t\}_{t=1}^T \mid X;\theta_w^{(0)})\Big]}{\mathbb{P}(\{y_t\}_{t=1}^T \mid X;\theta_w^{(0)})}$$

Using this notation, we follow Hu and Schennach (2008) and define for $j = 1,\ldots,J-1$

$$\frac{d\ell(y,X;\theta_w^{(0)})}{dw_j}[\rho^{d(n)}] = \left(\frac{d\ell(y,X;\theta_w^{(0)})}{dw_j}[\rho_0^{d(n)}],\ldots,\frac{d\ell(y,X;\theta_w^{(0)})}{dw_j}[\rho_{d(n)}^{d(n)}]\right)^\top$$

and

$$\frac{d\ell(y,X;\theta_w^{(0)})}{d\beta} = \left(\frac{\partial\ell(y,X;\theta_w^{(0)})}{\partial\beta_1^{(0)}},\ldots,\frac{\partial\ell(y,X;\theta_w^{(0)})}{\partial\beta_{d_\beta}^{(0)}}\right)^\top$$

$$\frac{d\ell(y,X;\theta_w^{(0)})}{d\theta_w}[\rho^{d(n)}] = \left(\left(\frac{d\ell(y,X;\theta_w^{(0)})}{d\beta}\right)^\top, \left(\frac{d\ell(y,X;\theta_w^{(0)})}{dw_1}[\rho^{d(n)}]\right)^\top,\ldots,\right.$$
$$\left.\left(\frac{d\ell(y,X;\theta_w^{(0)})}{dw_{J-1}}[\rho^{d(n)}]\right)^\top\right)^\top$$

Letting

$$\Omega_{d(n)} = \mathbb{E}\left[\left(\frac{d\ell(y_i,X_i;\theta_w^{(0)})}{d\theta_w}[\rho^{d(n)}]\right)\left(\frac{d\ell(y_i,X_i;\theta_w^{(0)})}{d\theta_w}[\rho^{d(n)}]\right)^\top\right] \tag{A.11}$$

we follow Hu and Schennach (2008) in making the following assumption.

Assumption A-7 *(Local behavior of objective function) The smallest eigenvalue of* $\Omega_{d(n)}$ *is bounded away from zero and* $\|\rho_d^{d(n)}\|_\infty < \infty$ *for* $d = 0,\ldots,d(n)$ *uniformly in* $d(n)$.

As mentioned in Hu and Schennach (2008), the first part of Assumption A-7 is a typical assumption in the context of series estimation. Since $\mathbb{X}$ is compact, the second part of the assumption is trivially satisfied for all common choices of series bases. Intuitively, the first part of Assumption A-7 ensures that the Fisher information matrix is positive definite uniformly over the sequence of series spaces as $d(n) \to \infty$. By the Fisher information matrix identity, this implies that the geometry of the space $\{\mathcal{L}(\Pi_n\theta_w^{(0)}) - \mathcal{L}(\theta_w) : \theta_w \in \mathcal{N}_{0n}\}$ is determined by a weighted inner product that induces a weighted Euclidean distance where $\Pi_n\theta$ denotes the projection of θ on $\Theta_{d(n)}^{(w)}$. If the projection error goes to zero sufficiently fast, this allows one to conclude that the local geometry of the centered objective function space is characterized by the Fisher norm $\|\cdot\|$ as defined in the previous section. More concretely, for all $\theta_w \in \mathcal{N}_{0n}$

$$\mathcal{L}(\theta_w^{(0)}) - \mathcal{L}(\theta_w) = \frac{1}{2}\|\theta_w - \theta_w^{(0)}\|^2(1+o(1)) \tag{A.12}$$

We make an additional assumption that ensures sufficiently fast convergence of the approximation error of $\|\Pi_n\theta_0^{(0)} - \theta_w^{(0)}\|_{c,\infty}$. As in Hu and Schennach (2008), in our setting, we can bound the remainder terms in (A.12) most easily in terms of the stronger consistency norm $\|\cdot\|_{c,\infty}$ so we require additional assumptions on the convergence rate in this stronger norm. A more careful argument may be able to relax these assumptions.

Assumption A-8 *(Approximation in consistency norm)* $\|\Pi_n\theta_w^{(0)} - \theta_w^{(0)}\|_{c,\infty} = O(d(n)^{-\eta/d_X}) = o(n^{-1/4})$.

Assumption A-8 is equivalent to requiring $\|\Pi_{j,n} w_j^{(0)} - w_j^{(0)}\|_\infty = O(d(n)^{-\eta/d_X}) = o(n^{-1/4})$ for $j = 1, \ldots, J-1$ where $\Pi_{j,n}$ is the projection onto $\mathcal{W}_{j,d(n)}$. Under the typical smoothness conditions that we discussed below Assumption A-2, Assumption A-8 is satisfied.

Next, we assume that the Riesz representer can be approximated well in the Fisher norm.

Assumption A-9 *(Approximation of Riesz Representer) For all* $\lambda \in \mathbb{R}^{d_\beta}$ *with* $\|\lambda\|_E = 1$, *there is* $v_n^*(\lambda) = \begin{pmatrix} v_\beta^*(\lambda) \\ -(\Pi_n \mu^*) v_\beta^*(\lambda) \end{pmatrix} \in \Theta_{d(n)}^{(w)} - \{\theta_w^{(0)}\}$ *such that* $\|v_n^*(\lambda) - v^*(\lambda)\| = o(n^{-1/4})$.

As mentioned by Ai and Chen (2003), one may use a different sieve space than $\Theta_{d(n)}^{(w)}$ if this has better approximation properties. Based on Shen (1997), this, however, requires that the following oracle inequality still holds for the local alternative $\theta_w^*(\hat{\theta}_{w,n}, \varepsilon_n) = (1 - \varepsilon_n)\hat{\theta}_{w,n} + \varepsilon_n(v^*(\lambda) + \theta_w^{(0)})$ with $\varepsilon_n = o(n^{-1/2})$:

$$\mathcal{L}_n(\hat{\theta}_{w,n}) \geq \mathcal{L}_n(\Pi_n \theta_w^*(\hat{\theta}_{w,n}, \varepsilon_n)) - O_p(\varepsilon_n^2)$$

If Π_n is the projection onto $\Theta_{d(n)}^{(w)}$, this inequality holds by the definition of the estimator as an approximate sieve maximum likelihood estimator.

We make one final assumption that ensures that Assumption 23 in the Online Appendix of Hu and Schennach (2008) is satisfied.

Assumption A-10 *(Higher-order term)* $F_{jt}(\cdot)$ *is four-times continuously differentiable.*

Assumption A-10 is satisfied in standard settings, for instance, when the component distributions are probit or logit models. We conclude this section with the following result.

Theorem A.4.3 *Under the Assumptions of Theorem A.4.2 and Assumptions A-6 to A-10,* $\sqrt{n}(\hat{\beta}_n - \beta^{(0)}) \xrightarrow{d} N(0, (V^*)^{-1})$.

Proof. The proof follows from the same arguments as the proof of Theorem 3 in the Online Appendix of Hu and Schennach (2008). Their proof requires the existence of various dominating functions. Some of them we have shown to exist in the proof of Theorem 4.1. The remaining relevant assumptions we still have to argue to hold are their Assumptions 20 and 23. We will do so now. Throughout the proof, we let $K = \dim(\tilde{x}_{it})$. Additionally, in this proof, we interpret the absolute value of a pathwise derivative in direction $\mathbf{1}$ as the sum of the absolute values of the corresponding partial derivative terms. This is similar to Hu and Schennach (2008).

A20 : In our notation, Assumption 20 requires that there exists some measurable function $h_1(y, X)$ with $\mathbb{E}[h_1(y_i, X_i)^2] < \infty$ such that for any $\overline{\theta}_w \in \mathcal{N}_0$

$$\left| \frac{\frac{\partial \mathbb{P}(y|X;\overline{\theta}_w + \mathbf{1}\tau)}{\partial \tau}\Big|_{\tau=0}}{\mathbb{P}(y \mid X; \overline{\theta}_w)} \right|^2 + \left| \frac{\frac{\partial^2 \mathbb{P}(y|X;\overline{\theta}_w + \mathbf{1}\tau)}{\partial^2 \tau}\Big|_{\tau=0}}{\mathbb{P}(y \mid X; \overline{\theta}_w)} \right| \leq h_1(y, X) \tag{A.13}$$

where $\mathbf{1} \in \mathbb{R}^{d_\beta+J-1}$ is a vector of ones. We show that the terms can be bounded individually. First, from the arguments in the proof of Theorem 4.1, we have $\mathbb{P}(y \mid X; \overline{\theta}_w) \geq \min\{\underline{f}, 1-\overline{f}\}^T$, and from our previous discussions in combination with Jensen's inequality

$$\left| \frac{\partial \, \mathbb{P}(y \mid X; \overline{\theta}_w + \mathbf{1}\tau)}{\partial \tau} \bigg|_{\tau=0} \right|^2 \leq 2 + 2KT^2 C_F^2 C_X^2$$

Next, some straightforward but tedious calculus gives[27]

$$\left| \frac{\partial^2 \, \mathbb{P}(y \mid X; \overline{\theta}_w + \mathbf{1}\tau)}{\partial^2 \tau} \bigg|_{\tau=0} \right| \leq 4 + 4(\sqrt{K} T C_F C_X) + T(C_F'' C_X^2 K + (T-1) C_F^2 C_X^2 K)$$

where C_F'' is an upper bound on $|\frac{\partial^2}{\partial s^2} F_{jt}(s)|_{s=x_t^\top \beta_{j,t} + \alpha_{j,t}}|$ uniformly over $\mathbb{X}$ and $\mathcal{B}$. C_F'' is finite as $\mathbb{X}$ and $\mathcal{B}$ are compact and $\frac{\partial^2}{\partial s^2} F_{jt}(s)$ is assumed to be continuous. We therefore conclude that (A.13) is satisfied with $h_1(y, X) = \frac{6+4\sqrt{K}TC_FC_X + 2KT^2C_F^2C_X^2 + T(C_F''C_X^2K + (T-1)C_F^2C_X^2K)}{\min\{\underline{f}, 1-\overline{f}\}^{2T}}$ which is clearly square integrable. Thus Assumption 20 is satisfied in our setting.

A23 : Assumption 23 in Hu and Schennach (2008) assumes that for all $\theta_w \in \mathcal{N}_{0n}$, there exists a measurable function $h_2(y, X)$ with $E[|h_2(y_i, X_i)|] < \infty$ such that

$$\left| \frac{\partial^4}{\partial \tau^2} \ell(y, X; \overline{\theta}_w + \tau(\theta_w - \theta_w^{(0)})) \bigg|_{\tau=0} \right| \leq h_2(y, X) \|\theta_w - \theta_w^{(0)}\|_{c,\infty}^4$$

uniformly over all $\overline{\theta}_w \in \mathcal{N}_0$. To this end, we note that

$$\begin{aligned}
&\left| \frac{\partial^4}{\partial \tau^4} \log\left(\mathbb{P}(y \mid X; \overline{\theta}_w + \tau(\theta_w - \theta_w^{(0)}))\right) \bigg|_{\tau=0} \right| \\
&\leq \Bigg(\left| \frac{6}{\min\{\underline{f}, 1-\overline{f}\}^{4T}} \left(\frac{\partial}{\partial \tau} \mathbb{P}(y \mid X; \overline{\theta}_w + \tau \mathbf{1}) \Big|_{\tau=0} \right)^4 \right| \\
&\quad + \left| \frac{12}{\min\{\underline{f}, 1-\overline{f}\}^{3T}} \left(\frac{\partial}{\partial \tau} \mathbb{P}(y \mid X; \overline{\theta}_w + \tau \mathbf{1}) \Big|_{\tau=0} \right)^2 \frac{\partial^2}{\partial \tau^2} \mathbb{P}(y \mid X; \overline{\theta}_w + \tau \mathbf{1}) \Big|_{\tau=0} \right| \\
&\quad + \left| \frac{3}{\min\{\underline{f}, 1-\overline{f}\}^{2T}} \left(\frac{\partial^2}{\partial \tau^2} \mathbb{P}(y \mid X; \overline{\theta}_w + \tau \mathbf{1}) \Big|_{\tau=0} \right)^2 \right| \\
&\quad + \left| \frac{4}{\min\{\underline{f}, 1-\overline{f}\}^{2T}} \left(\frac{\partial}{\partial \tau} \mathbb{P}(y \mid X; \overline{\theta}_w + \tau \mathbf{1}) \Big|_{\tau=0} \right) \frac{\partial^3}{\partial \tau^3} \mathbb{P}(y \mid X; \overline{\theta}_w + \tau \mathbf{1}) \Big|_{\tau=0} \right| \\
&\quad + \left| \frac{1}{\min\{\underline{f}, 1-\overline{f}\}^{T}} \frac{\partial^4}{\partial \tau^4} \mathbb{P}(y \mid X; \overline{\theta}_w + \tau \mathbf{1}) \Big|_{\tau=0} \right| \Bigg) \|\theta_w - \theta_w^{(0)}\|_{c,\infty}^4
\end{aligned}$$

[27] Since the derivation is neither interesting nor providing any intuition, we exclude it. If desired, it is available upon request.

It remains to argue that the respective derivatives can be bounded. To avoid tedious calculus, we observe (abusing the notation a bit) that

$$\frac{\partial}{\partial w_{j^*}} L_j(X,w) = \begin{cases} L_j(X;w)(1-L_j(X;w)) & \text{if } j^* = j \\ -L_j(X;w)L_{j^*}(X;w) & \text{otherwise} \end{cases}$$

so that $L_j(X;w)$ is four-times continuously differentiable in $(w_1(X),\ldots,w_{J-1}(X))$ (at a fixed X) with all derivatives involving $L_j(X;w)$ and sums or products thereof. Hence, all these derivatives can be bounded above in absolute value. Combining this with the fact that $F_{jt}(\cdot)$ is four-times continuously differentiable and that $\mathcal{B}$ as well as $\mathbb{X}$ are compact, we observe that there exists a $C<\infty$ such that $|\frac{\partial^p}{\partial s^p}F_{jt}(s)|_{s=x_t^\top\beta_{j,t}+\alpha_{j,t}}| < C$ uniformly over $\mathbb{X}$ and $\mathcal{B}$ for $p\in\{0,1,2,3,4\}$. Since all of the above derivatives involve only finite sums of products of these terms and these terms are individually bounded above uniformly over $\Theta^{(w)}$ and $\Theta^{(w)}_{d(n)}$, we conclude that there exists some finite constant C^* such that

$$\left|\frac{\partial}{\partial\tau}\,\mathbb{P}(y\mid X;\overline{\theta}_w+\tau\mathbf{1})\Big|_{\tau=0}\right| \le C^* \;;\quad \left|\frac{\partial^2}{\partial\tau^2}\,\mathbb{P}(y\mid X;\overline{\theta}_w+\tau\mathbf{1})\Big|_{\tau=0}\right| \le C^*$$
$$\left|\frac{\partial^3}{\partial\tau^3}\,\mathbb{P}(y\mid X;\overline{\theta}_w+\tau\mathbf{1})\Big|_{\tau=0}\right| \le C^* \;;\quad \left|\frac{\partial^4}{\partial\tau^4}\,\mathbb{P}(y\mid X;\overline{\theta}_w+\tau\mathbf{1})\Big|_{\tau=0}\right| \le C^*$$

uniformly over $\Theta^{(w)}$ and $\Theta^{(w)}_{d(n)}$. Hence, Assumption 23 in Hu and Schennach (2008) is satisfied.

The remainder of the proof follows from the same arguments as the proof of Theorem 3 in Hu and Schennach (2008) in combination with Shen (1997). These arguments imply that for every $\lambda\neq 0$

$$\begin{aligned}\sqrt{n}\lambda^\top(\hat{\beta}_n-\beta^{(0)}) = \sqrt{n}\langle v^*(\lambda),\hat{\theta}_{w,n}-\theta_w^{(0)}\rangle &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{d\ell(y_i,X_i;\theta_w^{(0)})}{d\theta_w}[v^*(\lambda)] + o_p(1)\\ &= \lambda^\top(V^*)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n D_{\mu^*}(y_i,X_i)^\top + o_p(1)\end{aligned}$$

Since $\mathbb{E}\left[\left(\frac{d\ell(y_i,X_i;\theta_w^{(0)})}{d\theta_w}[v^*(\lambda)]\right)^2\right] = \|v^*(\lambda)\|^2 = \lambda^\top(V^*)^{-1}\lambda$, the CLT in combination with the Cramér-Wold device now implies the desired result. We omit further details here and conclude the proof. □

A.4.3.2 Asymptotic normality of $\hat{\pi}_j$

We derive the limiting distribution of $\sqrt{n}(\hat{\pi}_j - \pi_j^{(0)})$. In the proof, the following functional plays a crucial role:

$$f_{\pi_j}(\theta_w) = \mathbb{E}[\pi_j(X_i) - \pi_j^{(0)}(X_i)] = \begin{cases} \mathbb{E}\left[\frac{\exp(w_j(X_i))}{1+\sum_{j=1}^{J-1}\exp(w_j(X_i))} - \pi_j^{(0)}(X_i)\right] & \text{if } j = 1, \ldots, J-1 \\ \mathbb{E}\left[\frac{1}{1+\sum_{j=1}^{J-1}\exp(w_j(X_i))} - \pi_j^{(0)}(X_i)\right] & \text{otherwise} \end{cases}$$

for $\theta_w \in \Theta_{d(n)}^{(w)}$. In the proof, we argue that $f_{\pi_j}(\theta_w)$ can be approximated by the following linear functional

$$f'_{\pi_j,\theta_w^{(0)}}[\theta_w - \theta_w^{(0)}] = \begin{cases} \mathbb{E}\left[\pi_j^{(0)}(X_i)(1-\pi_j^{(0)}(X_i))(w_j(X_i) - w_j^{(0)}(X_i)) \right. & \text{if } j = 1, \ldots, J-1 \\ \left. -\pi_j^{(0)}(X_i)\sum_{j^* \neq j}^{J-1}\pi_{j^*}^{(0)}(X_i)(w_{j^*}(X_i) - w_{j^*}^{(0)}(X_i))\right] & \\ -\mathbb{E}\left[\pi_J^{(0)}(X_i)\sum_{j=1}^{J-1}\pi_j^{(0)}(X_i)(w_j(X_i) - w_j^{(0)}(X_i))\right] & \text{otherwise} \end{cases}$$

in an appropriate sense and where $\sum_{j^* \neq j}^{J-1}$ denotes the sum over $j^* \in \{1, \ldots, J-1\} \setminus \{j\}$. On a heuristic level, we will argue that for all $j = 1, \ldots, J$ $f'_{\pi_j,\theta_w^{(0)}}[\theta_w - \theta_w^{(0)}]$ is bounded on $(\overline{\boldsymbol{V}}, \|\cdot\|)$ so there exists a Riesz representer $v_{\pi_j}^* \in \overline{\boldsymbol{V}}$ such that, for all $\theta_w \in \Theta_0^{(w)}$, $f'_{\pi_j,\theta_w^{(0)}}[\theta_w - \theta_w^{(0)}] = \langle v_{\pi_j}^*, \theta_w - \theta_w^{(0)}\rangle$. In a future version of this paper, it may be interesting to see whether the Riesz representer can be explicitly characterized. Although we do not study the estimation of the Riesz representer in this paper, we note that one may follow the arguments in, for instance, Chen and Liao (2015) to estimate the Riesz representer.

We make the following assumption, which mirrors Assumption A-9 in the previous section.

Assumption A-11 *(Approximation of Riesz representer) For all $j = 1, \ldots, J$, there is $v_{n,\pi_j}^* \in \Theta_{d(n)}^{(w)} - \{\theta_w^{(0)}\}$ such that $\|v_{n,\pi_j}^* - v_{\pi_j}^*\| = o(n^{-1/4})$.*

The same remarks below Assumption A-9 apply to Assumption A-11. To control the empirical process term in the asymptotic representation of the estimator, we make the following standard assumption.

Assumption A-12 *(Donsker property of parameter space) For the parameter spaces of the component weights, $2\eta > d_X$.*

To control the linearization error of the above functional, we assume:

Assumption A-13 *(Linearization error) The maximal eigenvalue of $\mathbb{E}[\rho^{d(n)}(X)\rho^{d(n)}(X)^\top]$ is bounded above uniformly over $d(n)$.*

Assumption A-13 is satisfied when $\{\rho^d(X)\}_{d=0}^{\infty}$ is orthonormal on $\mathbb{X}$ with respect to Lebesgue measure and the density of X_i is bounded above and greater than 0 almost surely. Both are standard assumptions in the literature.

Lastly, following Chen and Pouzo (2015), we define $\boldsymbol{V}$ as the linear span of $\Theta_0^{(w)} - \{\theta_w^{(0)}\}$, which can be endowed with the consistency norm $\|\cdot\|_{c,\infty}$ or the Fisher norm $\|\cdot\|$. We recall that $\overline{\boldsymbol{V}}$ is the closure of $\boldsymbol{V}$ under $\|\cdot\|$. We define a sieve analog of $\overline{\boldsymbol{V}}$ with $\overline{\boldsymbol{V}}_{d(n)} = \mathbb{R}^{d_\beta} \times \{\gamma_{1,n}^\top \rho^{d(n)}(\cdot) : \gamma_{1,n} \in \mathbb{R}^{d(n)}\} \times \cdots \times \{\gamma_{J-1,n}^\top \rho^{d(n)}(\cdot) : \gamma_{J-1,n} \in \mathbb{R}^{d(n)}\}$. Generally, the choice of basis functions does not need to be the same as in the original series space $\Theta_{d(n)}^{(w)}$ as long as $\Omega_{d(n)}$ (as defined in Equation (A.11)) based on the other set of basis functions has eigenvalues that are bounded away from 0 uniformly over $d(n)$. For notational convenience, we choose the same set of basis functions as in $\Theta_{d(n)}^{(w)}$. The next assumption in combination with Lemma 3.3(2) in Chen and Pouzo (2015) implies that $f'_{\pi_j,\theta_w^{(0)}}[v]$ is bounded on $(\boldsymbol{V}, \|\cdot\|)$, which ensures the existence of the Riesz representer $v_{\pi_j}^* \in \overline{\boldsymbol{V}}$.[28] Assumption A-14 is weak.

Assumption A-14 *(Existence Riesz representer)* $\{\overline{\boldsymbol{V}}_d\}_{d=0}^{\infty}$ *is dense in* $(\boldsymbol{V}, \|\cdot\|_{c,\infty})$.

We conclude this section with the following theorem.

Theorem A.4.4 *Under the Assumptions of Theorem A.4.2, Assumptions A-6 to A-8 and Assumptions A-10 to A-14,* $\sqrt{n}(\hat{\pi}_j - \pi_j^{(0)}) \xrightarrow{d} N(0, \Pi_j)$ *for* $j = 1, \ldots, J$ *with* $\Pi_j = \mathbb{E}\left[\left\{\pi_j(X) - \mathbb{E}[\pi_j^{(0)}(X)] + \frac{d\ell(y,X;\theta_w^{(0)})}{d\theta_w}[v_{\pi_j}^*]\right\}^2\right]$.

A direct corollary of the proof of Theorem A.4.4 is the following.

Corollary A.4.4.1 *Given the Assumptions of Theorem A.4.2, Assumptions A-7 and A-13, and that for all* $j = 1, \ldots, J-1$, *there exists* $\Pi_n w_j^{(0)} \in \mathcal{W}_{j,d(n)}$ *such that* $\|\Pi_n w_j^{(0)} - w^{(0)}\|_{L_2(P_0)} = o(n^{-1/4})$, *then* $\sum_{j=1}^{J-1} \|\hat{w}_j - w_j^{(0)}\|_{L_2(P_0)} = o_p(n^{-1/4})$ *and* $\sum_{j=1}^{J-1} \|\hat{\pi}_j - \pi_j^{(0)}\|_{L_2(P_0)} = o_p(n^{-1/4})$.

