Nonparametric panel data models with interactive fixed effects[*]

Joachim Freyberger[‡]

Department of Economics, University of Wisconsin - Madison

November 3, 2012

Abstract

This paper studies nonparametric panel data models with multidimensional, unobserved individual effects when the number of time periods is fixed. I focus on models where the unobservables have a factor structure and enter an unknown structural function nonadditively. A key distinguishing feature of the setup is to allow for the various unobserved individual effects to impact outcomes differently in different time periods. When individual effects represent unobserved ability, this means that the returns to ability may change over time. Moreover, the models allow for heterogeneous marginal effects of the covariates on the outcome. The first set of results in the paper provides sufficient conditions for point identification when the outcomes are continuously distributed. These results lead to identification of marginal and average effects. I provide further point identification conditions for discrete outcomes and a dynamic model with lagged dependent variables as regressors. Using the identification conditions, I present a nonparametric sieve maximum likelihood estimator and study its large sample properties. In addition, I analyze flexible semiparametric and parametric versions of the model and characterize the asymptotic distribution of these estimators. Monte Carlo experiments demonstrate that the estimators perform well in finite samples. Finally, in an empirical application, I use these estimators to investigate the relationship between teaching practice and student achievement. The results differ considerably from those obtained with commonly used panel data methods.

[*] I am very grateful to my advisor Joel Horowitz as well as Ivan Canay and Elie Tamer for their excellent advice, constant support, and many helpful comments and discussions. I have also received valuable feedback from Matt Masten, Konrad Menzel, Diane Schanzenbach, Arek Szydlowski, Alex Torgovitsky, and seminar particatipants at Northwestern University. I thank Jan Bietenbeck for sharing his data. Financial support from the Robert Eisner Memorial Fellowship is gratefully acknowledged.

[‡] Email: j.freyberger@u.northwestern.edu. Comments are very welcome.

1 Introduction

This paper is about identification and estimation of panel data models with multidimensional, unobserved individual effects. In particular, I study models based on the outcome equation

$$Y_{it} = g_t\left(X_{it}, \lambda_i' F_t + U_{it}\right), \qquad i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{1}$$

where $X_{it} \in \mathbb{R}^{d_x}$ is a random vector of explanatory variables, Y_{it} is a scalar outcome variable, and U_{it} is a scalar idiosyncratic disturbance term. The structural functions g_t are unknown. Both $\lambda_i \in \mathbb{R}^R$ and $F_t \in \mathbb{R}^R$ are unobserved vectors of unknown dimension R. I study models with both continuously and discretely distributed outcomes. The explanatory variables X_{it} can be continuous or discrete and may depend on the individual effects λ_i. In this paper, T is fixed while $n \to \infty$.

Nonlinear and nonparametric panel data models have recently received much attention.[1] All of these models are based on special cases of the general outcome equation

$$Y_{it} = g_t\left(X_{it}, \lambda_i, U_{it}\right), \qquad i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{2}$$

where U_{it} and λ_i may be infinite dimensional. With $X_i = (X_{i1}, \ldots, X_{iT})$ and $x = (x_1, \ldots, x_T)$, most of these models share the feature that

$$E\left[Y_{it} \mid X_i = x, \lambda_i = \lambda_1\right] < E\left[Y_{it} \mid X_i = x, \lambda_i = \lambda_2\right]$$

implies

$$E\left[Y_{is} \mid X_i = x, \lambda_i = \lambda_1\right] < E\left[Y_{is} \mid X_i = x, \lambda_i = \lambda_2\right],$$

for all $s \neq t$ such that $x_s = x_t$. Consequently, the ranking of individuals with the same observed characteristics, based on their expected outcome, cannot change over t without a change in observables X_{it}. This condition usually follows because either it is assumed that λ_i is a scalar and that $E\left[Y_{it} \mid X_i, \lambda_i\right]$ is strictly increasing in λ_i for all t, or because it is assumed that $E\left[Y_{it} \mid X_i = x, \lambda_i = \lambda\right] = E\left[Y_{is} \mid X_i = x, \lambda_i = \lambda\right]$ for $s \neq t$ such that $x_s = x_t$. In contrast, models based on (1) do not impose such assumptions, but instead allow for multidimensional unobserved individual effects, which may affect the outcome variable Y_{it} differently for different t. When individual effects represent unobserved abilities, this means that both the returns to the various

[1] I discuss the related literature in the next section.

abilities, as well as the relative importance of each ability on the outcome, can change over t. Another important feature of model (1) is that it allows for heterogeneous marginal effects of the covariates on the outcome, which implies, for example, that returns to education may depend on unobserved ability.[2]

The flexibility of model (1) is important in many applications. For example, suppose that we are interested in the relationship between teacher characteristics and student achievement. Assume that each student i takes T tests, such as subject specific tests, and Y_{it} is the outcome of test t for student i. The vector X_{it} contains explanatory variables such as student, classroom, and teacher characteristics. A linear fixed effects model, which is based on the outcome equation

$$Y_{it} = X_{it}'\beta + \lambda_i + U_{it}, \qquad i = 1, \ldots, n, \ t = 1, \ldots, T,$$

is often used in this setting. For example, Dee (2007) analyzes whether assignment to a same-gender teacher has an influence on student achievement. Clotfelter, Ladd, and Vigdor (2010) and Lavy (2011) investigate the relationship between teacher credentials and student achievement and teaching practice and student achievement, respectively. In a linear fixed effects model λ_i is a scalar and represents unobserved ability of student i. Loosely speaking, this model assumes that if student i and j have the same observed characteristics and student i is better in subject t, say Mathematics, then student i must also be better in subject s, say English. In contrast, the vector λ_i in model (1) accounts for different dimensions of unobserved abilities and F_t represents the importance of each ability for test t. Hence, the model allows some students to have abilities such that they have a higher expected outcome in Mathematics, while others may have a higher expected outcome in English, without changes in observables. Furthermore, the impact of a teacher on students may differ for students with different abilities, which is ruled out in linear models. The main object of interest in this setting could then be marginal effects of teacher characteristics on students' test outcomes for different levels of students' abilities.

Other applications of model (1) include estimating returns to education or the effect of union membership on wages. In these examples, t represents time and the outcome Y_{it} is wage at time t of person i. The covariates X_{it} may include years of education, experience, and union membership. The vector λ_i represents different unobserved abilities of person i, and F_t represents the price of these abilities at time t. Model (1) also applies to macroeconomic situations. For example, assume

[2] Equation (1) becomes a model when combined with assumptions in later sections. For simplicity I refer to the outcome equation as model (1) in the introduction, because the structure is one of the models' main features.

that Y_{it} is output of country i in period t and X_{it} contains input variables such as labor and capital. The vector F_t denotes common shocks, such as technology shocks or a financial crisis, while λ_i represents the heterogeneous impacts of the common shocks on country i.[3]

This paper presents sufficient conditions for point identification of model (1) when T is fixed and $n \to \infty$. All parameters of the model are point identified up to normalizations under the assumptions in this paper. In models with this specific structure of the unobservables, the vector F_t is usually referred to as the factors, while λ_i are called the loadings. The factor structure of the unobservables is commonly called interactive fixed effects because of the interaction of λ_i and F_t. Since T is fixed and $n \to \infty$, I treat F_t as a vector of constants and λ_i as a vector of random variables.[4] The identified parameters include the functions g_t as well as the number of factors R, the factors F_t, and the distributions of λ_i and U_{it} conditional on the observed covariates. Although T is fixed, I require that $T \geq 2R + 1$. This condition means that for a given T, only factor models with less than or equal to $T/2 - 1$ factors are point identified under the assumptions I provide. These identification results imply identification of average effects as well as marginal effects, which are often the primary objects of interest in applications. I first consider continuously distributed outcomes, in which case the functions g_t are assumed to be strictly increasing in the second argument. The identification strategy follows two main steps. First, I use results from the measurement error literature to show that the distribution of (Y_i, λ_i) is identified up to some nonunique features, such as any one to one transformation of λ_i. In the second step, I establish uniqueness of all parameters by combining arguments from linear factor models and single index models with unknown link functions. The set of assumptions provided in this case, rules out that X_{it} contains lagged dependent variables and that the outcome Y_{it} is discrete. Therefore, I discuss several extensions to the above model where some assumptions are relaxed to accommodate these cases. The cost is strengthening other assumptions. Most importantly, I require λ_i to be discretely distributed if the outcomes are discrete, and T needs to be larger if X_{it} contains past outcomes.

After providing sufficient conditions for identification, I present a nonparametric sieve maximum likelihood estimator, which can be used to estimate the structural functions g_t as well as the factors and the conditional densities of λ_i and U_{it} consistently. The estimator requires estimating objects which might be high dimensional in applications, such as the conditional density of λ_i. Therefore, in addition to a fully nonparametric estimator, this paper also provides a flexible

[3] For more examples in economics where factor models are applicable see Bai (2009) and references therein.

[4] I use the terminology interactive fixed effects, although I make some assumptions about the distribution of λ_i. Graham and Powell (2012) provide a discussion on the difference between *fixed effects* and *correlated random effects.*

semiparametric estimator. In this setup, I reduce the dimensionality of the estimation problem by assuming a location and scale model for the conditional distributions. The structural functions can be nonparametric, semiparametric, or parametric. In the latter two cases, many of the parameters of interest are finite dimensional. I show that the estimators of the finite dimensional parameters are $\sqrt{n}$ consistent and asymptotically normally distributed, which yields confidence intervals for these parameters. I also describe an easy to implement fully parametric estimator. Finally, I show how the null hypothesis that the model has R factors can be tested against the alternative that the model has more than R factors, and how this result can be used to consistently estimate the number of factors. I provide Monte Carlo simulation results which demonstrate that the semiparametric estimator performs well in finite samples.

In an empirical application, I use the semiparametric estimator to investigate the relationship between teaching practice and student achievement. The outcome variables Y_{it} are different mathematics and science test scores for each student i. The main regressors are a measure of traditional teaching practice and a measure of modern teaching practice for each class a student attends. These measures are constructed using students' answers to questions about class activities. Traditional teaching practice is associated with lecture based classes with an emphasis on memorizing definitions and formulas. Modern teaching practice is associated with cooperative group work and justification of answers. The main objects of interest in this application are marginal effects of teaching practice, on mathematics and science test scores, for different levels of students' abilities. Using a standard linear fixed effects model, I find a positive relationship between traditional teaching practice and test outcomes in both mathematics and science. I then estimate model (1) with two factors and obtain substantially different results. I still find a positive relationship between traditional teaching practice and mathematics test scores, but a positive relationship between modern teaching practice and science test scores. Furthermore, the structural functions are significantly nonlinear. In particular, the magnitude of the relationship between teaching practice and test outcomes is higher for students with low abilities than for students with high abilities.

It should be noted that there are potential costs of identifying all features of model (1). In particular, certain objects, such as average marginal effects, may be identified under weaker assumptions. I leave these questions for future research and instead focus on point identification of all parameters. This approach has the advantage that it leads to identification of many interesting objects in applications, such as marginal effects for different levels of abilities.

The remainder of the paper is organized as follows. The next section connects this paper to related literature on linear factor models, nonparametric panel data models, and measurement error models. Section 3 deals with identification of model (1) with continuous outcomes and without lagged dependent variables as regressors. Section 4 extends the arguments to allow for lagged dependent variables and discrete outcomes. Section 5 discusses different ways to estimate the model, including the number of factors. Section 6 and 7 present Monte Carlo results and the empirical application, respectively. Finally, Section 8 concludes. All proofs are contained in the appendix.

2 Related literature

This paper is related to a vast literature on linear factor models, nonlinear and nonparametric panel data models, and measurement error models. Linear factor models are well understood and provide a way to deal with multidimensional unobserved heterogeneity. The model usually is

$$Y_{it} = X_{it}'\beta + \lambda_i' F_t + U_{it}, \qquad i = 1, \ldots, n, \ t = 1, \ldots, T.$$

The theoretical econometrics literature on linear factor models includes Holtz-Eakin, Newey, and Rosen (1988), Ahn, Lee, and Schmidt (2001), Bai and Ng (2002), Bai (2003), Andrews (2005), Pesaran (2006), Bonhomme and Robin (2008), Bai (2009), Ahn, Lee, and Schmidt (2010), Moon and Weidner (2010), Bai and Ng (2011), and Bai (2012). Some papers (e.g. Bai 2009) let $n \rightarrow \infty$ and $T \rightarrow \infty$ while others (e.g. Ahn et al. 2010) have T fixed and $n \rightarrow \infty$, as in this paper. Estimating the number of factors is considered by Bai and Ng (2002). Nonlinear additive factor models of the form

$$Y_{it} = g\left(X_{it}\right) + \lambda_i' F_t + U_{it}, \qquad i = 1, \ldots, n, \ t = 1, \ldots, T$$

have been studied recently by Huang (2010) and Su and Jin (2012), in a setup where $n \rightarrow \infty$ and $T \rightarrow \infty$. The drawback of linear models is that they impose homogeneous marginal effects. In my application this means that the influence of teachers on students is identical for students with different abilities. Moreover, the analysis in these papers is tailored to the linear model. For example, Bai (2009) estimates the factors using the method of principal components.

Factor models have been used in several applications. Related to the application in this paper, Carneiro, Hansen, and Heckman (2003) use five test scores to estimate a linear factor model with two factors. Heckman, Stixrud, and Urzua (2006) use a linear factor model to explain labor market and behavioral outcomes. Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010) estimate the evolution of cognitive and noncognitive skills with factor models. Williams, Heckman, and Schennach (2010) study a model where the first stage is a linear factor model and the estimated factor scores are used in a nonparametric second stage. They use their model to estimate the technology of skill formation.

Many recent papers in the nonparametric panel data literature with continuously distributed outcomes are related to the model I consider.[5] Evdokimov (2010, 2011) provides identification results for nonlinear models with a scalar heterogeneity term. Arellano and Bonhomme (2012) analyze a random coefficients model which allows for multidimensional unobserved heterogeneity. Other nonnested setups include Chernozhukov, Fernandez-Val, Hahn, and Newey (2012), Graham and Powell (2012), and Hoderlein and White (2012) who mainly focus on identification and estimation of average marginal effects and quantile effects. In all these papers, the ranking of individuals based on their mean or median outcome cannot change over t without changes in observable regressors. Bester and Hansen (2009) are also concerned with average marginal effects but do not impose this assumption. Instead they restrict the conditional distribution of λ_i. Altonji and Matzkin (2005) require an external variable which they construct in a panel data model by restricting the conditional distribution of λ_i.

Nonlinear panel data models with discrete outcomes generally need a different treatment. For example Chamberlain (2010) shows that in binary outcome panel data models, point identification fails in case the support of the regressors is bounded and the disturbance is not logistic distributed. Honoré and Tamer (2006) demonstrate lack of point identification in a similar model with lagged dependent variables. Williams (2011) derives partial identification results for panel data models with discrete outcomes and shows that the identified set converges to a point as $T \to \infty$. The identification strategy used in my paper yields point identification of the distribution of $(Y_i, \lambda_i) \mid X_i$ with discrete outcomes, provided that λ_i has a discrete distribution as well.

The identification strategy with continuously distributed outcomes is related to Hu and Schennach (2008) and Cunha, Heckman, and Schennach (2010), because it relies on an eigendecomposition of a linear operator, but the arguments differ in important steps. Hu and Schennach (2008) study a nonparametric measurement error model with instrumental variables. The connection to the factor model is that λ_i can be seen as unobserved regressors. A subset of the outcomes represents the observed and mismeasured regressors, while another subset of outcomes serves as

[5] Arellano and Bonhomme (2011) provide a recent survey on nonlinear panel data models.

instruments. Cunha, Heckman, and Schennach (2010) apply results in Hu and Schennach (2008) to identify a nonparametric factor model similar to model (2). One of their identifying assumptions fixes a measure of location of the distribution of a subset of outcomes given λ_i.[6] Such measures generally do not exist in model (1) when all functions g_t are unknown. Instead, I use the relation between Y_{it} and λ_i delivered by (1), combined with arguments from linear factor models and single index models with unknown link functions. Hence, the two models are nonnested. Furthermore, the different sets of outcomes, which represent regressors and instruments, are interchangeable which allows me to show that $T = 2R + 1$ is sufficient for identification in model (1). The structure of the model also leads to more primitive sufficient conditions for some of the high level assumptions.[7]

Similarly, in the case of discrete λ_i, the identification strategy is related to the one of Hu (2008) who is concerned with a measurement error model with one discrete mismeasured regressor. The identification strategy with lagged dependent variables is related to Hu and Shum (2012) and Sasaki (2012) who use arguments related to the approach of Hu and Schennach (2008). Again, I do not impose one of their main identifying assumptions, but instead use the additional structure of the factor model. Shiu and Hu (2011) study a dynamic panel data model with covariates which requires certain conditions on the process of X_{it}. The assumptions in all these papers are nonnested with the assumptions I present. I complement these papers by focusing on the factor structure of the error terms but by using different conditions which allow for different interesting models.

3 Identification of static factor model with continuous outcomes

This section is about identification of a model based on (1) with continuously distributed outcome variables Y_{it} and continuously distributed λ_i. I first introduce important notation and state the assumptions. Afterwards, I discuss the assumptions and show that they are sufficient for identification. To simplify the notation, I first assume that the number of factors R is known. In Section 3.3 I show how the number of factors can be identified.

3.1 Assumptions, definitions, and notation

As stated in the introduction, I assume in this section that the structural functions g_t are strictly increasing in the second argument. Define the inverse function $h_t(Y_{it}, X_{it}) \equiv g_t^{-1}(Y_{it}, X_{it})$. Then

[6] The corresponding assumption in Hu and Schennach (2008) fixes a measure of location of the distribution of the measurement error.

[7] These assumptions are invertibility of integral operators. Lemma 1 provides sufficient conditions.

equation (1) can be written as

$$h_t\left(Y_{it}, X_{it}\right) = \lambda_i' F_t + U_{it}, \qquad i = 1, \ldots, n,\ t = 1, \ldots, T. \tag{3}$$

A necessary condition for point identification is $T \geq 2R+1$.[8] To simplify the notation I assume that $T = 2R+1$, but the extension to a larger T is straightforward. In fact, the assumptions with a larger T are weaker, as discussed below.

Assumption S1. R is known and $T = 2R+1$.

I now introduce some important definitions and notation followed by the remaining assumptions. For each t, let $\mathcal{X}_t \subseteq \mathbb{R}^K$ and $\mathcal{Y}_t \subseteq \mathbb{R}$ be the supports of X_{it} and Y_{it}, respectively. Let $\Lambda \subseteq \mathbb{R}^R$ be the support of λ_i. Define $X_i = (X_{i1}, \ldots, X_{iT})$ and define Y_i and U_i analogously. Let $\mathcal{X} \equiv \cup_{t=1}^T \mathcal{X}_t$ and $\mathcal{Y} \equiv \cup_{t=1}^T \mathcal{Y}_t$ be the supports of X_i and Y_i, respectively. Define the vector of the last R outcomes

$$Z_{i1} \quad \equiv \quad \left(Y_{i(R+2)}, \ldots, Y_{i(2R+1)}\right).$$

Let $K \equiv \{k_1, k_2, \ldots, k_R\} \subset \{1, 2, \ldots, R+1\}$ be a set of any R integers between 1 and $R+1$ with $k_1 < k_2 < \ldots < k_R$. Let $k_{R+1} \equiv \{1, 2, \ldots, R+1\} \setminus K$ be the remaining integer. Define

$$\begin{aligned} Z_{iK} \quad &\equiv \quad \left(Y_{ik_1}, \ldots, Y_{ik_R}\right) \qquad \text{and} \\ Z_{ik_{R+1}} \quad &\equiv \quad Y_{ik_{R+1}}. \end{aligned}$$

For example, if $R = 2$ and $T = 5$, then $Z_{i1} = (Y_{i4}, Y_{i5})$ and Z_{iK} can be (Y_{i1}, Y_{i2}) or (Y_{i1}, Y_{i3}) or (Y_{i2}, Y_{i3}). The scalar $Z_{ik_{R+1}}$ is the remaining outcome which is neither contained in Z_{i1} nor in Z_{iK}. Let $\mathcal{Z}_1 \subseteq \mathbb{R}^R$ and $\mathcal{Z}_K \subseteq \mathbb{R}^R$ be the supports of Z_{i1} and Z_{iK}, respectively.

The conditional probability mass or density function of any random variable $W \mid V$ is denoted by $f_{W|V}(w; v)$ and the marginal probability density (or mass) function by $f_W(w)$. Let $F_{W|V}(w; v)$ and $F_W(w)$ be the cumulative distribution functions of $W \mid V$ and W, respectively. The α-quantile of $W \mid V$ is denoted by $Q_\alpha[W \mid V]$. The median, $Q_{1/2}[W \mid V]$, is denoted by $M[W \mid V]$. A random variable W is complete for V if for all real measurable functions m such that $E[|m(W)|] < \infty$

$$E[m(W) \mid V] = 0 \text{ a.s. implies that } m\left(W\right) = 0 \text{ a.s.}$$

[8] This is shown by Carneiro, Hansen, and Heckman (2003) in a linear factor model with covariance restrictions. Related arguments can be used here to establish that point identification fails if $T < 2R+1$.

W is bounded complete for V if the implication holds for any bounded function m.

Let F be the $R \times T$ matrix containing all factors and write it as

$$F = \begin{pmatrix} F_1 & F_2 & \cdots & F_T \end{pmatrix} = \begin{pmatrix} F^1 & F^2 & F^3 \end{pmatrix} \tag{4}$$

where F^1 is $R \times R$, F^2 is $R \times 1$, and F^3 is $R \times R$. Let $I_{R\times R}$ denote the $R \times R$ identity matrix.

This section focuses on the continuous case. Therefore, I make the following assumption.

Assumption S2. $f_{Y_{i1},\ldots,Y_{iT},\lambda_i|X_i}(y_1,\ldots,y_T,\lambda;x)$ is bounded on $\mathcal{Y}_1\times\cdots\times\mathcal{Y}_T\times\Lambda\times\mathcal{X}$ and continuous in $(y_1,\ldots,y_T,\lambda) \in \mathcal{Y}_1\times\cdots\times\mathcal{Y}_T\times\Lambda$ for all $x \in \mathcal{X}$. All marginal and conditional densities are bounded.

Next, I present several additional assumptions. In addition to assuming that h_t is strictly increasing in the first argument, I require location and scale normalizations. Second, I use moment restrictions, independence assumptions, and completeness conditions. Finally, just as in linear factor models, the factors and loadings are only identified up to a transformation.

Assumption S3.

(i) h_t is strictly increasing and differentiable in its first argument.

(ii) Let $\bar{y}_t = M[Y_{it}]$. There exist $\bar{x}_t \in \mathcal{X}_t$ such that $h_t(\bar{y}_t,\bar{x}_t) = 0$ for all $t = R+2,\ldots,2R+1$ and $\frac{\partial h_t(\bar{y}_t,\bar{x}_t)}{\partial y} = 1$ for all $t = 1,\ldots,T$.

Assumption S4. $M[U_{it} \mid X_i,\lambda_i] = 0$ for all $t = 1,\ldots,T$. If $R = 1$, $Q_\alpha[U_{it} \mid X_i,\lambda_i]$ is independent of λ_i for all α in a neighborhood of $1/2$.

Assumption S5. $U_{i1},\ldots,U_{iT}$ are jointly independent conditional on λ_i and X_i.

Assumption S6. $F^3 = I_{R\times R}$ and F^1 has full rank.

Assumption S7. The $R \times R$ covariance matrix of λ_i has full rank conditional on X_i.

Assumption S8. Z_{i1} is bounded complete for Z_{iK} for any R integers $K \subset \{1,\ldots,R+1\}$ conditional on X_i. Moreover, λ_i is bounded complete for Z_{i1} conditional on X_i.

The location normalizations in Assumption S3 are needed because without them one could add a constant c to λ_{ir} as well as cF_{tr} to h_t without affecting equality (3). The location of U_{it} is fixed by Assumption S4. Similarly, the scale normalizations are needed because otherwise for

$t = T - R + 1, \ldots, T$ one can multiply h_t, λ_i and U_{it} by a constant c while for all other t one can multiply h_t, F_t and U_{it} by c. The assumption that the standardizations occur at the median of Y_{it} is only used for a minor part of the main identification theorem as discussed below.

Assumption S4 implies that while the regressors can arbitrarily depend on λ_i, they are strictly exogenous with respect to U_{it}. This assumption rules out, for example, that X_{it} contains lagged dependent variables. Moreover, for $R = 1$, the model would not be identified without the second part of Assumption S4. To see this, let $B(\lambda)$ be any strictly increasing function and let $F_t = 1$ for all t. Then

$$Y_{it} = g_t\left(X_{it}, B^{-1}\left(\tilde{\lambda}_i + \tilde{U}_{it}\right)\right)$$

where $\tilde{\lambda}_i = B(\lambda_i)$ and $\tilde{U}_{it} = B\left(U_{it} + \lambda_i\right) - B(\lambda_i)$ and $\tilde{U}_{it}$ satisfies the median restriction. Assumption S5 is strong but independence is hard to avoid in nonadditive models. Although the unobservables $\lambda_i' F_t + U_{it}$ are correlated over t, the assumption says that any dependence is due to λ_i. Autoregressive U_{it} are thus ruled out. However, note that the assumptions do not require that U_{it} and X_{it} are independent, nor that U_{it} and λ_i are independent. Hence, heteroskedasticity is permitted.

