Semiparametric Estimation of Nonlinear Panel Data Models with Generalized Random Effects*

Martin Weidner‡

January 4, 2011

Abstract

This paper considers non-linear panel data models with individual effects and a large number of time periods T. For these models it is well-known that a fixed effect estimation approach results in an incidental parameter bias of order $1/T$. We show that under appropriate assumptions this incidental parameter bias can be substantially reduced if instead of a fixed effect approach one estimates the distribution of the individual effects jointly with the parameter of interest by maximum likelihood, thereby treating the individual effect distribution non-parametrically. The convergence rate of the incidental parameter bias in this approach is shown to be only limited by the smoothness properties of the true individual effect distribution. To allow inference on this distribution we make a "generalized random effect" assumption, which requires the cross-sectional units to be partitioned into groups and imposes a random effect assumption in each group. In Monte Carlo simulations we consider the dynamic binary choice model, and we find the finite sample properties of our estimator to be in accordance with the asymptotic results.

1 Introduction

This paper explores a new approach to higher order bias correction in non-linear panel data models under an asymptotics where both the number of cross-sectional units N and the number of timer periods T become large (referred to simply as

*I am grateful to Stéphane Bonhomme, Gary Chamberlain, Iván Fernández-Val, Cheng Hsiao, Hyungsik Roger Moon, Geert Ridder, and Matthew Shum for helpful comments and discussions.

‡Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089-0253, Email: mweidner@usc.edu, The latest version of this paper is available at http://www-scf.usc.edu/∼mweidner.

"large T asymptotics" in the following). Instead of estimating the individual effects, which parameterize the unobserved heterogeneity of each individual, as additional parameters of the model (fixed effect approach), we propose to instead consistently estimate the distribution of these individual effects conditional on the regressors in a non-parametric way. Our main findings are, firstly, that as T grows the incidental parameter bias of the parameters of interest vanishes at a rate that is determined by the convergence rate of the estimator for the individual effect distribution (under the Hellinger distance); and secondly, that these convergence rates can be much higher than the ones obtained from the existing bias reduction techniques in the fixed effect approach. This improvement, however, also comes at a cost: in order to allow for consistent non-parametric inference on the individual effect distribution, we need to impose assumptions that restrict the dependence of the individual effects ond the regressors, and that require this distribution to be sufficiently smooth. Our results are particularly important for applications in which the number of time periods T is modestly large, while the number of cross-sectional units N is much larger than T. In such a scenario the existing bias correction techniques may be insufficient to achieve a reduction of the incidental parameter bias to a level where it can be ignored relative to the standard error of the estimator, so that imposing additional restrictions in order to further reduce the bias becomes a very attractive option.

We are now going to present a brief overview of the existing literature in order to then give a more detailed comparison with the methods developed in this paper. In general, the possibility to control for unobserved heterogeneity is a very attractive feature of panel data analysis. While there are well-established techniques for handling the unobserved heterogeneity in linear panel data models this issue is still a serious econometric challenge for many non-linear models. Broadly speaking, one can distinguish three different approaches in the literature to meet this challenge:

Firstly, there is the "classic" panel data literature that considers point identification and point estimation under an asymptotic where the number of time periods T remains constant, while the cross-sectional size N goes to infinity. Obtaining a consistent point estimator for the parameters of interest at fixed T is most desirable and can indeed be achieved for some non-linear models. However, at fixed T a non-linear panel data model may not be point identified, or may not possess a $\sqrt{N}$-consistent estimator, as discussed by Chamberlain (2010) for the binary choice model. Furthermore, an incidental parameter problem (Neyman and Scott (1948), see e.g. Lancaster (2000) for a review) usually appears in fixed T estimation of non-linear panel data models since the number of incidental parameters (individual

effects) grows with the sample size. Resolving this problem usually requires a model specific augmentation of standard estimation procedures like maximum likelihood. We refer e.g. to Chamberlain (1984) and Arellano and Honoré (2001) for reviews of this branch of the literature.