As the proof of Corollary A.4.4.1 is part of the proof of Theorem A.4.4, we omit a standalone proof. The second part of the claim follows from the Lipschitz continuity of the link function; similar to the proof of Theorem 4.1. It is important to note that Corollary A.4.4.1 can be applied to other settings as well and therefore has interesting implications that go beyond the scope of this paper. In particular, Corollary A.4.4.1 provides conditions under which $n^{1/4}$- convergence under the Fisher norm can be translated into $n^{1/4}$-convergence under the stronger $L_2(P_0)$-norm. This result is helpful and important when proving the asymptotic normality of functionals of the infinite-dimensional parameter $w^{(0)}$.

[28] When $f'_{\pi_j,\theta_w^{(0)}}[v]$ is bounded on $(\boldsymbol{V}, \|\cdot\|)$, there exists a unique extension of $f'_{\pi_j,\theta_w^{(0)}}[v]$ from $(\boldsymbol{V}, \|\cdot\|)$ to $(\overline{\boldsymbol{V}}, \|\cdot\|)$ that is bounded on $(\overline{\boldsymbol{V}}, \|\cdot\|)$, which implies the existence of the Riesz representer.

Proof. We rewrite

$$\sqrt{n}(\hat{\pi}_j - \pi_j^{(0)}) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \pi_j^{(0)}(X_i) - \mathbb{E}[\pi_j^{(0)}(X_i)]\right) + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n(\hat{\pi}_j(X_i) - \mathbb{E}[\hat{\pi}_j(X_i)]) - \frac{1}{\sqrt{n}}\sum_{i=1}^n(\pi_j^{(0)}(X_i) - \mathbb{E}[\pi_j^{(0)}(X_i)])}_{(I)} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n(\mathbb{E}[\hat{\pi}_j(X_i)] - \mathbb{E}[\pi_j^{(0)}(X_i)])}_{(II)}$$

where the expectations treat $\hat{\pi}_j(\cdot)$ as a fixed function and $\pi_j^{(0)} = \mathbb{E}[\pi_j^{(0)}(X_i)]$.

First, we show that $(I) = o_p(1)$. To do so, we rewrite

$$(I) = \frac{1}{\sqrt{n}}\sum_{i=1}^n(L_j(X_i;\hat{w}) - \mathbb{E}[L_j(X_i;\hat{w})]) - \frac{1}{\sqrt{n}}\sum_{i=1}^n(L_j(X_i;w^{(0)}) - \mathbb{E}[L_j(X_i;w^{(0)})])$$

Consider the class of functions $\mathcal{Q}_j := \{L_j(X;w) : w \in \mathcal{W} = \Lambda_M^\eta(\mathbb{X}) \times \cdots \times \Lambda_M^\eta(\mathbb{X})\}$ for $j = 1,\ldots,J$. For any $w_1, w_2 \in \mathcal{W}$, we have $|L_j(X,w_1) - L_j(X,w_2)| \leq \sum_{j=1}^{J-1}|w_{1,j}(X) - w_{2,j}(X)|$ from a mean-value theorem along the same lines as in the final part of the proof of Theorem 4.1. Upon noticing that under Assumption A-12 $\mathcal{W}$ is Donsker by Corollary 2.7.2 and Theorem 2.5.6 in van der Vaart and Wellner (2023), we therefore conclude that $\mathcal{Q}_j$ is Donsker by Theorem 2.10.8 in van der Vaart and Wellner (2023) for all $j = 1,\ldots,J$. At the same time, Theorem 4.1 implies that $\|\hat{\pi}_j - \pi_j^{(0)}\|_{L_2(P_0)} = o_p(1)$. Now, Lemma 19.24 in van der Vaart (1998) implies $(I) = o_p(1)$.

Next, we turn to (II) and, using Assumption E-1, rewrite

$$(II) = \underbrace{\sqrt{n}f'_{\pi_j,\theta_w^{(0)}}[\hat{\theta}_{w,n} - \theta_w^{(0)}]}_{(II.A)} + \underbrace{\sqrt{n}(\mathbb{E}[\hat{\pi}_j(X) - \pi_j^{(0)}(X)] - f'_{\pi_j,\theta_w^{(0)}}[\hat{\theta}_{w,n} - \theta_w^{(0)}])}_{(II.B)}$$

We first argue that $(II.B) = o_p(1)$. To this end, we fix X, w_1, and w_2, and consider the function $q_j(X,w_1,w_2;\tau) = L_j(X;w_2 + \tau(w_1 - w_2))$ as a function of τ. We observe that $f'_{\pi_j,\theta_w^{(0)}}[\hat{\theta}_{w,n} - \theta_w^{(0)}] = \mathbb{E}[\frac{\partial}{\partial\tau}q_j(X,\hat{w},w^{(0)};\tau)|_{\tau=0}]$ so that $(II.B) = \mathbb{E}[\hat{\pi}_j(X) - \pi_j^{(0)}(X) - \frac{\partial}{\partial\tau}q_j(X,\hat{w},w^{(0)};\tau)|_{\tau=0}]$ for $j = 1,\ldots,J$. A second-order Taylor expansion of $q_j(X;1)$ around 0 yields at a fixed X

$$\hat{\pi}_j(X) - \pi_j^{(0)}(X) - \frac{\partial}{\partial\tau}q_j(X,\hat{w},w^{(0)};\tau)|_{\tau=0} = \frac{\partial^2}{\partial\tau^2}q_j(X,\hat{w},w^{(0)};\tau)\Big|_{\tau=\tilde{\tau}}$$

for some $\tilde{\tau} \in (0,1)$. Some calculus in combination with standard inequalities and the fact that $L_j(X;w^{(0)} + \tilde{\tau}(w - w^{(0)})) \in [0,1]$ shows that there exists some finite constant C such

that

$$\left|\frac{\partial^2}{\partial\tau^2}q_j(X,\hat{w},w^{(0)};\tau)\Big|_{\tau=\tilde{\tau}}\right| \leq C\sum_{j=1}^{J-1}(\hat{w}_j(X)-w_j^{(0)}(X))^2$$

for all $j=1,\ldots,J$. Thus, letting $\Pi_n w_j^{(0)} = \gamma_{j,n}^{(0)\top}\rho^{d(n)}(X)$ denote the projection of $w_j^{(0)}$ onto $\Theta_{d(n)}^{(w)}$, we have

$$\begin{aligned}
|(II.B)| &\leq \sqrt{n}C\sum_{j=1}^{J-1}\mathbb{E}[(\hat{w}_j(X_i)-w_j^{(0)}(X_i))^2] \\
&\leq 2\sqrt{n}C\sum_{j=1}^{J-1}\mathbb{E}[((\hat{\gamma}_{j,n}-\gamma_{j,n}^{(0)})^\top\rho^{d(n)}(X_i))^2] + 2\sqrt{n}C\sum_{j=1}^{J-1}\|\Pi_n w_j^{(0)}-w_j^{(0)}\|_{L_2(P_0)}^2 \\
&\leq 2\sqrt{n}C\sum_{j=1}^{J-1}\mathbb{E}[((\hat{\gamma}_{j,n}-\gamma_{j,n}^{(0)})^\top\rho^{d(n)}(X_i))^2] + 2\sqrt{n}C\|\Pi_n\theta_w^{(0)}-\theta^{(0)}\|_{c,\infty}^2
\end{aligned}$$

Under Assumption A-8, $\|\Pi_n\theta_w^{(0)}-\theta^{(0)}\|_{c,\infty}^2 = o(n^{-1/2})$ so $2\sqrt{n}C\|\Pi_n\theta_w^{(0)}-\theta^{(0)}\|_{c,\infty}^2 = o(1)$. For the first term, we let $\hat{\xi}_n = (\hat{\beta}^\top,\hat{\gamma}_{1,n}^\top,\ldots,\hat{\gamma}_{J-1,n}^\top)^\top$ and $\xi_n^{(0)} = (\beta^{(0)\top},\gamma_{1,n}^{(0)\top},\ldots,\gamma_{J-1,n}^{(0)\top})^\top$ so that $\|\hat{\theta}_{w,n}-\Pi_n\theta_w^{(0)}\|^2 = (\hat{\xi}_n-\xi_n^{(0)})^\top\Omega_{d(n)}(\hat{\xi}_n-\xi_n^{(0)})$ with $\Omega_{d(n)}$ defined in equation (A.11). Under Assumption A-7 there exists a positive finite constant c that does not depend on the sample size such that $c\|\hat{\xi}_n-\xi_n^{(0)}\|_E^2 \leq \|\hat{\theta}_w-\Pi_n\theta_w^{(0)}\|^2$. At the same time, under Assumption A-13

$$\begin{aligned}
\sqrt{n}\sum_{j=1}^{J-1}\mathbb{E}[((\hat{\gamma}_{j,n}-\gamma_{j,n}^{(0)})^\top\rho^{d(n)}(X_i))^2] &= \sqrt{n}\sum_{j=1}^{J-1}(\hat{\gamma}_{j,n}-\gamma_{j,n}^{(0)})^\top\,\mathbb{E}[\rho^{d(n)}(X_i)\rho^{d(n)}(X_i)^\top](\hat{\gamma}_{j,n}-\gamma_{j,n}^{(0)}) \\
&\leq C\sqrt{n}\sum_{j=1}^{J-1}\|\hat{\gamma}_{j,n}-\gamma_{j,n}^{(0)}\|_E^2 \leq C\sqrt{n}\|\hat{\xi}_n-\xi_n^{(0)}\|_E^2 \\
&\leq C\sqrt{n}\|\hat{\theta}_{w,n}-\Pi_n\theta_w^{(0)}\|^2 \\
&\leq C\sqrt{n}\|\hat{\theta}_{w,n}-\theta_w^{(0)}\|^2 + C\sqrt{n}\|\theta_w^{(0)}-\Pi_n\theta_w^{(0)}\|^2 \\
&\leq C\sqrt{n}\|\hat{\theta}_{w,n}-\theta_w^{(0)}\|^2 + C\sqrt{n}\|\theta_w^{(0)}-\Pi_n\theta_w^{(0)}\|_{c,\infty}^2 = o_p(1)
\end{aligned}$$

where C is a finite constant that may change from line to line and does not depend on n. The final equality follows from Assumption A-8 and Theorem A.4.2. We therefore conclude that $(II.B) = o_p(1)$.

We now turn to $(II.A)$. We aim to use Theorem 1 in Shen (1997). To this end, we first observe that equation (4.2) in Shen (1997) is trivially satisfied because $f'_{\pi_j,\theta_w^{(0)}}[\theta_w-\theta_w^{(0)}]$ is linear in $\theta_w-\theta_w^{(0)}$. Second, we have to argue that $f'_{\pi_j,\theta_w^{(0)}}[v]$ is bounded on $(\boldsymbol{V},\|\cdot\|)$. We do so using Lemma 3.3(2) in Chen and Pouzo (2015). To do so, we have to check (i) $\{\overline{\boldsymbol{V}}_d\}_{d=0}^\infty$ is dense in $(\boldsymbol{V},\|\cdot\|_{c,\infty})$, (ii) the sieve Riesz representer $\tilde{v}_{n,\pi_j}^* \in \overline{\boldsymbol{V}}_{d(n)}$, which we define below, is

bounded on $(\overline{\boldsymbol{V}}_{d(n)}, \|\cdot\|)$ uniformly over $d(n)$, and (iii) $f'_{\pi_j,\theta_w^{(0)}}[v]$ is bounded on $(\boldsymbol{V}, \|\cdot\|_{c,\infty})$. (i) is satisfied by Assumption A-14. To show (ii), we abuse the notation a bit and let $\xi_n = (\beta^\top, \gamma_{1,n}^\top, \ldots, \gamma_{J-1,n}^\top)^\top$ for $v \in \overline{\boldsymbol{V}}_{d(n)}$ with $v_n(X) = (\beta^\top, \gamma_{1,n}^\top \rho^{d(n)}(X), \ldots, \gamma_{J-1,n}^\top \rho^{d(n)}(X))^\top$ so that $\|v\|^2 = \xi_n^\top \Omega_{d(n)} \xi_n$. Then, for $j = 1, \ldots, J-1$, we consider the sieve Riesz representer $\tilde{v}^*_{n,\pi_j} \in \overline{\boldsymbol{V}}_{d(n)}$ defined by (see equation (3.6.) in Chen and Pouzo (2015))

$$\begin{aligned}
\|\tilde{v}^*_{n,\pi_j}\|^2 = & \sup_{v \in \overline{\boldsymbol{V}}_{d(n)}: \|v\| \neq 0} \Big(\mathbb{E}\Big[\pi_j^{(0)}(X_i)(1-\pi_j^{(0)}(X_i))\gamma_{j,n}^\top \rho^{d(n)}(X_i) \\
& - \pi_j^{(0)}(X_i) \sum_{j^* \neq j}^{J-1} \pi_{j^*}^{(0)}(X_i) \gamma_{j^*,n}^\top \rho^{d(n)}(X_i)\Big]\Big)^2 / (\xi_n^\top \Omega_{d(n)} \xi_n) \\
\leq & C \sup_{v \in \overline{\boldsymbol{V}}_{d(n)}: \|v\| \neq 0} \left(\sum_{j=1}^{J-1} \gamma_{j,n}^\top \mathbb{E}[\rho^{d(n)}(X_i)\rho^{d(n)}(X_i)^\top] \gamma_{j,n} \right) / (\xi_n^\top \Omega_{d(n)} \xi_n) \\
\leq & C \sup_{\gamma_n = \in \mathbb{R}^{(J-1)d_X}: \|\gamma\|_E \neq 0} \sum_{j=1}^{J-1} \|\gamma_{j,n}\|_E^2 / \|\gamma_n\|_E^2 = C
\end{aligned}$$

where C varies from inequality to inequality and does not depend on n. The first inequality follows from multiple applications of Jensen's inequality and the fact that $\pi_j^{(0)}(X_i) \in (0,1)$. The second inequality follows from Assumptions A-7 and A-13 and the fact that $\|\xi_n\|_E^2 \geq \|\gamma_n\|_E^2$ where $\gamma_n = (\gamma_{1,n}^\top, \ldots, \gamma_{J-1,n}^\top)^\top$. As a similar argument applies to $\|v^*_{\pi_J}\|$, we conclude that the sieve Riesz representer is bounded uniformly over $d(n)$ for all $j = 1, \ldots, J$. For (iii), we observe

$$\begin{aligned}
& \sup_{v \in \boldsymbol{V}: \|v\|_{c,\infty} \neq 0} \frac{|f'_{\pi_j,\theta_w^{(0)}}[v]|}{\|v\|_{c,\infty}} \\
\leq & \sup_{v \in \boldsymbol{V}: \|v\|_{c,\infty} \neq 0} C \frac{\sum_{j=1}^{J-1} \mathbb{E}[|w_j(X_i) - w_j^{(0)}(X_i)|]}{\sum_{j=1}^{J} \sum_{t=1}^{T} \|\beta - \beta^{(0)}\|_E^2 + \sum_{j=1}^{J-1} \|w_j - w_j^{(0)}\|_\infty} \leq C
\end{aligned}$$

where C varies from line to line. The first inequality follows from multiple applications of Jensen's inequality and $\pi_j^{(0)}(\cdot) \in (0,1)$. The second inequality follows from setting $\sum_{j=1}^{J} \sum_{t=1}^{T} \|\beta - \beta^{(0)}\|_E^2$ to zero and the fact that $\|\cdot\|_{L_1(P_0)} \leq \|\cdot\|_\infty$. Hence, (iii) is satisfied. Therefore, Lemma 3.3(2) in Chen and Pouzo (2015) implies that $f'_{\pi_j,\theta_w^{(0)}}[v]$ is bounded on $(\boldsymbol{V}, \|\cdot\|)$.

Now, to apply Theorem 1 in Shen (1997), we have to check Conditions A to D of the paper. This can be done by following identical arguments as in the proofs of Theorem A.4.3 above and Theorem 3 in the Online Appendix of Hu and Schennach (2008). We therefore conclude that the assertion of Theorem 1 in Shen (1997) applies, that is,

$$(II.A) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[v^*_{\pi_j}] + o_p(1)$$

Therefore,

$$\sqrt{n}(\hat{\pi}_j - \pi_j^{(0)}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\pi_j^{(0)}(X_i) - \mathbb{E}[\pi_j^{(0)}(X_i)] + \frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[v_{\pi_j}^*] \right) + o_p(1)$$

where $\mathbb{E}\left[\pi_j^{(0)}(X_i) - \mathbb{E}[\pi_j^{(0)}(X_i)] + \frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[v_{\pi_j}^*]\right] = 0$ and $\Pi_j \leq 1 + \|v_{\pi_j}^*\|^2 < \infty$ where the inequality follows from $\pi_j^{(0)}(X) \in (0,1)$ and the definition of the Fisher norm. Now, the conclusion of Theorem A.4.4 follows from a standard CLT for iid data. This concludes the proof. □

A.4.3.3 Asymptotic normality of $\widehat{\text{AME}}$

In this section, we focus on two objects of interest.

$$\widehat{\text{AME}}_{j,t} = \frac{1}{n\hat{\pi}_j} \sum_{i=1}^{n} \hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \hat{\tilde{\beta}}_{j,t}) \hat{\beta}_{j,t} \; ; \quad \widehat{\text{AME}}_j = \frac{1}{T} \sum_{t=1}^{T} \widehat{\text{AME}}_{j,t}$$

with population analogs

$$\text{AME}_{j,t} = \frac{\mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) \beta_{j,t}^{(0)}]}{\pi_j^{(0)}} = \mathbb{E}\left[\frac{\partial}{\partial x_{it}} \mathbb{P}_t(y_{it} = 1 \mid x_{it}, g_i = j) \mid g_i = j \right]$$

$$\text{AME}_j = \frac{1}{T} \sum_{t=1}^{T} \text{AME}_{j,t}$$

where $\pi_j^{(0)} = \mathbb{E}[\pi_j^{(0)}(X_i)]$. As the asymptotic normality result for $\widehat{\text{AME}}_j$ follows immediately once we have established such a result for $\widehat{\text{AME}}_{j,t}$, we focus on the latter estimator for the most part of this section. The following functional plays a crucial role in the proof:

$$f_{\text{AME}_{j,t}}(\theta_w) = \mathbb{E}\left[\left(\pi_j(X_i) - \pi_j^{(0)}(X_i) \right) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) \right]$$

for $\theta_w \in \Theta_{d(n)}^{(w)}$. In the proof, we argue that $f_{\text{AME}_{j,t}}(\theta_w)$ can be approximated (in an appropriate sense) by the following linear functional: For $j = 1, \ldots, J-1$

$$\begin{aligned} f'_{\text{AME}_{j,t}, \theta_w^{(0)}}[\theta_w - \theta_w^{(0)}] = \; & \mathbb{E}\left[F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) \left[\pi_j^{(0)}(X_i)(1 - \pi_j^{(0)}(X_i))(w_j(X_i) - w_j^{(0)}(X_i)) \right.\right. \\ & \left.\left. - \pi_j^{(0)}(X_i) \sum_{j^* \neq j}^{J-1} \pi_{j^*}^{(0)}(X_i)(w_{j^*}(X_i) - w_{j^*}^{(0)}(X_i)) \right] \right] \end{aligned}$$

where $\sum_{j^* \neq j}^{J-1}$ denotes the sum over $j^* \in \{1, \ldots, J-1\} \setminus \{j\}$ and

$$f'_{\text{AME}_{J,t}, \theta_w^{(0)}}[\theta_w - \theta_w^{(0)}] = -\mathbb{E}\left[F'_{Jt}(\tilde{x}_{it}^\top \tilde{\beta}_{J,t}^{(0)}) \left[\pi_J^{(0)}(X_i) \sum_{j=1}^{J-1} \pi_j^{(0)}(X)(w_j(X_i) - w_j^{(0)}(X_i)) \right] \right]$$

Recalling Fact 3 from Appendix A.4, we know that $|F'_{jt}(\tilde{x}_t^\top \tilde{\beta}_{j,t}^{(0)})| \leq C_F$ uniformly over $\mathbb{X}$. It is therefore easy to see that for all $j = 1, \ldots, J$

$$|f'_{\text{AME}_{j,t},\theta_w^{(0)}}[\theta_w - \theta_w^{(0)}]| \leq C_F |f'_{\pi_j,\theta_w^{(0)}}[\theta_w - \theta_w^{(0)}]|$$

Since the latter functional is bounded on $(\boldsymbol{V}, \|\cdot\|)$ for $j = 1, \ldots, J$, we conclude that $f'_{\text{AME}_{j,t},\theta_w^{(0)}}[v]$ is bounded on $(\boldsymbol{V}, \|\cdot\|)$ for all j, t so that, for all j, t, there exists a Riesz representer $v^*_{\text{AME}_{j,t}} \in \overline{\boldsymbol{V}}$ such that $f'_{\text{AME}_{j,t},\theta_w^{(0)}}[v] = \langle v, v^*_{\text{AME}_{j,t}} \rangle$ for all $v \in \overline{\boldsymbol{V}}$. Analogously to Assumptions A-9 and A-11, we make the following assumption.

Assumption A-15 *(Approximation of Riesz representer) For all $j = 1, \ldots, J$ and $t = 1, \ldots, T$, there is $v^*_{n,AME_{j,t}} \in \Theta_{d(n)}^{(w)} - \{\theta_w^{(0)}\}$ such that $\|v^*_{n,AME_{j,t}} - v^*_{AME_{j,t}}\| = o(n^{-1/4})$.*

The same remarks as for Assumptions A-9 and A-11 apply here. We are ready to state our final theorem.

Theorem A.4.5 *Under the Assumptions of Theorem A.4.2 and Assumptions A-6 to A-15, $\sqrt{n}(\widehat{AME}_{j,t} - AME_{j,t}) \xrightarrow{d} N(0, A_{j,t})$ with $A_{j,t} = \mathbb{E}\left[IF_{\widehat{AME}_{j,t}}(y_i, X_i) IF_{\widehat{AME}_{j,t}}(y_i, X_i)^\top\right]$ for all $j = 1, \ldots, J$ and $t = 1, \ldots, T$ and with $IF_{\widehat{AME}_{j,t}}(y_i, X_i)$ defined in equation (A.15).*

An estimator for $A_{j,t}$ may be constructed via a sample analog estimator with the Riesz representers being as estimated as discussed in Ackerberg et al. (2012) and Chen and Liao (2015). A direct corollary of Theorem A.4.5 is the following.

Corollary A.4.5.1 *Under the Assumptions of Theorem A.4.5, $\sqrt{n}(\widehat{AME}_j - AME_j) \xrightarrow{d} N(0, A_j)$ with $A_j = \mathbb{E}\left[\left(T^{-1}\sum_{t=1}^T IF_{\widehat{AME}_{j,t}}(y_i, X_i)\right)\left(T^{-1}\sum_{t=1}^T IF_{\widehat{AME}_{j,t}}(y_i, X_i)\right)^\top\right]$ for all $j = 1, \ldots, J$.*

Since the proof of Corollary A.4.5.1 is immediate given Theorem A.4.5, we omit the proof. We now prove Theorem A.4.5.