Assumption S6 is a normalization which is needed because for any $R \times R$ invertible matrix H

$$\lambda_i' F_t = \lambda_i' H H^{-1} F_t = (H'\lambda_i)'(H^{-1}F_t) = \tilde{\lambda}_i' \tilde{F}_t.$$

Although the linear combination $\lambda_i' F_t$ can be identified, the factors and loadings can only identified up to a transformation since $\lambda_i' F_t = \tilde{\lambda}_i' \tilde{F}_t$. Hence R^2 restrictions on the factors and loadings are needed to identify a certain transformation, which I impose by assuming that $F^3 = I_{R \times R}$. This normalization corresponds to $H = F^3$ above and implicitly assumes that F^3 is invertible.[9]

Assumption S7 is a rank condition which rules out that some element of λ_i is a linear combination of the other elements. Furthermore, all constant elements of λ_i, and thus time trends, are absorbed by the function h_t.

Assumption S8 is a bounded completeness condition. Completeness conditions are often used in nonparametric instrumental variable models, in which case the regressor is required to be complete for the instrument. This condition is a generalization of the rank condition in linear instrumental variable regressions. The first part of Assumption S8 therefore says that Z_{iK} serves as an instrument

[9] In linear factor models it is often assumed that the factors are orthogonal and have length 1 and that the covariance matrix of the loadings is diagonal. This is convenient in the linear model because, when estimating the factors by the method of principal components, the estimates are orthogonal.

for Z_{i1}. The second part ensures that λ_i does not contain too much information relative to Z_{i1} which, for example, rules out that λ_i is continuously distributed while Z_{i1} is discrete. In some models this assumption can be hard to interpret.[10] However, here since g_t is strictly increasing, Assumption S8 is solely an assumption about the distribution of λ_i and U_{it} given X_i and the values of F_t. The next lemma provides lower level sufficient conditions for this assumption to be satisfied.

Lemma 1. Assume that any $R \times R$ submatrix of F has full rank and that λ_i has support on $\mathbb{R}^R$ conditional on X_i. Moreover, assume that the characteristic function of U_{it} is nonvanishing on $(-\infty, \infty)$ for all t, and that $U_i \perp\!\!\!\perp \lambda_i$. Then Assumptions S3(i) and S5 imply Assumption S8.

A nonvanishing characteristic function holds for many standard distributions such as the normal family, the t-distribution, or the gamma distribution. However, it does not hold for all distributions, for instance uniform and triangular distributions. As an important special case, the lemma shows that if λ_i and U_i are normally distributed and independent, and the covariance matrix of λ_i as well as any $R \times R$ submatrix of F have full rank, then Assumption S8 holds.

If $T > 2R + 1$, then Assumption S8 only needs to hold for $R + 1$ different sets of R integers $K = \{k_1, k_2, \ldots, k_R\}$. Also only for one these sets the full rank part of Assumption S6 has to hold.

3.2 Identification of g_t and F_t and the conditional distributions of λ_i and U_{it}

In this section I outline the main arguments for identifying g_t, the factors, as well as the distribution of $(U_i, \lambda_i) \mid X_i$. Afterwards, I state the main identification theorem. The formal proof is given in the appendix. In the next subsection I prove identification of the number of factors.

To use the scale and location normalizations define

$$\tilde{\mathcal{X}} \equiv \{(x_1, \ldots, x_T) \in \mathcal{X} : x_t = \bar{x}_t \text{ for all } t = R + 2, \ldots, 2R + 1\}.$$

where $\bar{x}_t$ is defined in Assumption S3. The role of this set is explained below. In the appendix, it is shown that Assumption S5 implies an operator equivalence result of the form

$$L_{1,k_{R+1},K} = L_{1,\lambda} D_{k_{R+1},\lambda} L_{\lambda,K},$$

where $L_{1,k_{R+1},K}$, $L_{1,\lambda}$, and $L_{\lambda,K}$ are linear integral operators and $D_{k_{R+1},\lambda}$ can be seen as a diagonal operator. The operator on the left hand side only depends on the population distribution of the

[10] Canay, Santos, and Shaikh (2012) show that the assumption is not testable under commonly used restrictions.

observables, while all operators on the right hand side depend on the joint distribution of Y_i and λ_i. Assumption S5 also yields a second operator equivalence result

$$L_{1,K} = L_{1,\lambda} L_{\lambda,K}.$$

The operator on the left hand side again depends on the population distribution of the observables only. Assumption S8 implies that the inverse of $L_{1,\lambda}$ exists and can be applied from the left. Therefore

$$L_{1,\lambda}^{-1} L_{1,K} = L_{\lambda,K}$$

and in turn

$$L_{1,k_{R+1},K} = L_{1,\lambda} D_{k_{R+1},\lambda} L_{1,\lambda}^{-1} L_{1,K}.$$

Assumption S8 also ensures that the right inverse of $L_{1,K}$ exists which means that

$$L_{1,k_{R+1},K} L_{1,K}^{-1} = L_{1,\lambda} D_{k_{R+1},\lambda} L_{1,\lambda}^{-1}. \tag{5}$$

In a paper on measurement error models, Hu and Schennach (2008) obtain a similar operator equality. They show that the right hand side is an eigenvalue-eigenfunction decomposition of the operator on the left hand side and that such a decomposition is unique up to three nonunique features. It is shown in the appendix that, conditional on $X_i \in \tilde{\mathcal{X}}$, these nonunique features cannot arise in model (1) under the assumptions provided in Section 3.1. To do so, I combine arguments from linear factor models and single index models with unknown link functions. The most important assumptions which are used to establish uniqueness are the factor structure, the normalizations, the moments conditions, and monotonicity of the structural functions. Furthermore, I use that the outcomes contained in Z_{iK} are interchangeable, which ensures that $T = 2R + 1$ is sufficient for identification. The left hand side of the operator equality (5) only depends on the population distribution of the observables. Uniqueness of the decomposition thus ensures that the operators $L_{1,\lambda}$ and $D_{k_{R+1},\lambda}$ are identified. It can then be shown that $L_{\lambda,K}$ is also identified. Identification of these integral operators is in this case equivalent to identification of $f_{Y_i,\lambda_i|X_i}$. It then follows that F_t is identified. Finally, under additional assumptions, it is shown that g_t and the distribution of $U_i, \lambda_i \mid X_i$ are identified.

The previous arguments lead to the following theorem which is one of the main results in this paper. The formal proof is given in the appendix.

Theorem 1. Assume that Assumptions S1 - S8 hold. Then $f_{Y_i,\lambda_i|X_i}(s,\lambda;x)$ is identified for all $s \in \mathcal{Y}$, $\lambda \in \Lambda$, and $x \in \tilde{\mathcal{X}}$. Moreover, F_t is identified. Assume in addition that either λ_i has support on $\mathbb{R}^R$ or that $U_i \perp\!\!\!\perp \lambda_i$. Then the functions g_t as well as the distribution of $(U_i, \lambda_i) \mid X_i = x$ are identified for all $x \in \tilde{\mathcal{X}}$.

Remark 1. Without the additional assumption that either λ_i has support on $\mathbb{R}^R$ or that $U_i \perp\!\!\!\perp \lambda_i$, the functions g_t and h_t are identified on a subset of the support of $\lambda_i' F_t + U_{it}$ and the support of Y_{it}, respectively. For example, $g_t(x_t, e_t)$ is identified for all e_t such that $e_t = \lambda' F_t$ for some $\lambda \in \Lambda$. The normalization at the median of Y_{it} in Assumption S3 is only used to identify F_t in cases where g_t is not identified for all values on the support.

Theorem 1 shows identification for all $x \in \tilde{\mathcal{X}}$ which is a strict subset of the support of X_i. After identifying the structural functions and distributions for all $x \in \tilde{\mathcal{X}}$, these quantities can be shown to be identified for $x \notin \tilde{\mathcal{X}}$. To do so for any $(\tilde{x}_1, \ldots, \tilde{x}_T) \in \tilde{X}$ take $(x_{R+1}, \ldots, x_{2R+1})$ such that

$$f_{X_{i(R+1)},\ldots,X_{i(2R+2)}|X_{i1},\ldots,X_{iR}}(x_{R+1}, \ldots, x_{2R+1}; \tilde{x}_1, \ldots, \tilde{x}_R) > 0.$$

Since $h_t(\bar{y}_t, \tilde{x}_t)$ is identified for all $t = 1, \ldots, R$, in the proof of the theorem the roles of the different periods t can be switched. In particular, instead of using a normalization at $\bar{x}_t$ for $t = R+2, \ldots, 2R+1$, the values $\tilde{x}_t$ for $t = 1, \ldots, R$ can take this role. It follows that for these $(x_{R+1}, \ldots, x_{2R+1})$, the function h_t is identified. This process can be iterated.

Hence, in the most favorable case, if $f_{X_i}(x) > 0$ for all $x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_T$, the functions h_t are identified for all $x_t \in \mathcal{X}_t$. On the other hand, if $X_{it} = x_i$ or $X_{it} = x_t$ for all i and t, h_t is only identified at $\bar{x}_t$ for all t. This is a standard problem in panel data models. If the regressor does not change over t, it is not possible to distinguish between the effect of X_{it} and the effect of λ_i on the outcome, without restricting the dependence between X_{it} and λ_i. Moreover, the function h_t depends on t. Thus, if $X_{it} = x_t$, a change in X_{it} cannot be distinguished from a change in the function. A similar problem occurs with time fixed effects in linear panel data models. There are many intermediate cases where h_t is identified for all $x_t \in X_t$. This is, for example, the case my empirical application (see Section 7.3 for the details), where neither $f_{X_i}(x) > 0$ for all $x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_T$ nor $X_{it} = x_i$ or $X_{it} = x_t$ for all i and t. Finally, if the structural functions are parametric, which is likely to be assumed in applications, it is easier to identify the model. For instance, if h_t is linear in a scalar X_{it}, identification at two points implies identification for all $x_t \in \mathcal{X}_t$.

3.3 Identification of the number of factors

To identify the number of factors let $\tilde{R} \in \left\{1, \ldots, \left\lfloor \frac{T}{2} \right\rfloor\right\}$, where $\left\lfloor \frac{T}{2} \right\rfloor$ is the largest integer smaller than or equal to $\frac{T}{2}$, and define

$$\tilde{Z}_{i1} \equiv \left(Y_{i(T-\tilde{R}+1)}, \ldots, Y_{iT}\right) \qquad \text{and} \qquad \tilde{Z}_{i2} \equiv \left(Y_{i1}, \ldots, Y_{i\tilde{R}}\right).$$

It can be shown that under the previous assumptions and $\lambda_i \perp\!\!\!\perp U_i$, $\tilde{Z}_{i1}$ is bounded complete for $\tilde{Z}_{i2}$ only if $\tilde{R} \leq R$. Hence, R is the largest integer $\tilde{R}$, less than or equal to $\frac{T}{2}$, such that $\tilde{Z}_{i1}$ is bounded complete for $\tilde{Z}_{i2}$. Since this condition only involves the data, it implies the following theorem.

Theorem 2. Assume that Assumptions S2, S3, S5, and S8 hold, that $\lambda_i \perp\!\!\!\perp U_i$, and that $T \geq 2R+1$. Then the number of factors, R, is identified.

Remark 2. More lengthy arguments than those in the proof of Theorem 2 can be used to show that the number of factors is identified without the completeness assumption. This result is important for estimating the number of factors, because there are no known conditions under which the completeness assumption is testable as shown by Canay, Santos, and Shaikh (2012).

3.4 Functionals invariant to normalizations

Although a few normalization assumptions are needed in Theorem 1, many potential objects of interest are invariant to these normalizations. Define $C_{it} \equiv \lambda_i' F_t$. Next let $Q_\alpha[C_{it} \mid X_i = x]$ and $Q_\alpha[U_{it} \mid X_i = x]$ be the conditional α-quantile of C_{it} and U_{it}, respectively. Fix any t and let $\tilde{x}_t \in \mathcal{X}_t$ as well as $x \in \mathcal{X}$ such that the previous identification results hold. Appendix A.4 shows that the following functionals are invariant to the normalizations in this paper.

1. Function values at quantiles of unobservables:

$$g_t\left(\tilde{x}_t, Q_{\alpha_1}\left[C_{it} \mid X_i = x\right] + Q_{\alpha_2}\left[U_{it} \mid X_i = x\right]\right)$$

   or

$$g_t\left(\tilde{x}_t, Q_\alpha\left[C_{it} + U_{it} \mid X_i = x\right]\right).$$

2. Average function values:

$$\int g_t\left(\tilde{x}_t, e\right) dF_{C_{it}+U_{it} \mid X_i = x}(e)$$

or

$$\int g_t\left(\tilde{x}_t, e + Q_\alpha\left[U_{it} \mid X_i\right]\right) dF_{C_{it}|X_i=x}(e)\,.$$

It then immediately follows that also differences of function values or differences of average function values are invariant to these normalizations. These quantities can be used to answer important policy questions. For example one could answer questions about the effect of a change in class size on test outcomes for students with different levels of abilities.

4 Further identification results

Assumptions S4 and S5 rule out that X_{it} contains lagged dependent variables. In this section I relax these assumptions to accommodate this case. To do so, I must strengthen other assumptions. Furthermore, I discuss the case of discretely distributed heterogeneity, in which case Y_{it} may also be discretely distributed.

4.1 Dynamic factor model with continuous outcomes

First rewrite equation (1) to

$$Y_{it} = g_t\left(Y_{i(t-1)}, X_{it}, \lambda_i' F_t + U_{it}\right), \qquad i = 1, \ldots, n,\ t = 1, \ldots, T. \tag{6}$$

I assume for simplicity that R is known and that there is only one lagged dependent variable. Several lagged dependent variables can be incorporated using similar arguments to the ones presented below. The main difference to the static model is that in the static case, periods t were interchangeable, which is not the case with lagged dependent variables. Therefore, T needs to be larger for the model to be identified.

Using lagged dependent variables requires adapting the arguments of Section 3. As explained in Section 2, similar arguments, in models nonnested with the one considered here, have been used by Shiu and Hu (2011), Hu and Shum (2012), and Sasaki (2012). I assume for simplicity that there are no (strictly exogenous) regressors X_{it}. Using these regressors simply requires to make all assumptions conditional on X_i just as in Section 3. Hence, the outcome equation is

$$\begin{aligned} Y_{it} &= g_t\left(Y_{i(t-1)}, \lambda_i' F_t + U_{it}\right), && i = 1, \ldots, n,\ t = 1, \ldots, T, \quad \text{or} \\ h_t\left(Y_{it}, Y_{i(t-1)}\right) &= \lambda_i' F_t + U_{it}, && i = 1, \ldots, n,\ t = 1, \ldots, T. \end{aligned}$$

The identification strategy requires that $T \geq 2R + \left\lceil \frac{R}{2} \right\rceil + 3$ where $\lceil A \rceil$ is the smallest integer larger than or equal to A. To simplify the exposition, I assume that $T = 3R + 3$. In the appendix it is explained how the model can be identified, under modified assumptions, with smaller T. In this section let $K = \{k_1, k_2, \ldots, k_R\}$ be R integers between $2R+3$ and $3R+3$ with $k_1 < k_2 < \ldots < k_R$. Also define $Z_{iK} \equiv (Y_{ik_1}, \ldots, Y_{ik_R})$ and $Z_{i1} \equiv (Y_{i1}, \ldots, Y_{iR})$. Notice that the definitions of Z_{i1} and Z_{iK} are slightly different to the ones used in Section 3. As before, $I_{R\times R}$ is the $R \times R$ identity matrix and F^3 is the matrix of factors from the last R periods.

Assumption L1. R is known and $T = 3R + 3$.

Assumption L2. $f_{Y_{i1},\ldots,Y_{iT},\lambda_i}(y_1, \ldots, y_T, \lambda)$ is bounded on $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_T \times \Lambda$ and continuous in $(y_1, \ldots, y_T, \lambda) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_T \times \Lambda$. All marginal and conditional densities are bounded.

Assumption L3.

(i) h_t is strictly increasing and differentiable in its first argument.

(ii) Let $\bar{y}_t = M[Y_{it}]$. For all $t = 2R+4, \ldots, 3R+3$ it holds that $h_t(\bar{y}_t, \bar{y}_{t-1}) = 0$ and for all $t = 1, \ldots, T$ it holds that $\frac{\partial h_t(\bar{y}_t, \bar{y}_{t-1})}{\partial y} = 1$.

Assumption L4. $M\left[U_{it} \mid Y_{i1}, \ldots, Y_{i(t-1)}, \lambda\right] = 0$ for all $t = 2, \ldots, T$. If $R = 1$, for all α in a neighborhood of $1/2$, $Q_\alpha\left[U_{it} \mid Y_{i1}, \ldots, Y_{i(t-1)}, \lambda_i\right]$ is independent of λ_i.

Assumption L5. For all $t \geq 2$, $(U_{iT}, \ldots, U_{it}) \perp\!\!\!\perp (Y_{i(t-1)}, \ldots, Y_{i1}) \mid \lambda_i$.

Assumption L6. For any R integers $K \subset \{2R+3, 2R+4, \ldots, 3R+3\}$, Z_{iK} is bounded complete for Z_{i1} given $Y_{i(R+1)}$ and $Y_{i(2R+2)}$. Moreover, λ_i is bounded complete for Z_{iK} given $Y_{i(2R+2)}$. If $R = 1$, Y_{i6} is bounded complete for Y_{i1} given $Y_{i(R+1)}$, $Y_{i(2R+2)}$ and $Y_{i(2R+3)}$.

Assumption L7. For all $\lambda_1 \neq \lambda_2$ and for all s_{2R+2}, there exist $s_{2R+1}, \ldots, s_{R+1}$ such that

$$\begin{aligned} &f_{Y_{i(2R+1)},\ldots,Y_{i(R+2)} \mid Y_{i(2R+2)},Y_{i(R+1)},\lambda_i}(s_{2R+1}, \ldots, s_{R+2}; s_{2R+2}, s_{R+1}, \lambda_1) \\ &\qquad \neq f_{Y_{i(2R+1)},\ldots,Y_{i(R+2)} \mid Y_{i(2R+2)},Y_{i(R+1)},\lambda_i}(s_{2R+1}, \ldots, s_{R+2}; s_{2R+2}, s_{R+1}, \lambda_2). \end{aligned}$$

Assumption L8. $F^3 = I_{R\times R}$.

Assumption L9. The $R \times R$ covariance matrix of λ_i has full rank conditional on $Y_{i1}, \ldots, Y_{iT}$.

The intuition for the certain number T required is that different sets of past outcomes serve as instruments for future outcomes. However, the sets are limited due to the dynamic structure. Hence, T needs to be larger compared to the static case. Assumption L3 is the same normalizing assumption made in the static case. Assumption L4 is a location normalization for U_{it}. It also says that the median of future values of U_{it} are median independent of λ_i and past values of Y_{it}. Again, an additional restriction for $R = 1$ is needed. Assumption L5 relaxes the previous assumption that U_{it} and U_{is} are independent given all regressors X_i. Assumption L6 is a completeness assumption similar as before: Past outcomes need to serve as instruments for future outcomes and λ_i cannot have too much variation relative to Y_{it}. By Assumption L5, the dependence between Z_{iK} and Z_{i1} can only come through λ_i. Hence, I implicitly make an assumption about the joint distribution of $\lambda_i, Y_{i1}, \ldots, Y_{iR}$. If, for example, $\lambda_i \perp\!\!\!\perp (Y_{i1}, \ldots, Y_{iR})$, this assumption fails. Assumption L7 is similar. It says that given $Y_{i(2R+2)}$, λ_i still affects $Y_{i(2R+1)}, \ldots, Y_{i(R+2)}$ for some $Y_{i(R+1)}$. This assumption fails, for example, if $\lambda_i \perp\!\!\!\perp (Y_{i(2R+1)}, \ldots, Y_{i(R+2)}) \mid (Y_{i(2R+2)}, Y_{i(R+1)})$ but is not strong in general since $Y_{i(2R+1)}, \ldots, Y_{i(R+2)}$ is of the same dimension as λ_i. It is mainly rules out that λ_i is a deterministic function of Y_{it} for some t. The last three assumptions together say that Y_{it} has to be affected by λ_i but cannot be explained perfectly by it. Appendix B.2 explains these assumptions in more detail using a particular example. The main result of this section now follows.

Theorem 3. Let Assumptions L1 - L9 hold. Then $f_{Y_i,\lambda_i}(s, \lambda)$ is identified for all $s \in \mathcal{Y}$ and $\lambda \in \Lambda$. Moreover, F_t is identified for all $t \geq 2$. Assume in addition that either λ_i has support on $\mathbb{R}^R$ or that $U_i \perp\!\!\!\perp \lambda_i$. Then for all $t \geq 2$ the functions g_t as well as the distribution of $(U_{i2}, \ldots, U_{iT}, \lambda_i)$ are identified.

The conclusions are similar to the static case. The main difference is that F_1, g_1 and the distribution of U_{i1} are not identified because Y_{i0} is not observed.

4.2 Factor models with discrete heterogeneity

Above I assumed that Y_i and λ_i are continuously distributed. If λ_i is continuously distributed and Y_i is discrete, then the distribution of $(Y_i, \lambda_i) \mid X_i$ is in general not nonparametrically point identified. This is shown by Honoré and Tamer (2006) and Chamberlain (2010) in dynamic and static binary choice models, respectively. In this section, I focus on nonparametric identification of the distribution of $(Y_i, \lambda_i) \mid X_i$ when λ_i is discrete and Y_{it} is either discrete or continuous. I examine the static case, but the dynamic case can be analyzed analogously using arguments from

Section 4.1.

Many of the assumptions below are similar to the ones in Section 3. However, some of the previous assumptions are too strong when λ_i is discretely distributed. One reason is that the completeness assumption (Assumption S8) implies that Y_{it} and λ_i have the same number of points of support, which is stronger than needed. The assumptions below can be satisfied if the number of points of support of λ_i is less than or equal to the number of points of support of Y_{it}. This distinction could be interesting in applications when, for example, Y_{it} is continuously distributed but one only wants to allow for a high type and a low type and thus, λ_i is binary.

Let $\text{supp}(\lambda_{ir}) = \{\lambda_r^1, \ldots, \lambda_r^S\}$ be the support of the rth element of λ_i. Assume that S, the number of points of support, is known and is the same for all $r = 1, \ldots, R$. Let $\text{supp}(Y_{it})$ be the support of Y_{it} and let $|\text{supp}(Y_{it})|$ be the number of points of support, which could be infinity. Again

$$Y_{it} = g_t\left(X_{it}, \lambda_i' F_t + U_{it}\right), \qquad i = 1, \ldots, n,\ t = 1, \ldots, T. \tag{7}$$

The points of support of λ_i do not have to be known. The results in this section show identification of the distribution of $(Y_i, \lambda_i) \mid X_i$ where $\lambda_{ir} \in \{\lambda_r^1, \ldots, \lambda_r^S\}$, which implies identification of marginal effects for different quantiles of λ_i. The assumptions are as follows.

Assumption D1. R is known and $T = 2R + 1$.

Assumption D2. $U_{i1}, \ldots, U_{iT}$ are jointly independent conditional on λ_i and X_i.

Assumption D3. $S < \infty$ is known. Moreover, $\text{supp}(\lambda_i) = \text{supp}(\lambda_{i1}) \times \cdots \times \text{supp}(\lambda_{iR})$. Assume that Y_{it} has the same number of points of support for each t, and $S \leq |\text{supp}(Y_{it})|$.

Let $A_1, \ldots, A_S$ be a partition of the support of Y_{it} such that for all $a_{s_1} \in A_{s_1}$ and $a_{s_2} \in A_{s_2}$ with $s_1 < s_2$ it holds that $a_{s_1} < a_{s_2}$. Define Z_{i1}, Z_{iK}, and $Z_{ik_{R+1}}$ as in Section 3.1. Let $M = R \times S$. Let $C_1, \ldots, C_M$ be a partition of the support of Z_{iK} (and thus also of Z_{i1}) such that $C_m = A_{m_1} \times \cdots \times A_{m_R}$. For example if $Y_{it} \in \{0, 1\}$ and $S = 2$, then $A_1 = \{0\}$ and $A_2 = \{1\}$ as well as $C_1 = \{0, 0\}$, $C_2 = \{0, 1\}$, $C_3 = \{1, 0\}$, and $C_4 = \{1, 1\}$.

Conditional on X_i and for all $m_K, m_1 \in \{1, \ldots, M\}$ define

$$P_{m_K, m_1} \equiv P\left(Z_{iK} \in C_{m_K}, Z_{i1} \in C_{m_1}\right).$$

Let $L_{1,K}$ be the $M \times M$ matrix containing the probabilities such that m_K increases over rows while

m_1 increases over columns. That is

$$L_{1,K} \equiv \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,M} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ P_{M,1} & P_{M,2} & \cdots & P_{M,M} \end{pmatrix}.$$

Let $\lambda^1, \ldots, \lambda^M$ be an ordering of all points of support of λ_i. Let $L_{1,\lambda}$ be the $M \times M$ matrix containing

$$P_{l,m_1} = P\left(Z_{i1} \in C_{m_1} \mid \lambda_i = \lambda^l\right)$$

with l increasing over rows and m_1 increasing over columns.

Assumption D4. $L_{1,K}$ is invertible for any ordering K conditional on X_i. $L_{1,\lambda}$ is invertible conditional on X_i.

Assumption D5. $P\left(Y_{it} \in A_S \mid \lambda_i\right)$ is strictly increasing in $\lambda_i' F_t$.

Assumption D6. $F^3 = I_{R\times R}$ and F^1 has full rank, where F^1 and F^3 are defined in equation (4).

The assumptions are similar to the ones in the continuous case. The assumption that Y_{it} has the same support for all t is not needed but used to simplify the notation. Invertibility of the matrix $L_{1,K}$ is analogous to a completeness assumption and is similar to identification conditions in nonparametric instrumental variable models with discrete instruments and discrete regressors. Invertibility of $L_{1,\lambda}$ implies that λ_i has at most as many points of support as Z_{i1}. While for $R = 1$, Assumption D6 is just a normalization, it is important to notice that $F^3 = I_{R\times R}$ is not just a normalization if $R > 1$. The reason is that for all t, $\lambda_i' F_t$ has up to R^2 points of support while λ_i only has R points of support. The assumption is used for the ordering of the eigenvectors in the eigendecomposition. It can easily be replaced by different assumptions in specific models as illustrated in Section C.2. These assumptions lead to the following theorem.

Theorem 4. Assume that Assumptions D1 - D6 hold. Then, conditional on $X_i \in \mathcal{X}$,

$$P\left(Y_{i1} \in B_1, \ldots, Y_T \in B_T, \lambda_i = \lambda^m\right)$$

is identified for all $B_t \subseteq \mathcal{Y}_t$ for all $t = 1, \ldots, T$ and $m \in \{1, \ldots, M\}$.

5 Estimation

In this section, I describe how the static factor model with continuous outcomes and continuous heterogeneity can be estimated. I first discuss estimation of the structural functions and the distribution of $(U_i, \lambda_i) \perp\!\!\!\perp X_i$ for a known number of factors. There are many papers with similar estimation problems. Most closely related are the papers by Ai and Chen (2003), Hu and Schennach (2008), and Carroll, Chen, and Hu (2010). In Section 5.2 I show how the null hypothesis that the model has R factors can be tested against the alternative that the model has more than R factors, and how this result can be used to consistently estimate the number of factors.

5.1 Estimation with a known number of factors

First notice that by Assumptions S3 and S5, the density of Y_i given X_i can be written as

$$f_{Y_{i1},\ldots,Y_{iT}|X_i}(y;x) = \int \prod_{t=1}^{T} f_{U_{it}|X_i,\lambda_i}(h_t\left(y_t, x_t\right) - \lambda' F_t; x, \lambda) h_t'\left(y_t, x_t\right) f_{\lambda_i|X_i}(\lambda; x) d\lambda,$$

where $h_t'\left(y_t, x_t\right)$ denotes the derivative with respect to the first argument. Section 3 establishes that there are unique functions $f_{U_{it}|X_i,\lambda_i}$, h_t, and $f_{\lambda_i|X_i}$ as well as vectors F_t such that the conditional density can be written in this way. Due to this uniqueness result, estimation can be based on the sieve maximum likelihood method. The idea of sieve estimators is that the unknown functions are replaced by a finite dimensional approximation, which becomes more accurate as the sample size increases.[11] Although I show that a completely nonparametric maximum likelihood estimator is consistent, such an estimator is likely to be unattractive in applications due to the high dimensionality of the estimation problem. For example, the unknown function $f_{\lambda_i|X_i}(\lambda; x)$ is a function of $R+Td_x$ arguments where d_x denotes the dimension of X_{it}. Furthermore, the nonparametric rates of convergence in the strong norm (introduced below) can be very slow because the model nests nonparametric deconvolution problems of densities which can have a logarithmic rate of convergence.[12] Therefore, in practice a more convenient approach is a semiparametric estimator. The semiparametric estimator discussed in this paper reduces the dimensionality of the estimation problem by assuming a location and scale model for the conditional distributions. The structural functions can be nonparametric, semiparametric, or parametric. In the latter two cases, many parameters of interest are finite dimensional. I show that these estimated finite dimensional parameters are $\sqrt{n}$

[11] For an overview of sieve estimators see Chen (2007).

[12] For related setups see for example Fan (1991), Delaigle, Hall, and Meister (2008), and Evdokimov (2010).

consistent and asymptotically normally distributed.

In the following subsections, I present three kinds of estimators. First, I discuss consistency of a fully nonparametric sieve estimator. Next, I present the semiparametric estimator. Finally, I describe a fully parametric estimator and its asymptotic distribution.

5.1.1 Fully nonparametric estimator

I prove nonparametric consistency in a very general setup. All assumptions are listed in the appendix. In this section I outline the main assumptions and discuss how the estimator can be implemented. I first impose smoothness restrictions on the unknown functions. To do so, for any d-dimensional multi-index a define $|a| \equiv \sum_{j=1}^{d} a_j$. For any $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ denote the $|a|$-th derivative of a function $\eta : \mathbb{R}^d \to \mathbb{R}$ by

$$\nabla^a \eta(z) = \frac{\partial^{|a|}}{\partial z_1^{a_1} \cdots \partial z_d^{a_d}} \eta(z),$$

where $\nabla^a \eta(z) = \eta(z)$ if $a_j = 0$ for all $j = 1, \ldots, d$. For any $\gamma > 0$, let $\underline{\gamma}$ be the largest integer strictly smaller than γ. Denote by $\Lambda^\gamma(\mathcal{Z})$ the Hölder space with smoothness γ. These are all functions such that the first $\underline{\gamma}$ derivatives are bounded and the $\underline{\gamma}$-th derivative is Hölder continuous with exponent $\gamma - \underline{\gamma}$. That is, for all $\eta \in \Lambda^\gamma(\mathcal{Z})$ it holds that

$$\max_{|a| \leq \bar{\gamma}} \sup_{z \in \mathcal{Z}} |\nabla^a \eta(z)| < \infty \quad \text{and}$$

$$\max_{|a| = \bar{\gamma}} |\nabla^a \eta(z_1) - \nabla^a \eta(z_2)| \leq const ||z_1 - z_2||_E^{\gamma - \underline{\gamma}},$$

where $|| \cdot ||_E$ denotes the Euclidean norm. Define the norm

$$||\eta||_{\Lambda^\gamma} \equiv \max_{|a| \leq \bar{\gamma}} \sup_{z \in \mathcal{Z}} |\nabla^a \eta(z)| + \max_{|a| = \bar{\gamma}} \sup_{z_1 \neq z_2} \frac{|\nabla^a \eta(z_1) - \nabla^a \eta(z_2)|}{||z_1 - z_2||_E^{\gamma - \underline{\gamma}}}.$$

Next, define $||\eta||_{\Lambda^{\gamma,\omega}} \equiv ||\tilde{\eta}||_{\Lambda^\gamma}$ where $\tilde{\eta}(z) = \eta(z)\omega(z)$ and ω is a smooth, positive, and bounded weight function. Precise assumptions about the weight function are given in the appendix. In many cases $\omega(z) = 1$ or $\omega(z) = (1 + ||z||_E^2)^{-\varsigma/2}$ with $\varsigma > 0$, but different weight functions are possible. Moreover, denote the corresponding weighted Hölder space by $\Lambda^{\gamma,\omega}(\mathcal{Z})$. Finally, let

$$\Lambda_c^{\gamma,\omega}(\mathcal{Z}) \equiv \{\eta \in \Lambda^{\gamma,\omega}(\mathcal{Z}) : ||\eta||_{\Lambda^{\gamma,\omega}} \leq c < \infty\}.$$

In the remainder of this section I assume that all components of X_{it} are continuous. Discrete regressors can be handled by splitting the sample into subgroups depending to the values of the regressors. Define

$$\theta_0 = (h_{10}, \ldots, h_0, f_{10}, \ldots, f_{T0}, f_{\lambda 0}, F_0) = \left(h_1, \ldots, h_T, f_{U_{i1}|X_i,\lambda_i}, \ldots, f_{U_{iT}|X_i,\lambda_i}, f_{\lambda_i|X_i}, F\right).$$

The notation h_{t0} is now used to highlight that these values correspond to the true structural functions. In this section I assume that $U_i \perp\!\!\!\perp \lambda_i$ and that λ_i has support on $\mathbb{R}^R$. In many applications these seem to be reasonable assumptions and they nest many important special cases such as normally distributed λ_i. I also assume that Z_{i1} is bounded complete for Z_{iK} and that Z_{iK} is bounded complete for Z_{i1} for any ordering K conditional on X_i. I make these assumptions to facilitate imposing the identification conditions given in Section 3.1. The parameter space needs to reflect these conditions to ensure that the population objective function given below has a unique maximum. All assumptions, except the completeness assumption S8, are easily imposable. With the additional assumptions, it follows that λ_i is bounded complete for Z_{i1} conditional on X_i. As a consequence, completeness does not have to be imposed as an additional constraint on the densities. The reason is that even without this constraint, a solution to the population problem given below corresponds to the true density of $Y_i \mid X_i$. This density satisfies that Z_{i1} is bounded complete for Z_{iK} and that Z_{iK} is bounded complete for Z_{i1} by assumption. These completeness conditions then imply that λ_i is bounded complete for Z_{i1}. Without either making these additional assumptions or imposing completeness as a constraint, the identification arguments do not exclude the case where in addition to the true densities, $f_{10}, \ldots, f_{T0}, f_{\lambda 0}$, there are other densities, which yield the same distribution of $Y_i \mid X_i$, but λ_i is not bounded complete for Z_{i1}.

Next define the function spaces

$$\begin{aligned}
\mathcal{H}_t &\equiv \left\{\eta_t \in \Lambda_c^{\gamma_1,\omega_1}(\mathcal{Y}_t, \mathcal{X}_t) \text{ for some } \gamma_1 > 2 : \frac{\partial}{\partial y_t}\eta_t(y_t, x_t) \geq \varepsilon \text{ for some } \varepsilon > 0 \text{ and } S3 \text{ holds}\right\} \\
\mathcal{F}_t &\equiv \left\{\eta_t \in \Lambda_c^{\gamma_2,\omega_2}(\mathcal{U}_t, \mathcal{X}) \text{ for some } \gamma_2 > 1 : \int_{\mathcal{U}_t} \eta_t(u,x)du = 1,\ \eta_t(u,x) \geq 0, \text{ and } S4 \text{ holds}\right\} \\
\mathcal{F}_\lambda &\equiv \left\{\eta \in \Lambda_c^{\gamma_3,\omega_3}(\Lambda, \mathcal{X}) \text{ for some } \gamma_3 > 1 : \int_{\Lambda} \eta_t(\lambda,x)d\lambda = 1,\ \eta(\lambda,x) \geq 0, \text{ and } S7 \text{ holds}\right\}.
\end{aligned}$$

I assume that $h_{t0} \in \mathcal{H}_t$ and $f_{t0} \in \mathcal{F}_t$ for all $t = 1, \ldots, T$ as well as $f_{\lambda 0} \in \mathcal{F}_\lambda$ for appropriate weight functions given in the appendix.

The factors are assumed to lie in the set

$$\mathcal{V} \equiv \left\{ \tilde{F} \in \tilde{\mathcal{V}} \subset \mathbb{R}^{T\times R} : \tilde{\mathcal{V}} \text{ is compact and } S6 \text{ holds} \right\}. \tag{8}$$

Now define $\Theta \equiv \mathcal{H}_1 \times \cdots \times \mathcal{H}_T \times \mathcal{F}_1 \times \cdots \times \mathcal{F}_T \times \mathcal{F}_\lambda \times \mathcal{V}$. An element in the parameter space is denoted by $\theta = \left(\tilde{h}_1, \ldots, \tilde{h}_T, f_1, \ldots, f_T, f_\lambda, \tilde{F}\right)$ with $\theta_0 \in \Theta$. Define $W_i \equiv (Y_i, X_i)$ and

$$l\left(\theta, W_i\right) \equiv \log \int \prod_{t=1}^{T} f_t(\tilde{h}_t\left(Y_{it}, X_{it}\right) - \lambda' \tilde{F}_t; X_i, \lambda) \tilde{h}'_t\left(Y_{it}, X_{it}\right) f_\lambda(\lambda; X_i) d\lambda.$$

It holds that

$$\theta_0 = \arg\max_{\theta\in\Theta} E\left[l\left(\theta, W_i\right)\right]$$

and identification implies that θ_0 is the unique maximizer. Define the norm

$$||\theta||_s = \sum_{t=1}^{T} \left(||\tilde{h}_t||_{\infty,\tilde{\omega}_1} + ||f_t||_{\infty,\tilde{\omega}_2} + ||\tilde{F}_t||_E \right) + ||f_\lambda||_{\infty,\tilde{\omega}_3},$$

where $||\eta||_{\infty,\tilde{\omega}_k} \equiv \sup_{z\in\mathcal{Z}} |\eta(z)\tilde{\omega}_k(z)|$, and $\tilde{\omega}_k$ are weight functions similar to ω_k. The weight functions are defined in the appendix. They satisfy $\tilde{\omega}_k(z)/\omega_k(z) \to 0$ as $||z||_E \to \infty$ when the functions have unbounded support. Consistency is proved in the norm $||\cdot||_s$.

The estimator is implemented using the method of sieves. In particular let Θ_n be a growing finite dimensional sieve space which is dense in Θ. Commonly used are linear sieves such as Hermite polynomials for densities and splines or polynomials for the structural functions. In this case let $\{\phi_j(y, x)\}_{j=1}^{J_n}$ be a sequence of basis functions such as splines or polynomial and define

$$\mathcal{H}_{n,t} = \left\{ \eta_t \in \mathcal{H}_t : \eta_t = \sum_{j=1}^{J_n} a_j \phi_j(y, x) \text{ for some } (a_1, \ldots, a_{J_n}) \in A \subset \mathbb{R}^{J_n} \right\}.$$

Similar sieve spaces, $\mathcal{F}_{n,t}$ and $\mathcal{F}_{n,\lambda}$, can be defined for the densities. The assumptions imply that J_n increases as n increases. More details on suitable sieve spaces are provided in the appendix. Define $\Theta_n \equiv \mathcal{H}_{n,1} \times \cdots \times \mathcal{H}_{n,T} \times \mathcal{F}_{n,1} \times \cdots \times \mathcal{F}_{n,T} \times \mathcal{F}_{n,\lambda} \times \mathcal{V}$. The estimator of θ_0 is

$$\hat{\theta} = \arg\max_{\theta\in\Theta_n} \sum_{i=1}^{n} \log \int \prod_{t=1}^{T} f_t(\tilde{h}_t\left(Y_{it}, X_{it}\right) - \lambda' \tilde{F}_t; X_i, \lambda) \tilde{h}'_t\left(Y_{it}, X_{it}\right) f_\lambda(\lambda; X_i) d\lambda.$$

The following theorem is proved in the appendix.

Theorem 5. Let Assumptions E1 - E10 in the Appendix hold. Then

$$||\hat{\theta} - \theta_0||_s \xrightarrow{p} 0.$$

Given the assumptions, consistency follows from Theorem 3.1 in combination with Condition 3.5M in Chen (2007). Different parameter spaces and different choices of norms are possible. The reason for using a weighted Hölder space is that it allows for unbounded support and unbounded functions. Moreover, since λ_i has support on $\mathbb{R}^R$, the functions h_t have unbounded derivatives if Y_{it} has compact support. In all of these cases assuming that $||h_t||_{\Lambda^\gamma} \leq c$ is not reasonable. A weighted Hölder norm accommodates these cases because the weight function down weights the tails of the function and its derivatives. The choice also guarantees that the parameter space is compact with respect to the norm $|| \cdot ||_s$. Other popular function spaces are Sobolev spaces used for example by Gallant and Nychka (1987), Newey and Powell (2003), and Sasaki (2012). With unbounded support, these function spaces imply that the functions and their derivatives converge to 0 as the argument diverges. This is not be reasonable in my setting because the structural functions may, for example, be linear and conditional densities commonly lie outside the Sobolev spaces used in the aforementioned papers.[13] The costs of the weighted Hölder norm is that I prove convergence in the norm $|| \cdot ||_s$, which also down weights the tails of the functions. The norm $|| \cdot ||_s$ implies convergence in different, easier to interpret, norms. For example $||\eta_n - \eta||_{\infty,\omega} = o(1)$ implies $\sup_{z \in \bar{\mathcal{Z}}} |\eta_n(z) - \eta(z)| = o(1)$ for any bounded set $\bar{\mathcal{Z}}$ on which the weight function is strictly positive. It is also easy to show that $||\eta_n - \eta||_{\infty,\omega} = o(1)$ implies $\sup_{z \in \bar{\mathcal{Z}}_n} |\eta_n(z) - \eta(z)| = o(1)$ where $\bar{\mathcal{Z}}_n$ is a bounded but growing set as long as the set increases slow enough. Finally, if

$$\int_{-\infty}^{\infty} \omega^{-1}(z) f_Z(z) dz < \infty$$

it also holds that

$$\begin{aligned}
\int_{-\infty}^{\infty} |\eta_n(z) - \eta(z)| f_Z(z) dz &= \int_{-\infty}^{\infty} |\eta_n(z) - \eta(z)| \omega(z) \omega^{-1}(z) f_Z(z) dz \\
&\leq ||\eta_n - \eta||_{\infty,\omega_1} \int_{-\infty}^{\infty} \omega^{-1}(z) f_Z(z) dz \\
&= o(1).
\end{aligned}$$

[13] Newey and Powell (2003) assume that the tails of the functions they estimate are known up to a finite dimensional parameter vector to allow for unbounded functions.

5.1.2 Semi-parametric estimator

Many different semiparametric approaches are possible in this setting. In this section, I describe the approach I use in the application in Section 7. Detailed assumptions are listed in the appendix. Here, I focus on the main assumptions, alternative implementations, and the main idea of the proof.

First I reduce the dimensionality of the optimization problem by assuming a location and scale model for the conditional distribution of λ_i. In particular, I assume that

$$F_{\lambda_i|X_i}(\lambda; x_i) = F_\lambda(\Sigma^{-1}(x_i, \beta_1)\,(\lambda - \mu(x_i, \beta_2)))$$

for a positive definite matrix $\Sigma(x_i, \beta_1) \in \mathbb{R}^{R\times R}$ and a vector $\mu(x_i, \beta_2) \in \mathbb{R}^R$. The matrix $\Sigma(x_i, \beta_1)$ and the vector $\mu(x_i, \beta_2)$ are assumed to be known up to finite dimensional parameter vectors β_1 and β_2. The distribution function F_λ is unknown and its derivative is denoted by f_λ. Furthermore, I assume that U_{it} is independent of X_i and that the distribution of U_{it} is unknown. An alternative to this assumption is to model the dependence between U_{it} and X_i to allow for heteroskedasticity. Hence, both the distribution of $\lambda_i \mid X_i$ and the distribution of $U_i \mid X_i$ are semiparametric. An alternative to a scale and location model is to assume that the support of X_i can be partitioned into G groups and that $f_{\lambda_i|X_i}$ is the same for all X_i in group g. This approach is used by Weidner (2011) in a panel data model where $T \to \infty$.

The structural functions can be parametric, semiparametric, or nonparametric depending on the application. If the functions are parametric, I assume that $h_t(Y_{it}, X_{it}) = h(Y_{it}, X_{it}; \beta_{3t})$ where h is known up to the finite dimensional parameter β_{3t}, and $h(Y_{it}, X_{it}; \beta_{3t})$ is strictly increasing in the first argument. One possibility is to set $h(Y_{it}, X_{it}; \beta_{3t}) = h(Y_{it}, \beta_{3t1}) - X_{it}'\beta_{3t2}$ where $h(Y_{it}, \beta_{3t1})$ is a monotone transformation such as the Box-Cox transformation. It is also possible to specify the function semiparametrically or nonparametrically, which leads to similar asymptotic properties of the finite dimensional parameters of the model. For instance, in the previous example one could assume that the transformation function is unknown, but strictly increasing and satisfies the smoothness restrictions above.

Define $\beta \equiv (\beta_1, \beta_2, \beta_{31}, \ldots, \beta_{3T}, F)'$ and assume that $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_\beta}$. Also let $\alpha \equiv (f_1, \ldots, f_T, f_\lambda)$ and $\theta \equiv (\alpha, \beta)$. Denote the true parameter value by $\theta_0 \equiv (\alpha_0, \beta_0)$. The density functions are assumed to satisfy similar smoothness assumptions as in the previous section.

In addition to the various finite dimensional parameters, only T one-dimensional densities as well as one R-dimensional density function have to be estimated. Just as in the previous section, these densities are estimated using the method of sieves. The norm is similar to the one used in the previous section. The norm for the finite dimensional parameters is the standard Euclidean norm. Therefore, the estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$, which is the maximizer of the log-likelihood function, is computationally almost identical to the one described in the previous section. The main difference is that there are less sieve terms and more finite dimensional parameters to maximize over.

The appendix provides assumptions which ensure that the finite dimensional parameter vector is estimated at a $\sqrt{n}$ rate, and is asymptotically normally distributed. The arguments and proofs are very similar to the ones in Ai and Chen (2003) and Carroll, Chen, and Hu (2010) among others. The idea of the proof is to first introduce a weaker norm, namely the Fisher norm. Then, I show that $\hat{\theta}$ can be estimated at a rate faster than $n^{-1/4}$ under this norm. Using this fast rate of convergence in the weak norm and consistency in the strong norm, it can be shown that $\hat{\beta}$ converges to β_0 at a rate of $n^{-1/2}$ and is asymptotically normal. In order to achieve this rate, the length of the sieve has to be chosen in such a way that it balances the bias and the variance. Formally, I obtain the following theorem.

Theorem 6. Assume that Assumptions E5 and E11 - E21 hold. Then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, (V^*)^{-1}\right)$$

where V^* is defined in equation (13) in the appendix.

Ackerberg, Chen, and Hahn (2012) provide a consistent estimator of the covariance matrix $(V^*)^{-1}$, which is easy to implement.

5.1.3 Parametric estimator

Finally, given the results in the previous section, it is straightforward to estimate the model completely parametrically. The only additional assumption needed is that the densities $f_{U_{it}}$ and f_{λ_i} are known up to a finite dimensional parameter vector. For example, one could assume that λ_i and U_i are normally distributed, where the mean and the covariance of λ_i is a parametric function of X_i and the variance of U_{it} is an unknown constant. This parameterization, which I also use in the application, satisfies all identification assumptions. Various other parameterizations are of course also possible. Consistency and asymptotic normality of the fully parametric maximum likelihood estimator follows from standard arguments, such as those in Newey and McFadden (1986).

5.2 Hypothesis testing about the number of factors

In this section, I show how the null hypothesis that the model has R_1 factors can be tested against the alternative that the model has more than R_1 factors. I also discuss how such a test can be used to estimate the number of factors consistently.

As in the previous section, I assume that $\lambda_i \perp\!\!\!\perp U_i$. Let R_1 and R_2 be integers such that $R_1 < R_2 \leq \frac{1}{2}(T-1)$. Without normalizations of the factors, an R_1-factor model is just a special case of an R_2-factor model with $F_{tr} = 0$ for all t and $r = R_1 + 1, \ldots, R_2$. In both cases, the conditional distribution of Y_i given X_i can be written as

$$f_{Y_{i1},\ldots,Y_{iT}|X_i}(y;x) = \int \prod_{t=1}^{T} f_{U_{it}|X_i,\lambda_i}(h_t(y_t,x_t) - \lambda' F_t; x, \lambda) h_t'(y_t,x_t) f_{\lambda_i|X_i}(\lambda; x) d\lambda,$$

where $\lambda_i \in \mathbb{R}^{R_2}$. Without further normalizations there can be several functions $f_{U_{it},\lambda_i|X_i}$ and vectors F_t which yield the same conditional distribution of Y_i given X_i.