Proof. Letting $\tilde{x}_{it} = (1, x_{it}^\top)^\top$ and $\hat{\tilde{\beta}}_{j,t} = (\hat{\alpha}_{j,t}, \hat{\beta}_{j,t}^\top)^\top$, we rewrite

$$\begin{aligned}
\widehat{\text{AME}}_{j,t} = & \underbrace{\frac{1}{n\pi_j^{(0)}} \sum_{i=1}^n \hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) \beta_{j,t}^{(0)}}_{(I)} + \underbrace{\frac{1}{n\hat{\pi}_j} \sum_{i=1}^n \hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \hat{\tilde{\beta}}_{j,t})(\hat{\beta}_{j,t} - \beta_{j,t}^{(0)})}_{(II)} \\
& \underbrace{- \frac{1}{\tilde{\pi}_j^2}(\hat{\pi}_j - \pi_j^{(0)}) \frac{1}{n} \sum_{i=1}^n \hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \hat{\tilde{\beta}}_{j,t}) \beta_{j,t}^{(0)}}_{(III)} \\
& \underbrace{+ \frac{1}{n\pi_j^{(0)}} \sum_{i=1}^n \hat{\pi}_j(X_i)[F'_{jt}(\tilde{x}_{it}^\top \hat{\tilde{\beta}}_{j,t}) - F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \beta_{j,t}^{(0)}}_{(IV)}
\end{aligned}$$

where $\tilde{\pi}_j$ is an intermediate value. Given the asymptotic normality results for $\hat{\pi}_j$ and $\hat{\beta}$, term (I) is the most difficult and requires some more work. To see this, we argue that the asymptotic distributions of the other terms are driven by the asymptotic distribution of $\hat{\pi}_j$ and $\hat{\beta}$.

Term (II) From the arguments in the consistency proofs and Theorem A.4.3

$$\sqrt{n}(II) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{jt}^{(0)})]}{\pi_j^{(0)}} \text{IF}_{\hat{\beta}_{j,t}}(y_i, X_i) + o_p(1) \tag{A.14}$$

where $\text{IF}_{\hat{\beta}_{j,t}}(y_i, X_i)$ is equal to the rows in $(V^*)^{-1} D_{\mu^*}(y_i, X_i)^\top$ associated with $\hat{\beta}_{j,t} - \beta_{j,t}^{(0)}$ as defined in the proof of and the discussion leading up to Theorem A.4.3.

Term (III) From the arguments in the consistency proofs, the fact that $\tilde{\pi}_j$ is an intermediate value, and Theorem A.4.4, we have

$$\begin{aligned}\sqrt{n}(III) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\pi_j^{(0)}(X_i) - \mathbb{E}[\pi_j^{(0)}(X_i)] + \frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta} [v_{\pi_j}^*] \right) \\ \times \frac{\mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})]}{(\pi_j^{(0)})^2} \beta_{j,t}^{(0)} + o_p(1)\end{aligned}$$

Term (IV) We rewrite

$$\begin{aligned}(IV) = &\underbrace{\frac{1}{n\pi_j^{(0)}} \sum_{i=1}^{n} \pi_j^{(0)}(X_i) [F'_{jt}(\tilde{x}_{it}^\top \hat{\tilde{\beta}}_{j,t}) - F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \beta_{j,t}^{(0)}}_{(IV.A)} \\ &+ \underbrace{\frac{1}{n\pi_j^{(0)}} \sum_{i=1}^{n} (\hat{\pi}_j(X_i) - \pi_j^{(0)}(X_i)) [F'_{jt}(\tilde{x}_{it}^\top \hat{\tilde{\beta}}_{j,t}) - F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \beta_{j,t}^{(0)}}_{(IV.B)}\end{aligned}$$

where a Taylor expansion gives

$$\begin{aligned}\sqrt{n}(IV.A) &= \beta_{j,t}^{(0)} \frac{1}{n\pi_j^{(0)}} \sum_{i=1}^{n} \pi_j^{(0)}(X_i) F''_{jt}(\tilde{x}_{it}^\top \underline{\tilde{\beta}}_{j,t}) \tilde{x}_{it}^\top \sqrt{n}(\hat{\tilde{\beta}}_{j,t} - \tilde{\beta}_{j,t}^{(0)}) \\ &= \beta_{j,t}^{(0)} \frac{1}{\pi_j^{(0)}} \mathbb{E}[\pi_j^{(0)}(X_i) F''_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) \tilde{x}_{it}^\top] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \text{IF}_{\hat{\tilde{\beta}}_{j,t}}(y_i, X_i) + o_p(1)\end{aligned}$$

with intermediate value $\underline{\tilde{\beta}}_{j,t}$ and $\text{IF}_{\hat{\tilde{\beta}}_{j,t}}(y_i, X_i)$ is equal to the rows in $(V^*)^{-1} D_{\mu^*}(y_i, X_i)^\top$ associated with $\hat{\tilde{\beta}}_{j,t} - \tilde{\beta}_{j,t}^{(0)}$ as defined in the proof of and the discussion leading up to Theorem A.4.3. The remainder converges to 0 by noting that the conditions of Lemma 4.3 in Newey and McFadden (1994) are satisfied for $\frac{1}{n} \sum_{i=1}^{n} \pi_j^{(0)}(X_i) F''_{jt}(\tilde{x}_{it}^\top \underline{\tilde{\beta}}_{j,t}) \tilde{x}_{it}^\top$ – the existence of an integrable dominating function follows from compactness of $\mathcal{B}$ and $\mathbb{X}$ in combination with the fact that $F_{jt}(\cdot)$ is twice continuously differentiable.

Next, we consider some arbitrary entry of $(IV.B)$ with $(IV.B)_k$ for $k = 1, \ldots, K-1$. From a similar Taylor expansion argument, we have

$$\sqrt{n}(IV.B)_k \leq \|\hat{\pi}_j - \pi_j^{(0)}\|_\infty \|\sqrt{n}(\hat{\tilde{\beta}}_{j,t} - \tilde{\beta}_{j,t}^{(0)})\|_E \frac{|\beta_{j,t,k}^{(0)}|}{n\pi_j^{(0)}} \sum_{i=1}^n |F''_{jt}(\tilde{x}_{it}^\top \underline{\tilde{\beta}}_{j,t})| \|\tilde{x}_{it}\|_E = o_p(1)$$

where the last equality follows from $\|\hat{\pi}_j - \pi_j^{(0)}\|_\infty = o_p(1)$, $\|\sqrt{n}(\hat{\tilde{\beta}}_{j,t} - \tilde{\beta}_{j,t}^{(0)})\|_E = O_p(1)$, and an application of Lemma 4.3 in Newey and McFadden (1994) so that $\frac{|\beta_{j,t,k}^{(0)}|}{n\pi_j^{(0)}} \sum_{i=1}^n |F''_{jt}(\tilde{x}_{it}^\top \underline{\tilde{\beta}}_{j,t})| \|\tilde{x}_{it}\| = O_p(1)$.

We therefore have

$$\begin{aligned}
\sqrt{n}(\widehat{\text{AME}}_{j,t} - \text{AME}_{j,t}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Bigg\{ & \frac{\hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})}{\pi_j^{(0)}} \beta_{j,t}^{(0)} - \text{AME}_{j,t} \\
& + \frac{\mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{jt}^{(0)})]}{\pi_j^{(0)}} \sqrt{n}(\hat{\beta}_{j,t} - \beta_{j,t}^{(0)}) \\
& - \sqrt{n}(\hat{\pi}_j - \pi_j^{(0)}) \frac{\mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})]}{(\pi_j^{(0)})^2} \beta_{j,t}^{(0)} \\
& + \beta_{j,t}^{(0)} \frac{1}{\pi_j^{(0)}} \mathbb{E}[\pi_j^{(0)}(X_i) F''_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) \tilde{x}_{it}^\top] \sqrt{n}(\hat{\tilde{\beta}}_{j,t} - \tilde{\beta}_{j,t}^{(0)}) \Bigg\} + o_p(1)
\end{aligned}$$

where the first term is still to be analyzed. To this end, we observe

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})}{\pi_j^{(0)}} \beta_{j,t}^{(0)} - \text{AME}_{j,t} \\
= & \frac{1}{\pi_j^{(0)}} \Bigg(\underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(\hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) - \mathbb{E}[\pi_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \Big)}_{(V)} \Bigg) \beta_{j,t}^{(0)}
\end{aligned}$$

We rewrite (V)

$$\begin{aligned}
(V) = & \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) - \mathbb{E}[\pi_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \Big)}_{(V.A)} \\
& + \Bigg\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(\hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) - \mathbb{E}[\hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \Big) \\
& - \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) - \mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \Big) \Bigg\}
\end{aligned}$$

$$+ \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbb{E}[\hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] - \mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \right)}_{(V.C)}$$

We denote the empirical process term in curly brackets with $(V.B)$. Upon noticing that $|F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t})| \leq C_F$, the same arguments we used to show that $(I) = o_p(1)$ in the proof of Theorem A.4.4 imply that $(V.B) = o_p(1)$. Turning to $(V.C)$, we rewrite using Assumption E-1

$$\begin{aligned}(V.C) = & \underbrace{\sqrt{n} f'_{\mathrm{AME}_{j,t},\theta_w^{(0)}}[\hat{\theta}_{w,n} - \theta_w^{(0)}]}_{(V.C.1)} \\ & + \underbrace{\sqrt{n}(\mathbb{E}[\hat{\pi}_j(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) - \pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] - f'_{\mathrm{AME}_{j,t},\theta_w^{(0)}}[\hat{\theta}_{w,n} - \theta_w^{(0)}])}_{(V.C.2)}\end{aligned}$$

Again using that $|F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t})| \leq C_F$, the identical arguments as used for term $(II.B)$ in the proof of Theorem A.4.4 imply that $(V.C.2) = o_p(1)$. For $(V.C.1)$, we would like to use Theorem 1 in Shen (1997). To this end, we notice that Equation (4.2) in Shen (1997) is trivially satisfied as $f'_{\mathrm{AME}_{j,t},\theta_w^{(0)}}[\theta_w - \theta_w^{(0)}]$ is linear in $\theta_w - \theta_w^{(0)}$. Also, we have already argued that $f'_{\mathrm{AME}_{j,t},\theta_w^{(0)}}[v]$ is bounded on $(\boldsymbol{V}, \|\cdot\|)$ because, under the assumptions, we have shown $f'_{\pi_j,\theta_w^{(0)}}[v]$ to be bounded on $(\boldsymbol{V}, \|\cdot\|)$ in the proof of Theorem A.4.4. It remains to check Conditions A to D of Shen (1997). This can be done in the same way as in the proofs of Theorem A.4.3 and Theorem 3 in the Online Appendix of Hu and Schennach (2008). We therefore conclude that Theorem 1 of Shen (1997) applies and implies

$$(V.C.1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[v^*_{\mathrm{AME}_{j,t}}] + o_p(1)$$

Collecting all terms, we have

$$\begin{aligned}\sqrt{n}(\widehat{\mathrm{AME}}_{j,t} - \mathrm{AME}_{j,t}) = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg\{ \Bigg(\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) - \mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})] \\ & + \frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta_w}[v^*_{\mathrm{AME}_{j,t}}] \Bigg) \frac{1}{\pi_j^{(0)}} \beta_{j,t}^{(0)} \\ & + \frac{\mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{jt}^{(0)})]}{\pi_j^{(0)}} \mathrm{IF}_{\hat{\beta}_{j,t}}(y_i, X_i) \\ & - \left(\pi_j^{(0)}(X_i) - \mathbb{E}[\pi_j^{(0)}(X_i)] + \frac{d\ell(y_i, X_i; \theta_w^{(0)})}{d\theta}[v^*_{\pi_j}] \right) \\ & \times \frac{\mathbb{E}[\pi_j^{(0)}(X_i) F'_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)})]}{(\pi_j^{(0)})^2} \beta_{j,t}^{(0)}\end{aligned}$$

$$+ \beta_{j,t}^{(0)} \frac{1}{\pi_j^{(0)}} \mathbb{E}[\pi_j^{(0)}(X_i) F''_{jt}(\tilde{x}_{it}^\top \tilde{\beta}_{j,t}^{(0)}) \tilde{x}_{it}^\top] \mathrm{IF}_{\hat{\tilde{\beta}}_{j,t}}(y_i, X_i) \Bigg\} + o_p(1)$$

$$=: \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}_{\widehat{\mathrm{AME}}_{j,t}}(y_i, X_i) + o_p(1) \tag{A.15}$$

where $\mathbb{E}[\mathrm{IF}_{\widehat{\mathrm{AME}}_{j,t}}(y_i, X_i)] = 0$ and $\mathbb{E}[\|\mathrm{IF}_{\widehat{\mathrm{AME}}_{j,t}}(y_i, X_i)\|^2] < \infty$ since all Riesz representers have finite norm under $\|\cdot\|$ and all other terms can be bounded under our assumptions. Now, the claim of the theorem follows from a standard CLT for iid data. This concludes the proof. □

A.4.3.4 V^* is positive definite

Following Chen and Pouzo (2015), we define $\boldsymbol{V}$ as the linear span of $\Theta_0^{(w)} - \{\theta_w^{(0)}\}$, which can be endowed with the consistency norm $\|\cdot\|_{c,\infty}$ or the Fisher norm $\|\cdot\|$. We recall that $\overline{\boldsymbol{V}}$ is the closure of $\boldsymbol{V}$ under $\|\cdot\|$. We define a sieve analog of $\overline{\boldsymbol{V}}$ with $\overline{\boldsymbol{V}}_{d(n)} = \mathbb{R}^{d_\beta} \times \{v_{w_1}(\cdot) = \gamma_{1,n}^\top \rho^{d(n)}(\cdot) : \gamma_{1,n} \in \mathbb{R}^{d(n)}\} \times \cdots \times \{v_{w_{J-1}}(\cdot) = \gamma_{J-1,n}^\top \rho^{d(n)}(\cdot) : \gamma_{J-1,n} \in \mathbb{R}^{d(n)}\}$ where for notational convenience, we use the same notation for the basis as previously – the bases need not be the same. We recall the definition of $\Omega_{d(n)}$ from equation (A.11) and partition the matrix

$$\Omega_{d(n)} = \mathbb{E}\left[\left(\frac{d\ell(y, X; \theta_w^{(0)})}{d\theta_w}[\rho^{d(n)}]\right)\left(\frac{d\ell(y, X; \theta_w^{(0)})}{d\theta_w}[\rho^{d(n)}]\right)^\top\right] = \begin{pmatrix} \mathcal{I}_\beta & \mathcal{I}_{n,\beta w} \\ \mathcal{I}_{n,w\beta} & \mathcal{I}_{n,w} \end{pmatrix}$$

where $\rho^{d(n)}(\cdot)$ is the family of basis functions used in $\overline{\boldsymbol{V}}_{d(n)}$ and

$$\mathcal{I}_\beta = \mathbb{E}\left[\left(\frac{d\ell(y, X; \theta_w^{(0)})}{d\beta}\right)\left(\frac{d\ell(y, X; \theta_w^{(0)})}{d\beta}\right)^\top\right]$$

$$\mathcal{I}_{n,\beta w} = \mathbb{E}\left[\left(\frac{d\ell(y, X; \theta_w^{(0)})}{d\beta}\right)\left(\frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\rho^{d(n)}]\right)^\top\right]$$

$$\mathcal{I}_{n,w} = \mathbb{E}\left[\left(\frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\rho^{d(n)}]\right)\left(\frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\rho^{d(n)}]\right)^\top\right]$$

and $\mathcal{I}_{n,w\beta} = \mathcal{I}_{n,\beta w}^\top$ as well as

$$\frac{d\ell(y, X; \theta_w^{(0)})}{dw}[\rho^{d(n)}] = \left(\left(\frac{d\ell(y, X; \theta_w^{(0)})}{dw_1}[\rho^{d(n)}]\right)^\top, \ldots, \left(\frac{d\ell(y, X; \theta_w^{(0)})}{dw_{J-1}}[\rho^{d(n)}]\right)^\top\right)^\top$$

We have the following result.

Lemma A.4.6 *If (i) $\mathcal{I}_{n,w}$ is invertible for all $d(n)$, (ii) the minimal eigenvalue of $\mathcal{I}_\beta -$*

$\mathcal{I}_{n,\beta w}\mathcal{I}_{n,w}^{-1}\mathcal{I}_{n,w\beta}$ is bounded away from zero uniformly over $d(n)$, and (iii) $\{\overline{\boldsymbol{V}}\}_{d=0}^{\infty}$ is dense in $(\boldsymbol{V}, \|\cdot\|_{c,\infty})$, then $\|v^*(\lambda)\| < \infty$ for all λ and V^* is positive definite. If (i) and (ii) are replaced with the assumption that the minimal eigenvalue of $\Omega_{d(n)}$ is bounded away from zero and (iii) continues to hold, the same conclusion holds.

The first two assumptions in the lemma are standard maximum likelihood conditions. Intuitively, the second part requires that the Fisher information on $\beta^{(0)}$ is bounded away from zero. To use the language of the semiparametric efficiency literature, uniformly over the sieve spaces, $\beta^{(0)}$ is not allowed to lie in the nuisance tangent space uniformly over the sieve spaces. The third assumption is satisfied by many series spaces – see, for instance, the discussions in Chen (2007).

Proof. We start showing the claim under assumptions (i) to (iii). We recall the linear functional of interest: $f(\theta_w) = \lambda^\top \beta$. Following Chen and Liao (2015), we restrict the linear functional to $\theta_w \in \overline{\boldsymbol{V}}_{d(n)}$. We let $\xi_n = (\beta^\top, \gamma_{1,n}^\top, \ldots, \gamma_{J-1,n}^\top)^\top$ and consider the squared operator norm of the restricted functional:

$$\sup_{\theta_w \in \overline{\boldsymbol{V}}_{d(n)} \theta_w \neq 0} \frac{(\lambda^\top \beta)^2}{\|\theta_w\|^2} = \sup_{\theta_w \in \overline{\boldsymbol{V}}_{d(n)} \theta_w \neq 0} \frac{(\lambda^\top \beta)^2}{\xi_n^\top \Omega_{d(n)} \xi_n}$$

where the second equality follows from the definition of the norm and sieve space. We minimize the denominator with respect to $\gamma_n = (\gamma_{1,n}^\top, \ldots, \gamma_{J-1,n}^\top)^\top$ by noticing that

$$\xi_n^\top \Omega_{d(n)} \xi_n = \beta^\top \mathcal{I}_\beta \beta + 2\beta^\top \mathcal{I}_{n,\beta w} \gamma_n + \gamma_n^\top \mathcal{I}_{n,w} \gamma_n$$

The minimum is attained at $\gamma_n^* = \mathcal{I}_{n,w}^{-1} \mathcal{I}_{n,w\beta} \beta$ so that

$$\begin{aligned}
\sup_{\theta_w \in \overline{\boldsymbol{V}}_{d(n)} \theta_w \neq 0} \frac{(\lambda^\top \beta)^2}{\|\theta_w\|^2} &\leq \sup_{\theta_w \in \overline{\boldsymbol{V}}_{d(n)} \theta_w \neq 0} \frac{(\lambda^\top \beta)^2}{\beta^\top (\mathcal{I}_\beta - \mathcal{I}_{n,\beta w} \mathcal{I}_{n,w}^{-1} \mathcal{I}_{n,w\beta}) \beta} \\
&\leq \frac{\|\lambda_E\|^2 \|\beta\|_E^2}{\lambda_{min}(\mathcal{I}_\beta - \mathcal{I}_{n,\beta w} \mathcal{I}_{n,w}^{-1} \mathcal{I}_{n,w\beta}) \|\beta\|_E^2} \\
&= \frac{\|\lambda\|_E^2}{\lambda_{\min}(\mathcal{I}_\beta - \mathcal{I}_{n,\beta w} \mathcal{I}_{n,w}^{-1} \mathcal{I}_{n,w\beta})} =: C < \infty
\end{aligned}$$

where $\lambda_{\min}(A)$ denotes the minimal eigenvalue of a matrix A. We therefore conclude that $f(\theta_w)$ is a bounded functional on $(\overline{\boldsymbol{V}}_{d(n)}, \|\cdot\|)$ for all λ. Hence, by the Riesz representation theorem there exists a Riesz representer $v_n^*(\lambda) \in \overline{\boldsymbol{V}}_{d(n)}$ such that $f(\theta_w) = \langle v_n^*(\lambda), \theta_w \rangle$ for all $\theta_w \in \overline{\boldsymbol{V}}_{d(n)}$ and $\|v_n^*(\lambda)\| = \sup_{\theta_w \in \overline{\boldsymbol{V}}_{d(n)} \theta_w \neq 0} \frac{(\lambda^\top \beta)^2}{\|\theta_w\|^2} < C$. Under the assumptions of the lemma, C does not depend on $d(n)$ so we conclude that $\|v_n^*(\lambda)\|$ does not diverge as $d(n) \to \infty$.

To arrive at the assertion of the lemma, we now use Lemma 3.3(2) of Chen and Pouzo (2015) to argue that $\|v^*(\lambda)\| < \infty$. To this end, we recall the lemma using our notation: Assume $f(\theta_w) = \lambda^\top \beta$ is bounded on $(\boldsymbol{V}, \|\cdot\|_{c,\infty})$ and $\{\overline{\boldsymbol{V}}\}_{d=0}^{\infty}$ is dense in $(\boldsymbol{V}, \|\cdot\|_{c,\infty})$. If

$f(\theta_w) = \lambda^\top \theta_w$ is unbounded on $(\boldsymbol{V}, \|\cdot\|)$, then $\lim_{d(n)\to\infty} \|v_n^*(\lambda)\| = \infty$. Hence, if we can show that the assumptions hold and $\|v_n^*(\lambda)\| < C < \infty$ uniformly over $d(n)$, $f(\theta_w) = \lambda^\top \theta_w$ is bounded on $(\boldsymbol{V}, \|\cdot\|)$ and thus its unique extension to $(\overline{\boldsymbol{V}}, \|\cdot\|)$ is bounded on $(\overline{\boldsymbol{V}}, \|\cdot\|)$ – see, for instance, the discussion on page 1032 in Chen and Pouzo (2015). First, $\{\overline{\boldsymbol{V}}\}_{d=0}^{\infty}$ is dense in $(\boldsymbol{V}, \|\cdot\|_{c,\infty})$ by hypothesis. Second, $\|v_n^*(\lambda)\| < C < \infty$ uniformly over $d(n)$ by our previous argument. Third,

$$\sup_{\theta_w \neq 0} \frac{\lambda^\top \beta}{\|\theta_w\|_{c,\infty}} = \sup_{\theta_w \neq 0} \frac{\lambda^\top \beta}{\sum_{j=1}^{J-1} \|w_j\|_\infty + \sum_{j=1}^{J} \sum_{t=1}^{T} \|\beta_{j,t}\|_E} \leq \sup_{\beta \neq 0} \frac{\|\lambda\|_E \|\beta\|_E}{\sum_{j=1}^{J} \sum_{t=1}^{T} \|\beta_{j,t}\|_E} \leq \|\lambda\|_E$$

where the last inequality uses that the square root is subadditive. We conclude that $f(\theta_w) = \lambda^\top \beta$ is bounded on $(\boldsymbol{V}, \|\cdot\|_{c,\infty})$.

Combining all arguments and the fact that the choice of $\lambda \neq 0$ was arbitrary throughout, we conclude that for all λ $f(\theta_w) = \lambda^\top \theta_w$ is bounded on $(\boldsymbol{V}, \|\cdot\|)$ and the unique extension of $f(\theta_w)$ to $(\overline{\boldsymbol{V}}, \|\cdot\|)$ is bounded on $(\overline{\boldsymbol{V}}, \|\cdot\|)$. This implies that the Riesz representer $v^*(\lambda) \in \overline{\boldsymbol{V}}$ exists and $\|v^*(\lambda)\| < \infty$ for all λ. Additionally, we can conclude that V^* is positive definite.

When the eigenvalue $\Omega_{d(n)}$ is bounded away from zero uniformly over $d(n)$, we have

$$\sup_{\theta_w \in \overline{\boldsymbol{V}}_{d(n)} \theta_w \neq 0} \frac{(\lambda^\top \beta)^2}{\|\theta_w\|^2} \leq \frac{\|\lambda\|_E^2}{\lambda_{\min}(\Omega_{d(n)})} =: \tilde{C} < \infty$$

where $\tilde{C}$ does not depend on $d(n)$ so that the previous arguments apply again. This concludes the proof. □

A.5 Proofs for Section 7.2 and Appendix A.1.5

A.5.1 Proof of Theorem A.1.3

Proof. If not noted otherwise, any fixed covariate value is the one in Theorem A.1.3. From similar arguments as *Step 1* to *Step 3* of the proof of Theorem 3.2, we know that $\mathcal{Q}_{\tilde{t}}(\{x_{\tilde{t}}^{(r)}\}_{r=1}^{R})$, $\Pi(\{x_{t,k}\}_{t=1}^{T})$, $\mathcal{P}_{t'}(x_{t'})$, and $\mathcal{Q}_{t^*}(\{x_{t^*}^{(r')}\}_{r'=1}^{R'})$ are identified up to the same relabeling of the groups. To pin down the sign and scaling of the eigenvalues in this argument, we use that the entries in $\mathcal{Q}_{\tilde{t}}(\{x_{\tilde{t}}^{(r)}\}_{r=1}^{R})$ are non-negative and all columns sum up to R.

Next, define $X_{\{\tilde{t}\}}^{(r)} = (x_{\tilde{t}}^{(r)^\top}, \{x_t\}_{t\neq\tilde{t}}^\top)^\top \in \mathbb{X}$ for $r = 1, \ldots, R$ and $X_{\{t^*\}}^{(r')} = (x_{t^*}^{(r')^\top}, \{x_{it}\}_{t\neq t^*}^\top)^\top$ for $r' = 1, \ldots, R'$. Then

$$\begin{aligned}
\underline{\pi}(\{x_{t,k}\}_{t=1}^{T}) &= \mathcal{Q}_{\tilde{t}}(\{x_{\tilde{t}}^{(r)}\}_{r=1}^{R})^\dagger \mathbf{P}_{\{\tilde{t}\}}(\{X_{\{\tilde{t}\}}^{(r)}\}_{r=1}^{R}) \\
\underline{\pi}(\{x_{t,k}\}_{t=1}^{T}) &= \mathcal{Q}_{t^*}(\{x_{t^*}^{(r')}\}_{r'=1}^{R'})^\dagger \mathbf{P}_{\{t^*\}}(\{X_{\{t^*\}}^{(r')}\}_{r'=1}^{R'})
\end{aligned}$$

where $\mathbf{P}_{\{\tilde{t}\}}(\{X_{\{\tilde{t}\}}^{(r)}\}_{r=1}^{R}) = (\mathbb{P}(y_{i\tilde{t}} = 0 \mid X_i = X_{\{\tilde{t}\}}^{(1)}), \mathbb{P}(y_{i\tilde{t}} = 1 \mid X_i = X_{\{\tilde{t}\}}^{(1)}), \ldots, \mathbb{P}(y_{i\tilde{t}} = 0 \mid X_i = X_{\{\tilde{t}\}}^{(R)}), \mathbb{P}(y_{i\tilde{t}} = 1 \mid X_i = X_{\{\tilde{t}\}}^{(R)}))^\top$ and $\mathbf{P}_{\{t^*\}}(\{X_{\{t^*\}}^{(r')}\}_{r'=1}^{R'})$ analogously. Since

$\mathbf{P}_{\{\tilde{t}\}}(\{X^{(r)}_{\{\tilde{t}\}}\}_{r=1}^R)$ and $\mathbf{P}_{\{t^*\}}(\{X^{(r')}_{\{t^*\}}\}_{r'=1}^{R'})$ are known for any realization of their respective arguments, $\underline{\pi}(\{x_{t,k}\}_{t=1}^T)$ is identified at $x_{\tilde{t},k}$ for all values of $\{x_{t,k}\}_{t\neq\tilde{t}}$ and at $x_{t^*,k}$ for all values of $\{x_{t,k}\}_{t\neq t^*}$ up the same group relabeling as the other objects.

Letting $\{A_r\}_{r=1}^R = (A_1^\top, \ldots, A_R^\top)^\top$, we have for any $t \neq \tilde{t}$

$$\left\{ \begin{pmatrix} \mathbb{P}(y_{i\tilde{t}} = 0, y_{it} = 0 \mid X_i = X^{(r)}_{\{\tilde{t}\}}) & \mathbb{P}(y_{i\tilde{t}} = 0, y_{it} = 1 \mid X_i = X^{(r)}_{\{\tilde{t}\}}) \\ \mathbb{P}(y_{i\tilde{t}} = 1, y_{it} = 0 \mid X_i = X^{(r)}_{\{\tilde{t}\}}) & \mathbb{P}(y_{i\tilde{t}} = 1, y_{it} = 1 \mid X_i = X^{(r)}_{\{\tilde{t}\}}) \end{pmatrix} \right\}_{r=1}^R$$
$$= \mathbf{P}_{\{\tilde{t},t\}}(\{X^{(r)}_{\{\tilde{t}\}}\}_{r=1}^R) = \mathcal{Q}_{\tilde{t}}(\{x^{(r)}_{\tilde{t}}\}_{r=1}^R)\Pi(\{x_{t,k}\}_{t=1}^T)\mathcal{P}_t(x_t)^\top$$

where $\mathbf{P}_{\{\tilde{t},t\}}(\{X^{(r)}_{\{\tilde{t}\}}\}_{r=1}^R)$ is known at all values of its arguments that are in $\mathbb{X}$. At $X_i = X$, $\Pi(\{x_{t,k}\}_{t=1}^T)$ is invertible by Assumption I-3′ and, for $\{x^{(r)}_{\tilde{t}}\}_{r=1}^R$, $\mathcal{Q}_{\tilde{t}}(\{x^{(r)}_{\tilde{t}}\}_{r=1}^R)$ has full column rank by Assumption I-2′ so that

$$\mathcal{P}_t(x_t)^\top = \Pi(\{x_{t,k}\}_{t=1}^T)^{-1}\mathcal{Q}_{\tilde{t}}(\{x^{(r)}_{\tilde{t}}\}_{r=1}^R)^\dagger \mathbf{P}_{\{\tilde{t},t\}}(\{X^{(r)}_{\{\tilde{t}\}}\}_{r=1}^R) \tag{A.16}$$

Because $\mathbf{P}_{\{\tilde{t},t\}}(\{X^{(r)}_{\{\tilde{t}\}}\}_{r=1}^R)$ is known at all values of its arguments that are in $\mathbb{X}$, $\mathcal{P}_t(\tilde{x}_t)$ is identified at all $\tilde{x}_t \in \mathbb{X}_t$ for which there exists $\{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}$ such that $\Pi(x_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \tilde{x}_t)$ has full rank – we recall that $\Pi(x_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \tilde{x}_{t,k})$ is identified from the previous argument. If $\Pi(x_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \tilde{x}_{t,k})$ does not have full rank but $\pi_j(x_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \tilde{x}_{t,k}) > 0$ for some $j \in \{1, \ldots, J\}$, one can follow the same arguments as in *Step 5* of the proof of Theorem 3.2 to derive an analogue of equation (A.7) and show that $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified for all $s_t \in \{0, 1\}$ at all $\tilde{x}_t \in \mathbb{X}_t$ for which there exists $\{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}$ such that $\pi_j(x_{\tilde{t},k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,\tilde{t}\}}, \tilde{x}_{t,k}) > 0$. Again, the relabeling of the groups is consistent with the relabeling of the groups of the other objects.

A symmetric argument for $t \neq t^*$ yields that $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, g_i = j)$ is identified for all $s_t \in \{0, 1\}$ at $\tilde{x}_t \in \mathbb{X}_t$ for which there exists $\{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,t^*\}}$ such that $\pi_j(x_{t^*,k}, \{x^*_{\underline{t},k}\}_{\underline{t}\notin\{t,t^*\}}, \tilde{x}_{t,k}) > 0$.

The third part of the statement follows from the simple observation that if for some $\underline{t} \in \{1, \ldots, T\}$ and $\tilde{x}^{(r)}_{\underline{t}} \in \mathbb{X}_{\underline{t}}$ for $r = 1, \ldots, R$ with $\tilde{x}^{(r)}_{\underline{t},k} = \tilde{x}^{(s)}_{\underline{t},k}$ for all r, s, $\mathcal{Q}_{\underline{t}}(\{\tilde{x}^{(r)}_{\underline{t}}\}_{r=1}^R)$ has full column rank and is identified, then $\underline{\pi}(\tilde{x}_{\underline{t},k}, \{x_{t,k}\}_{t\neq\underline{t}}) = \mathcal{Q}_{\underline{t}}(\{\tilde{x}^{(r)}_{\underline{t}}\}_{r=1}^R)^\dagger \mathbf{P}_{\{\underline{t}\}}(\{X^{(r)}_{\{\underline{t}\}}\}_{r=1}^R)$ is identified at $\tilde{x}_{\underline{t},k}$ for all $\{x_{t,k}\}_{t\neq\underline{t}}$. An identical argument holds for any submodel $\mathcal{I} \subseteq \{1, \ldots, T\}$ and $\{\tilde{x}^{(r)}_t\}_{t\in\mathcal{I}}$ for $r = 1, \ldots, R$ with $\tilde{x}^{(r)}_{t,k} = \tilde{x}^{(s)}_{t,k}$ for all $t \in \mathcal{I}$ and r, s such that $\mathcal{Q}_{\mathcal{I}}(\{\{\tilde{x}^{(r)}_t\}_{t\in\mathcal{I}}\}_{r=1}^R)$ has full column rank and is identified.

This concludes the proof. The identification result can be strengthened along similar lines as in the proof of Theorem 3.2.

□

A.5.2 Proof of Theorem A.1.4

Proof. The proof follows directly by combining the arguments in the proofs of Theorem 3.3 and Theorem A.1.3. We therefore omit it. □

A.6 Proof for Section 7.4

A.6.1 Proof of Lemma 7.3

Proof. Using the notation from Appendix A.1.3, the first part of the lemma follows from

$$\begin{aligned}\operatorname{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) &= \operatorname{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top)\\ &\leq \min(\operatorname{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})), \operatorname{rank}(\Pi(X)), \operatorname{rank}(\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})))\\ &\leq J\end{aligned}$$

The second part follows from observing that by assumption $\operatorname{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)) = J$ so that Sylvester's inequality implies that $\operatorname{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) = \operatorname{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})^\top) \geq \operatorname{rank}(\mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\mathcal{I}_1})\Pi(X)) + \operatorname{rank}(\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\mathcal{I}_2})) - J = J$. As the former inequality continues to hold, $\operatorname{rank}(\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)) = J$. This concludes the proof. □

A.7 Proofs for Appendix A.1.7

A.7.1 Proof of Theorem A.1.5

Proof. If not noted otherwise, X and x_t refer to the fixed value of the covariates in the theorem. In the following, we sometimes condition on (X, s). This is to be understood as conditioning on $X_i = X$ and $y_{i0} = s$ if not noted otherwise. Throughout the proof, we leverage the following observation: For all $1 \leq t' < \tilde{t} \leq T$

$$\begin{aligned}\mathbb{P}(\{y_{it} = s_t\}_{t=t'}^{\tilde{t}} \mid X_i, \{y_{it}\}_{t=0}^{t'-1}) = &\sum_{j=1}^{J} \mathbb{P}(g_i = j \mid X_i, \{y_{it}\}_{t=0}^{t'-1})\mathbb{P}_{t'}(y_{it'} = s_{t'} \mid x_{it'}, y_{it'-1}, g_i = j)\\ &\times \prod_{t=t'+1}^{\tilde{t}} \mathbb{P}_t(y_{it} = s_t \mid x_{it}, y_{it-1} = s_{t-1}, g_i = j)\end{aligned}$$

where the equality follows from Assumption M-2.

We define $\mathcal{O}^\oplus = \{t+1 : t \in \mathcal{O}\}$ and $\underline{\mathcal{O}} = \mathcal{O} \cup \mathcal{O}^\oplus$. Our argument now proceeds in multiple steps.

1. *Identification given* $\{y_{it} = s\}_{t\in\mathcal{O}_1^\oplus\cup\mathcal{O}_2^\oplus\cup\{0\}}$ *and* $X_i = X$. We consider the identified

matrix

$$\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s) = \Big\{\Big\{ \mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{O}_1}, \{y_{it}=s\}_{t\in\mathcal{O}_1^{\oplus}}, \{y_{it}=s_t^*\}_{t\in\mathcal{O}_2}, \\ \{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}} \mid X, y_{i0}=s)\Big\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}}\Big\}_{\{s_t^*\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}}$$

where a row of $\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X_i,s)$ fixes $\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}$ and iterates through all $2^{|\mathcal{O}_2|}$ combinations of $\{s_t^*\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}$, whereas a column fixes $\{s_t^*\}_{t\in\mathcal{O}_2}$ and iterates through all $2^{|\mathcal{O}_1|}$ combinations of $\{s_t\}_{t\in\mathcal{O}_1}$. For $k=1,2$, we similarly define the identified matrix

$$\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2\cup\{T\},k}(X,s) = \Big\{\Big\{ \mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{O}_1}, \{y_{it}=s\}_{t\in\mathcal{O}_1^{\oplus}}, \{y_{it}=s_t^*\}_{t\in\mathcal{O}_2}, \{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}}, \\ y_{iT}=k-1 \mid X, y_{i0}=s)\Big\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}}\Big\}_{\{s_t^*\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}}$$

Some algebra yields

$$\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s) = \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\Pi(X,s)\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})^{\top} \tag{A.17}$$

$$\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2\cup\{T\},k}(X,s) = \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\mathcal{D}_{T,k}(x_T,s)\Pi(X,s)\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})^{\top} \tag{A.18}$$

where $\mathcal{D}_{T,k}(x_T,s) = \mathrm{diag}([\mathcal{P}_T(x_T,s)]_{k,\cdot})$ for $k=1,2$.

Under Assumptions I-2″ and I-3″, the arguments of the proof of Theorem 3.2 readily imply that $\mathcal{P}_T(x_T,s)$ is identified up to relabeling. However, since the columnwise sums of $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})$ are unknown, these arguments only identify $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})$ up to scaling, but up to the same permutation of the groups as in $\mathcal{P}_T(x_T,s)$, that is, these arguments identify $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})K$ for some diagonal matrix K with non-zero diagonal entries; the entries are non-zero as $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})K$ has full column rank. For now, we assume that K is identified so that $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})$ is identified up to the same relabeling of the groups as in $\mathcal{P}_T(x_T,s)$. We come back to the identification of K at the end of Step 6 below. There we shall see that $\Pi(X,s)$ is identified without rescaling so that the arguments of the proof of Lemma A.10.3 apply and K is identified.

Next, define

$$\begin{aligned}\mathbf{P}_{\mathcal{O}_1}(X,s) &= \{\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{O}_1}, \{y_{it}=s\}_{t\in\mathcal{O}_1^{\oplus}} \mid X,s)\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}} \\ &= \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\underline{\pi}(X,s) \\ &= \Big\{\sum_{j=1}^{J}\pi_j(X,s)\prod_{t\in\mathcal{O}_1}\mathbb{P}_t^{\otimes}(s_t,x_t,s,x_{t+1},s,j)\Big\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}}\end{aligned}$$

with $\mathbf{P}_{\mathcal{O}_1}(X,s)$ being identified. Under Assumption I-2″, we therefore identify $\underline{\pi}(X,s)$ up to the identical relabeling of the groups with $\underline{\pi}(X,s) = \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})^{\dagger}\mathbf{P}_{\mathcal{O}_1}(X,s)$

– where we use that K is identified.[29] Since $\mathbf{P}_{\mathcal{O}_1}(X,s)$ is identified for all $X \in \mathbb{X}$ and $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})$ is a function of x_t for $t \in \underline{\mathcal{O}_1}$, $\underline{\pi}(X,s)$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_1}}$ and s for all $\{x_t\}_{t\notin\underline{\mathcal{O}_1}}$.

Since $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})$, $\mathcal{P}_T(x_T,s)$, and $\underline{\pi}(X,s)$ are identified up to the same permutation of the groups, $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})^\top$ is identified up the same permutation with $\Pi(X,s)^{-1}\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})^\dagger\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s)$, where we use Assumptions I-2″ and I-3″. Importantly, the first part of Lemma A.10.3 implies that $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})$ is identified up to identical relabeling of the groups without rescaling issue.[30]

2. Changing $\{y_{it}\}_{t\in\mathcal{O}_1^{\oplus}\cup\{0\}}$, while keeping the rest fixed. Next, we change y_{it} from s to $\tilde{s}\neq s$ for $t\in\mathcal{O}_1^{\oplus}\cup\{0\}$ and, similar to before, define the following identified matrices

$$\begin{aligned}
&\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,\{y_{it}=\tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}\cup\{0\}},\{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}})\\
&=\Big\{\Big\{\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{O}_1},\{y_{it}=\tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}},\{y_{it}=s_t^*\}_{t\in\mathcal{O}_2},\\
&\quad\{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}}\mid X,y_{i0}=\tilde{s})\Big\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}}\Big\}_{\{s_t^*\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2\cup\{T\},k}(X,\{y_{it}=\tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}\cup\{0\}},\{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}})\\
&=\Big\{\Big\{\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{O}_1},\{y_{it}=\tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}},\{y_{it}=s_t^*\}_{t\in\mathcal{O}_2},\\
&\quad\{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}},y_{iT}=k-1\mid X,y_{i0}=\tilde{s})\Big\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}}\Big\}_{\{s_t^*\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}}
\end{aligned}$$

Next, recalling Assumption I-2″ we define

$$\begin{aligned}
&\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1}=\tilde{s};\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})\\
&=\mathcal{P}^{\otimes}_{\min(\mathcal{O}_2)}(x_{\min(\mathcal{O}_2)},\tilde{s},x_{\min(\mathcal{O}_2)+1},s)\overset{\text{col}}{\otimes}\mathcal{P}^{\otimes}_{\mathcal{O}_2\setminus\min(\mathcal{O}_2)}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2\setminus\min(\mathcal{O}_2)})
\end{aligned}\tag{A.19}$$

As before, we have

$$\begin{aligned}
&\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,\{y_{it}=\tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}\cup\{0\}},\{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}})\\
&=\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,\tilde{s},x_{t+1},\tilde{s}\}_{t\in\mathcal{O}_1})\Pi(X,\tilde{s})\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1}=\tilde{s};\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})^\top\\
&\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2\cup\{T\},k}(X,\{y_{it}=\tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}\cup\{0\}},\{y_{it}=s\}_{t\in\mathcal{O}_2^{\oplus}})\\
&=\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,\tilde{s},x_{t+1},\tilde{s}\}_{t\in\mathcal{O}_1})\mathcal{D}_{T,k}(x_T,s)\Pi(X,\tilde{s})\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1}=\tilde{s};\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})^\top
\end{aligned}$$

From the same arguments as in Step 1, we note that $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,\tilde{s},x_{t+1},\tilde{s}\}_{t\in\mathcal{O}_1})\tilde{K}$ for some diagonal matrix $\tilde{K}$ with non-zero diagonal entries, $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1}=\tilde{s};\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})$,

[29] More specifically, $K^{-1}\underline{\pi}(X,s)=K^{-1}\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})^\dagger\mathbf{P}_{\mathcal{O}_1}(X,s)$ is identified.

[30] This follows from the simple observation that $(K^{-1}\Pi(X,s))^{-1}(\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})K)^\dagger = \Pi(X,s)^{-1}\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})^\dagger$.

and $\mathcal{P}_T(x_T, s)$ are identified up to the same relabeling of the groups. Importantly, the relabeling is consistent with the labels of Step 1 as $\mathcal{P}_T(x_T, s)$ has distinct columns, that is, matching the columns of the identified $\mathcal{P}_T(x_T, s)$ in this step with the previous step aligns the labels. As in Step 1, we assume, for now, that $\tilde{K}$ is identified and move the exact discussion to the end of Step 6. Then, following the same arguments as in Step 1, $\underline{\pi}(X, \tilde{s})$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_1}}$ and $\tilde{s}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_1}}$.

3. Identification given $\{y_{it} = \tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}\cup\mathcal{O}_2^{\oplus}\cup\{0\}}$ and $X_i = X$. For $\tilde{s} \neq s$, an identical argument as in Step 1 based on $\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X, \tilde{s})$ and $\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2\cup\{T\},k}(X, \tilde{s})$ implies that $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})K^*$, $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_2})$, and $\mathcal{P}_T(x_T, \tilde{s})$ are identified up to the same relabeling of the groups. K^* is a diagonal scaling matrix with non-zero diagonals entries. To argue that the group labels agree with the labels in the previous steps, we let Δ be a $J \times J$ permutation matrix, that is, a matrix that contains a single 1 per column and row and zeros otherwise. In the current step, we identified $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})K^*\Delta$[31], while the previous argument identified $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})\tilde{K}$, where, without loss of generality, we set the permutation matrix to the identity matrix. We proceed to argue that Δ is identified. Since $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})$ has full column rank, we have

$$(\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})K^*\Delta)^{\dagger}\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})\tilde{K} = \Delta^{\top}(K^*)^{-1}\tilde{K}$$

where the LHS is identified. Since K^* and $\tilde{K}$ are diagonal matrices with non-zero entries diagonal entries and $\Delta^{\top}$ is a permutation matrix, the non-zero entries of $\Delta^{\top}(K^*)^{-1}\tilde{K}$ identify $\Delta^{\top}$ so that Δ is identified. Then, $(K^*)^{-1}\tilde{K}$ is identified, too. We conclude that $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})K^*$, $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_2})$, and $\mathcal{P}_T(x_T, \tilde{s})$ are identified up to the same relabeling of the groups as in the previous steps. As before, we treat K^* as known in the following and move the exact identification argument to the end of Step 6.

4. Collecting the current results. Up to identical relabeling

- $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_1})$ for $s \in \{0, 1\}$ (up to an identified rescaling matrix),
- $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$ for $\tilde{s} \in \{0, 1\}$[32] and $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$ for $s \in \{0, 1\}$,
- $\mathcal{P}_T(x_T, s)$ for $s \in \{0, 1\}$, and
- $\underline{\pi}(X, s)$ at $\{x_t\}_{t\in\underline{\mathcal{O}_1}}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_1}}$ and $s \in \{0, 1\}$ (up to an identified rescaling matrix)

are identified. The scaling matrices are shown to be identified at the end of Step 6.

[31]The order of scaling and permutation matrices is without loss of generality since for some other scaling matrix K_1, we have $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})\Delta K_1 = \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})K^*\Delta$ with $K^* := \Delta K_1\Delta^{\top}$, a diagonal matrix. We use that $\Delta^{\top} = \Delta^{-1}$.

[32]To see this, when $\tilde{s} = s$, then $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2}) = \mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$.

5. Identification of $\mathcal{P}_t(x_t, s)$ for $t \notin \underline{\mathcal{O}_1}$.[33] We note $\{1, \ldots, T\} \backslash \underline{\mathcal{O}_1} = \{\min(\mathcal{O}_2), \min(\mathcal{O}_2)+1, \ldots, T\}$ and start with period $t^* = \min(\mathcal{O}_2) = \max(\underline{\mathcal{O}_1})+1$ to subsequently iterate through the remaining periods until period T. Throughout this step, we let $t^* = \min(\mathcal{O}_2)$.