Now consider the following null hypothesis

$$H_0(R_1,R_2) : F_{tr} = 0 \text{ for all } t \text{ and } r = R_1+1,\ldots,R_2$$

versus

$$H_1(R_1,R_2) : F_{tr} \neq 0 \text{ for some } t \text{ and } r = R_1+1,\ldots,R_2.$$

Under the null hypothesis, the model is an R_1-factor model, while under the alternative, there are more than R_1 factors. As before, define $W_i \equiv (Y_i, X_i)$ and let

$$l(\theta, W_i) \equiv \log \int \prod_{t=1}^{T} f_t(\tilde{h}_t(Y_{it}, X_{it}) - \lambda' \tilde{F}_t; X_i, \lambda) \tilde{h}_t'(Y_{it}, X_{it}) f_\lambda(\lambda; X_i) d\lambda.$$

Furthermore, define

$$\phi(\theta, R_1, R_2) \equiv \left(\tilde{F}_{1(R_1+1)}, \ldots, \tilde{F}_{1R_2}, \ldots, \tilde{F}_{T(R_1+1)}, \ldots, \tilde{F}_{TR_2}\right).$$

The hypotheses above are equivalent to

$$H_0(R_1,R_2) : \phi(\theta_0, R_1, R_2) = 0 \quad \text{and} \quad H_1(R_1,R_2) : \phi(\theta_0, R_1, R_2) \neq 0.$$

$H_0(R_1, R_2)$ can be tested using a test statistic based on a scaled sample analog of

$$2\left(\sup_{\theta\in\Theta} E[l(\theta, W_i)] - \sup_{\theta\in\Theta:\phi(\theta,R_1,R_2)=0} E[l(\theta, W_i)]\right).$$

The intuition is that, under the null hypothesis, the difference above is equal to 0. Under the alternative, the difference is strictly positive because the maximum of the unconstrained problem is attained outside of the set where $\phi(\theta, R_1, R_2) = 0$. Furthermore, although not all features of the model are point identified, the value of the likelihood at θ_0 is identified. Define

$$LR(R_1, R_2) = 2\left(\sup_{\theta\in\Theta_n} \sum_{i=1}^{n} l(\theta, W_i) - \sup_{\theta\in\Theta_n:\phi(\theta,R_1,R_2)=0} \sum_{i=1}^{n} l(\theta, W_i)\right).$$

Chen, Tamer, and Torgovitsky (2011) show that, under the null hypothesis, $LR(R_1, R_2)$ converges in distribution to a supremum of a tight centered Gaussian process. They also prove that the quantiles of the asymptotic distribution can be approximated consistently using a weighted bootstrap. In a finite dimensional setup a similar result has been obtained by Lui and Shao (2003). Let $c_\alpha(R_1, R_2)$ denote the $1-\alpha$ quantile of the weighted bootstrap distribution. The test rejects the null hypothesis if and only if $LR(R_1, R_2) > c_\alpha(R_1, R_2)$. The following theorem is now a direct consequence of the results in Chen, Tamer, and Torgovitsky (2011).

Theorem 7. Assume that Assumptions S2 - S8 and Assumptions 1 - 4 in Chen, Tamer, and Torgovitsky (2011) hold. Also assume that $\lambda_i \perp\!\!\!\perp U_i$ and that $T \geq 2R+1$. Then:

(i) Under the null hypothesis $LR(R_1, R_2)$ converges in distribution to a supremum of a tight centered Gaussian process.

(ii) The likelihood ratio test has asymptotic size α:

$$P(LR(R_1, R_2) > c_\alpha(R_1, R_2) \mid H_0(R_1, R_2)) \to \alpha \text{ as } n \to \infty.$$

(iii) The likelihood ratio test is consistent against any fixed alternative:

$$P(LR(R_1, R_2) > c_\alpha(R_1, R_2) \mid H_1(R_1, R_2)) \to 1 \text{ as } n \to \infty.$$

Remark 3. The Assumptions in Chen, Tamer, and Torgovitsky (2011) are directly applicable to the setting in this paper but introducing them in detail would require a lot of notation and is therefore omitted. Similarly, a more precise description of the asymptotic distribution is omitted.

Next, I assume that the true number of factors is at most R^* where $R^* \leq \frac{1}{2}(T-1)$ is known.[14] Let $c_n(R_1, R^*)$ be a sequence of constants such that

$$\frac{c_n(R_1, R^*)}{n} \to 0 \quad \text{and} \quad \frac{c_n(R_1, R^*)}{\log(\log(n))} \to \infty.$$

It now follows from the fact that $LR(R_1, R^*)$ converges to a supremum of a tight Gaussian process under the null hypothesis and Theorem 1.3 in Dudley and Philipp (1983), which is a law of iterated logarithm for empirical processes, that

$$P(LR(R_1, R^*) > c_n(R_1, R^*) \mid H_0(R_1, R^*)) \to 0 \quad \text{and}$$
$$P(LR(R_1, R^*) > c_n(R_1, R^*) \mid H_1(R_1, R^*)) \to 1.$$

Let $LR(R^*, R^*) = 0$ and $c_n(R^*, R^*) > 0$. Then the estimated number of factors is

$$\hat{R} = \min\left\{R_1 \in \{0, \ldots, R^*\} : LR(R_1, R^*) < c_n(R_1, R^*)\right\}.$$

The following theorem is a direct consequence of the previous derivation.

Theorem 8. Assume that Assumptions S2 - S8 and Assumptions 1 - 4 in Chen, Tamer, and Torgovitsky (2011) hold. Also assume that $\lambda_i \perp\!\!\!\perp U_i$ and that $R \leq R^* \leq \frac{1}{2}(T-1)$. For any $R_1 = 0, \ldots, R^*-1$ let $c_n(R_1, R^*)$ be a sequence of constants that satisfies

$$\frac{c_n(R_1, R^*)}{n} \to 0 \quad \text{and} \quad \frac{c_n(R_1, R^*)}{\log(\log(n))} \to \infty.$$

Let $LR(R^*, R^*) = 0$ and $c_n(R^*, R^*) > 0$. Then $P(\hat{R} = R) \to 1$.

Remark 4. In practice, the sequence $c_n(R^*, R^*)$ needs to be chosen. This is a very important issue because for any given sample, the estimated number of factors directly follows from this sequence. Selecting $c_n(R^*, R^*)$ with desirable finite sample properties, similar to the ones proposed by Bai and Ng (2002) in linear factor models, is left for future research. In applications with small T, R^* might be quite small. For instance, in the application in this paper, since $T = 6$ the upper bound

[14] A similar assumption is made in linear factor models; see for example Bai and Ng (2002).

for R^* is 2. Hence, one might only be interested in testing whether $R = 1$ versus the alternative that $R = 2$. In these cases, a test based on Theorem 7 might be more appealing than estimating the number of factors consistently.

6 Monte Carlo simulation

In this section, I investigate the finite sample properties of the parametric and the semiparametric estimator. I consider two setups with different shapes of the structural functions. In the first setup, I am concerned with the finite sample properties of the finite dimensional parameters for various distributions of U_{it} and different sample sizes. In the second setup, I replicate the distribution of the data used in the next section and investigate finite sample properties of marginal effects.

6.1 Setup 1

I simulate the data from the model

$$\frac{Y_{it}^{\alpha} - 1}{\alpha} = \gamma + X_{it}'\beta + \lambda_i' F_t + U_{it}, \tag{9}$$

where $X_{it} \in \mathbb{R}^4$, $\lambda_i \in \mathbb{R}^2$, and $T = 6$. I assume that $X_{it1}, X_{it2} \sim U\left[0, 1\right]$ and $X_{it3}, X_{it4} \sim TN\left(0, 1\right)$, where $TN(0, 1)$ denotes the standard normal distribution truncated at -1 and 1. Furthermore, $\lambda_i \sim N\left(\mu(X_i, \theta), \Sigma\right)$ with

$$\Sigma = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix}, \quad \mu(X_i, \theta) = \begin{pmatrix} \bar{X}_{i\cdot 1}\theta_1 + \bar{X}_{i\cdot 3}\theta_3 \\ \bar{X}_{i\cdot 2}\theta_2 + \bar{X}_{i\cdot 4}\theta_4 \end{pmatrix}, \quad \text{and} \quad \bar{X}_{i\cdot k} = \frac{1}{T}\sum_{t=1}^{T} X_{itk}.$$

The interpretation is that both skills depend linearly on two of the covariates. I vary the distribution of U_{it}, which is either (i) Normal, (ii) Gamma with scale parameter 1 and shape parameter 5, (iii) Student's t with 5 degrees of freedom, or (iv) Logistic. Each distribution is standardized such that the median equals 0 and the standard deviation is 0.5. I assume that $U_i \perp\!\!\!\perp (X_i, \lambda_i)$. These four error distributions cover many interesting cases. The first and last are standard and commonly used distributions while the Gamma distribution is asymmetric and the Student's t distribution does not have exponentially decreasing tails. I assume that

$$\beta = \begin{pmatrix} 0.5 & 0.8 & -0.5 & 0.2 \end{pmatrix}', \quad \theta = \begin{pmatrix} 0.5 & -0.3 & 0.3 & 0.4 \end{pmatrix}', \quad \text{and}$$

$$F = \begin{pmatrix} 1 & 0 & -0.212 & 0.559 & 0.856 & 1.221 \\ 0 & 1 & 1.017 & 0.123 & -0.133 & 0.232 \end{pmatrix}'.$$

The value of α determines the shape of the structural function. Here $\alpha = 0.75$ which means that g_t is convex. In the next section $\alpha > 1$. I choose γ such that the right hand side is positive in every sample because the Box-Cox transformation is only valid in this case. One could also redefine the transformation to allow for negative values.

I estimate the model using both a fully parametric estimator and a semiparametric estimator. For both estimators, I assume that

$$\lambda_i = \mu(X_i, \theta) + \varepsilon_i,$$

where the function μ is known up to the finite dimensional parameter vector θ. I also assume in both cases that the functional form of the outcome equation (9) is known. When using the parametric estimator, I assume that the distribution of ε_i is known up to the covariance matrix, and the distribution of U_{it} is known up to the variance. In the semiparametric setting, both distributions are unknown. Whether or not these distributions are known is the only difference between the semiparametric and the parametric estimator. The sample size is either $n = 200$ or $n = 500$ which are both smaller than the sample size in the application. I assume that $T = 6$ which implies that the total number of observations, $n \times T$, is 1200 and 3000, respectively.

For the semiparametric estimator, I approximate the distribution of U_{it} using Hermite polynomials of length $J_{un} = 3$ for $n = 200$ and $J_{un} = 5$ for $n = 500$. The distribution of λ_i is estimated using a Tensor product of Hermite polynomials of length 3. See Chen (2007) for an expression of Hermite polynomials and other basis functions which could be used. With Hermite polynomials, the constraints that the distribution of U_{it} has a median of 0 and that all distributions are positive and integrate to 1 are linear and quadratic in the sieve coefficients. Hence, they are easy to implement. Alternatively, one could approximate the square root of the density using Hermite polynomials, which has the advantage that the resulting density is always positive. This alternative approach does not change my simulation results.

I approximate the integral over λ_i in the likelihood using Gauss-Hermite quadrature. This is a convenient choice because the weight function, e^{-x^2}, appears both in the expression of the Hermite polynomials and the normal density. Alternative options are, for example, Monte Carlo integration with quasi-random draws or Halton sequences. These approaches are possible even in the semiparametric case because Hermite polynomials are built around a normal density. My experience is that Gauss-Hermite quadrature leads to slightly better finite sample properties.

Table 1 provides finite sample properties for estimates of α, β_1, and θ_1. Different values of α yield very similar results and are therefore not reported. The parameters α, β_1, and θ_1 cover different aspects of the model. The parameter α measures the nonlinearity of the structural function, β_1 is one of the coefficients of the regressor, and θ_1 is a parameter of the distribution of $\lambda_i \mid X_i$. The results for the other parameters are very similar. I focus on the mean and the root mean square error (RMSE) of the estimated parameters. All results are based on 1000 Monte Carlo iterations. The fully parametric model has very good finite sample properties. For all values of α and all distributions of U_{it}, both the bias and the RMSE are small and the RMSE decreases with the sample size. According to the asymptotic approximation, the RMSE with $n = 200$ should be $\sqrt{5/2} \approx 1.58$ times larger than the RMSE with $n = 500$, which is roughly true in the parametric case. In the semiparametric case, this is only approximately the case if U_{it} is either logistically or normally distributed. Clearly, the fully parametric estimator has better finite sample properties compared to the semiparametric estimator. However, the semiparametric estimator performs quite well for all distributions even in relatively small samples. The semiparametric estimator achieves its best results for the normal distribution. This is not surprising since the Hermite polynomials are built around a normal density. The estimator performs well even in the case of an asymmetric error distribution if the size of the sieve is long enough as in the case of the Gamma distribution. With $n = 200$ and $J_{un} = 3$, the estimates are substantially biased. This bias mostly disappears with $n = 500$ and $J_{un} = 5$. The Student-t and logistic distribution have fatter tails, which results in more variation of the estimates relative to the normal distribution.

6.2 Setup 2

The second setup mimics the data and the model used in the application. Now

$$\frac{Y_{it}^{\alpha} - 1}{\alpha} = \gamma + X_{it}\beta_t + Z_{it}\delta_t + \lambda_i' F_t + U_{it},$$

where $X_{it}, Z_{it} \in \mathbb{R}$, $\lambda_i \in \mathbb{R}^2$, and $T = 6$. Moreover, $X_{it} = X_{i1}$ for all $t = 1, 2, 3$ and $X_{it} = X_{i4}$ for all $t = 4, 5, 6$. The same holds for Z_{it}. I assume that X_{it} and Z_{it} have truncated normal distributions. The means, the covariance matrix, and the cutoffs are chosen such that the distributions closely mimic the empirical distributions of the teaching practice measures in the application.[15] The

[15] The regressors X_{it} and Z_{it} correspond to the traditional and modern teaching practice measure, respectively. In the application $t = 1, 2, 3$ belongs to mathematics and $t = 4, 5, 6$ belongs to science test scores. Nonparametric

Table 1: Mean and root-MSE for α, β_1, and θ_1

	$\alpha = 0.75$		$\beta_1 = 0.5$		$\theta_1 = 0.5$	
	Mean	RMSE	Mean	RMSE	Mean	RMSE
$U_{it} \sim$ Normal - SPMLE						
$n = 200$	0.748	0.049	0.505	0.098	0.497	0.253
$n = 500$	0.750	0.040	0.505	0.075	0.515	0.160
$U_{it} \sim$ Normal - MLE						
$n = 200$	0.746	0.046	0.500	0.094	0.487	0.238
$n = 500$	0.748	0.033	0.501	0.066	0.503	0.148
$U_{it} \sim$ Gamma - SPMLE						
$n = 200$	0.583	0.173	0.306	0.202	0.320	0.232
$n = 500$	0.694	0.066	0.426	0.091	0.427	0.146
$U_{it} \sim$ Gamma - MLE						
$n = 200$	0.742	0.046	0.493	0.085	0.515	0.230
$n = 500$	0.745	0.032	0.494	0.057	0.491	0.142
$U_{it} \sim$ Student's t - MMLE						
$n = 200$	0.784	0.109	0.598	0.330	0.612	0.502
$n = 500$	0.776	0.089	0.576	0.251	0.589	0.296
$U_{it} \sim$ Student's t - MLE						
$n = 200$	0.741	0.062	0.498	0.114	0.499	0.254
$n = 500$	0.745	0.048	0.499	0.091	0.505	0.162
$U_{it} \sim$ Logistic - SMLE						
$n = 200$	0.784	0.060	0.531	0.121	0.559	0.283
$n = 500$	0.752	0.044	0.507	0.082	0.510	0.163
$U_{it} \sim$ Logistic - MLE						
$n = 200$	0.741	0.053	0.504	0.104	0.523	0.248
$n = 500$	0.748	0.036	0.499	0.069	0.497	0.153

The number of Monte Carlo simulations is 1000. The true value of α, β_1, and θ_1 are 0.75, 0.5, and 0.5 respectively. The distribution of U_{it} is approximated using Hermite polynomials of length $J_{un} = 3$ for $n = 200$ and $J_{un} = 5$ for $n = 500$. The distribution of λ_i is approximated using a tensor product of Hermite polynomials of length 3.

sample size is $n = 835$ as in the application. I also assume that $\beta_t = \beta_1$ for $t = 1, 2, 3$ and $\beta_t = \beta_4$ for $t = 4, 5, 6$, and I make the same assumption for δ_t. As before, $\lambda_i \sim N(\mu(X_i, Z_i, \theta), \Sigma)$ but now

$$\Sigma = \begin{pmatrix} 2.21 & 1.47 \\ 1.47 & 2.21 \end{pmatrix},$$

which are the point estimates from the empirical application in Section 7. I also assume that $\mu(X_i, Z_i, \theta)$ is a quadratic function of X_{i1}, X_{i4}, Z_{i1}, and Z_{i4}. Notice that the correlation between the two skills is roughly 0.67. I set $\alpha = 1.18$ as well as

$$\beta = \begin{pmatrix} 0.29 & 0.29 & 0.29 & -0.20 & -0.20 & -0.20 \end{pmatrix}, \quad \text{and}$$

$$\delta = \begin{pmatrix} -0.18 & -0.18 & -0.18 & 0.42 & 0.42 & 0.42 \end{pmatrix},$$

which are the point estimates from the empirical application. Notice that the function g_t is concave. The values of θ are also set at the point estimates and so is

$$F = \begin{pmatrix} 1.00 & 0.93 & 0.80 & 0.03 & 0.00 & 0.11 \\ 0.00 & 0.06 & 0.04 & 0.98 & 1.00 & 0.88 \end{pmatrix}.$$

I assume that $U_{it} \sim N\left(0, \sigma_t^2\right)$ with the vector of standard deviations being

$$\sigma = \begin{pmatrix} 0.23 & 0.40 & 0.88 & 0.37 & 0.36 & 0.50 \end{pmatrix}',$$

which are again the point estimates in the application.

I investigate finite sample properties of estimated marginal effects. There are four marginal effects I consider. First there are the effects of a change of X_{it} on the outcome for $t = 1, 2, 3$ and for $t = 4, 5, 6$. In the application, these are the marginal effects of a change in traditional teaching practice on the mathematics scores ($t = 1, 2, 3$) and the science scores ($t = 4, 5, 6$), respectively. Second there are the effects of a change of Z_{it} on the outcome for $t = 1, 2, 3$ and for $t = 4, 5, 6$. In the application, these are the marginal effects of a change in modern teaching practice on the mathematics scores and the science scores, respectively. The results are based on 1000 Monte Carlo simulations. As in the previous subsection, the unknown distributions of the unobservables are approximated using Hermite polynomials.

identification in this setup is shown in Section 7.3.

Table 2 shows the means of the estimated marginal effects for different estimation methods. All marginal effects are evaluated at the median values of X_{it}, Z_{it}, and of the unobservables. The first line contains the true marginal effects, which are very close the estimated marginal effects in the application. The second line shows the marginal effects when a standard linear fixed effects model is used for estimation. For $t = 1, 2, 3$, the estimated marginal effects have the right signs but the estimates are considerably biased downwards in absolute value. For $t = 4, 5, 6$, the estimates are far from the true values. It appears as if X_{it} has a positive effect on the outcome, although the true effect is negative. Both the parametric model and the semiparametric two factor model yield estimated marginal effects that are close to the true values, with the parametric estimator performing better on average.

Table 2: Mean of estimated marginal effects

	$t = 1, 2, 3$		$t = 4, 5, 6$	
	X_{it}	Z_{it}	X_{it}	Z_{it}
True marginal effect	0.186	-0.112	-0.132	0.269
Linear fixed effects	0.074	-0.009	0.044	0.020
Parametric - two factor	0.181	-0.108	-0.134	0.270
Semiparametric - two factors	0.172	-0.099	-0.113	0.235

Table 3 displays the root means square error (RMSE) of the estimates. The estimates from the linear model have a large RMSE compared to the two factor models. The difference is especially large for $t = 4, 5, 6$. The parametric and the semiparametric two factor models yield very similar results.

Table 3: RMSE of estimated marginal effects

	$t = 1, 2, 3$		$t = 4, 5, 6$	
	X_{it}	Z_{it}	X_{it}	Z_{it}
Linear fixed effects	0.116	0.107	0.178	0.250
Parametric - two factor	0.062	0.066	0.068	0.090
Semiparametric - two factors	0.063	0.063	0.069	0.091

7 Application

In this section I use the static factor model to investigate the relationship between teaching practice and student achievement. I use test scores of students from the Trends in International Mathematics and Science Study (TIMSS) as outcome variables. This study is linked to a student questionnaire, which allows me to calculate a measure of modern teaching practice and a measure of traditional teaching practice for each class a student attends. Modern teaching practice is associated with group work and reasoning, while traditional teaching practice is based on lectures and memorizing. The two measures are defined below. A standard linear fixed effects model controls for a scalar unobserved ability term, which has the same effect on all outcomes. This assumption means, loosely speaking, that if two students, A and B, have the same observed characteristics and student A is better in subject 1, then student A must also be better in subject 2. Furthermore, the impact of the teaching practice on student achievement is, by assumption, the same for all levels of unobserved ability. These assumptions can be relaxed by using the factor model described in this paper.

Linear fixed effects models are often used in similar settings to control for unobserved ability. Bietenbeck (2011) and Lavy (2011) study the relationship between teaching practice and student achievement. Dee (2007) analyzes whether the teacher's gender has an influence of student achievement. Clotfelter, Ladd, and Vigdor (2010) investigate the relationship between teacher credentials and student achievement. In all these papers, the T dimension in the panel are different subjects. Aucejo (2011) studies teacher effectiveness and allows for student-teacher interactions with a scalar student fixed effect. Although a nonlinear factor model, which controls for multiple abilities, seems attractive in these settings, it should be noted that such a model is only applicable if $T \geq 5$.

In this application, I make use the definitions of traditional and modern teaching practice by Bietenbeck (2011). The main difference in my paper is the model used to estimate the parameters of interest. Moreover, I use different test scores as outcome variables and the sample differs slightly as explained below.

7.1 Data

TIMSS is an international assessment of mathematics and science knowledge of fourth and eighth-grade students. It is carried out every four years. I make use of the 2007 sample of eighth-grade students in the United States. This sample consists of 7377 students in 235 schools. Each student attends a mathematics and an integrated science class with different teachers in each of the two classes for almost all students. I exclude students which cannot be linked to their teachers, students in classes with less than three students, as well as observations with missing values in the teaching practice variables or control variables (defined below). This reduction leaves 4642 students in 182 schools. This is the estimation sample of Bietenbeck (2011), who provides more details on the data. I further restrict myself to white, American born students between the age of 13.5 and 15.5 with English as their first language. I also restrict the sample to schools with an enrollment between 100 and 600 students, where parents' involvement is not reported to be very low, and where less than 75% of the students receive free lunch. The resulting sample consists of 835 male and 935 female students in 99 schools with 144 mathematics and 161 science teachers.[16]

In addition to the overall test score for mathematics and science, the TIMSS contains test scores for different cognitive domains of the tests which are mathematics knowing, applying and reasoning, as well as science knowing, applying and reasoning. I use these six test scores as the dependent variables Y_{it}, where i denotes a student and t denotes a test. Hence, $T = 6$ which allows me to estimate a factor model with two factors. The main regressors are the measures of teaching practice. To construct these, I use the student questionnaire where students are asked questions about how often they do certain activities in class. The answers are on a four point scale with 1 corresponding to *never*, 2 to *some lessons*, 3 to *about half of the lessons*, and 4 to *every or almost every lesson.* The exact questions about class activities, which are used to construct the measures, are listed in Table 6. These particular questions are used because they can be unambiguously matched to recommendations on teaching practices in Zemelman, Daniels, and Hyde (2005). These recommendations are based on a survey of the standards movement in teaching practices literature and categorize teaching methods as either to be increased or to be decreased. In Table 6, questions belonging to traditional teaching are the ones labeled to be decreased in Zemelman et al. (2005), while questions belonging to modern teaching are labeled to be increased.[17] For each student and each class I compute the mean response of the answers in the modern and traditional category. I then calculate class means of these averages, excluding the student's own response. These class means are the measures of traditional and modern teaching practice faced by student i in the mathematics and science class.[18] These measures therefore range from 1 to 4. In addition to these teaching measures, I control for class size, hours spent in class, teacher experience, whether a teacher is certified in the field, and the gender of the teacher.

[16] All results are very similar in different samples, such as the sample without conditioning on the school variables.

[17] See Bietenbeck (2011) for more details on the teaching practice measures and background literature.

[18] The results are very similar when I include the student's own response when constructing the teaching measures.

7.2 Model and implementation

The results reported in this paper are based on the outcome equation

$$\frac{Y_{it}^{\alpha}-1}{\alpha} = \gamma + X_{it}^{mod}\beta_t^{mod} + X_{it}^{trad}\beta_t^{trad} + Z_{it}'\gamma_t + \lambda_i' F_t + U_{it}, \tag{10}$$

where $t = 1, 2, 3$ corresponds to the mathematics scores (knowing, applying, reasoning) and $t = 4, 5, 6$ to the science scores (knowing, applying, reasoning). The scalars X_{it}^{mod} and X_{it}^{trad} are the modern and traditional teaching practice measures of the classes which student i attends. I assume that

$$\beta_1^{trad} = \beta_2^{trad} = \beta_3^{trad} \quad \text{and} \quad \beta_4^{trad} = \beta_5^{trad} = \beta_6^{trad}.$$

I make an analogous assumption for β_t^{mod}. Hence, I allow traditional and modern teaching practice to have a different impact on mathematics and science subjects, but the same impact across cognitive domains. The vector Z_{it} includes class size, hours spent in class, teacher experience, whether a teacher is certified in the field, and the gender of the teacher. I assume that

$$\lambda_i = \mu(X_i, \theta) + \varepsilon_i,$$

where $\varepsilon_i \perp\!\!\!\perp X_i$ and μ is a quadratic function of X_i^{mod} and X_i^{trad} but does not include interactions of these regressors.[19] I also assume that $U_i \perp\!\!\!\perp (\lambda_i, X_i)$. I implemented various specifications, none of which change the main conclusions. These different specifications include adding nonlinear terms of X_i and Z_i to the right hand side of equation (10), assuming that the covariance matrix of ε_i is a parametric function of X_i, allowing λ_i to depend on Z_i, and allowing β to differ across all t.