1. Period $t^* = \min(\mathcal{O}_2)$: For any $s \in \{0,1\}$, define the identified matrix

$$\begin{aligned}\mathbf{P}_{\mathcal{O}_1 \cup \{t^*\}}(X, s) &= \{\mathbb{P}(\{y_{it} = s_t\}_{t \in \mathcal{O}_1}, \{y_{it} = s\}_{t \in \mathcal{O}_1^{\oplus}}, y_{it^*} = s_{t^*} \mid X, s)\}_{\{s_t\}_{t \in \mathcal{O}_1 \cup \{t^*\}} \in \{0,1\}^{|\mathcal{O}_1|+1}} \\ &= \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \Pi(X, s) \mathcal{P}_{t^*}(x_{t^*}, s)^\top\end{aligned}$$

   Under Assumptions I-2″ and I-3″, it follows that

$$\mathcal{P}_{t^*}(x_{t^*}, s)^\top = \Pi(X, s)^{-1} \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1})^\dagger \mathbf{P}_{\mathcal{O}_1 \cup \{t^*\}}(X, s)$$

   is identified. Additionally, $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1})$ has full column rank for $s \in \{0,1\}$ and is not a function of $\{x_t\}_{t \notin \underline{\mathcal{O}_1}}$, whereas $\Pi(X, s)$ is identified at $\{x_t\}_{t \in \underline{\mathcal{O}_1}}$ for all $\{x_t\}_{t \notin \underline{\mathcal{O}_1}}$ and $s \in \{0,1\}$, and $\mathbf{P}_{\mathcal{O}_1 \cup \{t^*\}}(X, s)$ is identified for all $(X^\top, s)^\top \in \mathbb{X} \times \{0,1\}$. Combining these observations with the arguments of *Step 5* in the proof of Theorem 3.2, we conclude that for $s_{t^*} \in \{0,1\}$ and all $j \in \{1, \ldots, J\}$ $\mathbb{P}_{t^*}(y_{it^*} = s_{t*} \mid x_{it^*} = \tilde{x}_{t^*}, y_{it^*-1} = s, g_i = j)$ is identified at all $(x_{t^*}^\top, s) \in \mathbb{X}_{t^*} \times \{0,1\}$ for which there exists $\{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \notin \underline{\mathcal{O}_1} \cup \{t^*\}}$ so that $\pi_j(\{x_t\}_{t \in \underline{\mathcal{O}_1}}, \tilde{x}_{t^*}, \{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \notin \underline{\mathcal{O}_1} \cup \{t^*\}}, s) > 0$. The group labels are identical to the group labels in Step 4. In particular, this implies that, for $s \in \{0,1\}$, $\mathcal{P}_{t^*}(x_{t^*}, s)$ is identified at the value x_{t^*} fixed in the Theorem at which Assumptions I-2″ and I-3″ hold.

2. Period $t^* + 1 = \min(\mathcal{O}_2) + 1$: We recall that from part (v) of Assumption I-2″, for $s \in \{0,1\}$ we have $\mathbb{P}_{t^*}(y_{it^*} = 1 \mid x_{it^*} = x_{t^*}, y_{it^*-1} = s, g_i = j) \in (0,1)$ for all $j = 1, \ldots, J$. For $k = 1, 2$, we define the identified matrix

$$\begin{aligned}&\mathbf{P}_{\mathcal{O}_1 \cup \{t^*, t^*+1\}}(X, s, k) \\ &= \{\mathbb{P}(\{y_{it} = s_t\}_{t \in \mathcal{O}_1}, \{y_{it} = s\}_{t \in \mathcal{O}_1^{\oplus}}, y_{it^*} = k-1, y_{it^*+1} = s_{t^*+1} \mid X, s)\}_{\{s_t\}_{t \in \mathcal{O}_1 \cup \{t^*+1\}} \in \{0,1\}^{|\mathcal{O}_1|+1}} \\ &= \left(\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \overset{\text{col}}{\otimes} [\mathcal{P}_{t^*}(x_{t^*}, s)]_{k,\cdot}^\top\right) \Pi(X, s) \mathcal{P}_{t^*+1}(x_{t^*+1}, k-1)^\top\end{aligned}$$

   where $[\mathcal{P}_{t^*}(x_{t^*}, s)]_{k,\cdot}$ is the k-th row of $\mathcal{P}_{t^*}(x_{t^*}, s)$ as a $J \times 1$ matrix, which is identified for $k = 1, 2$ and $s \in \{0,1\}$ from the previous step. By Lemma A.10.2, $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \overset{\text{col}}{\otimes} [\mathcal{P}_{t^*}(x_{t^*}, s)]_{k,\cdot}^\top$, which is a function of $\{x_t\}_{t \in \underline{\mathcal{O}_1} \cup \{t^*\}}$, has full column rank for all $k = 1, 2$ and $s \in \{0,1\}$. Thus, for $k = 1, 2$ and $s \in \{0,1\}$

$$\mathcal{P}_{t^*+1}(x_{t^*+1}, k-1)^\top = \Pi(X, s)^{-1} \left(\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \overset{\text{col}}{\otimes} [\mathcal{P}_{t^*}(x_{t^*}, s)]_{k,\cdot}^\top\right)^\dagger$$

[33] For this step, it is actually irrelevant that the scaling matrices K, $\tilde{K}$ and K^* are not identified as they cancel in the relevant expressions. Nevertheless, to keep the notation light, we assume them to be known.

$$\times \mathbf{P}_{\mathcal{O}_1 \cup \{t^*, t^*+1\}}(X, s, k)$$

is identified. Also, since $\mathbf{P}_{\mathcal{O}_1 \cup \{t^*, t^*+1\}}(X, s, k)$ is identified for all $(X^\top, s, k) \in \mathbb{X} \times \{0, 1\} \times \{1, 2\}$ and $\Pi(X, s)$ is identified at $\{x_t\}_{t \in \underline{\mathcal{O}_1}}$ for all $\{x_t\}_{t \notin \underline{\mathcal{O}_1}}$ and $s \in \{0, 1\}$, analogous arguments as in *Step 5* of the proof of Theorem 3.2 imply that for all $j = 1, \ldots, J$, $\mathbb{P}_{t^*+1}(y_{it^*+1} = s_{t^*+1} \mid x_{it^*+1} = \tilde{x}_{t^*+1}, y_{it^*} = \tilde{s}, g_i = j)$ is identified for $s_{t^*+1} \in \{0, 1\}$ at all $(x_{t^*+1}^\top, \tilde{s}) \in \mathbb{X}_{t^*+1} \times \{0, 1\}$ for which there exists $\{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \geq t^*+2}$ and $s \in \{0, 1\}$ such that $\pi_j(\{x_t\}_{t \in \underline{\mathcal{O}_1} \cup \{t^*\}}, \tilde{x}_{t^*+1}, \{\tilde{x}_{\tilde{t}}\}_{\tilde{t} \geq t^*+2}, s) > 0$.[34] The group labels are identical to the group labels in Step 4.

3. Period $\tilde{t} \in \{\min(\mathcal{O}_2) + 2, \ldots, T - 1\}$: Without loss of generality, we assume $T - 1 \geq \min(\mathcal{O}_2)+2$, otherwise this step is redundant. For any $s \in \{0, 1\}$, it holds that $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, y_{it-1} = s, g_i = j) \in (0, 1)$ for all $j = 1, \ldots, J$ and $t \in \{t^*, \ldots, \tilde{t} - 1\}$.[35] Next, we define

$$\begin{aligned}
&\mathbf{P}_{\mathcal{O}_1 \cup \{t^*, \ldots, \tilde{t}\}}(X, s, k) \\
&= \{\mathbb{P}(\{y_{it} = s_t\}_{t \in \mathcal{O}_1}, \{y_{it} = s\}_{t \in \mathcal{O}_1^{\oplus} \cup \{t^*, \ldots, \tilde{t}-2\}}, y_{i\tilde{t}-1} = k - 1, y_{i\tilde{t}} = s_{\tilde{t}} \mid X, s)\}_{\{s_t\}_{t \in \mathcal{O}_1 \cup \{\tilde{t}\}} \in \{0,1\}^{|\mathcal{O}_1|+1}} \\
&= \left(\mathcal{P}_{\mathcal{O}_1}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \overset{\text{col}}{\otimes} \left(\odot_{t \in \{t^*, \ldots, \tilde{t}-2\}} \left[\mathcal{P}_t(x_t, s)\right]_{s+1,\cdot}\right)^\top \overset{\text{col}}{\otimes} \left[\mathcal{P}_{\tilde{t}-1}(x_{\tilde{t}-1}, s)\right]_{k,\cdot}^\top\right) \\
&\quad \times \Pi(X, s) \mathcal{P}_{\tilde{t}}(x_{\tilde{t}}, k - 1)^\top
\end{aligned}$$

where $\odot$ denotes the Hadamard product. Iteratively applying Lemma A.10.2 implies that $\mathcal{P}_{\mathcal{O}_1}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \overset{\text{col}}{\otimes} \left(\odot_{t \in \{t^*, \ldots, \tilde{t}-2\}} \left[\mathcal{P}_t(x_t, s)\right]_{s+1,\cdot}\right)^\top \overset{\text{col}}{\otimes} \left[\mathcal{P}_{\tilde{t}-1}(x_{\tilde{t}-1}, s)\right]_{k,\cdot}^\top$ has full column rank for $k = 1, 2$ and $s \in \{0, 1\}$. Additionally, applying the identification argument iteratively implies that $\mathcal{P}_{\mathcal{O}_1}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \overset{\text{col}}{\otimes} \left(\odot_{t \in \{t^*, \ldots, \tilde{t}-2\}} \left[\mathcal{P}_t(x_t, s)\right]_{s+1,\cdot}\right)^\top \overset{\text{col}}{\otimes} \left[\mathcal{P}_{\tilde{t}-1}(x_{\tilde{t}-1}, s)\right]_{k,\cdot}^\top$ is identified for $k = 1, 2$ and $s \in \{0, 1\}$. Now, the same arguments as for period $t^* + 1$ imply that for any $j \in \{1, \ldots, J\}$ $\mathbb{P}_{\tilde{t}}(y_{i\tilde{t}} = s_{\tilde{t}} \mid x_{i\tilde{t}} = \tilde{x}_{\tilde{t}}, y_{i\tilde{t}-1} = \tilde{s}, g_j)$ is identified at all $(x_{\tilde{t}}^\top, \tilde{s}) \in \mathbb{X}_{\tilde{t}} \times \{0, 1\}$ for which there exists $\{\tilde{x}_t\}_{t > \tilde{t}}$ and $s \in \{0, 1\}$ such that $\pi_j(\{x_t\}_{t \in \underline{\mathcal{O}_1} \cup \{t^*, \ldots, \tilde{t}-1\}}, \tilde{x}_{\tilde{t}}, \{\tilde{x}_t\}_{t > \tilde{t}}, s) > 0$. The group labels are identical to the group labels in Step 4.

4. Period T: The same argument as in the previous step applies. However, since there is no time period larger than T, we have that $\mathbb{P}_T(y_{iT} = s_t \mid x_{iT} = \tilde{x}_T, y_{iT-1} = \tilde{s}, g_i = j)$ is identified at all $(x_T^\top, \tilde{s}) \in \mathbb{X}_T \times \{0, 1\}$ such that $\pi_j(\{x_t\}_{\{1, \ldots, T-1\}}, \tilde{x}_T, s) > 0$ for some $s \in \{0, 1\}$. While this identification argument suggests that we cannot vary covariates

[34] We note that this result is slightly stronger than the result, we state in the theorem itself. Specifically, the statement in the theorem sets $\tilde{s}$ equal to s, which is not required.

[35] This follows from Assumption I-2″ (v). However, it is easy to see that our argument only requires that there exists some $s_{t-1} \in \{0, 1\}$ such that $\mathbb{P}_t(y_{it} = 1 \mid x_{it} = x_t, y_{it-1} = s_{t-1}, g_i = j) \in (0, 1)$ for all j, but not that s_{t-1} is the same for all t. However, to keep notation as light as possible, we do so here. An adaption to the case with varying s_{t-1} is straightforward.

of periods different than T to ensure that $\pi_j(\{x_t\}_{\{1,\ldots,T-1\}}, \tilde{x}_T, s) > 0$, this is only an artifact of the current argument. For instance, if there exists $\{\tilde{x}_t\}_{t\in\{t^*,\ldots,T-1\}}$ such that for $\{s_t\}_{t=t^*-1}^{T-1} \in \{0,1\}^{T-t^*}$, $\prod_{t=t^*}^{T-1} \mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, y_{it-1} = s_{t-1}, g_i = j) \in (0,1)$ is identified and $\pi_j(\{x_t\}_{t\in\underline{\mathcal{O}}_1}, \{\tilde{x}_t\}_{t\in\{t^*,\ldots,T-1\}}, \tilde{x}_T, s) > 0$, then the same argument as in the previous step applies and implies that $\mathbb{P}_T(y_{iT} = s_t \mid x_{iT} = \tilde{x}_T, y_{iT-1} = s, g_i = j)$ is identified. A similar argument also applies for the time periods $\tilde{t} > t^*$ in the previous steps.

These arguments in particular imply that for $t \geq t^*$, $\mathcal{P}_t(x_t, s)$ is identified for all $s \in \{0,1\}$ at the at the value x_t fixed in the theorem at which Assumptions I-2″ and I-3″ hold.

For $k \in \mathbb{N}$, we define $\underline{\pi}(X, \{s_t\}_{t=0}^k) = (\mathbb{P}(g_i = 1 \mid X_i = X, \{y_{it} = s_t\}_{t=0}^k), \ldots, \mathbb{P}(g_i = J \mid X_i = X, \{y_{it} = s_t\}_{t=0}^k))^\top$ and $\Pi(X, \{s_t\}_{t=0}^k) = \text{diag}(\underline{\pi}(X, \{s_t\}_{t=0}^k))$ for the remainder of this proof.

6. An intermediate step – Identification of $\mathbb{P}(g_i = j \mid X_i = X, \{y_{it} = s_t\}_{t=1}^{\max(\underline{\mathcal{O}}_1)})$. Let $\{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)} \in \{0,1\}^{|\underline{\mathcal{O}}_1|+1}$ and define the identified matrix

$$\begin{aligned}\mathbf{P}_{\mathcal{O}_2}(X, s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)}) &= \{\mathbb{P}(\{y_{it} = s_t\}_{t\in\mathcal{O}_2}, \{y_{it} = s\}_{t\in\mathcal{O}_2^\oplus} \mid X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)})\}_{\{s_t\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}} \\ &= \mathcal{P}_{\mathcal{O}_2}^{\otimes}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}_{\min(\mathcal{O}_2)-1}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})\underline{\pi}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)})\end{aligned}$$

where we use that $\min(\mathcal{O}_2) - 1 = \max(\underline{\mathcal{O}}_1)$. Since $\mathcal{P}_{\mathcal{O}_2}^{\otimes}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}_{\min(\mathcal{O}_2)-1}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$, which we defined in equation (A.19), has full column rank for $\tilde{s}_{\min(\mathcal{O}_2)-1} \in \{0,1\}$, is identified for $\tilde{s}_{\min(\mathcal{O}_2)-1} \in \{0,1\}$ (Step 3), and is not a function of $\{x_t\}_{t\notin\underline{\mathcal{O}}_2}$, while $\mathbf{P}_{\mathcal{O}_2}(X, s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)})$ is identified for all $(X^\top, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)})^\top \in \mathbb{X} \times \{0,1\}^{|\underline{\mathcal{O}}_1|+1}$, we conclude that $\underline{\pi}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)})$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}}_2}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}}_2}$ and $\{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)} \in \{0,1\}^{|\underline{\mathcal{O}}_1|+1}$ with

$$\underline{\pi}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)}) = \mathcal{P}_{\mathcal{O}_2}^{\otimes}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}_{\min(\mathcal{O}_2)-1}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})^\dagger \mathbf{P}_{\mathcal{O}_2}(X, s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)})$$

where the group labels are identical to the group labels in Step 4.

The previous discussion implies that $\underline{\pi}(X, \{s_t\}_{t=0}^k)$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}}_2}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}}_2}$ and $\{s_t\}_{t=0}^k \in \{0,1\}^{k+1}$ up to identical relabeling of the groups for $k = 0, \ldots, \max(\underline{\mathcal{O}}_1)$. To see this, we fix some $j \in \{1, \ldots, J\}$ and consider

$$\begin{aligned}&\mathbb{P}(g_i = j \mid X_i = X, \{y_{it} = s_t\}_{t=0}^k) \\ = &\sum_{\{\tilde{s}_t\}_{t\in\{k+1,\ldots,\max(\underline{\mathcal{O}}_1)\}}\in\{0,1\}^{|\underline{\mathcal{O}}_1|-k}} \mathbb{P}(g_i = j, \{\tilde{s}_t\}_{t\in\{k+1,\ldots,\max(\underline{\mathcal{O}}_1)\}} \mid X_i = X, \{y_{it} = s_t\}_{t=0}^k) \\ = &\sum_{\{\tilde{s}_t\}_{t\in\{k+1,\ldots,\max(\underline{\mathcal{O}}_1)\}}\in\{0,1\}^{|\underline{\mathcal{O}}_1|-k}} \mathbb{P}(g_i = j \mid X_i = X, \{y_{it} = s_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)})\end{aligned}$$

$$\times \mathbb{P}(\{\tilde{s}_t\}_{t\in\{k+1,\ldots,\max(\underline{\mathcal{O}_1})\}} \mid X_i = X, \{y_{it} = s_t\}_{t=0}^{k})$$

where $\mathbb{P}(g_i = j \mid X_i = X, \{y_{it} = s_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})})$ is identified at the required points by the previous argument and $\mathbb{P}(\{\tilde{s}_t\}_{t\in\{k+1,\ldots,\max(\underline{\mathcal{O}_1})\}} \mid X_i = X, \{y_{it} = s_t\}_{t=0}^{k})$ is a conditional distribution of the observed variables and thus identified at the required points.

Identification of K, $\tilde{K}$ and K^*. The arguments in this step solely require $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}_{\min(\mathcal{O}_2)-1}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$ to be identified for $\tilde{s}_{\min(\mathcal{O}_2)-1} \in \{0,1\}$. Hence, irrespective of the scaling matrices K, $\tilde{K}$ and K^*, the arguments of this step imply that $\pi(X, s_0)$ is identified for $s_0 \in \{0,1\}$ – and this **not** up to rescaling. Then, following analogous arguments as in the proof of Lemma A.10.3 for the case when $\underline{\pi} = \underline{\pi}^*$, we conclude that all scaling matrices are identified. For instance, $\tilde{K}$ is identified via the following equation

$$\begin{aligned}
&\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X, \{y_{it} = \tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}\cup\{0\}}, \{y_{it} = s\}_{t\in\mathcal{O}_2^{\oplus}}) \\
&= \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})\tilde{K}\tilde{K}^{-1}\Pi(X, \tilde{s})\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})^{\top} \\
&\Rightarrow \tilde{K}^{-1} = (\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})\tilde{K})^{\dagger}\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X, \{y_{it} = \tilde{s}\}_{t\in\mathcal{O}_1^{\oplus}\cup\{0\}}, \{y_{it} = s\}_{t\in\mathcal{O}_2^{\oplus}}) \\
&\quad \times (\mathcal{P}^{\otimes}_{\mathcal{O}_2}(y_{i\min(\mathcal{O}_2)-1} = \tilde{s}; \{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})^{\top})^{\dagger}\Pi(X, \tilde{s})^{-1}
\end{aligned}$$

7. Identification of $\mathcal{P}_t(x_t, s)$ for $t \in \underline{\mathcal{O}_1}$. We recall that $\underline{\mathcal{O}_2}^c = \underline{\mathcal{O}_1} \cup \{T\}$ and define

$$\mathcal{P}_t^{(\tilde{s})}(x_t, s) = (\mathbb{P}_t(y_{it} = \tilde{s} \mid x_{it} = x_t, y_{it-1} = s, g_i = 1), \ldots, \mathbb{P}_t(y_{it} = \tilde{s} \mid x_{it} = x_t, y_{it-1} = s, g_i = J))^{\top}$$

for $\tilde{s} \in \{0,1\}$. We start with period $t^* = \max(\underline{\mathcal{O}_1}) = \min(\mathcal{O}_2) - 1$ to subsequently iterate through the remaining periods until period 1. We treat period T as a special case in the next step. Throughout this step, we let $t^* = \max(\underline{\mathcal{O}_1})$. This step is similar to Step 5 previously. To keep the notation light and make the proof more readable, we establish our arguments by first focusing on the case when $\pi_j(\cdot) > 0$ for all $j = 1, \ldots, J$. We will subsequently note that this is not required as can be easily established using the arguments from *Step 5* in the proof of Theorem 3.2.

1. Period $t^* = \max(\underline{\mathcal{O}_1})$: We define

$$\begin{aligned}
&\mathbf{P}_{\mathcal{O}_2\cup\{t^*\}}(X, s, \tilde{s}, \{s_t^*\}_{t=0}^{t^*-2}) \\
&= \{\mathbb{P}(\{y_{it} = s_t\}_{t\in\mathcal{O}_2}, \{y_{it} = s\}_{t\in\mathcal{O}_2^{\oplus}\cup\{t^*\}} \mid X, y_{it^*-1} = \tilde{s}, \{y_{it} = s_t^*\}_{t=0}^{t^*-2})\}_{\{s_t\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}} \\
&= \mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})\Pi(X, y_{it^*-1} = \tilde{s}, \{y_{it} = s_t^*\}_{t=0}^{t^*-2})\mathcal{P}_{t^*}^{(s)}(x_{t^*}, \tilde{s})
\end{aligned}$$

Given the arguments in Step 2, $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$ is identified and has full column rank for $s \in \{0,1\}$. Additionally, $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$ is not a function of $\{x_t\}_{t\notin\underline{\mathcal{O}_2}}$, whereas $\mathbf{P}_{\mathcal{O}_2\cup\{t^*\}}(X, s, \tilde{s}, \{s_t^*\}_{t=0}^{t^*-2})$ is identified for all $(X^{\top}, s, \tilde{s}, \{s_t^*\}_{t=0}^{t^*-2})^{\top} \in$

$\mathbb{X} \times \{0,1\}^{t^*+1}$ and $\Pi(X, y_{it^*-1} = \tilde{s}, \{y_{it} = s_t^*\}_{t=0}^{t^*-2})$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_2}}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_2}}$ and $(\tilde{s}, \{s_t^*\}_{t=0}^{t^*-2})$ from the arguments in Step 6. Thus, for any $(\tilde{x}_{t^*}^\top, \tilde{s})^\top \in \mathbb{X}_{t^*} \times \{0,1\}$ for which there exists $\{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{t^*\}}$ and $\{\underline{s}_t\}_{t=0}^{t^*-2}$ such that $\Pi(\{x_t\}_{t\in\underline{\mathcal{O}_2}}, \tilde{x}_{t^*}, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{t^*\}}, y_{it^*-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{t^*-2})$ has full rank, $\mathcal{P}_{t^*}(\tilde{x}_{t^*}, \tilde{s})$ is identified with $(\mathcal{P}_{t^*}^{(0)}(\tilde{x}_{t^*}, \tilde{s})^\top, \mathcal{P}_{t^*}^{(1)}(\tilde{x}_{t^*}, \tilde{s})^\top)^\top$ where for $s \in \{0,1\}$

$$\begin{aligned}\mathcal{P}_{t^*}^{(s)}(\tilde{x}_{t^*}, \tilde{s}) = \Pi(\{x_t\}_{t\in\underline{\mathcal{O}_2}}, \tilde{x}_{t^*}, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{t^*\}}, y_{it^*-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{t^*-2})^{-1} \\ \times \mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})^\dagger \mathbf{P}_{\mathcal{O}_2\cup\{t^*\}}(X, s, \tilde{s}, \{\underline{s}_t\}_{t=0}^{t^*-2})\end{aligned}$$

and the group labels are identical to the group labels in Step 4. Given Assumption I-3″, we note that $\mathcal{P}_{t^*}(x_{t^*}, \tilde{s})$ is identified for $\tilde{s} \in \{0,1\}$, which we make use of subsequently. An adaption of the above argument, analogous to *Step 5* in the proof of Theorem 3.2, implies that for any $s_{t^*} \in \{0,1\}$ and $j \in \{1, \ldots, J\}$, $\mathbb{P}_{t^*}(y_{it^*} = s_{t^*} \mid x_{it^*} = \tilde{x}_{t^*}, y_{it^*-1} = \tilde{s}, g_i = j)$ is identified at all $(\tilde{x}_{t^*}^\top, \tilde{s})^\top \in \mathbb{X}_{t^*} \times \{0,1\}$ for which there exists $\{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{t^*\}}$ and $\{\underline{s}_t\}_{t=0}^{t^*-2}$ such that $\pi_j(\{x_t\}_{t\in\underline{\mathcal{O}_2}}, \tilde{x}_{t^*}, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{t^*\}}, y_{it^*-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{t^*-2}) > 0$.

2. Period $\tilde{t} \in \{1, \ldots, t^*-1\}$: If $t^* - 1 \leq 0$, we are done so that we assume without loss of generality that $t^* - 1 \geq 1$. We consider the identified matrix

$$\begin{aligned}&\mathbf{P}_{\mathcal{O}_2\cup\{\tilde{t},\ldots,t^*\}}(X, s, \tilde{s}, \{s_t^*\}_{t=0}^{\tilde{t}-2}) \\ &= \{\mathbb{P}(\{y_{it} = s_t\}_{t\in\mathcal{O}_2}, \{y_{it} = s\}_{t\in\mathcal{O}_2^{\oplus}\cup\{\tilde{t},\ldots,t^*\}} \mid X, y_{i\tilde{t}-1} = \tilde{s}, \{y_{it} = s_t^*\}_{t=0}^{\tilde{t}-2})\}_{\{s_t\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}} \\ &= \left(\mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2}) \overset{\text{col}}{\otimes} \left(\odot_{t\in\{\tilde{t}+1,\ldots,t^*\}}\mathcal{P}_t^{(s)}(x_t, s)\right)^\top\right) \\ &\quad \times \Pi(X, y_{i\tilde{t}-1} = \tilde{s}, \{y_{it} = s_t^*\}_{t=0}^{\tilde{t}-2})\mathcal{P}_{\tilde{t}}^{(s)}(x_{\tilde{t}}, \tilde{s})\end{aligned}$$

where, abusing the notation, we drop $\{y_{it} = s_t^*\}_{t=0}^{\tilde{t}-2}$ from the above when $\tilde{t} = 1$.
We focus on period $\tilde{t} = t^* - 1$. An inductive argument argument applies to the remaining periods. From the previous point, $\mathcal{P}_t^{(s)}(x_t, s)$ is identified for $s \in \{0,1\}$ and is a vector containing no zeros by Assumption I-2″ (v) so that from Lemma A.10.2 $\mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2}) \overset{\text{col}}{\otimes} \left(\odot_{t\in\{\tilde{t}+1,\ldots,t^*\}}\mathcal{P}_t^{(s)}(x_t, s)\right)^\top = \mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2}) \overset{\text{col}}{\otimes} \mathcal{P}_{t^*}^{(s)}(x_{t^*}, s)^\top$ has full column rank for $s \in \{0,1\}$. Additionally, $\mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2}) \overset{\text{col}}{\otimes} \mathcal{P}_{t^*}^{(s)}(x_{t^*}, s)^\top$ is identified and not a function of $\{x_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{\tilde{t}+1,\ldots,t^*\}} = \{x_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{t^*\}}$, while, from Step 6, $\Pi(X, y_{i\tilde{t}-1} = \tilde{s}, \{y_{it} = s_t^*\}_{t=0}^{\tilde{t}-2})$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_2}}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_2}}$ and $\{s_t\}_{t=0}^{\tilde{t}-1} = \{s_t\}_{t=0}^{t^*-2}$, and $\mathbf{P}_{\mathcal{O}_2\cup\{\tilde{t},\ldots,t^*\}}(X, s, \tilde{s}, \{s_t^*\}_{t=0}^{\tilde{t}-2})$ is identified for all $(X^\top, s, \tilde{s}, \{s_t^*\}_{t=0}^{\tilde{t}-2})^\top$. Hence, $\mathcal{P}_{\tilde{t}}(\tilde{x}_{\tilde{t}}, \tilde{s})$ is identified at all $(\tilde{x}_{\tilde{t}}^\top, \tilde{s}) \in \mathbb{X}_{\tilde{t}} \times \{0,1\}$ for which there exists $\{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{\tilde{t},\tilde{t}+1,\ldots,t^*\}}$ and $\{\underline{s}_t\}_{t=0}^{\tilde{t}-2}$ such that $\Pi(\{x_t\}_{t\in\underline{\mathcal{O}_2}\cup\{\tilde{t}+1,\ldots,t^*\}}, \tilde{x}_{\tilde{t}}, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{\tilde{t},\tilde{t}+1,\ldots,t^*\}}, y_{i\tilde{t}-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{\tilde{t}-2})$ has full rank. Specifically, $\mathcal{P}_{\tilde{t}}(\tilde{x}_{\tilde{t}}, \tilde{s})$ is identified with $(\mathcal{P}_{\tilde{t}}^{(0)}(\tilde{x}_{\tilde{t}}, \tilde{s})^\top, \mathcal{P}_{\tilde{t}}^{(1)}(\tilde{x}_{\tilde{t}}, \tilde{s})^\top)^\top$ where

for $s \in \{0,1\}$

$$\mathcal{P}_{\tilde{t}}^{(s)}(\tilde{x}_{\tilde{t}}, \tilde{s}) = \Pi(\{x_t\}_{t\in\underline{\mathcal{O}_2}\cup\{\tilde{t}+1,\ldots,t^*\}}, \tilde{x}_{\tilde{t}}, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{\tilde{t},\tilde{t}+1,\ldots,t^*\}}, y_{i\tilde{t}-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{\tilde{t}-2})^{-1}$$
$$\times \left(\mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2}) \overset{\text{col}}{\otimes} \left(\odot_{t\in\{\tilde{t}+1,\ldots,t^*\}}\mathcal{P}_t^{(s)}(x_t, s)\right)^\top\right)^\dagger$$
$$\times \mathbf{P}_{\mathcal{O}_2\cup\{\tilde{t},\ldots,t^*\}}(X, s, \tilde{s}, \{\underline{s}_t\}_{t=0}^{\tilde{t}-2})$$

Applying the argument inductively implies an identical identification result for all $\tilde{t} \in \{1,\ldots,t^*-1\}$. The group labels are identical to the group labels in Step 4. Again, an adaption of the above argument, analogous to *Step 5* in the proof of Theorem 3.2, implies that for any $s_{\tilde{t}} \in \{0,1\}$ and $j \in \{1,\ldots,J\}$, $\mathbb{P}_{\tilde{t}}(y_{i\tilde{t}} = s_{\tilde{t}} \mid x_{i\tilde{t}} = \tilde{x}_{\tilde{t}}, y_{i\tilde{t}-1} = \tilde{s}, g_i = j)$ is identified at all $(\tilde{x}_{\tilde{t}}^\top, \tilde{s}) \in \mathbb{X}_{\tilde{t}} \times \{0,1\}$ for which there exists $\{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{\tilde{t},\tilde{t}+1,\ldots,t^*\}}$ and $\{\underline{s}_t\}_{t=0}^{\tilde{t}-2}$ such that $\pi_j(\{x_t\}_{t\in\underline{\mathcal{O}_2}\cup\{\tilde{t}+1,\ldots,t^*\}}, \tilde{x}_{\tilde{t}}, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{\tilde{t},\tilde{t}+1,\ldots,t^*\}}, y_{i\tilde{t}-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{\tilde{t}-2}) > 0$.

We note that the final period $T \notin \{1,\ldots,t^*\} \cup \underline{\mathcal{O}_2}$ so that x_T may be varied in $\pi_j(\cdot)$ for all $\tilde{t} \in \{1,\ldots,t^*-1\}$ to ensure that $\pi_j(\{x_t\}_{t\in\underline{\mathcal{O}_2}\cup\{\tilde{t}+1,\ldots,t^*\}}, \tilde{x}_{\tilde{t}}, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{\tilde{t},\tilde{t}+1,\ldots,t^*\}}, y_{i\tilde{t}-1} = \tilde{s}, \{y_{it} = \underline{s}_t\}_{t=0}^{\tilde{t}-2}) > 0$.

8. Identification of $\mathcal{P}_T(x_T, s)$. We extend the identification result of $\mathcal{P}_T(x_T, s)$ from Step 5. Specifically, we note that

$$\mathbf{P}_{\mathcal{O}_2\cup\{T\}}(X, s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})$$
$$= \{\mathbb{P}(\{y_{it} = s_t\}_{t\in\mathcal{O}_2\cup\{T\}}, \{y_{it} = s\}_{t\in\mathcal{O}_2^\oplus} \mid X, y_{i\max(\underline{\mathcal{O}_1})} = s, \{y_{it} = \tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})\}_{\{s_t\}_{t\in\mathcal{O}_2\cup\{T\}}\in\{0,1\}^{|\mathcal{O}_2|+1}}$$
$$= \mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})\Pi(X, y_{i\max(\underline{\mathcal{O}_1})} = s, \{y_{it} = \tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})\mathcal{P}_T(x_T, s)^\top$$

where $\mathbf{P}_{\mathcal{O}_2\cup\{T\}}(X, s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})$ is identified for all $(X^\top, s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})^\top$ and $\mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})$ is identified and has full column rank for $s \in \{0,1\}$ and is not a function of $\{x_t\}_{t\notin\underline{\mathcal{O}_2}}$. From Step 6, $\Pi(X, y_{i\max(\underline{\mathcal{O}_1})} = s, \{y_{it} = \tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_2}}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_2}}$ and $(s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})$. Since $T \notin \underline{\mathcal{O}_2}$, we conclude that $\mathcal{P}_T(\tilde{x}_T, s)$ is identified at $(\tilde{x}_T^\top, s) \in \mathbb{X}_T \times \{0,1\}$ for which there exists $\{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{T\}}$ and $\{\underline{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1}$ such that $\Pi(\{x_t\}_{t\in\underline{\mathcal{O}_2}}, \tilde{x}_T, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{T\}}, y_{i\max(\underline{\mathcal{O}_1})} = s, \{y_{it} = \underline{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})$ has full rank. Specifically,

$$\mathcal{P}_T(\tilde{x}_T, s)^\top = \Pi(\{x_t\}_{t\in\underline{\mathcal{O}_2}}, \tilde{x}_T, \{\tilde{x}_t\}_{t\notin\underline{\mathcal{O}_2}\cup\{T\}}, y_{i\max(\underline{\mathcal{O}_1})} = s, \{y_{it} = \underline{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})^{-1}$$
$$\times \mathcal{P}_{\mathcal{O}_2}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})^\dagger \mathbf{P}_{\mathcal{O}_2\cup\{T\}}(X, s, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})-1})$$

Again, the group labels are identical to the group labels in Step 4. Similar to before, an adaption of the above argument, analogous to *Step 5* in the proof of Theorem 3.2, implies that for any $s_T \in \{0,1\}$ and $j \in \{1,\ldots,J\}$, $\mathbb{P}_T(y_{iT} = s_T \mid x_{iT} = \tilde{x}_T, y_{iT-1} = s, g_i = j)$ is

identified at all $(\tilde{x}_T^\top, s) \in \mathbb{X}_T \times \{0,1\}$ for which there exists $\{\tilde{x}_t\}_{t \notin \underline{\mathcal{O}}_2 \cup \{T\}}$ and $\{\underline{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)-1}$ such that $\pi_j(\{x_t\}_{t \in \underline{\mathcal{O}}_2}, \tilde{x}_T, \{\tilde{x}_t\}_{t \notin \underline{\mathcal{O}}_2 \cup \{T\}}, y_{i \max(\underline{\mathcal{O}}_1)} = s, \{y_{it} = \underline{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}}_1)-1}) > 0$.

9. Identification of $\underline{\pi}(X, s)$. Take any submodel $\mathcal{O} \subseteq \mathcal{T}^{(\text{odd})}$ of adjacent periods and $\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}}$ such that for $\tilde{s} \in \{0,1\}$ $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}})$ has full column rank and is identified (up to identical relabeling) given 1., 2., or 3.. Let $\{\tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2} \in \{0,1\}^{\min(\mathcal{O})-1}$, $\tilde{X} = (\tilde{x}_1^\top, \ldots, \tilde{x}_T^\top)^\top$, and define the identified matrix

$$
\begin{aligned}
&\mathbf{P}_{\mathcal{O}}(\tilde{X}, \tilde{s}, \{\tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2}) \\
&= \{\mathbb{P}(\{y_{it} = s_t\}_{t \in \mathcal{O}}, \{y_{it} = \tilde{s}\}_{t \in (\mathcal{O}^\oplus \setminus \max(\mathcal{O}^\oplus))} \mid \tilde{X}, y_{i \min(\mathcal{O})-1} = \tilde{s}, \{y_{it} = \tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2})\}_{\{s_t\}_{t \in \mathcal{O}} \in \{0,1\}^{|\mathcal{O}|}} \\
&= \mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}}) \underline{\pi}(\tilde{X}, y_{i \min(\mathcal{O})-1} = \tilde{s}, \{y_{it} = \tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2}))
\end{aligned}
$$

where we abuse the notation and drop the conditioning set $\{\tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2}$ whenever $\min(\mathcal{O}) - 2 < 0$ and $\{y_{it} = \tilde{s}\}_{t \in (\mathcal{O}^\oplus \setminus \max(\mathcal{O}^\oplus))}$ when $\mathcal{O}^\oplus$ is a singleton. Hence,

$$
\underline{\pi}(\tilde{X}, y_{i \min(\mathcal{O})-1} = \tilde{s}, \{y_{it} = \tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2})) = \mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}})^\dagger \mathbf{P}_{\mathcal{O}}(\tilde{X}, \tilde{s}, \{\tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2})
$$

Since $\mathcal{P}_{\mathcal{O}}^{\otimes^-}(\{\tilde{x}_t, \tilde{s}, \tilde{x}_{t+1}, \tilde{s}\}_{t \in \mathcal{O}})$ is not a function of $\{\tilde{x}_t\}_{t \notin (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$, but has full column rank for $\tilde{s} \in \{0,1\}$, and $\mathbf{P}_{\mathcal{O}}(\tilde{X}, \tilde{s}, \{\tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2})$ is identified for all values in the support of its argument, we conclude that $\underline{\pi}(\tilde{X}, y_{i \min(\mathcal{O})-1} = \tilde{s}, \{y_{it} = \tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2}))$ is identified at $\{\tilde{x}_t\}_{t \in (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ for all $\{x_t\}_{t \notin (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ and $(\tilde{s}, \{\tilde{s}_t\}_{t=0}^{\min(\mathcal{O})-2})$. Following the arguments of Step 6, it therefore follows that $\underline{\pi}(\tilde{X}, y_{i0} = \tilde{s})$ is identified at $\{\tilde{x}_t\}_{t \in (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ for all $\{\tilde{x}_t\}_{t \notin (\underline{\mathcal{O}} \setminus \max(\underline{\mathcal{O}}))}$ and $\tilde{s} \in \{0,1\}$ where the group labels are identical to the group labels in Step 4.

This concludes the proof. We note that, as discussed in the previous proofs, we may strengthen the identification result by applying the same steps above to the identified values. □

A.7.2 Proof of Theorem A.1.6

Proof. The proof proceeds in multiple steps, which are similar to the steps in the proof of Theorems 3.3 and A.1.5.

Step 1. We define the identified matrix

$$
\begin{aligned}
\mathbf{P}_{\mathcal{O}_1 \cup \mathcal{O}_2}(X, s)^- = \Big\{ \Big\{ &\mathbb{P}(\{y_{it} = s_t\}_{t \in \mathcal{O}_1}, \{y_{it} = s\}_{t \in \mathcal{O}_1^\oplus}, \{y_{it} = s_t^*\}_{t \in \mathcal{O}_2}, \\
&\{y_{it} = s\}_{t \in \mathcal{O}_2^\oplus \setminus \{T+1\}} \mid X, y_{i0} = s) \Big\}_{\{s_t\}_{t \in \mathcal{O}_1} \in \{0,1\}^{|\mathcal{O}_1|}} \Big\}_{\{s_t^*\}_{t \in \mathcal{O}_2} \in \{0,1\}^{|\mathcal{O}_2|}}
\end{aligned}
$$

and observe

$$
\begin{aligned}
\mathbf{P}_{\mathcal{O}_1 \cup \mathcal{O}_2}(X, s)^- &= \mathcal{P}_{\mathcal{O}_1}^{\otimes}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_1}) \Pi(X, s) \mathcal{P}_{\mathcal{O}_2}^{\otimes^-}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_2})^\top \\
\mathbf{P}_{\mathcal{O}_1 \cup \mathcal{O}_2}(\tilde{X}, s)^- &= \mathcal{P}_{\mathcal{O}_1}^{\otimes}(\{\tilde{x}_t, s, \tilde{x}_{t+1}, s\}_{t \in \mathcal{O}_1}) \Pi(\tilde{X}, s) \mathcal{P}_{\mathcal{O}_2}^{\otimes^-}(\{x_t, s, x_{t+1}, s\}_{t \in \mathcal{O}_2})^\top
\end{aligned}
$$

$$\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{X},s)^- = \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\Pi(\underline{X},s)\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{\underline{x}_t,s,\underline{x}_{t+1},s\}_{t\in\mathcal{O}_2})^\top$$
$$\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{\tilde{X}},s)^- = \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{\tilde{x}_t,s,\tilde{x}_{t+1},s\}_{t\in\mathcal{O}_1})\Pi(\underline{\tilde{X}},s)\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{\underline{x}_t,s,\underline{x}_{t+1},s\}_{t\in\mathcal{O}_2})^\top$$

where $\tilde{X} = (\{\tilde{x}_t,\tilde{x}_{t+1}\}^\top_{t\in\mathcal{O}_1},\{x_t,x_{t+1}\}^\top_{t\in\mathcal{O}_2\setminus\{T\}},x_T^\top)^\top$, $\underline{X} = (\{x_t,x_{t+1}\}^\top_{t\in\mathcal{O}_1},\{\underline{x}_t,\underline{x}_{t+1}\}^\top_{t\in\mathcal{O}_2\setminus\{T\}},\underline{x}_T^\top)^\top$, and $\underline{\tilde{X}} = (\{\tilde{x}_t,\tilde{x}_{t+1}\}^\top_{t\in\mathcal{O}_1},\{\underline{x}_t,\underline{x}_{t+1}\}^\top_{t\in\mathcal{O}_2\setminus\{T\}},\underline{x}_T^\top)^\top$. Under Assumptions $\tilde{\text{I}}$-4′ and $\tilde{\text{I}}$-3′, we have

$$\begin{aligned}
&\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\tilde{X},s)^-)^\dagger\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{\tilde{X}},s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{X},s)^-)^\dagger\\
&= \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\Pi(X,s)\Pi(\tilde{X},s)^{-1}\Pi(\underline{\tilde{X}},s)\Pi(\underline{X},s)^{-1}\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})^\dagger\\
&\Rightarrow \mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\tilde{X},s)^-)^\dagger\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{\tilde{X}},s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{X},s)^-)^\dagger\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\\
&= \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\Pi(X,s)\Pi(\tilde{X},s)^{-1}\Pi(\underline{\tilde{X}},s)\Pi(\underline{X},s)^{-1}
\end{aligned}$$

where $\Pi(X,s)\Pi(\tilde{X},s)^{-1}\Pi(\underline{\tilde{X}},s)\Pi(\underline{X},s)^{-1}$ is a diagonal matrix with distinct and positive diagonal entries. Hence, up to joint permutation, the diagonal entries of $\Pi(X,s)\Pi(\tilde{X},s)^{-1}\Pi(\underline{\tilde{X}},s)\Pi(\underline{X},s)^{-1}$ are identified as the eigenvalues of $\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\tilde{X},s)^-)^\dagger\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{\tilde{X}},s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{X},s)^-)^\dagger$ and the columns of $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})K$ for some scaling matrix K – a diagonal matrix with non-zero entries – are identified by a set of corresponding eigenvectors of $\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\tilde{X},s)^-)^\dagger\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{\tilde{X}},s)^-(\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(\underline{X},s)^-)^\dagger$. Below, we argue that similar arguments as in the proof of Theorem A.1.5 imply that K is identified. Before doing so, we recall some notation from the proof of Theorem A.1.5

$$\begin{aligned}
\mathbf{P}_{\mathcal{O}_1}(X,s) &= \{\mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{O}_1},\{y_{it}=s\}_{t\in\mathcal{O}_1^\oplus})\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}}\\
&= \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\underline{\pi}(X,s)
\end{aligned}$$

where $\mathbf{P}_{\mathcal{O}_1}(X,s)$ is identified. It follows that $K^{-1}\underline{\pi}(X,s)$ is identified up to identical relabeling of the groups at $\{x_t\}_{t\in\underline{\mathcal{O}_1}}$ and s for all $\{x_t\}_{t\notin\underline{\mathcal{O}_1}}$ with

$$K^{-1}\underline{\pi}(X,s) = (\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})K)^\dagger\mathbf{P}_{\mathcal{O}_1}(X,s)$$

Hence, once K is identified $\underline{\pi}(X,s)$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_1}}$ and s for all $\{x_t\}_{t\notin\underline{\mathcal{O}_1}}$ up to the identical relabeling of the groups. Last, we note that $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})^\top$ is identified up to the identical relabeling of the groups by

$$\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_2})^\top = (K^{-1}\Pi(X,s))^{-1}(\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})K)^\dagger\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X,s)^-$$

$\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{\underline{x}_t,s,\underline{x}_{t+1},s\}_{t\in\mathcal{O}_2})^\top$ is similarly identified up to identical relabeling.