In the most general model I estimate, the distributions of U_{it} and ε_i are unspecified and estimated with Hermite polynomials of order 5 and a tensor product of Hermite polynomials of order 3, respectively. Changing the order does not affect the results much. I also estimate different parametric models assuming that ε_i and U_{it} are normally distributed with an unknown covariance matrix and an unknown variance, respectively. In particular, I estimate a parametric two factor model, as well as a parametric one factor model and a parametric one factor model with $F_t = 1$. In all of these models, I use the functional form assumptions above. The estimates are compared to the ones obtained from a linear fixed effects model. All integrals are approximated using Gauss-Hermite quadrature.

[19] With this assumption, λ_i become correlated random effects instead of fixed effects.

7.3 Nonparametric identification

Although a semiparametric model is used in this application, the structural functions are nonparametrically identified, under the assumptions of Theorem 1. Nonparametric identification might not be immediately obvious from Theorem 1 because $Z_{it} = Z_{i1}$ for $t = 1, 2, 3$ and $Z_{it} = Z_{i4}$ for $t = 4, 5, 6$. The same is true for X_{it}^{mod} and X_{it}^{trad}. To simplify the notation, define $X_{it} \equiv \left(X_{it}^{mod}, X_{it}^{trad}, Z_{it}\right)$. First let $X_{i5} = X_{i6} = \bar{x}_6$ as in Theorem 1. Hence also $X_{i4} = \bar{x}_6$. If $f_{X_{i1},X_{i6}}(x_1, \bar{x}_6) > 0$ for all $x_1 \in \mathcal{X}_1$, it follows immediately from Theorem 1 that h_t is identified for all $x_t \in \mathcal{X}_t$ and $t = 1, 2, 3$. Now assume that for some $\bar{x}_1$ it holds that $f_{X_{i1},X_{i6}}(\bar{x}_1, x_6) > 0$ for all $x_6 \in \mathcal{X}_6$. Then, just as in the discussion after Theorem 1, we can switch roles of $t = 1, 2, 3$ and $t = 4, 5, 6$ and identify h_t for all $x_t \in \mathcal{X}_t$ and $t = 4, 5, 6$.

To achieve these identification results, it is not necessary that $f_{X_{i1},X_{i6}}(x_1, \bar{x}_6) > 0$ for all $x_1 \in \mathcal{X}_1$. For example, assume that all components of X_{it} are continuously distributed. Furthermore, assume that for all $\bar{x}$ and some $\delta > 0$ it holds that $f_{X_{i1},X_{i6}}(x_1, \bar{x}) > 0$ and $f_{X_{i1},X_{i6}}(\bar{x}, x_6) > 0$ for all $x_1, x_6 \in [\bar{x} - \delta, \bar{x} + \delta]$. This assumption is reasonable in this application. It says that if the traditional teaching measure in the mathematics class has the value $\bar{x}$, then all values in the interval $[\bar{x} - \delta, \bar{x} + \delta]$ are possible values for the traditional teaching measure in the science class. A similar statement has to be true for the modern teaching measure. With these assumptions, Theorem 1 yields identification of h_t for all $x_t \in [\bar{x} - \delta, \bar{x} + \delta]$ and $t = 1, 2, 3$. Switching roles of $t = 1, 2, 3$ and $t = 4, 5, 6$ shows identification of h_t for all $x_t \in [\bar{x} - 2\delta, \bar{x} + 2\delta]$ and $t = 4, 5, 6$. Switching roles again, it then follows that h_t is identified for all $x_t \in [\bar{x} - 3\delta, \bar{x} + 3\delta]$ and $t = 1, 2, 3$. Since this process can be iterated, the functions h_t are identified for all t and for all $x_t \in \mathcal{X}_t$.

Next to identifying the structural functions, the distribution of $\lambda_i \mid X_i$ is identified using similar arguments. However, note that this conditional distribution is not identified for all $x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_6$. The reason is that $X_{i1} = X_{i2} = X_{i3}$ and $X_{i4} = X_{i5} = X_{i6}$ for all i and thus, the distribution of $\lambda_i \mid X_i$ is only identified for values of X_i for which these equalities hold.

7.4 Results

Table 4 contains the estimation results for the sample of 835 boys.[20] The table shows marginal effects of an increase in teaching practice evaluated at the median levels of the regressors and the

[20] Since each student only answers a random sample of test questions, the TIMSS data set contains five imputed values. I use the first imputed value, but the results based on the other ones are similar. The standard errors do not account for imputation of missing values. Furthermore, all issues caused by the fact that the teaching measures are generated regressors are ignored.

unobservables. In a linear model, these marginal effects are equal to the four coefficients β_t^{mod} and β_t^{trad} with $t = 1$ and $t = 4$. First consider the results of the linear fixed effects model in the first row. Using this model, I find a positive relationship between traditional teaching practice and student achievement in both subjects. The relationship between modern teaching practice and science scores is also positive but insignificant. These results are in line with findings in Bietenbeck (2011) and Lavy (2011). Bietenbeck (2011) mainly finds a positive relationship between traditional teaching practice and student achievement while Lavy (2011) finds evidence of positive effects of both modern and traditional elements. I standardized Y_{it} and the teaching practice measures to have a standard deviation of 1. Hence, the economic interpretation is that a one standard deviation increase of tradition practice is associated with a 0.085 standard deviation increase in the mathematics test score.

Table 4: Marginal effects teaching practice for boys

	Math scores		Science scores	
	Traditional	Modern	Traditional	Modern
Linear fixed effects	0.085***	-0.002	0.040*	0.027
Parametric - one factor - $F_t = 1$	0.087***	-0.005	0.045***	0.027**
Parametric - one factor	0.126***	-0.032*	0.093***	-0.015
Parametric - two factors	0.188***	-0.114**	-0.134**	0.265***
Semiparametric - two factors	0.165***	-0.137**	-0.145**	0.236***

The symbols *, **, and *** denote significance at 10%, 5%, and 1% level respectively.

The next line shows marginal effects for a parametric one factor model with $F_t = 1$. The marginal effects, evaluated at median values, are very similar to the linear model. This is expected since the models are very similar. Notice that in the linear model, the relation between λ_i and X_i is not modeled. Hence, a large difference in these marginal effects could be due to misspecification of this relationship while the similarity suggests that the relationship is well specified. Allowing F_t to vary produces different marginal effects as shown in the third line. The results still suggest that traditional teaching practices are associated with better test scores in both subjects.

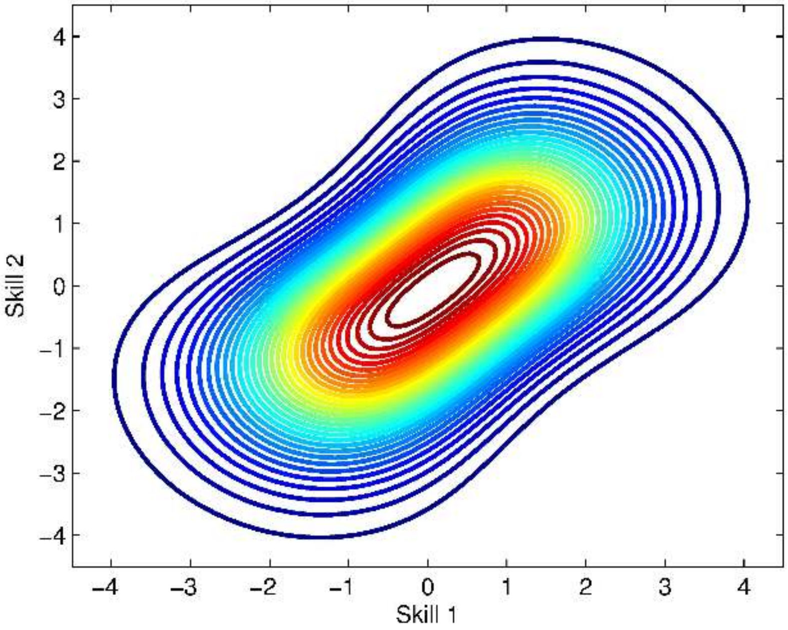
A parametric two factor model yields very different results. I still find a positive relationship between traditional teaching practice and mathematics scores, but a positive relationship between modern teaching practice and science scores. The two other marginal effects are significantly negative. Finally, the semiparametric model yields similar results as the parametric two factor model. Results for the sample of 935 girls are given in Appendix E and are qualitatively similar.

The semiparametric model also provides an estimate of the distribution of λ_i. Figure 1 shows level curves of this distribution at the median level of X_i. The correlation between the two skills is around 60%. In the parametric model, the estimated correlation is around 67%. It is also interesting to look at the estimated factors from the semiparametric model which are

$$\hat{F} = \begin{pmatrix} 1.00 & 0.94 & 0.81 & 0.04 & 0.00 & 0.13 \\ 0.00 & 0.06 & 0.04 & 0.97 & 1.00 & 0.88 \end{pmatrix}.$$

The rows correspond to the two different factors and columns to the different subject. The first column belongs to mathematics knowing, and columns $2-6$ to mathematics applying, mathematics reasoning, science knowing, science applying and science reasoning, respectively. The mathematics subjects have more weights on the first skill, while science subjects have more weight on the second skill. Notice that two numbers are exactly 0 and two are exactly 1, which corresponds to a particular normalization.

Figure 1: Level curves skill distribution



Using the results from Section 5.2, I can test the parametric one factor model against the parametric two factor model, and the semiparametric one factor model against the semiparametric two factor model. In both cases, the null hypothesis is a one factor model and the alternative is a two factor model. Both null hypotheses are rejected at any significance level greater than 0.5%.

The conclusions one can draw from using a nonlinear model are illustrated in Table 5. This table contains marginal effects, using the semiparametric two factor model, for different values of the skill distribution. High skills means being two standard deviations above the median value in both dimensions, and low skills means being two standard deviations below. The estimated coefficient measuring nonlinearity, α, is equal to 1.221 with a standard error of 0.062. Hence, the estimate is significantly different from 1. In the table it can be observed that the marginal effects are larger in absolute value for students with lower abilities. This means that the teaching method seems to have a larger influence on students with low ability. The difference in the marginal effects between the high and the low skilled students is more than 20%.

Table 5: Marginal effects teaching practice for boys

	Math scores		Science scores	
	Traditional	Modern	Traditional	Modern
Low skill	0.185	-0.154	-0.163	0.265
Median skill	0.165	-0.137	-0.145	0.236
High skill	0.154	-0.128	-0.135	0.220

These marginal effects are based on the semiparametric two factor model. They are evaluated at the median of the observables.

8 Conclusion

In this paper, I consider identification and estimation of nonparametric panel data models with multidimensional, unobserved individual effects. The models have two key distinguishing features. First, the setup allows for the various unobserved individual effects to impact outcomes differently in different time periods. Individual effects often represent unobserved ability, in which case the model permits the returns to ability to change over time. Second, the models allows for heterogeneous marginal effects of the covariates on the outcome, which implies, for example, that returns to education may depend on unobserved ability.

I present nonparametric point identification conditions for all parameters of the models. These parameters include the structural functions as well as the number of factors, the factors themselves, and the distributions of the unobservables, λ_i and U_{it}, conditional on the regressors. The identification results imply identification of average effects as well as marginal effects, which are often the main objects of interest in applications. I consider different settings with both discrete and continuous outcomes as well as a dynamic model with lagged dependent variables as regressors. After providing sufficient conditions for identification, I present a nonparametric maximum likelihood estimator which allows estimating the structural functions, the factors, and the conditional distribution of (λ_i, U_i) consistently. The estimator requires estimating functions which might be high dimensional in applications, such as the conditional density of λ_i. Therefore, in addition to a fully nonparametric estimator, this paper provides a flexible semiparametric estimator. In this setup, the dimensions of the conditional densities are reduced by modeling them as a location and scale model. The structural functions can be nonparametric, semiparametric, or parametric. In the latter two cases, many parameters of interest are finite dimensional. I show that these estimated finite dimensional parameters are $\sqrt{n}$ consistent and asymptotically normally distributed. An easy to implement fully parametric estimator is also described. Finally, I discuss how the null hypothesis that the model has R factors can be tested against the alternative that the model has more than R factors, and how this result can be used to estimate the number of factors consistently. I show in Monte Carlo experiments that the semiparametric estimator performs well in finite samples.

I use the semiparametric estimator to investigate the relationship between teaching practice and student achievement. The outcome variables Y_{it} are different mathematics and science test scores for each student i. The main regressors are a measure of traditional teaching practice and a measure of modern teaching practice for each class a student attends. These measures are constructed using students' answers to questions about class activities. A linear fixed effects estimator is commonly used in related applications. This model controls for a scalar unobserved ability term, which has the same effect on all outcomes. Furthermore, the impact of the teaching practice on student achievement is, by assumption, the same for all levels of unobserved ability. With this model, I find a positive relationship between traditional teaching practice and test outcomes in both mathematics and science. I then estimate a nonlinear two factor model and obtain substantially different results. I still find a positive relationship between traditional teaching practice and mathematics test scores, but a positive relationship between modern teaching practice and science test scores. Furthermore, the structural functions are significantly nonlinear. In particular, the magnitude of the relationship between teaching practice and test outcomes is higher for students with low abilities than for students with high abilities.

Appendices

A Proofs and examples static continuous factor model

A.1 Proof of Lemma 1

First condition on $X_i \in \mathcal{X}$ and suppress X_i. Now notice that Z_{i1} and Z_{iK} are independent conditional on λ_i. Therefore

$$\begin{aligned} P\left(Z_{i1} < s_1, Z_{iK} < s_K\right) &= E_{\lambda_i}\left[P\left(Z_{i1} < s_1 \mid \lambda_i\right) P\left(Z_{iK} < s_K \mid \lambda_i\right)\right] \\ &= \int P\left(Z_{i1} < s_1 \mid \lambda_i = \lambda\right) P\left(Z_{iK} < s_K \mid \lambda_i = \lambda\right) f_{\lambda_i}(\lambda) d\lambda. \end{aligned}$$

Moreover,

$$f_{Z_{i1},Z_{iK}}(s_1, s_K) = \int f_{Z_{i1}|\lambda_i}(s_1;\lambda) f_{Z_{iK}|\lambda_i}(s_K;\lambda) f_{\lambda_i}(\lambda) d\lambda$$

It follows that for any bounded function m such that $E[|m(Z_{i1})|] < \infty$

$$\int f_{Z_{i1},Z_{iK}}(s_1, s_K) m(s_1) ds_1 = \int f_{Z_{iK},\lambda_i}(s_K, \lambda) \left(\int f_{Z_{i1}|\lambda_i}(s_1;\lambda) m(s_1) ds_1\right) d\lambda.$$

From Theorem 2.1 in Mattner (1993) and Theorem 2.1 in D'Haultfoeuille (2011) it now follows that Z_{i1} is bounded complete for λ_i. Furthermore, Proposition 2.4 in D'Haultfoeuille (2011) implies that λ_i is (bounded) complete for Z_{iK} and that λ_i is (bounded) complete for Z_{i1}. Hence, by the previous equality, Z_{i1} is bounded complete for Z_{iK}. □

A.2 Proof of Theorem 1

First condition on $X_i \in \mathcal{X}$ such that $X_{it} = \bar{x}_t$ for all $t = R+2, \ldots, 2R+1$. To simplify the notation, I suppress X_{it} in the function h_t in the proof. Define

$$\mathcal{L}^R = \left\{m : \mathbb{R}^R \to \mathbb{R} : \int_{\mathbb{R}^R} |m(v)| dv < \infty\right\}$$

and

$$\mathcal{L}^R_{bnd} = \left\{m \in \mathcal{L}^R : \sup_{v \in \mathbb{R}^R} |m(v)| < \infty\right\}.$$

Furthermore define

$$\begin{aligned} \mathcal{L}^R(\mathcal{Z}_1) &\equiv \left\{m : \mathbb{R}^R \to \mathbb{R} : \int_{\mathbb{R}^R} |m(v)| f_{Z_{i1}}(v) dv < \infty\right\} \\ \mathcal{L}^R_{bnd}(\mathcal{Z}_1) &\equiv \left\{m \in \mathcal{L}^R(\mathcal{Z}_1) : \sup_{v \in \mathbb{R}^R} |m(v)| < \infty\right\} \end{aligned}$$

as well as $\mathcal{L}^R(\mathcal{Z}_K)$, $\mathcal{L}^R_{bnd}(\mathcal{Z}_K)$, $\mathcal{L}^R(\Lambda)$, and $\mathcal{L}^R_{bnd}(\Lambda)$ analogously. Now let $s_{k_{R+1}} \in \mathbb{R}$ be a fixed constant and define the operators

$$\begin{aligned}
&L_{1,k_{R+1},K} : \mathcal{L}^R_{bnd}(\mathcal{Z}_K) \to \mathcal{L}^R_{bnd} \\
&\qquad\qquad \left(L_{1,k_{R+1},K}m\right)(s_1, s_{k_{R+1}}) \equiv \int f_{Z_{i1},Z_{iK},Z_{ik_{R+1}}}(s_1, s_K, s_{k_{R+1}})m(s_K)ds_K \\
&L_{1,K} : \mathcal{L}^R_{bnd}(\mathcal{Z}_K) \to \mathcal{L}^R_{bnd} \\
&\qquad\qquad \left(L_{1,K}m\right)(s_1) \equiv \int f_{Z_{i1},Z_{iK}}(s_1, s_K)m(s_K)ds_K \\
&L_{1,\lambda} : \mathcal{L}^R_{bnd} \to \mathcal{L}^R_{bnd} \\
&\qquad\qquad \left(L_{1,\lambda}m\right)(s_1) \equiv \int f_{Z_{i1}|\lambda_i}(s_1;\lambda)m(\lambda)d\lambda \\
&L_{\lambda,K} : \mathcal{L}^R_{bnd}(\mathcal{Z}_K) \to \mathcal{L}^R_{bnd} \\
&\qquad\qquad \left(L_{\lambda,K}m\right)(\lambda) \equiv \int f_{Z_{iK},\lambda_i}(s_K, \lambda)m(s_K)ds_K \\
&D_{k_{R+1},\lambda} : \mathcal{L}^R_{bnd}(\Lambda) \to \mathcal{L}^R_{bnd}(\Lambda) \\
&\qquad\qquad \left(D_{k_{R+1},\lambda}m\right)(s_{k_{R+1}}, \lambda) \equiv f_{Z_{ik_{R+1}}|\lambda_i}(s_{k_{R+1}};\lambda)m(\lambda).
\end{aligned}$$

The operator $L_{1,k_{R+1},K}$ should be seen as a mapping from $\mathcal{L}^R_{bnd}(\mathcal{Z}_K)$ to $\mathcal{L}^R_{bnd}$ for a fixed value $s_{k_{R+1}}$. Changing the value of $s_{k_{R+1}}$ gives a different mapping. With these definitions for any $m \in \mathcal{L}^R_{bnd}(\mathcal{Z}_K)$

$$\begin{aligned}
&\left(L_{1,K,k_{R+1}}m\right)(s_1, s_{k_{R+1}}) \\
&\qquad = \int f_{Z_{i1},Z_{iK},Z_{ik_{R+1}}}(s_1, s_K, s_{k_{R+1}})m(s_K)ds_K \\
&\qquad = \int \left(\int f_{Z_{i1}|\lambda_i}(s_1;\lambda)f_{Z_{iK}|\lambda_i}(s_K;\lambda)f_{Z_{ik_{R+1}}|\lambda_i}(s_{k_{R+1}};\lambda)f_\lambda(\lambda)d\lambda\right)m(s_K)ds_K \\
&\qquad = \int f_{Z_{i1}|\lambda_i}(s_1;\lambda)f_{Z_{ik_{R+1}}|\lambda_i}(s_{k_{R+1}};\lambda)\left(\int f_{Z_{iK},\lambda_i}(s_K,\lambda)m(s_K)ds_K\right)d\lambda \\
&\qquad = \int f_{Z_{i1}|\lambda_i}(s_1;\lambda)f_{Z_{ik_{R+1}}|\lambda_i}(s_{k_{R+1}};\lambda)\left(L_{\lambda,K}m\right)(\lambda)d\lambda \\
&\qquad = \int f_{Z_{i1}|\lambda_i}(s_1;\lambda)\left(D_{k_{R+1},\lambda}L_{\lambda,K}m\right)(s_{k_{R+1}},\lambda)d\lambda \\
&\qquad = \left(L_{1,\lambda}D_{k_{R+1},\lambda}L_{\lambda,K}m\right)(s_1, s_{k_{R+1}}).
\end{aligned}$$

Similarly,

$$\left(L_{1,K}m\right)(s_1) \quad = \quad \left(L_{1,\lambda}L_{\lambda,K}m\right)(s_1).$$

Since these equalities hold for all functions $m \in \mathcal{L}^R_{bnd}(\mathcal{Z}_K)$ it follows that

$$L_{1,k_{R+1},K} = L_{1,\lambda}D_{k_{R+1},\lambda}L_{\lambda,K}$$

and

$$L_{1,K} = L_{1,\lambda} L_{\lambda,K}.$$

By Assumption S8, $L_{1,\lambda}$ is invertible and the inverse can be applied from the left. Therefore

$$L_{1,\lambda}^{-1} L_{1,K} = L_{\lambda,K},$$

which implies that

$$L_{1,k_{R+1},K} = L_{1,\lambda} D_{k_{R+1},\lambda} L_{1,\lambda}^{-1} L_{1,K}.$$

Lemma 1 of Hu and Schennach (2008) and Assumption S8 imply that $L_{1,K}$ has a right inverse which is densely defined on $\mathcal{L}_{bnd}^{R}$. Therefore,

$$L_{1,k_{R+1},K} L_{1,K}^{-1} = L_{1,\lambda} D_{k_{R+1},\lambda} L_{1,\lambda}^{-1}.$$

The operator on the left hand side depends on the population distribution of the observables only. Hence, it can be considered known. Hu and Schennach (2008) deal with the same type of operator equality in a measurement error setup. They show that the operator on the left hand side is bounded and its domain can therefore be extended to $\mathcal{L}_{bnd}^{R}$. They also show that the right hand side is an eigenvalue-eigenfunction decomposition of the known operator $L_{1,k_{R+1},K} L_{1,K}^{-1}$. The eigenfunctions are $f_{Z_{i1}|\lambda_i}(s_1;\lambda)$ with corresponding eigenvalues $f_{Z_{ik_{R+1}}|\lambda_i}(s_{k_{R+1}};\lambda)$. Each λ indexes an eigenfunction and an eigenvalue. The eigenfunctions are functions of s_1 and $s_{k_{R+1}}$ is a fixed constant. Hu and Schennach (2008) show that this decomposition is unique up to three nonunique features:

1. Scaling: Multiplying each eigenfunction by a constant yields a different eigenvalue-eigenfunction decomposition belonging to the same operator $L_{1,k_{R+1},K} L_{1,K}^{-1}$.

2. Eigenvalue degeneracy: If two or more eigenfunctions share the same eigenvalue, any linear combination of these eigenfunctions are also eigenfunctions. Then several different eigenvalue-eigenfunction decompositions belong to the same operator $L_{1,k_{R+1},K} L_{1,K}^{-1}$.

3. Ordering: For $\tilde{\lambda}$ that satisfies $\tilde{\lambda} = B(\lambda)$ for any one to one transformation $B : \mathbb{R}^R \to \mathbb{R}^R$

$$L_{1,\lambda} D_{k_{R+1},\lambda} L_{1,\lambda}^{-1} = L_{\tilde{\lambda},1} D_{k_{R+1},\tilde{\lambda}} L_{\tilde{\lambda},1}^{-1}.$$

These conditions are very similar to conditions for nonuniqueness of an eigendecomposition of a square matrix. While for matrices the order of the columns of the matrix that contains the eigenvectors is not fixed, with operators any one to one transformation of λ leads to an eigendecomposition. I show next that, given the assumptions of the theorem, the scaling and the ordering are fixed by assumption. Furthermore, all eigenvalues are unique. It then follows that there are unique operators $L_{1,\lambda}$ and $D_{k_{R+1},\lambda}$ such that $L_{1,k_{R+1},K} L_{1,K}^{-1} = L_{1,\lambda} D_{k_{R+1},\lambda} L_{1,\lambda}^{-1}$.