Step 2. For $\tilde{s}\neq s$, we define

$$\begin{aligned}
&\mathcal{P}^{\otimes}_{\mathcal{O}_1}(y_{i\max(\mathcal{O}_1)+1}=\tilde{s};\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\\
&= \mathcal{P}^{\otimes}_{\mathcal{O}_1\setminus\max(\mathcal{O}_1)}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1\setminus\max(\mathcal{O}_1)}) \overset{\text{col}}{\otimes} \mathcal{P}^{\otimes}_{\max(\mathcal{O}_1)}(x_{\max(\mathcal{O}_1)},s,x_{\max(\mathcal{O}_1)+1},\tilde{s})
\end{aligned}$$

where $\mathcal{P}^{\otimes}_{\mathcal{O}_1\setminus\max(\mathcal{O}_1)}(\{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1\setminus\max(\mathcal{O}_1)})^\top := \mathbf{1}_J \in \mathbb{R}^J$ if $\mathcal{O}_1 \setminus \max(\mathcal{O}_1) = \emptyset$, and the identified matrix

$$\begin{aligned}
&\mathbf{P}^{-}_{\mathcal{O}_1\cup\mathcal{O}_2}(X, \{y_{it}=s\}_{t\in\mathcal{O}_1^{\oplus}\setminus\{\max(\mathcal{O}_1)+1\}}, \{y_{it}=\tilde{s}\}_{t\in(\mathcal{O}_2^{\oplus}\cup\{\max(\mathcal{O}_1)+1\})\setminus\{T+1\}})\\
&= \Big\{\Big\{ \mathbb{P}(\{y_{it}=s_t\}_{t\in\mathcal{O}_1}, \{y_{it}=s\}_{t\in\mathcal{O}_1^{\oplus}\setminus\{\max(\mathcal{O}_1)+1\}}, \{y_{it}=s_t^*\}_{t\in\mathcal{O}_2},\\
&\quad \{y_{it}=\tilde{s}\}_{t\in(\mathcal{O}_2^{\oplus}\cup\{\max(\mathcal{O}_1)+1\})\setminus\{T+1\}} \mid X, y_{i0}=s)\Big\}_{\{s_t\}_{t\in\mathcal{O}_1}\in\{0,1\}^{|\mathcal{O}_1|}}\Big\}_{\{s_t^*\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}}
\end{aligned}$$

For $X^* = (x_1^{*\top}, \ldots, x_T^{*\top})^\top$, we then have

$$\begin{aligned}
&\mathbf{P}^{-}_{\mathcal{O}_1\cup\mathcal{O}_2}(X^*, \{y_{it}=s\}_{t\in\mathcal{O}_1^{\oplus}\setminus\{\max(\mathcal{O}_1)+1\}}, \{y_{it}=\tilde{s}\}_{t\in(\mathcal{O}_2^{\oplus}\cup\{\max(\mathcal{O}_1)+1\})\setminus\{T+1\}})\\
&= \mathcal{P}^{\otimes}_{\mathcal{O}_1}(y_{i\max(\mathcal{O}_1)+1}=\tilde{s}; \{x_t^*, s, x_{t+1}^*, s\}_{t\in\mathcal{O}_1})\Pi(X^*, s)\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t^*, \tilde{s}, x_{t+1}^*, \tilde{s}\}_{t\in\mathcal{O}_2})^\top
\end{aligned}$$

This holds specifically for all $X^* \in \{X, \tilde{X}, \underline{X}, \underline{\tilde{X}}\}$ so that the same arguments as in *Step 1* imply that for some scaling matrix $\tilde{K}$ and up to identical permutation of the group labels the following quantities are identified:

- $\Pi(X,s)\Pi(\tilde{X},s)^{-1}\Pi(\underline{\tilde{X}},s)\Pi(\underline{X},s)^{-1}$
- $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(y_{i\max(\mathcal{O}_1)+1}=\tilde{s}; \{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})\tilde{K}$
- $\tilde{K}^{-1}\Pi(X^*,s)$ for $X^* \in \{X, \tilde{X}\}$
- $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t,\tilde{s},x_{t+1},\tilde{s}\}_{t\in\mathcal{O}_2})^\top$ and $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{\underline{x}_t,\tilde{s},\underline{x}_{t+1},\tilde{s}\}_{t\in\mathcal{O}_2})^\top$

By aligning the distinct diagonal entries of $\Pi(X,s)\Pi(\tilde{X},s)^{-1}\Pi(\underline{\tilde{X}},s)\Pi(\underline{X},s)^{-1}$ in this step with *Step 1*, we align the group labels of this step with the group labels of *Step 1*. Hence, all objects in this step and *Step 1* are identified up to identical relabeling of the groups. Also, we note that $\tilde{K}^{-1}K = \tilde{K}^{-1}\Pi(X,s)(K^{-1}\Pi(X,s))^{-1}$ since $\Pi(X,s)$ is by assumption invertible. As the RHS is identified, $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(y_{i\max(\mathcal{O}_1)+1}=\tilde{s}; \{x_t,s,x_{t+1},s\}_{t\in\mathcal{O}_1})K$ is identified up to identical relabeling of the groups.

Step 3. For $X^* \in \mathbb{X}$ and $\tilde{s} \neq s$, we have

$$\mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X^*,\tilde{s})^- = \mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t^*,\tilde{s},x_{t+1}^*,\tilde{s}\}_{t\in\mathcal{O}_1})\Pi(X^*,\tilde{s})\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t^*,\tilde{s},x_{t+1}^*,\tilde{s}\}_{t\in\mathcal{O}_2})^\top$$

This is especially true for $X^* \in \{X, \tilde{X}, \underline{X}, \underline{\tilde{X}}\}$ so that the same arguments as in *Step 1* imply that the following objects are identified up to identical relabeling of the groups and scaling:

- $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t,\tilde{s},x_{t+1},\tilde{s}\}_{t\in\mathcal{O}_1})K^*$
- $(K^*)^{-1}\underline{\pi}(X,\tilde{s})$ at $\{x_t\}_{t\in\underline{\mathcal{O}_1}}$ and $\tilde{s}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_1}}$
- $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t,\tilde{s},x_{t+1},\tilde{s}\}_{t\in\mathcal{O}_2})^\top$

Since $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t,\tilde{s},x_{t+1},\tilde{s}\}_{t\in\mathcal{O}_2})$ has full column rank, it has unique columns. Hence, one can align the group labels of this step with the group labels of the previous steps by aligning the

columns of $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_2})^\top$ in this step with the one of *Step 2*. Thus, all objects of this step are identified up to the same group relabeling as in the previous steps.

Step 4. We have the following **intermediate identification result**: The following quantities are identified up to identical group relabeling:

- $\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_1})$ is identified for $\tilde{s} \in \{0,1\}$ (up to an identified rescaling matrix).,
- $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, \tilde{s}, x_{t+1}, \tilde{s}\}_{t\in\mathcal{O}_2})$ is identified for $\tilde{s} \in \{0,1\}$, and
- $\underline{\pi}(X, s)$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_1}}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_1}}$ and $s \in \{0,1\}$ (up to an identified rescaling matrix).

Identification of rescaling matrices: We proceed to argue that K and K^* are identified. To this end, we use similar arguments as in *Step 6* in the proof of Theorem A.1.5. Specifically, we observe

$$\begin{aligned}\mathbf{P}_{\mathcal{O}_2}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})}) &= \{\mathbb{P}(\{y_{it} = s_t\}_{t\in\mathcal{O}_2}, \{y_{it} = \tilde{s}_{\max(\underline{\mathcal{O}_1})}\}_{t\in\mathcal{O}_2^{\oplus}\setminus\{T+1\}} \mid X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})})\}_{\{s_t\}_{t\in\mathcal{O}_2}\in\{0,1\}^{|\mathcal{O}_2|}} \\ &= \mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, \tilde{s}_{\max(\underline{\mathcal{O}_1})}, x_{t+1}, \tilde{s}_{\max(\underline{\mathcal{O}_1})}\}_{t\in\mathcal{O}_2})\underline{\pi}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})})\end{aligned}$$

for $\{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})} \in \{0,1\}^{|\underline{\mathcal{O}_1}|+1}$. As $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, \tilde{s}_{\max(\underline{\mathcal{O}_1})}, x_{t+1}, \tilde{s}_{\max(\underline{\mathcal{O}_1})}\}_{t\in\mathcal{O}_2})$ has full column rank and is identified up to identical relabeling for $\tilde{s}_{\max(\underline{\mathcal{O}_1})} \in \{0,1\}$ and $\mathbf{P}_{\mathcal{O}_2}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})})$ is identified for all $(X^\top, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})}) \in \mathbb{X} \times \{0,1\}^{|\underline{\mathcal{O}_1}|+1}$, we conclude that $\underline{\pi}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})})$ is identified at $\{x_t\}_{t\in\underline{\mathcal{O}_2}}$ for all $\{x_t\}_{t\notin\underline{\mathcal{O}_2}}$ and $\{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})} \in \{0,1\}^{|\underline{\mathcal{O}_1}|+1}$ with

$$\underline{\pi}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})}) = \mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, \tilde{s}_{\max(\underline{\mathcal{O}_1})}, x_{t+1}, \tilde{s}_{\max(\underline{\mathcal{O}_1})}\}_{t\in\mathcal{O}_2})^\dagger \mathbf{P}_{\mathcal{O}_2}(X, \{\tilde{s}_t\}_{t=0}^{\max(\underline{\mathcal{O}_1})})$$

where the group labels are identical to the ones in the previous steps. Now, the arguments in *Step 6* in the proof of Theorem A.1.5 imply that, up to identical relabeling of the groups, $\underline{\pi}(X, s)$ is identified for $s \in \{0,1\}$. This can now be used to identify K and K^*. For instance, K is identified via

$$K^{-1} = (\mathcal{P}^{\otimes}_{\mathcal{O}_1}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_1})K)^\dagger \mathbf{P}_{\mathcal{O}_1\cup\mathcal{O}_2}(X, s)^- (\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, s, x_{t+1}, s\}_{t\in\mathcal{O}_2})^\top)^\dagger \Pi(X, s)^{-1}$$

K^* is identified similarly.

Remainder of the proof. The remainder of the proof follows from similar steps as *Step 5 – Step 9* in the proof of Theorem A.1.5. Specifically, one may replace $\mathcal{O}_2^{\oplus}$ with $\mathcal{O}_2^{\oplus}\setminus\{T+1\}$ and $\mathcal{P}^{\otimes}_{\mathcal{O}_2}(\{x_t, y_{t-1}, x_{t+1}, y_{t+1}\}_{t\in\mathcal{O}_2})$ with $\mathcal{P}^{\otimes^-}_{\mathcal{O}_2}(\{x_t, y_{t-1}, x_{t+1}, y_{t+1}\}_{t\in\mathcal{O}_2})$. We omit the remaining details. This concludes the proof. □

A.7.3 Proof of Corollary A.1.6.1

Proof. The proof follows immediately from Theorem A.1.6. □

A.8 Proofs for Appendix A.1.8

A.8.1 Proof of Theorem A.1.7

Proof. Defining $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X)$ and $\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X)$ as in Section 3.2, now conditioning on $X = (x_0^\top,\ldots,x_T^\top)^\top$, we have

$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2}(X) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\overline{\mathcal{I}}_1})\Pi(X)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\overline{\mathcal{I}}_2})^\top$$
$$\mathbf{P}_{\mathcal{I}_1\cup\mathcal{I}_2\cup\{t'\},k}(X) = \mathcal{P}_{\mathcal{I}_1}(\{x_t\}_{t\in\overline{\mathcal{I}}_1})(X)\mathcal{D}_{t',k}(x_{t'},x_{t'-1})\Pi(X)\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\overline{\mathcal{I}}_2})^\top$$

where $\mathcal{D}_{t',k}(x_{t'},x_{t'-1}) = \operatorname{diag}([\mathcal{P}_{t'}(x_{t'},x_{t'-1})]_{k,\cdot})$ for $k=1,2$. Following the arguments in the proof of Theorem 3.2 in combination with Lemma A.10.1, we know that $\mathcal{P}_t(x_t,x_{t-1})$ for all $t=1,\ldots,T$ and $\underline{\pi}(X)$ are identified up to the same relabeling. Hence, $\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\overline{\mathcal{I}}_\ell})$ is identified up to the same relabeling for $\ell=1,2$.

Throughout the proof, we will use that given Assumption I-2‴ $\overline{\mathcal{I}}_1\subseteq\{0,1,\ldots,t'-1\}$ and $\overline{\mathcal{I}}_2\subseteq\{t',t'+1,\ldots,T\}$. It now follows:

1. *Identification of $\underline{\pi}(X)$ (I)*: For ℓ, define $\mathbf{P}_{\mathcal{I}_\ell}(X) = \mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\overline{\mathcal{I}}_\ell})\underline{\pi}(X) = \{\sum_{j=1}^J \pi_j(X)\prod_{t\in\mathcal{I}_\ell}\mathbb{P}_t(y_{it}=s_t \mid x_{it}=x_t, x_{it-1}=x_{t-1}, g_i=j)\}_{\{s_t\}_{t\in\mathcal{I}_\ell}\in\{0,1\}^{|\mathcal{I}_\ell|}}\in\mathbb{R}^{2^{|\mathcal{I}_\ell|}}$, which is identified. Then, under Assumption I-2‴, for $\ell=1,2$

$$\underline{\pi}(X) = \mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\overline{\mathcal{I}}_\ell})^\dagger\mathbf{P}_{\mathcal{I}_\ell}(X)$$

   $\mathcal{P}_{\mathcal{I}_\ell}(\{x_t\}_{t\in\overline{\mathcal{I}}_\ell})$ is a function of the covariates in the time periods $\overline{\mathcal{I}}_\ell$, while $\mathbf{P}_{\mathcal{I}_\ell}(X)$ is identified for all $X\in\mathbb{X}$. Hence, at $\{x_t\}_{t\in\overline{\mathcal{I}}_\ell}$, $\underline{\pi}(X)$ is identified for all $\{x_t\}_{t\notin\overline{\mathcal{I}}_\ell}$ up to the same relabeling of the groups as the other objects. That the relabeling is identical follows from the arguments in the proof of Theorem 3.2. We conclude

   - $\underline{\pi}(\{x_t\}_{t\in\overline{\mathcal{I}}_1},\{\tilde{x}_t\}_{t\notin\overline{\mathcal{I}}_1})$ is identified for all $\{\tilde{x}_t\}_{t\notin\overline{\mathcal{I}}_1}$ up to the same relabeling.
   - $\underline{\pi}(\{x_t\}_{t\in\overline{\mathcal{I}}_2},\{\tilde{x}_t\}_{t\notin\overline{\mathcal{I}}_2})$ is identified for all $\{\tilde{x}_t\}_{t\notin\overline{\mathcal{I}}_2}$ up to the same relabeling.

2. $t\in\{1,\ldots,t'-1\}$: We observe

$$\Big\{\mathbb{P}(\{y_{it'}=s_{t'}\}_{t'\in\mathcal{I}_2\cup\{t\}}\mid X)\Big\}_{\{s_{t'}\}_{t'\in\mathcal{I}_2\cup\{t\}}\in\{0,1\}^{|\mathcal{I}_2|+1}}$$
$$= \mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X) = \mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\overline{\mathcal{I}}_2})\Pi(X)\mathcal{P}_t(x_t,x_{t-1})^\top$$

   so that under Assumptions I-2‴ and I-3, we have

$$\mathcal{P}_t(x_t,x_{t-1})^\top = \Pi(X)^{-1}\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\overline{\mathcal{I}}_2})^\dagger\mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X) \tag{A.20}$$

   As $\mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X)$ is identified for all $X\in\mathbb{X}$, $\mathcal{P}_{\mathcal{I}_2}(\{x_t\}_{t\in\overline{\mathcal{I}}_2})$ is a function of $\{x_t\}_{t\in\overline{\mathcal{I}}_2}$, and, at $\{x_t\}_{t\in\overline{\mathcal{I}}_2}$, $\Pi(X)$ is identified for all $\{x_t\}_{t\notin\overline{\mathcal{I}}_2}$, the RHS is identified for all $\{x_t\}_{t\notin\overline{\mathcal{I}}_2}$ as long as $\Pi(X)$ is invertible. Noting that $t\notin\overline{\mathcal{I}}_2$ and $t-1\notin\overline{\mathcal{I}}_2$, we conclude that $\mathcal{P}_t(\tilde{x}_t,\tilde{x}_{t-1})$ is

identified at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which there exists $\{x^*_{t^*}\}_{t^* \notin \overline{\mathcal{I}}_2 \cup \{t,t-1\}}$ such that $\Pi(\{x_t\}_{t\in\overline{\mathcal{I}}_2}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1})$ has full rank. More so, the arguments around equation (A.7) in *Step 5* of the proof of Theorem 3.2 can be applied to equation (A.20). Doing so implies that for all $s_t \in \{0,1\}$ and $j \in \{1,\dots,J\}$, $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ is identified at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}$ such that $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_2}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_2\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1}) > 0$.

3. $t \in \{t'+1,\dots,T\}$: A symmetric argument as in the previous step now based on $\mathcal{I}_1$ implies that $\mathbb{P}_t(y_{it} = s_t \mid x_{it} = \tilde{x}_t, x_{it-1} = \tilde{x}_{t-1}, g_i = j)$ is identified for all $s_t \in \{0,1\}$ at $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top \in \mathbb{X}_t \times \mathbb{X}_{t-1}$ for which there exists $\{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_1\cup\{t,t-1\}}$ such that $\pi_j(\{x_t\}_{t\in\overline{\mathcal{I}}_1}, \{x^*_{t^*}\}_{t^*\notin\overline{\mathcal{I}}_1\cup\{t,t-1\}}, \tilde{x}_t, \tilde{x}_{t-1}) > 0$.
4. $t = t'$: For some submodel $\mathcal{I}$ with $t \notin \mathcal{I}$, we define $\mathbf{P}_{\mathcal{I}\cup\{t\}}(X)$ and $\mathbf{P}_{\mathcal{I}}(X)$ analogously to $\mathbf{P}_{\mathcal{I}_2\cup\{t\}}(X)$ and $\mathbf{P}_{\mathcal{I}_2}(X)$, respectively.
   Now, take any $(\tilde{x}_{t'}^\top, \tilde{x}_{t'-1}^\top)^\top \in \mathbb{X}_{t'} \times \mathbb{X}_{t'-1}$ for which there exists $\{\tilde{x}_{t^*}\}_{t^*\notin\{t',t'-1\}}$ and a submodel $\mathcal{I} \subseteq \{1,\dots,T\} \setminus \{t'\}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})$ has full column rank and is identified via argument 2. or 3. above, then $\Pi(\tilde{X})$ with $\tilde{X} = (\tilde{x}_0^\top, \dots, \tilde{x}_T^\top)^\top$ is identified via

$$\underline{\pi}(\tilde{X}) = \mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})^\dagger \mathbf{P}_{\mathcal{I}}(\tilde{X})$$

   Next, we let $\mathbf{J} = \{j \in \{1,\dots,J\} : \pi_j(\tilde{X}) > 0\}$ be the set of groups with positive component weight at $\tilde{X}$, which is identified, and define $[\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})]^\dagger_{\cdot,j\in\mathbf{J}}$ and $[\mathbf{P}_{t'}(\tilde{x}_{t'}, \tilde{x}_{t'-1})]_{\cdot,j\in\mathbf{J}}$ as the collection of the columns of the respective matrices associated with the column index $j \in \mathbf{J}$ as well as $[\Pi(\tilde{X})]_{j\in\mathbf{J},j\in\mathbf{J}}$ as the $|\mathbf{J}| \times |\mathbf{J}|$ submatrix of $\Pi(\tilde{X})$ containing the non-zero component weights on its diagonal.[36] Then,

$$[\mathbf{P}_{t'}(\tilde{x}_{t'}, \tilde{x}_{t'-1})]^\top_{\cdot,j\in\mathbf{J}} = [\Pi(\tilde{X})]^{-1}_{j\in\mathbf{J},j\in\mathbf{J}} [\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})]^\dagger_{\cdot,j\in\mathbf{J}} \mathbf{P}_{\mathcal{I}\cup\{t\}}(\tilde{X})$$

   where the RHS is identified and thus the LHS is identified up to the identical relabeling as the RHS, that is, for all $s_t \in \{0,1\}$ $\mathbb{P}_{t'}(y_{it'} = s_{t'} \mid x_{it'} = \tilde{x}_{t'}, x_{it'-1} = \tilde{x}_{t'-1}, g_i = j)$ is identified for all $j \in \mathbf{J}$.
5. *Identification of $\underline{\pi}(X)$ (II)*: For any submodel $\mathcal{I}$ and $\{\tilde{x}_t\}_{t\in\mathcal{I}}$ such that $\mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})$ has full column rank and is identified, we have $\underline{\pi}(\tilde{X}) = \mathcal{P}_{\mathcal{I}}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}})^\dagger \mathbf{P}_{\mathcal{I}}(\tilde{X})$ so that, following previous arguments, $\underline{\pi}(\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}}, \{x_t\}_{t\notin\overline{\mathcal{I}}})$ is identified at $\{\tilde{x}_t\}_{t\in\overline{\mathcal{I}}}$ for all $\{x_t\}_{t\notin\overline{\mathcal{I}}}$.

This concludes the proof.

We proceed to argue how the claim may be further strengthened. Specifically, once $\mathcal{P}_t(\tilde{x}_t, \tilde{x}_{t-1})$ has been identified for multiple values of $(\tilde{x}_t^\top, \tilde{x}_{t-1}^\top)^\top$, one can combine multiple time periods and find other covariate values X and submodels $\mathcal{I}$ such that $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\overline{\mathcal{I}}})$ is

[36] For a precise definition of these matrices, we refer to the proof of Theorem 3.2, see equation (A.7).

identified and has full column rank. Then, one can work through the arguments of the preceding steps with this submodel to improve on the identification result. □

A.8.2 Proof of Theorem A.1.8

Proof. The proof works through the same arguments as the proof of Theorem A.1.7 so that we omit it. □

A.8.3 Proof of Corollary A.1.8.1

Proof. The claim follows directly from Theorem A.1.8. □

A.9 Proofs for Appendix A.1.10

A.9.1 Proof of Lemma A.1.11

Proof. We begin with the first if and only if statement:

$\Rightarrow$: If $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ has full column rank, the assertion holds by definition for $\mathcal{I}^* = \mathcal{I}$.

$\Leftarrow$: Assume otherwise, that is, $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ does not have full column rank. Then, there exists $\gamma \neq 0 \in \mathbb{R}^J$ such that $\sum_{j=1}^J \gamma_j [\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})]_{\cdot,j} = 0$. Taking j^* with $\gamma_{j^*} \neq 0$, we have $[\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})]_{\cdot,j^*} = \sum_{j\neq j^*} \frac{\gamma_j}{-\gamma_{j^*}} [\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})]_{\cdot,j} = 0$. Since the columns of $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ add up to one, we have $\sum_{j\neq j^*} \frac{\gamma_j}{-\gamma_{j^*}} = 1$. Next, take any $\mathcal{I}^* \subseteq \mathcal{I}$. Adding up the respective rows of $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ gives $\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})$. It therefore follows that for any $\mathcal{I}^* \subseteq \mathcal{I}$, we have $[\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})]_{\cdot,j^*} = \sum_{j\neq j^*} \frac{\gamma_j}{-\gamma_{j^*}} [\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})]_{\cdot,j} = 0$, which concludes the proof of the first if and only if statement.

The second if and only if statement uses the following observation: There does not exist $\gamma \neq 0 \in \mathbb{R}^J$ and $j^* \in \{1,\dots,J\}$ such that $[\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})]_{\cdot,j^*} = \sum_{j\neq j^*} \frac{\gamma_j}{-\gamma_{j^*}} [\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})]_{\cdot,j}$ with $\sum_{j\neq j^*} \frac{\gamma_j}{-\gamma_{j^*}} = 1$ for all $\mathcal{I}^* \subseteq \mathcal{I}$ if and only if there does not exist $\gamma \neq 0$ such that $\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})\gamma = 0$ with $\sum_{j=1}^J \gamma_j = 0$ for all $\mathcal{I}^* \subseteq \mathcal{I}$.

$\Leftarrow$: Take $\mathcal{I}^* = \mathcal{I}$ and assume that there exists $\gamma \neq 0$ such that $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})\gamma = 0$. Since $[\mathcal{P}_t(x_t)]_{2,\cdot} = 1 - [\mathcal{P}_t(x_t)]_{1,\cdot}$, any entry in $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})\gamma$ may represented as follows

$$\sum_{j=1}^J \prod_{t\in\mathcal{I}_1} (1 - [\mathcal{P}_t(x_t)]_{1,j}) \prod_{t\in\mathcal{I}_2} [\mathcal{P}_t(x_t)]_{1,j}\gamma_j = 0$$

for some partition $\mathcal{I}_1$ and $\mathcal{I}_2$ of $\mathcal{I}$; when $\mathcal{I}_j = \emptyset$ for some $j = 1, 2$, then a product over this set is defined to be 1. Now, take $\mathcal{I}_1 = \emptyset$ and $\mathcal{I}_2 = \mathcal{I}$, then $\sum_{j=1}^J \prod_{t\in\mathcal{I}} [\mathcal{P}_t(x_t)]_{1,j}\gamma_j = 0$. Next, take $\mathcal{I}_1 = \{t^*\}$ and $\mathcal{I}_2 = \mathcal{I} \setminus \{t^*\}$, then, using the previous equality, $\sum_{j=1}^J (1 - [\mathcal{P}_{t^*}(x_t^*)]_{1,j}) \prod_{t\in\mathcal{I}_2} [\mathcal{P}_t(x_t)]_{1,j}\gamma_j = 0 \Leftrightarrow \sum_{j=1}^J \prod_{t\in\mathcal{I}_2} [\mathcal{P}_t(x_t)]_{1,j}\gamma_j = 0$. Continuing this argument and collecting the respective terms in the sums gives the rows of $V(\{x_t\}_{t\in\mathcal{I}})$ and thus implies that $V(\{x_t\}_{t\in\mathcal{I}})$ does not have full rank as there exists γ such that $V(\{x_t\}_{t\in\mathcal{I}})\gamma = 0$.