In now verify that these three nonunique features cannot occur in my model. First, the *scaling* is unambiguous because all eigenfunctions have to integrate to one. Second, I show that *eigenvalue degeneracy* cannot occur, by showing that linear combinations of the eigenfunctions $f_{Z_{i1}\lambda_i}(s_1;\bar{\lambda}_1)$ and $f_{Z_{i1}|\lambda_i}(s_1;\bar{\lambda}_2)$ cannot be eigenfunctions themselves when varying K and the value of $s_{k_{R+1}}$. To see this notice that the eigenfunctions neither depend on K nor on $s_{k_{R+1}}$. Now take all functions which are eigenfunctions of $L_{1,k_{R+1},K}L_{1,K}^{-1}$ for all K and all $s_{k_{R+1}}$. Then there can only be an eigenvalue degeneracy problem if two eigenfunctions share the same eigenvalue for all K and all $s_{k_{R+1}}$. But this means for all $k_{R+1} = 1, \dots, R+1$ and all $s_{k_{R+1}} \in \mathbb{R}$

$$f_{Z_{ik_{R+1}}|\lambda_i}(s_{k_{R+1}};\bar{\lambda}_1) = f_{Z_{ik_{R+1}}|\lambda_i}(s_{k_{R+1}};\bar{\lambda}_2)$$

and hence for any $t = 1, \dots, R+1$

$$M\left[Y_{it} \mid \lambda_i = \bar{\lambda}_1\right] = M\left[Y_{it} \mid \lambda_i = \bar{\lambda}_2\right] \qquad \text{or} \qquad g_t\left(\bar{\lambda}_1' F_t\right) = g_t\left(\bar{\lambda}_2' F_t\right).$$

But since F^1 has full rank it has to hold that $\bar{\lambda}_1' F_t \neq \bar{\lambda}_2' F_t$ for some t which is a contradiction since g_t is strictly increasing.

Third, I show that there is a unique *ordering* of eigenfunctions which coincides with $L_{1,\lambda}$. Let $\tilde{\lambda} = B(\lambda)$. Both $\tilde{\lambda}$ and λ have to be consistent with model (1). In particular, for $\tilde{\lambda}$ there has to exist a strictly functions $\tilde{h}_t$ as well as $\tilde{F}_t$ and $\tilde{U}_{it}$ such that for all $r = 1, \dots, R$

$$\tilde{h}_{T-r+1}\left(Y_{i(T-r+1)}\right) = \tilde{\lambda}_r + \tilde{U}_{i(T-r+1)}$$

and

$$\tilde{h}_{T-r+1}\left(M(Y_{i(T-r+1)} \mid \tilde{\lambda}_i = \tilde{\lambda})\right) = \tilde{\lambda}_r.$$

Since

$$M(Y_{i(T-r+1)} \mid \lambda_i = \lambda) = M(Y_{i(T-r+1)} \mid B(\lambda_i) = B(\lambda))$$

it follows that

$$g_{T-r+1}\left(\lambda_r\right) = \tilde{g}_{T-r+1}\left(B_r(\lambda)\right).^{21}$$

Hence

$$B_r(\lambda) = \tilde{h}_{T-r+1}\left(g_{T-r+1}\left(\lambda_r\right)\right)$$

which implies that $B_r(\cdot)$ is differentiable and $B_r(\lambda)$ only depends on λ_r. Similarly for all $t < R+2$

$$g_t\left(\sum_{r=1}^{R} F_{tr}\lambda_r\right) = \tilde{g}_t\left(\sum_{r=1}^{R} \tilde{F}_{tr} B_r(\lambda_r)\right)$$

[21] As before $\tilde{g}$ is the inverse function of $\tilde{h}$.

and hence

$$\frac{\partial g_t\left(\sum_{r=1}^{R} F_{tr}\lambda_r\right)}{\partial \lambda_r} = \frac{\partial \tilde{g}_t\left(\sum_{r=1}^{R} \tilde{F}_{tr}B_r(\lambda_r)\right)}{\partial \lambda_r}.$$

Now assume that $R > 1$. Then since all functions are strictly monotonic and differentiable, it follows that

$$\frac{F_{tr}}{F_{ts}} = \frac{\tilde{F}_{tr}B_r'(\lambda_r)}{\tilde{F}_{ts}B_r'(\lambda_s)},$$

which implies that $B_r'(\lambda_r) = C_r$ for all r.[22] If $R = 1$, then in a neighborhood of $\alpha = 0.5$

$$g_T\left(\lambda + Q_\alpha(U_{iT})\right) = \tilde{g}_T\left(B(\lambda) + Q_\alpha(\tilde{U}_{iT})\right)$$

or

$$B\left(\lambda + Q_\alpha(U_{iT})\right) = B(\lambda) + Q_\alpha(\tilde{U}_{iT}).$$

Differentiating with respect to α and λ yields $B'(\lambda) = C$. Thus in all cases only linear transformations of λ_r can lead to consistent observationally equivalent models. Then for $r = 1, \ldots, R$

$$g_{T-r+1}\left(\lambda_r\right) = \tilde{g}_{T-r+1}\left(C_r\lambda_r + d_r\right).$$

The previous line can be rewritten to

$$\tilde{h}_{T-r+1}\left(y_r\right) = C_r h_{T-r+1}\left(y_r\right) + d_r.$$

where $y_r \equiv g_{T-r+1}\left(\lambda_r\right)$. But since at $\bar{y}_r$ (recall that $X_{it} = \bar{x}_t$ for $t = R+2, \ldots, 2R+1$)

$$\tilde{h}_{T-r+1}'\left(\bar{y}_r\right) = h_{T-r+1}'\left(\bar{y}_r\right) = 1$$

it has to hold that $C_r = 1$. Finally,

$$\tilde{h}_{T-r+1}\left(\bar{y}_r\right) = h_{T-r+1}\left(\bar{y}_r\right) = 0$$

which implies that $d_r = 0$. Therefore $B(\lambda) = \lambda$.

Since none of the three nonunique features can occur due to the assumptions and the structure of the model, $L_{1,\lambda}$ and $D_{k_{R+1},\lambda}$ are identified. By the relation

$$L_{1,\lambda}^{-1}L_{1,K} = L_{\lambda,K}$$

it also holds that $L_{\lambda,K}$ is identified. The operator being identified is the same as the kernel being identified. Hence $f_{Y_i,\lambda_i}(s,\lambda)$ is identified for all $s \in \mathbb{R}^T$ and $\lambda \in \Lambda$.

In the last step, I use one of the additional assumptions to show that g_t is identified, which then

[22] Here I use that $F_{tr} \neq 0$ for some t.

implies that $f_{U_i,\lambda_i}(u,\lambda)$ is identified. If λ_r has support on $\mathbb{R}$ for all $r = 1,\ldots,R$, then since

$$M\left[Y_{i(T-r+1)} \mid \lambda_i = \lambda\right] = g_{T-r+1}\left(\lambda_r\right)$$

and since f_{Y_i,λ_i} is identified, g_{T-r+1} is identified for all $r = 1,\ldots,R$. Then also for all $t < R+2$

$$M\left[Y_{it} \mid \lambda_i = \lambda\right] = g_t\left(\lambda' F_t\right).$$

If $R = 1$, then g_t is identified up to scale, which is fixed by Assumption 3. If $R > 1$, taking ratios of derivatives with respect to different elements of λ identifies $\frac{F_{tr}}{F_{ts}}$ for all $r,s = 1,\ldots,R$. Hence, again g_t is identified up to scale which is fixed. Therefore, g_t and F_t are identified. Finally if $U_i \perp\!\!\!\perp \lambda_i$, then for all $r = 1,\ldots,R$

$$P\left(Y_{T-r+1} < s \mid \lambda_i = \lambda\right) = F_{U_{i(T-r+1)}}\left(h_{T-r+1}(s) - \lambda_r\right).$$

It follows that

$$\frac{\frac{\partial P(Y_{T-r+1}<s|\lambda_i=\lambda)}{\partial s}}{\frac{\partial P(Y_{T-r+1}<s|\lambda_i=\lambda)}{\partial \lambda_r}} = h'_{T-r+1}(s).$$

Since $h_{T-r+1}\left(M\left[Y_{i(T-r+1)} \mid \lambda_i = \lambda\right]\right) = \lambda_r$, the location is fixed and $h_{T-r+1}(s)$ is identified for all $s \in \mathcal{Y}_{T-r+1}$. Similarly, h_t and F_t are identified for all $t = 1,\ldots,R+1$ using in addition the scale normalization in Assumption S3.

If neither λ_i has support on $\mathbb{R}^R$ nor $U_i \perp\!\!\!\perp \lambda_i$, take λ^* such that $\bar{y}_t = M\left[Y_{it} \mid \lambda_i = \lambda^*\right]$. Then, since $h_t(M\left[Y_{it} \mid \lambda_i = \lambda\right]) = \lambda' F_t$, we can differentiate with respect to λ, evaluate at $\lambda = \lambda^*$, use the scale and location normalization, and identify F_t. Then also $h_t(M\left[Y_{it} \mid \lambda_i = \lambda\right])$ is identified for all y_t such that $y_t = M\left[Y_{it} \mid \lambda_i = \lambda\right]$ for some λ. □

A.3 Proof of Theorem 2

Let $\tilde{R} \leq \frac{T}{2}$ be a positive integer. Define

$$\tilde{Z}_{i1} \equiv \left(Y_{i(T-\tilde{R}+1)},\ldots,Y_{iT}\right) \quad \text{and} \quad \tilde{Z}_{i2} \equiv \left(Y_{i1},\ldots,Y_{i\tilde{R}}\right).$$

Let $s_1, s_2 \in \mathbb{R}^{\tilde{R}}$. If $\tilde{R} > R$ there exits $\gamma \in \mathbb{R}^{\tilde{R}}$ such that

$$\begin{pmatrix} F_{T-\tilde{R}+1} & \ldots & F_T \end{pmatrix}\gamma = 0.$$

For this vector γ it holds that

$$\begin{pmatrix} h_{T-\tilde{R}+1}\left(Y_{i(T-\tilde{R}+1)}\right) & \ldots & h_T\left(Y_{iT}\right)\end{pmatrix}'\gamma = \begin{pmatrix} U_{i(T-\tilde{R}+1)} & \ldots & U_{iT}\end{pmatrix}'\gamma.$$

Define

$$\tilde{h}\left(\tilde{Z}_{i1}\right) \equiv \left(h_{T-\tilde{R}+1}\left(Y_{i(T-\tilde{R}+1)}\right) \quad \ldots \quad h_T\left(Y_{iT}\right)\right)'\gamma$$

and

$$\bar{h}\left(\tilde{Z}_{i1}\right) = \tilde{h}\left(\tilde{Z}_{i1}\right)\mathbf{1}\left(|\tilde{h}\left(\tilde{Z}_{i1}\right)| \leq 1\right) - E\left[\tilde{h}\left(\tilde{Z}_{i1}\right)\mathbf{1}\left(|\tilde{h}\left(\tilde{Z}_{i1}\right)| \leq 1\right)\right].$$

Then under the additional assumption that $\lambda_i \perp\!\!\!\perp U_i$, it follows that

$$E\left[\bar{h}\left(\tilde{Z}_{i1}\right) \mid \lambda_i\right] = 0.$$

As a consequence $\tilde{Z}_{i1}$ is not bounded complete for λ_i. Now since

$$\int f_{\tilde{Z}_{i1},\tilde{Z}_{i2}}(s_1,s_2)\bar{h}(s_1)ds_1 = \int f_{Z_{i2},\lambda_i}(s_2,\lambda)\left(\int f_{Z_{i1}|\lambda_i}(s_1;\lambda)\bar{h}(s_1)ds_1\right) = 0,$$

it follows that $\tilde{Z}_{i1}$ is bounded complete for $\tilde{Z}_{i2}$ only if $\tilde{R} \leq R$. □

A.4 Functionals invariant to normalizations

In this section I show that quantiles of unobservables as well as average function values do not depend on the normalizations. To see that these objects are invariant to the normalizations in Assumptions S3, S4, and S6 first notice that C_{it} is invariant to the normalization $F^3 = I_{R\times R}$. Now recall that the other normalizations (Assumptions S3 and S4) are needed because for any $a_t, c_t \in \mathbb{R}$, $b \in \mathbb{R}^R$ and $d_t > 0$ such that $a_t = b'F_t + c_t$ we can write for all t

$$\frac{h_t\left(Y_{it}, X_{it}\right) + a_t}{d_t} = \frac{\left(\lambda_i' + b'\right)F_t}{d_t} + \frac{U_{it} + c_t}{d_t} \Leftrightarrow \tilde{h}_t\left(Y_{it}, X_{it}\right) = \tilde{\lambda}_i'\tilde{F}_t + \tilde{U}_{it}$$

where

$$\tilde{h}_t\left(Y_{it}, X_{it}\right) = \frac{h_t\left(Y_{it}, X_{it}\right) + a_t}{d_t} \quad \text{and} \quad \tilde{U}_{it} = \frac{U_{it} + c_t}{d_t}.$$

Furthermore, for $r = 1, \ldots, R$

$$\tilde{\lambda}_{ir} = \frac{\lambda_{ir} + b_r}{d_{T-r+1}} \quad \text{and} \quad \tilde{F}_{T-r+1} = F_{T-r+1}$$

and for $t = 1, \ldots, R+1$ and $r = 1, \ldots, R$

$$\tilde{F}_{tr} \quad = \quad F_{tr}\frac{d_{T-r+1}}{d_t}.$$

Hence $\tilde{F}^3 = I_{R\times R}$ is satisfied. It then follows that

$$Q_\alpha\left[\tilde{U}_{it} \mid X_i\right] = \frac{Q_\alpha\left[U_{it} \mid X_i\right] + c_t}{d_t} \quad \text{and} \quad Q_\alpha\left[\tilde{C}_{it} \mid X_i\right] = \frac{Q_\alpha\left[C_{it} \mid X_i\right] + b'F_t}{d_t}.$$

As a consequence for $\tilde{x}_t \in \mathcal{X}_t$ and $x \in \mathcal{X}$ it holds that

$$\begin{aligned}
&\tilde{g}_t\left(\tilde{x}_t, Q_{\alpha_1}\left[\tilde{C}_{it} \mid X_i = x\right] + Q_{\alpha_2}\left[\tilde{U}_{it} \mid X_i = x\right]\right) \\
&\qquad = g_t\left(\tilde{x}_t, \left(Q_{\alpha_1}\left[\tilde{C}_{it} \mid X_i = x\right] + Q_{\alpha_2}\left[\tilde{U}_{it} \mid X_i = x\right]\right) d_t - a_t\right) \\
&\qquad = g_t\left(\tilde{x}_t, \left(\frac{Q_{\alpha_1}\left[C_{it} \mid X_i\right] + b' F_t}{d_t} + \frac{Q_{\alpha_2}\left[U_{it} \mid X_i\right] + c_t}{d_t}\right) d_t - a_t\right) \\
&\qquad = g_t\left(\tilde{x}_t, Q_{\alpha_1}\left[C_{it} \mid X_i\right] + Q_{\alpha_2}\left[U_{it} \mid X_i\right] + b' F_t + c_t - a_t\right) \\
&\qquad = g_t\left(\tilde{x}_t, Q_{\alpha_1}\left[C_{it} \mid X_i\right] + Q_{\alpha_2}\left[U_{it} \mid X_i\right]\right).
\end{aligned}$$

Similarly, since

$$P\left(\tilde{C}_{it} + \tilde{U}_{it} < e \mid X_i = x\right) = P\left(C_{it} + U_{it} < e d_t - b' F_t - c_t \mid X_i = x\right)$$

it follows that

$$\begin{aligned}
\int \tilde{g}_t\left(\tilde{x}_t, e\right) dF_{\tilde{C}_{it}+\tilde{U}_{it} \mid X_i = x}(e) &= \int g_t\left(\tilde{x}_t, e\right) dF_{C_{it}+U_{it} \mid X_i = x}\left(e d_t - b' F_t - c_t\right) \\
&= \int \tilde{g}_t\left(\tilde{x}_t, \frac{e + b' F_t + c_t}{d_t}\right) dF_{C_{it}+U_{it} \mid X_i = x}(e) \\
&= \int g_t\left(\tilde{x}_t, \left(\frac{e + b' F_t + c_t}{d_t}\right) d_t - a_t\right) dF_{C_{it}+U_{it} \mid X_i = x}(e) \\
&= \int g_t\left(\tilde{x}_t, e + b' F_t + c_t - a_t\right) dF_{C_{it}+U_{it} \mid X_i = x}(e) \\
&= \int g_t\left(\tilde{x}_t, e\right) dF_{C_{it}+U_{it} \mid X_i = x}(e).
\end{aligned}$$

Identical arguments yield

$$\tilde{g}_t\left(\tilde{x}_t, Q_{\alpha}\left[\tilde{C}_{it} + \tilde{U}_{it} \mid X_i = x\right]\right) = g_t\left(\tilde{x}_t, Q_{\alpha}\left[C_{it} + U_{it} \mid X_i\right]\right) \quad \text{and}$$

$$\int \tilde{g}_t\left(\tilde{x}_2, e + Q_{\alpha}\left[\tilde{U}_{it} \mid X_i\right]\right) dF_{\tilde{C}_{it} \mid X_i = x}(e) = \int g_t\left(\tilde{x}_2, e + Q_{\alpha}\left[U_{it} \mid X_i\right]\right) dF_{C_{it} \mid X_i = x}(e).$$

B Proofs and examples dynamic continuous factor model

B.1 Proof of Theorem 3

Notice that by Assumption L5 it holds that for all $t \geq 2$,

$$f_{Y_{iT},\ldots,Y_{it} \mid \lambda_i, Y_{i(t-1)},\ldots,Y_{i1}} = f_{Y_{iT},\ldots,Y_{it} \mid \lambda_i, Y_{i(t-1)}}.$$

First assume that $R \geq 2$. Now recall that Z_{iK} is a vector outcomes containing R element of $Y_{i(3R+3)}, \ldots, Y_{i(2R+3)}$ and $Z_{i1} = (Y_{i1}, \ldots, Y_{iR})$. Furthermore, define $Z_{i2} = Y_{i(R+2)}, \ldots, Y_{i(2R+1)}$.

Let $s_1, s_2, s_K \in \mathbb{R}^R$ and $s_{R+1}, s_{2R+2} \in \mathbb{R}$. Then,

$$\begin{aligned}
&f_{Z_{i1},Y_{i(R+1)},Z_{i2},Y_{i(2R+2)},Z_{iK}}(s_1, s_{R+1}, s_2, s_{2R+2}, s_K) \\
&\qquad = \int f_{Z_{i1},Y_{i(R+1)},Z_{i2},Y_{i(2R+2)},Z_{iK}|\lambda_i}(s_1, s_{R+1}, s_2, s_{2R+2}, s_K; \lambda) f_{\lambda_i}(\lambda) d\lambda \\
&\qquad = \int f_{Z_{iK}|Y_{i(2R+2)},\lambda_i}(s_K; s_{2R+2}, \lambda) f_{Y_{i(2R+2)},Z_{i2}|Y_{i(R+1)},\lambda_i}(s_{2R+2}, s_2; s_{R+1}, \lambda) \\
&\qquad\quad \times f_{Y_{i(R+1)},Z_{i1},\lambda_i}(s_{R+1}, s_1, \lambda) d\lambda \\
&\qquad = \int f_{Z_{iK}|Y_{i(2R+2)},\lambda_i}(s_K; s_{2R+2}, \lambda) f_{Z_{i2}|Y_{i(2R+2)},Y_{i(R+1)},\lambda_i}(s_2; s_{2R+2}, s_{R+1}, \lambda) \\
&\qquad\quad \times f_{Y_{i(2R+2)}|Y_{i(R+1)},\lambda_i}(s_{2R+2}; s_{R+1}, \lambda) f_{Y_{i(R+1)},Z_{i1},\lambda_i}(s_{R+1}, s_1, \lambda) d\lambda.
\end{aligned}$$

Integrating over s_2 yields in addition

$$\begin{aligned}
&f_{Z_{i1},Y_{i(R+1)},Y_{i(2R+2)},Z_{iK}}(s_1, s_{R+1}, s_{2R+2}, s_K) \\
&\qquad = \int f_{Z_{iK}|Y_{i(2R+2)},\lambda_i}(s_K; s_{2R+2}, \lambda) \\
&\qquad\quad \times f_{Y_{i(2R+2)}|Y_{i(R+1)},\lambda_i}(s_{2R+2}; s_{R+1}, \lambda) f_{Y_{i(R+1)},Z_{i1},\lambda_i}(s_{R+1}, s_1, \lambda) d\lambda.
\end{aligned}$$

Now for any fixed $s_{2R+2}, \ldots, s_{R+1}$ define, just as in the static case, the integral operators

$$\begin{aligned}
(L_{K,2,1}m)(s_K, s_{2R+2}, s_2, s_{R+1}) &\equiv \int f_{Z_{iK},Y_{i(2R+2)},Z_{i2},Y_{i(R+1)},Z_{i1}}(s_K, s_{2R+2}, s_2, s_{R+1}, s_1) m(s_1) ds_1 \\
(L_{K,1}m)(s_K, s_{2R+2}, s_{R+1}) &\equiv \int f_{Z_{iK},Y_{i(2R+2)},Y_{i(R+1)},Z_{i1}}(s_K, s_{2R+2}, s_{R+1}, s_1) m(s_1) ds_1 \\
(L_{K,\lambda}m)(s_K, s_{2R+2}) &\equiv \int f_{Z_{iK}|Y_{i(2R+2)},\lambda_i}(s_K; s_{2R+2}, \lambda) m(\lambda) d\lambda \\
(L_{\lambda,1}m)(s_{R+1}, \lambda) &\equiv \int f_{Y_{i(R+1)},Z_{i1},\lambda_i}(s_{R+1}, s_1, \lambda) m(s_1) ds_1 \\
(D_{2,2R+2,R+1}m)(s_2; s_{2R+2}, s_{R+1}, \lambda) &\equiv f_{Z_{i2}|Y_{i(2R+2)},Y_{i(R+1)},\lambda_i}(s_2; s_{2R+2}, s_{R+1}, \lambda) m(\lambda) \\
(D_{2R+2,R+1}m)(\lambda)(s_{2R+2}; s_{R+1}, \lambda) &\equiv f_{Y_{i(2R+2)}|Y_{i(R+1)},\lambda_i}(s_{2R+2}; s_{R+1}, \lambda) m(\lambda).
\end{aligned}$$

The operators are defined on similar function spaces as in the static case. These definitions yield, similar to the static case, the operator equalities

$$L_{K,2,1} = L_{K,\lambda} D_{2,2R+2,R+1} D_{2R+2,R+1} L_{\lambda,1}$$

and

$$L_{K,1} = L_{K,\lambda} D_{2R+2,R+1} L_{\lambda,1}.$$

By Assumption L6, the operator $L_{K,\lambda}$ has a left inverse for any s_{2R+2} and $L_{K,1}$ has a right inverse for any s_{2R+2} and s_{R+1}. Therefore, similar as before,

$$L_{K,2,1}L_{K,1}^{-1} = L_{K,\lambda}D_{2,2R+2,R+1}L_{K,\lambda}^{-1}.$$

This is an eigendecomposition just as in the static case with bounded eigenvalues due to Assumption L2. The eigenfunctions are $f_{Z_{iK}|Y_{i(2R+2)},\lambda_i}(s_K; s_{2R+2}, \lambda)$. These are functions of s_K and each λ indexes a different eigenfunction. Remember that s_{2R+2} is fixed. The corresponding eigenvalues are $f_{Z_{i2}|Y_{i(2R+2)},Y_{i(R+1)},\lambda_i}(s_2; s_{2R+2}, s_{R+1}, \lambda)$ where s_2, s_{2R+2}, and s_{R+1} are fixed. Again, just as in the static case, the eigenfunctions and eigenvalues are identified up to three nonunique features.