$\Rightarrow$: Assume there exists $\gamma \neq 0$ such that $V(\{x_t\}_{t\in\mathcal{I}})\gamma = 0$. Then, following a similar argument as in the previous part, this implies that there exists $\gamma \neq 0$ such that $\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})\gamma = 0$ with $\sum_{j=1}^{J}\gamma_j = 0$ for all $\mathcal{I}^* \subseteq \mathcal{I}$. □

A.10 Additional technical results

Lemma A.10.1 *For some submodel $\mathcal{I}$, let $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ have full column rank. Then, $\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*}) = \mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}}) \overset{col}{\otimes} \mathcal{P}_t(x_t)$ has full column rank with $\mathcal{I}^* = \mathcal{I}\cup\{t\}$ and $\mathcal{P}_t(x_t)$ as defined in Section 3.2.*

Proof. Assume that $\mathcal{P}_{\mathcal{I}^*}(\{x_t\}_{t\in\mathcal{I}^*})$ does not have full column rank, then there exists $\gamma \neq 0 \in \mathbb{R}^J$ such that $\sum_{j=1}^{J}\gamma_j[\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})]_{\cdot,j} \otimes [\mathcal{P}_t(x_t)]_{\cdot,j} = 0$. Since the columns of $\mathcal{P}_t(x_t)$ add up to 1, it follows that $\sum_{j=1}^{J}\gamma_j[\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})]_{\cdot,j} = 0$, that is, $\mathcal{P}_{\mathcal{I}}(\{x_t\}_{t\in\mathcal{I}})$ does not have full column rank. □

The next lemma is similar to the previous one, but even more immediate. It may be used to generalize the previous lemma.

Lemma A.10.2 *Let Q be a $q \times J$ matrix and v a $J \times 1$ vector with non-zero entries v_j for $j = 1,\ldots,J$. If Q has full column rank, then $Q \overset{col}{\otimes} v^\top$ has full column rank.*

Proof. Assume otherwise, that is, $Q \overset{\text{col}}{\otimes} v^\top$ does not have full rank. Then, there exists $\gamma \neq 0 \in \mathbb{R}^J$ such that $\sum_{j=1}^{J}[Q]_{\cdot,j}v_j\gamma_j = 0$. Since $v_j \neq 0$ for all $j = 1,\ldots,J$, this implies that Q does not have full rank. □

Lemma A.10.3 *We have the following set of equations:*

$$\mathbf{P}_1 = \mathcal{P}_1\underline{\pi}\ ;\quad \mathbf{P}_2 = \mathcal{P}_2\underline{\pi}^*\ ;\quad \mathbf{P}_{1,2} = \mathcal{P}_1\Pi\mathcal{P}_2^\top\ ;\quad \Pi = diag(\underline{\pi})$$

where $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_{1,2}$ are observed, and $\mathcal{P}_1$ and $\mathcal{P}_2$ have full column rank, while $\underline{\pi}$ contains only non-zero entries. Additionally, for some diagonal matrix K with non-zero diagonal entries we observe $\mathcal{P}_1^K = \mathcal{P}_1K$.

1. *If $\underline{\pi} \neq \underline{\pi}^*$: $\mathcal{P}_2$ and $\underline{\pi}^*$ are identified.*
2. *If $\underline{\pi} = \underline{\pi}^*$: K, $\mathcal{P}_1$, $\mathcal{P}_2$, and $\underline{\pi}$ are identified.*

Proof. We first observe that $\mathbf{P}_1 = \mathcal{P}_1KK^{-1}\underline{\pi} \Rightarrow K^{-1}\mathcal{P}_1^\dagger\mathbf{P}_1 = K^{-1}\underline{\pi}$ so that $K^{-1}\underline{\pi}$ can be treated as known. Hence, the following matrix can be treated as known

$$\Pi^K = \text{diag}(K^{-1}\underline{\pi}) = K^{-1}\text{diag}(\underline{\pi}) = K^{-1}\Pi$$

Next, we have

$$\mathbf{P}_{1,2} = \mathcal{P}_1KK^{-1}\Pi\mathcal{P}_2^\top = \mathcal{P}_1^K\Pi^K\mathcal{P}_2^\top \Rightarrow \mathcal{P}_2^\top = (\Pi^K)^{-1}(\mathcal{P}_1^K)^\dagger\mathbf{P}_{1,2}$$

that is, $\mathcal{P}_2$ is identified. Then, $\underline{\pi}^* = \mathcal{P}_2^\dagger \mathbf{P}_2$, which proves the first part of the claim.

Now, if $\underline{\pi}^* = \underline{\pi}$, Π is identified so that

$$\mathbf{P}_{1,2} = \mathcal{P}_1^K K^{-1} \Pi \mathcal{P}_2^\top \Rightarrow (\mathcal{P}_1^K)^\dagger \mathbf{P}_{1,2} (\mathcal{P}_2^\top)^\dagger \Pi^{-1} = K^{-1}$$

where the LHS is identified so that K^{-1} and thus K is identified. This implies that $\mathcal{P}_1 = \mathcal{P}_1^K K^{-1}$ is identified. This proves the second part of the claim. □

A.11 Additional simulation results

A.11.1 Weights – Time average

We consider a setting with $J = 3$ and $T = 4$ and keep the ordered probit model from Section 5.1 but change the index as follows

$$\begin{aligned}
g_i = 1 \text{ iff } & \ \zeta_0 + \zeta_1 \overline{X}_{i,1}^2 + \zeta_2 \log(\overline{X}_{i,2}^2 + 1) + \zeta_3 \overline{X}_{i,3} + u_i \leq 0 \\
g_i = 2 \text{ iff } & \ 0 < \zeta_0 + \zeta_1 \overline{X}_{i,1}^2 + \zeta_2 \log(\overline{X}_{i,2}^2 + 1) + \zeta_3 \overline{X}_{i,3} + u_i \leq \mu \\
g_i = 3 \text{ iff } & \ \zeta_0 + \zeta_1 \overline{X}_{i,1}^2 + \zeta_2 \log(\overline{X}_{i,2}^2 + 1) + \zeta_3 \overline{X}_{i,3} + u_i > \mu
\end{aligned}$$

where $u_i \sim N(0,1)$ as before and $\zeta = (0.5, -6, 7, 1)^\top$ and $\mu = 1$. Given this group assignment mechanism, the two groups are of similar size in expectation. The coefficients are now

$$\begin{aligned}
&\beta_1^{(0)} = (0.8, -0.4, -0.3)^\top \ ; \ \ \beta_2^{(0)} = (-1.3, -0.2, -1.0)^\top \ ; \ \ \beta_3^{(0)} = (1.2, 0.5, -0.1)^\top \\
&\alpha_1^{(0)} = (1, -1.5, -0.6, -0.2)^\top \ ; \ \ \alpha_2^{(0)} = (-0.7, 1.0, 0.7, 0.4)^\top \ ; \ \ \alpha_3^{(0)} = (-0.1, -0.4, 1.2, -0.8)^\top
\end{aligned}$$

Apart from these changes, the DGP remains as in Sections 5 and 5.1. We approximate the weight functions via a multinomial link function as discussed in Section 4.1 where we choose orthogonal Legendre polynomials up to order $d(n) = \lceil n^{1/7} \rceil$. Table A.4 contains the simulation results for the estimator of $\beta^{(0)}$ for $n \in \{1000, 2000\}$ and Table A.5 contains the results for the estimator of the group-specific AMEs for $n = 1000$. Table 4 in Section 5.1 contains the simulation results for the estimator of the group-specific AMEs for $n = 2000$.

Due to numerical issues, we drop some simulation runs either because the coefficients are unreasonably large (in absolute value greater than 10) or because the numerical optimization routine threw an error message. This numerical instability is reflected in the fact that the RMSE may drop by more than the expected factor of $1/\sqrt{2}$ when doubling the sample size from $n = 1000$ to $n = 2000$. The numerical issues disappear as the sample size increases or identification "becomes easier", that is, T increases. The lower part of a table indicates the number of successful and problematic simulation runs.

Table A.4: Simulation results for the coefficients. Weights depend on time averages, $J = 3$ and $T = 4$.

	$n = 1000$ and $T = 4$						$n = 2000$ and $T = 4$					
	Bias			**RMSE**			**Bias**			**RMSE**		
	Gr. 1	Gr. 2	Gr. 3	Gr. 1	Gr. 2	Gr. 3	Gr. 1	Gr. 2	Gr. 3	Gr. 1	Gr. 2	Gr. 3
α_1	0.046	-0.003	-0.003	0.143	0.150	0.104	0.026	-0.001	-0.001	0.094	0.095	0.069
α_2	-0.081	-0.014	-0.019	0.378	0.180	0.109	-0.030	-0.013	-0.011	0.128	0.117	0.072
α_3	-0.030	0.003	0.023	0.127	0.139	0.153	-0.013	-0.001	0.007	0.084	0.089	0.107
α_4	-0.001	-0.003	-0.024	0.088	0.130	0.127	0.005	-0.002	-0.010	0.060	0.088	0.081
β_1	0.018	0.005	0.035	0.111	0.181	0.145	0.010	-0.002	0.012	0.076	0.120	0.089
β_2	-0.022	0.008	0.020	0.137	0.140	0.117	-0.009	0.007	0.012	0.090	0.096	0.076
β_3	-0.007	-0.004	0.004	0.115	0.145	0.108	-0.003	0.000	0.002	0.075	0.098	0.072
Group assign.			83.35%						85.12%			
#sim			999						1000			
#num. p			1						1000			
#large coef.			0						0			

Table A.5: Simulation results for the AMEs. Weights depend on time averages and $J = 3$. The sample size of $n = 1000$ and $T = 4$.

	Gr. 1				Gr. 2				Gr. 3			
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 1$	$t = 2$	$t = 3$	$t = 4$
						Bias						
x_1	-0.0045	-0.0052	-0.0012	0.0025	0.0037	0.0011	0.0033	0.0031	0.0006	-0.0002	0.0024	0.0003
x_2	-0.0005	0.0009	-0.0031	-0.0058	0.0019	0.0014	0.0020	0.0020	0.0019	0.0016	0.0018	0.0015
x_3	0.0014	0.0018	0.0004	-0.0010	0.0006	-0.0011	0.0004	0.0001	0.0017	0.0017	0.0010	0.0014
						RMSE						
x_1	0.0274	0.0281	0.0304	0.0330	0.0297	0.0300	0.0302	0.0302	0.0277	0.0284	0.0308	0.0276
x_2	0.0304	0.0220	0.0398	0.0467	0.0326	0.0297	0.0322	0.0345	0.0311	0.0304	0.0223	0.0273
x_3	0.0270	0.0186	0.0347	0.0397	0.0271	0.0260	0.0269	0.0272	0.0315	0.0306	0.0216	0.0269
						Coverage						
x_1	0.9269	0.9089	0.9379	0.9299	0.9329	0.9259	0.9349	0.9349	0.9019	0.9119	0.9009	0.9229
x_2	0.9239	0.9049	0.9299	0.9099	0.9459	0.9469	0.9510	0.9479	0.9129	0.9179	0.9169	0.9239
x_3	0.9149	0.8949	0.9169	0.9179	0.9530	0.9439	0.9499	0.9570	0.9159	0.9159	0.9119	0.9199

A.11.2 Weights – Single covariate

X_i enters the component weights only through $\{x_{it,1}\}_{t=1}^2$. In particular, we consider as setting with $J = 2$ and $T = 2$. $\beta^{(0)}$ is the same as in the two-period setting in Section 5.1. The groups are generated with $g_i = \mathbb{1}\left(\zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) - u_i \geq 0\right) + 2\mathbb{1}\left(\zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) - u_i < 0\right)$ where $u_i \sim N(0, 1)$ independent from X_i as well as $\varepsilon_{j,it}$ for all j and t and $\zeta = (0.25, 1.2, -1.4)^\top$.

Table A.6 presents the simulation results for the estimators of the group-specific AMEs as well as the percentage of correctly assigned individuals based on the posterior group probability for $n \in \{1000, 2000\}$.

Table A.6: Simulation results for the AMEs. Weights depend on the first covariate only and $J = 2$.

	$n = 1000$ and $T = 2$				$n = 2000$ and $T = 2$			
	Gr. 1		Gr. 2		Gr. 1		Gr. 2	
	$t = 1$	$t = 2$	$t = 1$	$t = 2$	$t = 1$	$t = 2$	$t = 1$	$t = 2$
Bias								
x_1	-0.0047	-0.0026	0.0013	0.0033	-0.0010	-0.0005	0.0005	0.0021
x_2	0.0024	0.0011	-0.0021	-0.0014	-0.0002	-0.0004	-0.0002	0.0000
x_3	0.0015	0.0010	0.0009	0.0018	-0.0003	-0.0002	-0.0005	0.0006
RMSE								
x_1	0.0396	0.0298	0.0378	0.0268	0.0273	0.0210	0.0258	0.0184
x_2	0.0373	0.0277	0.0349	0.0265	0.0255	0.0192	0.0244	0.0187
x_3	0.0351	0.0251	0.0293	0.0260	0.0236	0.0170	0.0204	0.0175
Group assign.	89.89%				90.39%			

Next, we focus on settings with two time periods and $J = 3$. Identification is still achieved via the arguments in Section 7.2. While $\beta^{(0)}$ remains unchanged, the group membership is now generated as follows:

$$g_i = \begin{cases} 1 & \text{iff } \zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + u_i \leq 0 \\ 2 & \text{iff } 0 < \zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + u_i \leq \mu \\ 3 & \text{iff } \zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + u_i > \mu \end{cases}$$

where $u_i \sim N(0, 1)$ independent from all other variables in the model, $\zeta = (1, 2, -2)^\top$, and $\mu = 1.2$. In expectation, the groups are of similar size. Similar to Appendix A.11.1, we drop some simulation runs because either the coefficients are unreasonably large (in absolute value greater than 10) or the numerical optimization routine threw an error message. The numerical issues disappear as the sample size increases or identification "becomes easier", that is, T increases. Table A.7 contains the simulation results for the group-specific AMEs.

Last, we consider a setting with $J = 3$ and $T = 3$. While $\beta^{(0)}$ remains as in Section 5.2,

Table A.7: Simulation results for the AMEs. The weights depend on the first covariate only and $J = 3$.

	$n = 1000$ and $T = 2$						$n = 2000$ and $T = 2$					
	Gr. 1		Gr. 2		Gr. 3		Gr. 1		Gr. 2		Gr. 3	
	$t=1$	$t=2$	$t=1$	$t=2$	$t=1$	$t=2$	$t=1$	$t=2$	$t=1$	$t=2$	$t=1$	$t=2$
Bias												
x_1	-0.0215	-0.0228	0.0007	0.0012	-0.0266	-0.0332	-0.0154	-0.0125	0.0008	-0.0009	-0.0064	-0.0092
x_2	0.0000	0.0057	-0.0047	-0.0035	0.0166	0.0165	0.0021	0.0021	-0.0037	-0.0034	0.0148	0.0146
x_3	0.0118	0.0060	-0.0014	-0.0023	-0.0072	-0.0064	0.0045	0.0018	-0.0009	-0.0033	0.0011	0.0015
RMSE												
x_1	0.1060	0.0607	0.0777	0.0666	0.1086	0.0994	0.0769	0.0419	0.0496	0.0397	0.0666	0.0593
x_2	0.0627	0.0410	0.0671	0.0605	0.0682	0.0691	0.0386	0.0260	0.0452	0.0412	0.0466	0.0474
x_3	0.0601	0.0354	0.0577	0.0541	0.0617	0.0618	0.0382	0.0240	0.0364	0.0366	0.0391	0.0404
Group assign.			73.92%						78.04%			
#sim			987						1000			
#num. p			0						0			
#large coef.			13						0			

the group membership is now generated as follows:

$$g_i = \begin{cases} 1 & \text{iff } \zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + \zeta_3 x_{i3,1}^2 + u_i \leq 0 \\ 2 & \text{iff } 0 < \zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + \zeta_3 x_{i3,1}^2 + u_i \leq \mu \\ 3 & \text{iff } \zeta_0 + \zeta_1 \sin(x_{i1,1}) + \zeta_2 \log(x_{i2,1}^2 + 1) + \zeta_3 x_{i3,1}^2 + u_i > \mu \end{cases}$$

where $u_i \sim N(0,1)$ independent from all other variables in the model, $\zeta = (1, 2, -2, -0.2)^\top$, and $\mu = 1.2$. In expectation, the groups are of similar size. Table A.8 presents the simulation results for estimators of the group-specific AMEs.

Table A.8: Simulation results for the AMEs. The weights depend on the first covariate only and $J = 3$.

	$n = 1000$ and $T = 3$									$n = 2000$ and $T = 3$								
	Gr. 1			Gr. 2			Gr. 3			Gr. 1			Gr. 2			Gr. 3		
	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$
Bias																		
x_1	-0.0134	-0.0064	-0.0089	-0.0043	-0.0033	-0.0015	-0.0234	-0.0256	-0.0192	-0.0058	-0.0037	-0.0040	-0.0003	-0.0002	-0.0003	-0.0113	-0.0116	-0.0091
x_2	-0.0006	-0.0005	-0.0037	-0.0017	-0.0016	-0.0013	-0.0032	-0.0035	-0.0037	-0.0022	-0.0005	-0.0036	0.0016	0.0014	0.0015	0.0003	0.0007	-0.0004
x_3	0.0021	0.0002	0.0002	0.0028	0.0026	0.0044	-0.0084	-0.0085	-0.0054	0.0012	0.0006	0.0004	0.0013	0.0010	0.0010	-0.0016	-0.0016	-0.0010
RMSE																		
x_1	0.0458	0.0335	0.0482	0.0494	0.0414	0.0449	0.0663	0.0643	0.0500	0.0318	0.0225	0.0326	0.0297	0.0256	0.0268	0.0425	0.0387	0.0327
x_2	0.0390	0.0242	0.0431	0.0467	0.0434	0.0458	0.0450	0.0470	0.0314	0.0261	0.0149	0.0283	0.0282	0.0262	0.0280	0.0288	0.0303	0.0205
x_3	0.0372	0.0216	0.0404	0.0393	0.0362	0.0388	0.0446	0.0466	0.0307	0.0236	0.0134	0.0253	0.0249	0.0243	0.0248	0.0275	0.0288	0.0189
Group assign.					78.61%									82.47%				
#sim					999									1000				
#num. p					1									0				
#large coef.					0									0				

References

Ackerberg, D., X. Chen, and J. Hahn (2012). "A practical asymptotic variance estimator for two-step semiparametric estimators". In: *Review of Economics and Statistics* 94.2, pp. 481–498.

Ai, C. and X. Chen (2003). "Efficient estimation of models with conditional moment restrictions containing unknown functions". In: *Econometrica* 71.6, pp. 1795–1843.

— (2007). "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables". In: *Journal of Econometrics* 141.1, pp. 5–43.

Aliprantis, C.D. and K.C. Border (2006). *Infinite dimensional analysis: a hitchhiker's guide*. Springer.

Blischke, W.R. (1964). "Estimating the parameters of mixtures of binomial distributions". In: *Journal of the American Statistical Association* 59.306, pp. 510–528.

Carroll, R.J., X. Chen, and Y. Hu (2010). "Identification and estimation of nonlinear models using two samples with nonclassical measurement errors". In: *Journal of Nonparametric Statistics* 22.4, pp. 379–399.

Chen, X. (2007). "Large sample sieve estimation of semi-nonparametric models". In: *Handbook of Econometrics* 6, pp. 5549–5632.

Chen, X. and Z. Liao (2015). "Sieve semiparametric two-step GMM under weak dependence". In: *Journal of Econometrics* 189.1, pp. 163–186.

Chen, X., Y. Liu, S. Ma, and Z. Zhang (2023). "Causal Inference of General Treatment Effects using Neural Networks with A Diverging Number of Confounders". In: arXiv: 2009.07055v6.

— (2024). "Causal inference of general treatment effects using neural networks with a diverging number of confounders". In: *Journal of Econometrics* 238.1.

Chen, X. and D. Pouzo (2015). "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models". In: *Econometrica* 83.3, pp. 1013–1079.

Chen, X. and X. Shen (1998). "Sieve extremum estimates for weakly dependent data". In: *Econometrica* 66.2, pp. 289–314.

Freyberger, J. (2018). "Non-parametric panel data models with interactive fixed effects". In: *The Review of Economic Studies* 85.3, pp. 1824–1851.

Freyberger, J. and M.A. Masten (2019). "A practical guide to compact infinite dimensional parameter spaces". In: *Econometric Reviews* 38.9, pp. 979–1006.

Frühwirth-Schnatter, S., G. Celeux, and C.P. Robert (2019). *Handbook of Mixture Analysis*. Chapman and Hall/CRC.

Hu, Y. and S.M. Schennach (2008). "Instrumental variable treatment of nonclassical measurement error models". In: *Econometrica* 76.1, pp. 195–216.

Huang, M., R. Li, and S. Wang (2013). "Nonparametric mixture of regression models". In: *Journal of the American Statistical Association* 108.503, pp. 929–941.

Ichimura, H. (1993). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models". In: *Journal of Econometrics* 58.1-2, pp. 71–120.

Kasahara, H. and K. Shimotsu (2009). "Nonparametric identification of finite mixture models of dynamic discrete choices". In: *Econometrica* 77.1, pp. 135–175.

Leurgans, S.E., R.T. Ross, and R.B. Abel (1993). "A decomposition for three-way arrays". In: *SIAM Journal on Matrix Analysis and Applications* 14.4, pp. 1064–1083.

McLachlan, G.J. and D. Peel (2000). *Finite mixture models*. John Wiley & Sons.

Newey, W.K. (1997). "Convergence rates and asymptotic normality for series estimators". In: *Journal of Econometrics* 79.1, pp. 147–168.

Newey, W.K. and D. McFadden (1994). "Large sample estimation and hypothesis testing". In: *Handbook of Econometrics* 4, pp. 2111–2245.

Newey, W.K. and J.L. Powell (2003). "Instrumental variable estimation of nonparametric models". In: *Econometrica* 71.5, pp. 1565–1578.

Shen, X. (1997). "On methods of sieves and penalization". In: *The Annals of Statistics* 25.6, pp. 2555–2591.

Shen, X. and W.H. Wong (1994). "Convergence rate of sieve estimates". In: *The Annals of Statistics* 22.2, pp. 580–615.

Sidiropoulos, N.D. and R. Bro (2000). "On the uniqueness of multilinear decomposition of N-way arrays". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 14.3, pp. 229–239.

Socio-Economic Panel (SOEP) (2024). *Data for years 1984 – 2022 (SOEP-Core v39, EU Edition)*. DOI: 10.5684/soep.core.v39eu.

van de Geer, S.A. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A.W. and J.A. Wellner (2023). *Weak Convergence and Empirical Processes*. 2nd ed. Springer.

Wang, S., W. Yao, and M. Huang (2014). "A note on the identifiability of nonparametric and semiparametric mixtures of GLMs". In: *Statistics & Probability Letters* 93, pp. 41–45.

Wong, W.H. and T.A. Severini (1991). "On maximum likelihood estimation in infinite dimensional parameter spaces". In: *The Annals of Statistics* 19.2, pp. 603–632.