I now show that the three nonunique features cannot occur due to the factor structure. First, *scaling* is not ambiguous in this case because the eigenfunction are functions of s_K and integrate to 1. Second, notice that only the eigenvalues depend on s_2 and s_{R+1}. By Assumption L7 the eigenvalues are unique when considering the set of functions which are eigenfunctions for all s_2 and s_{R+1} and for a given s_{2R+2}. Hence, *eigenvalue degeneracy* cannot occur. Therefore, only the *ordering* ambiguity needs to be solved. Notice that the eigenvalues are the same for any vector Z_{iK}. The important difference to the static case is that here both the eigenfunctions and the eigenvalues depend on s_{2R+2}. Therefore, the ordering could depend on the value of s_{2R+2}. In the static case, the ordering does depend on the value of X_i, but the object of interest is the distribution of $\lambda_i \mid X_i$. In the dynamic case, one is not primarily interested in the distribution of $\lambda_i \mid Y_{i(2R+2)}$. Hence, I need to show that the ordering cannot depend on s_{2R+2}. As in the static case let $B(\cdot, s_{2R+2}) : \mathbb{R}^R \to \mathbb{R}^R$ be a one to one transformation for each s_{2R+2} and let $\tilde{\lambda}_i = B(\lambda_i, s_{2R+2})$ be a different ordering of the eigenvalues. Any such ordering yields eigenfunctions

$$f_{Z_{iK}|Y_{i(2R+2)},\tilde{\lambda}_i}(s_K; s_{2R+2}, B(\lambda, s_{2R+2}))$$

with the true ordering being

$$f_{Z_{iK}|Y_{i(2R+2)},\lambda_i}(s_K; s_{2R+2}, \lambda).$$

From the densities above, the density of $Y_{i(T-r+1)} \mid (Y_{i(T-r)}, Y_{i(2R+2)}, \tilde{\lambda}_i)$ is known for all $r = 1, \ldots, R$ up to the ordering ambiguity. But any such ordering needs to be consistent with the model. In particular, for all $r = 1, \ldots, R$

$$\begin{aligned} &M\left[Y_{T-r+1} \mid Y_{T-r} = s_{T-r}, Y_{2R+2} = s_{2R+2}, \lambda_i = \lambda\right] \\ &\qquad = M\left[Y_{T-r+1} \mid Y_{T-r} = s_{T-r}, Y_{2R+2} = s_{2R+2}, \tilde{\lambda}_i = B(\lambda, s_{2R+2})\right] \end{aligned}$$

and

$$g_{T-r+1}\left(s_{T-r}, \lambda_r\right) = \tilde{g}_{T-r+1}\left(s_{T-r}, B_r(\lambda, s_{2R+2})\right)$$

for some strictly increasing function $\tilde{g}_{T-r+1}$. Since the left hand side does not depend on s_{2R+2}, and since $\tilde{g}_{T-r+1}$ is strictly increasing, the ordering cannot depend on s_{2R+2}. Therefore for some $V : \mathbb{R}^R \to \mathbb{R}^R$, it holds that $B(\lambda, s_{2R+2}) = V(\lambda)$, where V_r only depends on λ_r, is differentiable, and strictly increasing. The remaining steps are now identical to the static case. In particular, it can be shown that $V(\lambda) = \lambda$ which leads to identification of f_{Y_i,λ_i}. It then follows that the remaining parameters are identified.

So far I assumed that $R \geq 2$. If $R = 1$, the eigenfunctions to not yield densities of $Y_{i(T-r+1)} \mid (Y_{i(T-r)}, Y_{i(2R+2)}, \lambda_i)$. Therefore, slightly different arguments are needed. For $Z_{iK} = Y_{i5}$ again

$$L_{K,2,1}L_{K,1}^{-1} \;=\; L_{K,\lambda}D_{2,2R+2,R+1}L_{K,\lambda}^{-1} \tag{11}$$

Moreover, for fixed any $s_5, \ldots, s_2 \in \mathbb{R}$ define

$$\begin{aligned}
\left(L_{6,2,1}m\right)(s_6,s_5,s_4,s_3,s_2) &\equiv \int f_{Y_{i6},Y_{i5},Y_{i4},Y_{i3},Y_{i2},Y_{i1}}(s_6,s_5,s_4,s_3,s_2,s_1)m(s_1)ds_1\\
\left(L_{6,1}m\right)(s_6,s_5,s_4,s_2) &\equiv \int f_{Y_{i6},Y_{i5},Y_{i4},Y_{i2},Y_{i1}}(s_6,s_5,s_4,s_2,s_1)m(s_1)ds_1\\
\left(L_{6,\lambda}m\right)(s_6,s_5,s_4) &\equiv \int f_{Y_{i6},Y_{i5}|Y_{i4},\lambda_i}(s_6,s_5;s_4,\lambda)m(\lambda)d\lambda.
\end{aligned}$$

Then in a similar way as before,

$$L_{6,2,1}L_{6,1}^{-1} = L_{6,\lambda}D_{2,2R+2,R+1}L_{6,\lambda}^{-1}. \tag{12}$$

Start with the eigendecomposition in (11). All eigenfunctions integrate to one and the eigenvalues are unique by Assumption L7 when considering the set of functions which are eigenfunctions for all s_2 and s_{R+1} and for a given s_{2R+2}. From this eigendecomposition we obtain

$$f_{Y_{i5}|Y_{i4},\tilde{\lambda}_i}(s_5; s_4, B(\lambda, s_4))$$

for an arbitrary one to one transformation $B(\lambda, s_4)$ as explained before. For this ordering of λ, we also obtain $f_{Y_{i6},Y_{i5}|Y_{i4},\tilde{\lambda}_i}(s_6, s_5; s_4, B(\lambda, s_4))$ up to scale from (12) because the eigendecompositions share the same eigenvalues. We only obtain the functions up to scale because they do not integrate to 1 since s_5 is fixed. However, the integral equals $f_{Y_{i5}|Y_{i4},\tilde{\lambda}_i}(s_5; s_4, B(\lambda, s_4))$ which is already known, so the scale is fixed. Now we can apply the previous arguments because from $f_{Y_{i6},Y_{i5}|Y_{i4},\tilde{\lambda}_i}(s_6, s_5; s_4, B(\lambda, s_4))$ we obtain the density of $Y_{i6} \mid (Y_{i5}, Y_{i4}, \tilde{\lambda}_i)$. Then we can show that the restrictions of the model imply that $B(\lambda, s_4)$ cannot depend s_4. Finally, we can use the arguments from the static case and prove identification of all parameters.

It is possible to identify the parameters when $T = 2R + \left\lceil \frac{R}{2} \right\rceil + 3$. This can be done by using very similar arguments as in the static case. In particular, when deriving the operator equalities, one of the diagonal operators is eliminated by integrating over s_2. This results in eigenfunctions being a function of s_2, namely $f_{Z_{i2}|Y_{i(2R+2)},Y_{i(R+1)},\lambda_i}(s_2; s_{2R+2}, s_{R+1}, \lambda)$. It is also possible to integrate out Y_{it} for different t in such a way that the resulting eigendecomposition has the same eigenvectors for any such t. This can only be done for certain values of t due to the dynamic structure and more completeness assumptions are needed. □

B.2 Example

To obtain more intuition for the assumptions, I verify the assumptions for some specific examples. First assume that

$$Y_{it} = \rho Y_{i(t-1)} + \lambda_i + U_{it}.$$

Also assume that Y_{i0} has been created by an infinite sequence of such a process with a fixed $0 < \rho < 1$. That is

$$Y_{it} = \sum_{j=0}^{\infty} \rho^j \left(\lambda_i + U_{i(t-j)}\right) = \frac{1}{1-\rho}\lambda_i + \sum_{j=0}^{\infty} \rho^j U_{i(t-j)}$$

Now assume that $\lambda_i \sim N\left(0, \sigma_\lambda^2\right)$ and that $U_{it} \sim N\left(0, \sigma_u^2\right)$ for all $t \geq 0$. Then

$$Y_{it} \sim N\left(0, \frac{\sigma_\lambda^2}{(1-\rho)^2} + \frac{\sigma_u^2}{1-\rho^2}\right).$$

Furthermore, for all $s < t$

$$cov\left(Y_{it}, Y_{is}\right) = cov\left(\frac{1}{1-\rho}\lambda_i + \sum_{j=0}^{\infty} \rho^j U_{i(t-j)}, \frac{1}{1-\rho}\lambda_i + \sum_{j=0}^{\infty} \rho^j U_{i(s-j)}\right) = \frac{1}{(1-\rho)^2}\sigma_\lambda^2 + \frac{\rho^{t-s}}{1-\rho^2}\sigma_u^2.$$

It follows that

$$\begin{pmatrix} Y_{i6} \\ Y_{i5} \\ Y_{i4} \\ Y_{i3} \\ Y_{i2} \\ Y_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \frac{\sigma_\lambda^2}{(1-\rho)^2} E_6 + \frac{\sigma_u^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}\right)$$

where E_6 denotes a 6×6 matrix of ones. Using the joint distribution it is easy to verify that $(Y_{i6}, Y_{i1}) \mid Y_{i4}, Y_{i2}$ and $(Y_{i5}, Y_{i1}) \mid Y_{i4}, Y_{i2}$ are normally distributed and that the covariance is not equal to 0 whenever $\sigma_\lambda^2 \neq 0$. The same holds for $(\lambda_i, Y_{i6}) \mid Y_{i4}$ and $(\lambda_i, Y_{i5}) \mid Y_{i4}$. Hence, Assumption L6 holds. Similarly, one can verify Assumption L7 using the normality of Y_i.

The arguments provided do not rely on ρ being the same for all t or all factors being one. For example assume that

$$Y_{it} = \alpha_t + \rho_t Y_{i(t-1)} + F_t \lambda_i + U_{it}.$$

and

$$Y_{i0} = \alpha_0 + \frac{1}{1-\rho_0}\lambda_i + U_{i0}$$

for some $\rho \in (-1,1)$ where $U_{i0} \sim N\left(0, \frac{\sigma_0^2}{1-\rho_0^2}\right)$ and $U_{it} \sim N\left(0, \sigma_t^2\right)$ for $t \geq 1$. Then if $F_t \neq 0$ for all t the assumptions can be verified in a similar way.

C Proofs and examples discrete factor models

C.1 Proof of Theorem 4

Conditional on X_i and for all $m_K, m_1 \in \{1, \ldots, M\}$ and some s define

$$P_{m_K,m_1,s} \equiv P\left(Z_{iK} \in C_{m_K}, Z_{i1} \in C_{m_1}, Z_{ik_{R+1}} \in A_s\right).$$

Define the matrix $L_{1,k_{R+1},K}$ as the $M \times M$ matrix containing these probabilities for a fixed s such that m_K increases over rows while m_1 increases over columns. That is

$$L_{1,k_{R+1},K} \equiv \begin{pmatrix} P_{1,1,s} & P_{1,2,s} & \cdots & P_{1,M,s} \\ P_{2,1,s} & P_{2,2,s} & \cdots & P_{2,M,s} \\ \vdots & \vdots & \ddots & \vdots \\ P_{M,1,s} & P_{M,2,s} & \cdots & P_{M,M,s} \end{pmatrix}.$$

Let $L_{1,K}$ be the $M \times M$ matrix containing the probabilities $P_{m_K,m_1} = P\left(Z_{iK} \in C_{m_K}, Z_{i1} \in C_{m_1}\right)$ such that m_K increases over rows only while m_1 increases over columns. Let $\lambda^1, \ldots, \lambda^M$ be an ordering of all points of support of λ_i. Let $L_{1,\lambda}$ be the $M \times M$ matrix containing $P_{l,m_1} = P\left(Z_{i1} \in C_{m_1} \mid \lambda_i = \lambda^l\right)$ with l increasing over rows and m_1 increasing over columns. Let $L_{\lambda,K}$ be the matrix containing $P_{m_K,l} = P\left(Z_{iK} \in C_{m_1}, \lambda_i = \lambda^l\right)$ with l increasing over columns and m_K over rows. Let $D_{k_{R+1},\lambda}$ be a diagonal matrix containing $P\left(Z_{ik_{R+1}} \in A_s \mid \lambda_i = \lambda^l\right)$ on the diagonal. Recall that s is fixed.

As in the continuous case, it is easy to show that the assumptions and definitions imply that

$$L_{1,k_{R+1},K} L_{1,K}^{-1} = L_{1,\lambda} D_{k_{R+1},\lambda} L_{1,\lambda}^{-1}.$$

The eigenvectors sum to 1 because they contain conditional probabilities. Just as in Section 3, the eigenfunctions are the same for all rotations K and any s. By Assumptions D3, D5, and full rank of F^1, the eigenvalues are unique, when considering all K and partitions of the support of $Z_{ik_{R+1}}$. The ordering of the eigenvectors follows from the assumption that $F^3 = I_{R\times R}$ and Assumption D5. Therefore, all elements of $L_{1,\lambda}$ as well as $D_{k_{R+1},\lambda}$ and $L_{K,\lambda}^{-1}$ are identified.

Now for all t let $\tilde{A}_{t,1}, \ldots, \tilde{A}_{t,S}$ be a partition of $\mathcal{Y}_t$ with $\tilde{A}_{t,s} = A_s$ for all $t = 1, \ldots, R$. Let $K = \{1, \ldots, R\}$. Define $\tilde{L}_{1,K}$, $\tilde{L}_{K,\lambda}$, and $\tilde{L}_{1,\lambda}$ analogously to $L_{1,K}$, $L_{K,\lambda}$, and $L_{1,\lambda}$ but using the sets $\tilde{A}_{t,s}$

instead of $A_{t,s}$. It then holds that $\tilde{L}_{K,\lambda}^{-1} = L_{K,\lambda}^{-1}$ and $\tilde{L}_{1,K}\tilde{L}_{K,\lambda}^{-1} = \tilde{L}_{1,\lambda}$. The left hand side is identified, hence the right hand side is identified for any $\tilde{A}_{t,s}$. Thus $P\left(Y_{it} \in \tilde{A}_{t,s} \mid \lambda_i = \lambda^m\right)$ is identified for all $t = R+2, \dots, T$ and all s and m. Since the partition is arbitrary, $P\left(Y_{it} \in B_t \mid \lambda_i = \lambda^m\right)$ is identified for all $t = R+2, \dots, T$. Similar argument yield $P\left(Y_{it} \in B_t, \lambda_i = \lambda^m\right)$ for all $t = 1, \dots, R+1$. □

C.2 Example

As an illustrative example assume that $Y_{it} \in [0,1]$, $R = 2$, $T = 5$, and $S = 2$. Let the points of support of λ_i be $\{(l_1, l_2), (l_1, h_2), (h_1, l_2), (h_1, h_2)\}$ with $l_1 < h_1$ and $l_2 < h_2$. This setup could interpreted as each person having two skills such as cognitive and noncognitive skills. For each skill, a person can either be of low type or of high type. Now define $A_1 = [0, 0.5]$ and $A_2 = (0.5, 1]$. Hence $C_1 = [0, 0.5] \times [0, 0.5]$, $C_2 = [0, 0.5] \times (0.5, 1]$, $C_3 = (0.5, 1] \times [0, 0.5]$ and $C_4 = (0.5, 1] \times (0.5, 1]$. Let $Z_{i1} = (Y_{i4}, Y_{i5})$ and $Z_{K1} = (Y_{i1}, Y_{i2})$. Let $P_{m_K, m_1} = P\left(Z_{iK} \in C_{m_K}, Z_{i1} \in C_{m_1}\right)$. Then

$$L_{1,K} = \begin{pmatrix} P_{1,1} & P_{1,2} & P_{1,3} & P_{1,4} \\ P_{2,1} & P_{2,2} & P_{2,3} & P_{2,4} \\ P_{3,1} & P_{3,2} & P_{3,3} & P_{3,4} \\ P_{4,1} & P_{4,2} & P_{4,3} & P_{4,4} \end{pmatrix}$$

and $L_{1,k_{R+1},K}$ is defined similarly but contains $P\left(Z_{iK} \in C_{m_K}, Z_{i1} \in C_{m_1}, Z_{ik_{R+1}} \in A_s\right)$ for a fixed s. The eigenvectors are contained in $L_{1,\lambda}$ which in this case is

$$L_{1,\lambda} = \begin{pmatrix} P\left(Z_{i1} \in C_1 \mid \lambda_i = \lambda^1\right) & P\left(Z_{i1} \in C_1 \mid \lambda_i = \lambda^2\right) & P\left(Z_{i1} \in C_1 \mid \lambda_i = \lambda^3\right) & P\left(Z_{i1} \in C_1 \mid \lambda_i = \lambda^4\right) \\ P\left(Z_{i1} \in C_2 \mid \lambda_i = \lambda^1\right) & P\left(Z_{i1} \in C_2 \mid \lambda_i = \lambda^2\right) & P\left(Z_{i1} \in C_2 \mid \lambda_i = \lambda^3\right) & P\left(Z_{i1} \in C_2 \mid \lambda_i = \lambda^4\right) \\ P\left(Z_{i1} \in C_3 \mid \lambda_i = \lambda^1\right) & P\left(Z_{i1} \in C_3 \mid \lambda_i = \lambda^2\right) & P\left(Z_{i1} \in C_3 \mid \lambda_i = \lambda^3\right) & P\left(Z_{i1} \in C_3 \mid \lambda_i = \lambda^4\right) \\ P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^1\right) & P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^2\right) & P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^3\right) & P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^4\right) \end{pmatrix}.$$

Notice that the eigenvectors sum to 1. Assumptions D5 and D6 imply that for all $m = 1, 2, 3$

$$P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^4\right) > P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^m\right)$$

and for all $m = 2, 3, 4$

$$P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^1\right) < P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^m\right).$$

Moreover,

$$P\left(Y_{i4} \in (0.5, 1] \mid \lambda_i = \lambda_2\right) > P\left(Y_{i4} \in (0.5, 1] \mid \lambda_i = \lambda_3\right)$$

or equivalently

$$P\left(Z_{i1} \in C_3 \mid \lambda_i = \lambda^2\right) + P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^2\right) > P\left(Z_{i1} \in C_3 \mid \lambda_i = \lambda^3\right) + P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^3\right).$$

To sum up, the data provides $L_{1,k_{R+1},K}L_{1,K}^{-1}$ for all rotations K and all s. There are exactly 4 vectors with elements summing to 1, which are eigenvectors of $L_{1,k_{R+1},K}L_{1,K}^{-1}$ for all rotations K and all s. The vector with the smallest fourth element belongs to λ^1 and the element with the largest fourth element belongs to λ^4. The other two vectors can be distinguished by the sum of their third and fourth element: the larger sum belongs to λ^2. The assumption that $F^3 = I_{R\times R}$ can in this case therefore be replaced with the weaker assumptions:

1. $P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^4\right) > P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^m\right)$ for all $m = 1, 2, 3$,
2. $P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^1\right) < P\left(Z_{i1} \in C_4 \mid \lambda_i = \lambda^m\right)$ for all $m = 2, 3, 4$, and
3. $P\left(Y_{i4} \in (0.5, 1] \mid \lambda_i = \lambda_2\right) > P\left(Y_{i4} \in (0.5, 1] \mid \lambda_i = \lambda_3\right).$

D Estimation

D.1 Fully nonparametric estimator

I first make the following three assumptions which strengthen Assumptions S5 and S8. These assumptions can be avoided if constraints on the unknown functions are imposed, which ensure that λ_i is bounded complete for Z_{i1}.

Assumption E1. $U_{i1}, \ldots, U_{iT}, \lambda_i$ are jointly independent conditional on $X_i \in \mathcal{X}$.

Assumption E2. Z_{i1} is bounded complete for Z_{iK} and Z_{iK} is bounded complete for Z_{i1} for any ordering K conditional on $X_i \in \mathcal{X}$.

Assumption E3. λ_i has support on $\mathbb{R}^R$.

Assumptions E1 - E3 imply that λ_i is complete for Z_{i1} which follows from Theorem 2.1 in Mattner (1993), Proposition 2.4. in D'Haultfoeuille (2011), and arguments in the proof of Lemma 1. Thus, as discussed in Section 5.1.1, all completeness assumptions are on distributions of observables only.

As in Section 5.1.1 let $W_i = (Y_i, X_i)$, $\theta = \left(\tilde{h}_1, \ldots, \tilde{h}_T, f_1, \ldots, f_T, f_\lambda, \tilde{F}\right)$ and

$$l\left(\theta, W_i\right) = \log \int \prod_{t=1}^{T} f_t(\tilde{h}_t\left(Y_{it}, X_{it}\right) - \lambda' \tilde{F}_t; X_i, \lambda) \tilde{h}_t'\left(Y_{it}, X_{i,}\right) f_\lambda(\lambda; X_i) d\lambda.$$

Moreover, recall the norm

$$||\theta||_s = \sum_{t=1}^{T} \left(||\tilde{h}_t||_{\infty,\tilde{\omega}_1} + ||f_t||_{\infty,\tilde{\omega}_2} + ||\tilde{F}_t||_E\right) + ||f_\lambda||_{\infty,\tilde{\omega}_3}.$$

Before, I state the smoothness assumptions let $\omega_{y_l,y_u}(y)$ be a bounded weight function on $[y_l, y_u]$ such that $\omega_{y_l,y_u}(y) > 0$ for all $y \in (y_l, y_u)$ and $\omega_{y_l,y_u}(y) \to 0$ as $y \to y_l$ and $y \to y_u$. Examples are

$\omega_{y_l,y_u}(y) = (y-y_l)^2(y_u-y)^2$ and $\omega_{y_l,y_u}(y) = (1+\Phi^{-1}(y-y_l)\Phi^{-1}(y_u-y))^{-2}$ where Φ^{-1} denotes the standard normal cdf. I assume for simplicity that either $\mathcal{Y}_t = \mathbb{R}$ or $\mathcal{Y}_t = [y_l, y_u]$, but cases where $\mathcal{Y}_t = [y_l, \infty)$ can be handled in a similar manner. All function spaces used in the assumptions are defined in Section 5.1.1.

Assumption E4.

(i) If $\mathcal{Y}_t = \mathbb{R}$, $h_{t0} \in \mathcal{H}_t$ with $\omega_1(y,x) = \left(1+||(y,x)||_E^2\right)^{-\varsigma_1/2}$, $\tilde{\omega}_1(y,x) = \left(1+||(y,x)||_E^2\right)^{-\tilde{\varsigma}_1/2}$, and $\tilde{\varsigma}_1 > \varsigma_1 > 0$. If $\mathcal{Y}_t = [y_l, y_u]$, $h_{t0} \in \mathcal{H}_t$ with $\omega_1(y,x) = \left(1+||x||_E^2\right)^{-\varsigma_1/2}\omega_{y_l,y_u}(y)$, $\tilde{\omega}_1(y,x) = \left(1+||x||_E^2\right)^{-\tilde{\varsigma}_1/2}\omega_{y_l,y_u}(y)$, and $\tilde{\varsigma}_1 > \varsigma_1 > 0$.

(ii) $f_{t0} = f_{U_{it}|X_i} \in \mathcal{F}_t$ for all $t = 1, \ldots, T$ with $\omega_2(u,x) = \left(1+||(u,x)||_E^2\right)^{-\varsigma_2/2}$. Also, $\tilde{\omega}_1(u,x) = \left(1+||(u,x)||_E^2\right)^{-\tilde{\varsigma}_2/2}$ and $\tilde{\varsigma}_2 > \varsigma_2 \geq 0$.

(iii) $f_{\lambda 0} = f_{\lambda_i|X_i} \in \mathcal{F}_\lambda$ with $\omega_3(\lambda,x) = \left(1+||(\lambda,x)||_E^2\right)^{-\varsigma_3/2}$. Also, $\tilde{\omega}_3(\lambda,x) = \left(1+||(\lambda,x)||_E^2\right)^{-\tilde{\varsigma}_3/2}$ and $\tilde{\varsigma}_3 > \varsigma_3 \geq 0$.

(iv) $F \in \mathcal{V}$.

Assumption E5. The data $\{(Y_i, X_i)\}_{i=1}^n$ are independent and identically distributed. Furthermore, $E\left[\sup_{\theta\in\Theta_n} |l(\theta, W_i)|\right]$ is bounded for all n.

Assumption E6. $E[l(\theta, W_i)]$ is continuous at θ_0 in Θ.

Assumption E7. $\Theta_n \subseteq \Theta_{n+1} \subseteq \Theta$ for all n. There exists a sequence $\pi_n\theta_0 \in \Theta_n$ such that $||\pi_n\theta_0 - \theta_0||_s \to 0$ as $n \to \infty$.

Assumption E8. The sieve spaces Θ_n are compact under $||\cdot||_s$.

Assumption E9. There are a finite $k > 0$ and a random variable $U_n(W_i)$ with $E[U_n(W_i)] < \infty$ such that

$$\sup_{\theta,\theta'\in\Theta_n:||\theta-\theta'||_s\leq\delta} |l(\theta, W_i) - l(\theta', W_i)| \leq \delta^k U_n(W_i).$$

Assumption E10. Let $N(\delta, \Theta_n, ||\cdot||_s)$ be the covering number without bracketing. Assume that $\log\left(N\left(\delta^{1/k}, \Theta_n, ||\cdot||_s\right)\right) = o(n)$ for all $\delta > 0$.

Assumption E4 restricts the parameter space. These restrictions lead to the definition of the norm $||\cdot||_s$ under which consistency is established. Different parameter spaces and different choices of norms are possible. The reason for using a weighted Hölder space is that it allows for unbounded support, unbounded functions, and unbounded derivatives. The choices also guarantee that the parameter space is compact with respect to the norm $||\cdot||_s$ (see Ai and Chen 2003). Assumption E5 assumes i.i.d. data and states that the objective is bounded for each n. Assumptions E6 and E9 impose a continuous population objective and a Lipschitz continuous log-likelihood, respectively. Assumptions E7, E8, and E10 place restrictions on the sieve space. Assumption E7 ensures that the sieve space grows as $n \to \infty$, while Assumption E10 states that the sieve space cannot grow too fast. See Chen (2007) and references therein for choices of sieves that satisfy these assumptions. For finite dimensional linear sieves, Assumption E10 is typically satisfied if $\dim(\Theta_n)/n \to 0$.

The most important aspects of the assumptions are uniqueness of the solution in the population (identification), compactness of the parameter space, and uniform convergence over the sieves space. As mentioned above, without Assumptions E1 - E3 one needs other conditions to ensure that the population maximizer over the parameter space is unique. Section 5.1.1 mentions other compact classes of functions and norms which are suitable in some applications. Uniform convergence depends on the choice of sieves and is in general not a necessary condition. See Bierens (2012) for an alternative approach.

I now restate Theorem 5. The theorem is a direct consequence of Theorem 3.1 in combination with Condition 3.5M in Chen (2007). Hence, the proof is omitted.

Theorem 5. Let Assumptions E1 - E10 hold. Then

$$||\hat{\theta} - \theta_0||_s \xrightarrow{p} 0.$$

In order to implement the estimator one needs to choose basis functions as well as the length of the sieve space. For density functions, orthogonal Hermite polynomials are a good choice (see Gallant and Nychka (1987) and Chen (2007)). One could also use spline-wavelets as discussed by Ai and Chen (2003). Using these basic functions, it is easy to impose conditions on the density functions such that they are positive and integrate to 1. An alternative approach is to approximate the square root of the density with a linear sieve. This approach has the advantage that the functions are positive by construction. In my simulations and in the application both approaches yield very similar results. Furthermore, the results in the application are not very sensitive to the number of terms in the sieve space. Finally notice that when evaluating the likelihood, one has to integrate over λ. In practice, solving the integral exactly is too time consuming. Therefore, one has to use a numerical approximation. Possible choices are quadrature rules or Monte Carlo integration, which are both easy to implement. The theory part assumes that the integral is evaluated exactly, or that the approximation error is small enough relative to the sampling error in the data. In my application and the simulations, I use Gauss Hermite quadrature rules.

D.2 Semi-parametric estimator

First, similar as in the nonparametric case, define the function spaces

$$\tilde{\mathcal{F}}_t \equiv \left\{ \eta_t \in \Lambda_c^{\gamma_2,\omega_2}\left(\mathcal{U}_t\right) \text{ for some } \gamma_2 > 1 : \int_{\mathcal{U}_t} \eta_t(u)du = 1 \text{ and } \eta_t(u) \geq 0 \text{ and } S4 \text{ holds} \right\}$$

and

$$\tilde{\mathcal{F}}_\lambda \equiv \left\{ \eta \in \Lambda_c^{\gamma_3,\omega_3}(\Lambda) \text{ for some } \gamma_3 > 1 : \int_\Lambda \eta_t(\lambda) d\lambda = 1 \text{ and } \eta(\lambda) \geq 0 \text{ and } S7 \text{ holds} \right\}.$$

The assumptions discussed in Section 5.1.2 can be summarized as follows.

Assumption E11. Assume that

$$f_{Y_i|X_i}(s;x) = \int \prod_{t=1}^{T} f_{t0}\left(h\left(s_t, x_i, \beta_{1t0}\right) - \left(\Sigma(x_i,\beta_{30})\lambda + \mu(x_i,\beta_{20})\right)' F_t\right) h'\left(s_t, x_t; \beta_{1t0}\right) f_{\lambda 0}\left(\lambda\right) d\lambda$$

where $f_{t0} \in \tilde{\mathcal{F}}_t$ with $\omega_2(u) = \left(1 + ||u||_E^2\right)^{-\varsigma_2/2}$ and $\varsigma_2 \geq 0$ and $f_{\lambda 0} \in \tilde{\mathcal{F}}_\lambda$ with $\omega_3(\lambda) = \left(1 + ||\lambda||_E^2\right)^{-\varsigma_3/2}$ and $\varsigma_3 \geq 0$. Furthermore, $\beta_0 = (\beta_{110}, \ldots, \beta_{1T0}, \beta_{20}, \beta_{30}, F)' \in \mathcal{B}$ where $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_\beta}$.

Although h_t is assumed to be a parametric function in this section, a similar proof can be used if it is semiparametric or nonparametric. Just as before, define $\theta = (f_1, \ldots, f_T, f_\lambda, \beta)$,

$$l(\theta, W_i) \equiv \log\left(\int \prod_{t=1}^{T} f_t \left(h\left(Y_{it}, X_{it}, \beta_{1t}\right) - \left(\Sigma(X_i,\beta_3)\lambda + \mu(X_i,\beta_2)\right)' \tilde{F}_t \right) h'\left(Y_{ii}, X_{it}; \beta_{1t}\right) f_\lambda\left(\lambda\right) d\lambda \right),$$

as well as the parameter space $\Theta = \tilde{\mathcal{F}}_1 \times \cdots \times \tilde{\mathcal{F}}_T \times \tilde{\mathcal{F}}_\lambda \times \mathcal{B}$.[23] Also recall that $\alpha \equiv (f_1, \ldots, f_T, f_\lambda)$. For simplicity, I now directly assume that the maximizer is unique. In the scale and location model, this implies that the model is not over parametrized. For example, either the location of f_λ has to be fixed, or $\mu(X_i, \beta_2)$ cannot have an intercept. Furthermore, it is clear that with these normalization, and the previous identification assumptions, the model is identified.

Assumption E12. There is a unique θ_0 such that

$$\theta_0 = (\alpha_0, \beta_0) = \arg\max_{\theta\in\Theta} E\left[l(\theta, W_i)\right].$$

Again, let Θ_n be a sieve space which is restricted in the assumptions below. The estimated parameter vector is

$$\hat{\theta} = (\hat{\alpha}, \hat{\beta}) = \arg\max_{\theta\in\Theta_n} \sum_{i=1}^{n} l(\theta, W_i).$$

Since consistency follows under the assumptions provided in the previous section, the goal now is to prove asymptotic normality of $\hat{\theta}$. I now present the main steps of the proof, which is very similar to the proof of Theorem 3.1 in Carroll, Chen, and Hu (2010). Some assumptions, such as differentiability, are already made in this outline, but are stated explicitly later.

[23] The parameter space Θ in this section and the previous section differ slightly due to the parametric components in this section. Nevertheless, they have the same name because they are conceptually the same object. The same holds for θ and $l(\theta, W_i)$.

I first prove that $\hat{\theta}$ converges to θ at a rate faster than $n^{-1/4}$ under a weaker norm $||\cdot||_2$. The norm is defined as

$$||v||_2 \equiv \sqrt{E\left[\left(\frac{dl(\theta, W_i)}{d\theta}[v]\right)^2\right]}$$

where

$$\frac{dl(\theta, W_i)}{d\theta}[v] \equiv \left.\frac{dl(\theta + \tau v, W_i)}{d\tau}\right|_{\tau=0}.$$

The corresponding inner product is

$$\langle v_1, v_2\rangle_2 \equiv E\left[\left(\frac{dl(\theta, W_i)}{d\theta}[v_1]\right)\left(\frac{dl(\theta, W_i)}{d\theta}[v_2]\right)\right].$$

Notice that the pathwise derivatives are linear in v. Let $\bar{V}$ denote the closure of linear space of $\Theta - \theta_0$ under the metric $||\cdot||_2$. Then $\left(\bar{V}, \langle\cdot,\cdot\rangle_2\right)$ is a Hilbert space and $\bar{V} = \bar{U} \times \mathbb{R}^{d_\beta}$ with $\bar{U}$ being the closure of the linear span of $\tilde{\mathcal{F}}_1 \times \cdots \times \tilde{\mathcal{F}}_T \times \tilde{\mathcal{F}}_\lambda - \alpha_0$. The directional derivatives can be written as

$$\begin{aligned}\frac{dl(\theta, W_i)}{d\theta}[\theta - \theta_0] &= \frac{dl(\theta, W_i)}{d\beta'}(\beta - \beta_0) + \frac{dl(\theta, W_i)}{d\alpha}[\alpha - \alpha_0] \\ &= \left(\frac{dl(\theta, W_i)}{d\beta'} - \frac{dl(\theta, W_i)}{d\alpha}[\mu]\right)(\beta - \beta_0)\end{aligned}$$

where

$$\frac{dl(\theta, W_i)}{d\alpha}[\mu] = \begin{pmatrix}\frac{dl(\theta,W_i)}{d\alpha}[\mu_1] & \ldots & \frac{dl(\theta,W_i)}{d\alpha}[\mu_{d_\beta}]\end{pmatrix}$$

and $\alpha - \alpha_0 = -\mu(\beta - \beta_0)$ with $\mu = (\mu_1, \ldots, \mu_{d_\beta})$. For any component $k = 1, \ldots, d_\beta$ define

$$\mu_k^* = \underset{\mu_k \in \bar{U}}{\arg\min}\, E\left[\left(\frac{dl(\theta, W_i)}{d\beta_j} - \frac{dl(\theta, W_i)}{d\alpha}[\mu_k]\right)^2\right].$$

Let $\mu^* = \left(\mu_1^*, \ldots, \mu_{d_\beta}^*\right)$ and

$$\frac{dl(\theta, W_i)}{d\alpha}[\mu^*] = \begin{pmatrix}\frac{dl(\theta,W_i)}{d\alpha}[\mu_1^*] & \ldots & \frac{dl(\theta,W_i)}{d\alpha}[\mu_{d_\beta}^*]\end{pmatrix}.$$

Define the $d_\beta \times d_\beta$ matrix

$$V^* = E\left[\left(\frac{dl(\theta, W_i)}{d\alpha}[\mu^*]\right)'\left(\frac{dl(\theta, W_i)}{d\alpha}[\mu^*]\right)\right]. \tag{13}$$

Now for any $\zeta \in \mathbb{R}^{d_\beta}\backslash\{0\}$ consider the linear functional $\zeta'\beta$. If the functional is bounded, it follows from the Riesz Representation Theorem that there exists a vector $v^*(\zeta)$ such that for all $\theta \in \bar{V}$ it holds that $\zeta'\beta = \langle \theta, v^*(\zeta) \rangle_2$. The squared norm of this functional is

$$
\begin{aligned}
\sup_{\theta \neq 0} \frac{|\zeta'\beta|^2}{E\left[\left(\left(\frac{dl(\theta,W_i)}{d\beta'} - \frac{dl(\theta,W_i)}{d\alpha}[\mu]\right)'\beta\right)^2\right]} &= \sup_{\mu \neq 0, \beta \neq 0} \frac{\beta'(\zeta\zeta')\beta}{E\left[\left(\left(\frac{dl(\theta,W_i)}{d\beta'} - \frac{dl(\theta,W_i)}{d\alpha}[\mu]\right)'\beta\right)^2\right]} \\
&= \sup_{\beta \neq 0} \frac{\beta'(\zeta\zeta')\beta}{\beta' V^* \beta} \\
&= \zeta'(V^*)^{-1}\zeta.
\end{aligned}
$$

The second equality follows from the definition of V^* and the last equality follows by noting that the supremum is attained at $\beta = (V^*)^{-1}\zeta$. Hence the functional is bounded if and only if V^* is positive-definite. In this case, the Riesz representer is $v^*(\zeta) = (v^*_\beta((\zeta)), v^*_\alpha((\zeta)))$ where $v^*_{\beta(\zeta)} = (V^*)^{-1}\zeta$ and $v^*_\alpha = -\mu^* v^*_\beta(\zeta)$. As implied by the Riesz Representation Theorem, it is easy to show that

$$
||v^*(\zeta)||_2^2 = \zeta'(V^*)^{-1}\zeta \quad \text{and} \quad \langle \theta, v^*(\zeta) \rangle_2 = \zeta'\beta.
$$

Next, it can be shown that

$$
\zeta'\left(\hat{\beta} - \beta_0\right) = \left\langle \hat{\theta} - \theta_0, v^*(\zeta) \right\rangle_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{dl(\theta, W_i)}{d\theta}[v^*(\zeta)] + o_p(n^{-1/2})
$$

and $E\left[\frac{dl(\theta,W_i)}{d\theta}[v^*(\zeta)]\right] = 0$. It then follows from the central limit theorem that

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{dl(\theta, W_i)}{d\theta}[v^*(\zeta)] \xrightarrow{d} N\left(0, E\left[\left(\frac{dl(\theta, W_i)}{d\theta}[v^*(\zeta)]\right)^2\right]\right).
$$

It was previously shown that

$$
E\left[\left(\frac{dl(\theta, W_i)}{d\theta}[v^*(\zeta)]\right)^2\right] = ||v||_2^2 = \zeta'(V^*)^{-1}\zeta.
$$

Therefore for all $\zeta \neq 0$ we get $\sqrt{n}\zeta'\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \zeta'(V^*)^{-1}\zeta\right)$ which implies by the Cramer Wold device that $\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, (V^*)^{-1}\right)$.

I now provide conditions for these arguments to be valid. Given consistency, the following proof focuses on a shrinking neighborhood of θ_0. Therefore, define the local parameter spaces

$$
\Theta_{0s} = \left\{\theta \in \Theta : ||\theta - \theta_0||_s = o(1), ||\theta||_s \leq c\right\}.
$$

and

$$
\Theta_{0sn} = \left\{\theta \in \Theta_n : ||\theta - \theta_0||_s = o(1), ||\theta||_s \leq c\right\}.
$$

I now make the following assumptions.

Assumption E13. For all $t = 1, \dots, T$, f_{t0} is approximated by a linear sieve of length J_{tn}, and $f_{\lambda 0}$ is approximated by a tensor product of a linear sieve of length $J_{\lambda n}$. Moreover, $||\pi_n f_{t0} - f_{t0}|| = O\left(J_{tn}^{-\gamma_2}\right) = o\left(n^{-1/4}\right)$ and $||\pi_n f_{\lambda 0} - f_{\lambda 0}|| = O\left(J_{\lambda n}^{-\gamma_3/R}\right) = o\left(n^{-1/4}\right)$.

Assumption E14. $\log\left(N\left(\varepsilon, \Theta_n, ||\cdot||\right)\right) \leq C \dim(\Theta_n) \log\left(\frac{\dim(\Theta_n)}{\varepsilon}\right)$ for all $\varepsilon > 0$ and a constant C.

Assumption E15. (i) Θ_{0s} is convex at θ_0 and $\beta_0 \in int(\mathcal{B})$; (ii) $l(W_i, \theta)$ is twice continuously pathwise differentiable with respect to $\theta \in \Theta_{0s}$.

Assumption E16. $\sup_{\tilde{\theta} \in \Theta_{0sn}} \sup_{\theta \in \Theta_{0sn}} \left| \frac{dl(W_i, \tilde{\theta})}{d\theta} \left[\frac{\theta - \theta_0}{||\theta - \theta_0||_s}\right] \right| \leq U_n\left(Z_i\right)$ for a random variable $U_n\left(Z_i\right)$ with $E\left[U_n\left(Z_i\right)^2\right] < \infty$.

Assumption E17. (i) $\sup_{\theta \in \Theta_{0s}: ||\theta||_s = 1} E\left(\frac{dl(W_i, \theta_0)}{d\theta}[\theta]\right)^2 \leq c < \infty$; (ii) uniformly over $\tilde{\theta} \in \Theta_{0sn}$ and $\theta \in \Theta_{0sn}$

$$-E\left[\frac{d^2 l(W_i, \tilde{\theta})}{d\theta d\theta'}[\theta - \theta_0, \theta - \theta_0]\right] = ||\theta - \theta_0||_2^2(1 + o(1)).$$

Assumptions E13 and E14 restricts the rates at which the number of sieve term diverge relative to the smoothness assumptions on the functions. For linear sieves these assumptions are typically satisfied if $\max\left\{J_{tn}^{-\gamma_2}, J_{\lambda n}^{-\gamma_3/R}\right\} = o\left(n^{-1/4}\right)$ (see Ai and Chen (2003) Proposition 3.2. and Chen (2007)). Assumption E15 is a smoothness assumption and restricts the finite dimensional parameter vector to be in the interior of the parameter space. Assumptions E16 and E17 imply that for all $\theta \in \Theta_{0s}$ it holds that $||\theta - \theta_0||_2 \leq \sqrt{c}||\theta - \theta_0||_s$ and that for all $\theta \in \Theta_{0sn}$ and for some finite positive constants c_1 and c_2

$$c_1||\theta - \theta_0||_2^2 \leq E[l(W_i, \theta_0) - l(W_i, \theta)] \leq c_2||\theta - \theta_0||_2^2.$$

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

Theorem E1. Let $\gamma \equiv \min\{\gamma_2, \gamma_3/R\} > 1/2$. Under Assumptions E5 and E11 - E17 with $J_{tn} = O\left(n^{\frac{1}{2\gamma_2+1}}\right)$ and $J_{\lambda n} = O\left(n^{\frac{1}{2\gamma_3/R+1}}\right)$

$$||\hat{\theta} - \theta_0||_2 = O_p\left(\max\left\{J_{tn}^{-\gamma_2}, J_{\lambda n}^{-\gamma_3/R}, \sqrt{\frac{J_{tn}}{n}\log(J_{tn})}, \sqrt{\frac{J_{\lambda n}\log(J_{\lambda n})}{n}}\right\}\right) = o_p(n^{-1/4}).$$

By the previous result I can now focus on the following local parameter spaces

$$\mathcal{N}_0 = \left\{\theta \in \Theta_{0s} : ||\theta - \theta_0||_2 = o(n^{-1/4})\right\}$$

and

$$\mathcal{N}_{0n} = \left\{\theta \in \Theta_{0sn} : ||\theta - \theta_0||_2 = o(n^{-1/4})\right\}.$$

The following assumptions are sufficient conditions for asymptotic normality of the finite dimensional parameter vector β.

Assumption E18. μ^* exists and V^* is positive definite.

Assumption E19. There exists $v_n^* \in \Theta_n - \pi_n \theta_0$ such that $||v_n^* - v^*||_2 = o(1)$ and $||v_n^* - v^*||_2 ||\hat{\theta} - \theta_0||_2 = o_p\left(\frac{1}{\sqrt{n}}\right)$.

Assumption E20. There exists a random variable $U_n(Z_i)$ with $E[U_n(Z_i)^2] < \infty$ and a nonnegative measurable function q with $\lim_{\delta \to 0} q(\delta) = 0$ such that for all $\theta \in \mathcal{N}_{0n}$

$$\sup_{\tilde{\theta} \in \mathcal{N}_0} \left| \frac{d^2 l\left(W_i, \tilde{\theta}\right)}{d\theta d\theta'} [\theta - \theta_0, v_n^*] \right| \leq U_n\left(W_i\right) q(||\theta - \theta_0||_s).$$

Assumption E21. Uniformly over $\tilde{\theta} \in \mathcal{N}_{0n}$ and $\theta \in \mathcal{N}_0$

$$E\left[\frac{d^2 l\left(W_i, \tilde{\theta}\right)}{d\theta d\theta'} [\theta - \theta_0, v_n^*] - \frac{d^2 l\left(W_i, \theta_0\right)}{d\theta d\theta'} [\theta - \theta_0, v_n^*]\right] = o\left(\frac{1}{\sqrt{n}}\right).$$

Assumption E18 is a necessary assumption for $\sqrt{n}$ estimation of β_0. Assumption E19 ensures that the sieve bias is negligible while Assumptions E20 and E21 control the reminder term. These assumptions are standard in nonparametric maximum likelihood estimation (see for example Carroll, Chen, and Hu (2010) or Ackerberg, Chen, and Hahn (2012)). Theorem 6 now follows.

Theorem 6. Assume that Assumptions E5 and E11 - E21 hold. Then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, (V^*)^{-1}\right).$$

The proof of this theorem follows the same steps as the proof of Theorem 3.1 in Carroll, Chen, and Hu (2010) assuming that their model is correctly specified. Hence, it is omitted.

E Additional tables application

Table 6: Questionnaire items used to construct teaching measures

	Mathematics
Traditional	We memorize formulas and procedures. We listen to the teacher give a lecture-style presentation.
Modern	We work together in small groups. We relate what we are learning in mathematics to our daily lives. We explain our answers. We give explanations about what we are studying. We decide on our own procedures for solving complex problems.
	Science
Traditional	We memorize science facts and principles. We listen to the teacher give a lecture-style presentation. We read our science textbooks and other resource materials.
Modern	We work in small groups on an experiment or investigation. We relate what we are learning in science to our daily lives. We design or plan an experiment or investigation. We make observations and describe what we see.

Table 7: Marginal effects teaching practice for girls

	Math scores		Science scores	
	Traditional	Modern	Traditional	Modern
Linear fixed effects	0.033	-0.004	0.067**	0.003
Parametric - one factor - $F_t = 1$	0.034***	-0.004	0.069***	0.004
Parametric - one factor	0.075***	-0.011	0.122***	-0.033**
Parametric - two factor	0.281***	-0.238***	-0.140**	0.315***
Semiparametric - two factors	0.279***	-0.260***	-0.131*	0.257***

References

Ackerberg, D., X. Chen, and J. Hahn (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *The Review of Economics and Statistics 94*(2), 481–498.

Ahn, S., Y. Lee, and P. Schmidt (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics 101*(2), 219–255.

Ahn, S., Y. Lee, and P. Schmidt (2010). Panel data models with multiple time-varying individual effects. Working paper.

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

Altonji, J. and R. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Andrews, D. (2005). Cross-section regression with common shocks. *Econometrica 73*(5), 1551–1585.

Arellano, M. and S. Bonhomme (2011). Nonlinear panel data analysis. *Annual Review of Economics 3*, 395–424.

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

Aucejo, E. (2011). Assessing the role of teacher-student interactions. Working paper.

Bai, J. (2003). Factor models of large dimensions. *Econometrica 71*(1), 135–171.

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

Bai, J. (2012). Fixed-effects dynamic panel models, a factor analytical method. *Econometrica, forthcoming*.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2011). Principal components estimation and identification of the factors. Working paper.

Bester, A. and C. Hansen (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business and Economic Statistics 27*(2), 235–250.

Bierens, H. (2012). Consistency and asymptotic normality of sieve estimators under weak and verifiable conditions. Working paper.

Bietenbeck, J. (2011). Teaching practices and student achievement: Evidence from TIMSS. Working paper.

Bonhomme, S. and J.-M. Robin (2008). Consistent noisy independent component analysis. *Journal of Econometrics 149*(1), 12–25.

Canay, I., A. Santos, and A. Shaikh (2012). On the testability of identification in some nonparametric models with endogeneity. Working paper.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International economic review 44*(2), 361–422.

Carroll, R. J., X. Chen, and Y. Hu (2010). Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. *Journal of Nonparametric Statistics 22*(4), 379–399.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica 78*(1), 159–168.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, Chapter 76, pp. 5550–5623. Elsevier.

Chen, X., E. Tamer, and A. Torgovitsky (2011). Sensitivity analysis in semiparametric likelihood models. Working paper.

Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2012). Average and quantile effects in nonseparable panel models. *Econometrica, forthcoming.*

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources 45*(3), 655–681.

Cunha, F. and J. J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

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

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources 42*(03), 528–554.

Delaigle, A., P. Hall, and A. Meister (2008). On deconvolution with repeated measurements. *The Annals of Statistics 36*(2), 665 – 685.

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

Dudley, R. and W. Philipp (1983). Invariance principles for sums of banach space valued random elements and empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 62*, 509–552.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. Working paper.

Evdokimov, K. (2011). Nonparametric identification of a nonlinear panel model with application to duration analysis with multiple spells. Working paper.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics 19*(3), 1257–1272.

Gallant, A. R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica 55*(2), 363–90.

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

Heckman, J. J., J. Stixrud, and S. Urzua (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics 24*(3), 411–482.

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica 56*(6), 1371–1395.

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*(3), 611–629.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*(1), 27–61.

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

Hu, Y. and M. Shum (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics, forthcoming*.

Huang, X. (2010). Nonparametric estimation in large panels with cross sectional dependence. Working paper.

Lavy, V. (2011). What makes an effective teacher? quasi-experimental evidence. Working paper.

Lui, X. and Y. Shao (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics 31*(3), 807–832.

Mattner, L. (1993). Some incomplete but boundedly complete location families. *The Annals of Statistics 21*(4), 2158–2162.

Moon, H. R. and M. Weidner (2010). Linear regression for panel with unknown number of factors as interactive fixed effects. Working paper.

Newey, W. and J. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K. and D. McFadden (1986). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4 of *Handbook of Econometrics*, Chapter 36, pp. 2111–2245. Elsevier.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*, 967–1012.

Sasaki, Y. (2012). Heterogeneity and selection in dynamic panel data. Working paper.

Shen, X. and H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics 22*(2), 580–615.

Shiu, J.-L. and Y. Hu (2011). Identification and estimation of nonlinear dynamic panel data models with unobserved covariates. Working paper.

Su, L. and S. Jin (2012). Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics 169*(1), 34–47.

Weidner, M. (2011). Semiparametric estimation of nonlinear panel data models with generalized random effects. Working paper.

Williams, B. (2011). A measurement model with discrete measurements and continuous latent variables. Working paper.

Williams, B., J. Heckman, and S. Schennach (2010). Nonparametric factor score regression with an application to the technology of skill formation. Working paper.

Zemelman, S., H. Daniels, and A. Hyde (2005). *Best practice: today's standards for teaching and learning in America's schools.* Heinemann.