Secondly, there is the "large T" panel data literature, which includes e.g. Phillips and Moon (1999), Hahn and Kuersteiner (2002), Lancaster (2002), Woutersen (2002), Hahn and Kuersteiner (2004), Hahn and Newey (2004), Carro (2007), Arellano and Bonhomme (2009), Fernandez-Val (2009), Bester and Hansen (2009), Dhaene and Jochmans (2010); a review is provided by Arellano and Hahn (2007). This literature considers an asymptotic where both panel dimensions N and T go to infinity. This large T asymptotics guarantees point identification of a very large class of models under weak regularity conditions, and provides an asymptotic solution to the incidental parameter problem. Namely, the (maximum likelihood) estimator is shown to have a bias of order $1/T$, which thus vanishes asymptotically, and bias correction techniques are discussed that augment the convergence rate of the bias further.

Finally, there are a few papers that acknowledge the fact that many non-linear panel data models are not point identified at fixed T and consequently discuss set identification (bound analysis) for the parameters of interest or for certain policy parameters like marginal effects. These include e.g. Chernozhukov, Hahn and Newey (2005), Honoré and Tamer (2006), Chernozhukov, Fernández-Val, Hahn and Newey (2009a) and Chernozhukov, Fernández-Val and Newey (2009b).

These three estimation approaches for non-linear panel data models should be viewed as complements rather than substitutes. If for fixed T a ($\sqrt{N}$-) consistent estimator is available for the particular model under consideration, then it probably should be used. If this is not the case, then chances are that the model may not be point identified at fixed T and in particular for small values of T one needs to consider inference using bound analysis. However, the above cited papers on set identification all point out that the bounds can be very tight and shrink rather rapidly as T grows. Thus, if T is sufficiently large one can safely ignore the fact that the model might only be set-identified and simply use the large T estimation methodology.

As mentioned above, the large T approach, which is also used in the present paper, is convenient since the alternative asymptotic $N, T \to \infty$ guarantees existence of a consistent point estimator for a large class of models, and provides an asymptotic solution to the incidental parameter problem. Various techniques are developed to decrease the order of the incidental parameter bias from $1/T$ to smaller orders (e.g.

to $1/T^2$). Whether the remaining bias is problematic depends on the relative size of N and T. The standard error of the estimator is of order $(NT)^{-1/2}$, i.e. it depends on both N and T, while the bias only depends on T. For applications where N is not too large relative to T one can thus ignore the bias relative to the standard error, but for very large values of N relative to T the bias dominates the standard error. In the latter case, getting additional data in the cross-sectional dimension, i.e. increasing N without increasing T, does not improve the estimator further, and in fact worsens the size properties of test statistics based on this biased estimator. The main research problem in the large T panel data literature is thus to find ways to decrease the incidental parameter bias further in order to make the estimation methodology applicable for a larger range of sample sizes N, T.

With the exceptions discussed below, most of the large T panel literature uses a "fixed effect" approach, in which the individual effects (which parameterize the cross-sectional heteroscedasticity) are themselves estimated as incidental parameters. In the present paper we consider a "random effect" approach, in which the distribution of the individual effects is estimated instead. While we allow for correlation between the individual effects and the observed regressors, we require some constraints on the structure of this correlation, since otherwise inference on the conditional distribution of the individual effects is infeasible due to the curse of dimensionality (large dimensional support of the conditioning variables, i.e. the regressors). For most of our results we need not specify the nature of these correlation constraints, so that they are applicable to various ways of reducing the dimensionality of the conditioning variables. As a concrete example for such a correlation constraint we discuss the "generalized random effect" assumption, which assumes that individuals can be grouped based on their observed regressor values and imposes independence between the individual effects and the regressors within each group. Apart from this generalized random effect assumption we impose no further parametric constraints on the individual effect distribution. We estimate the parameters of interest (parametric component) jointly with the individual effect distribution (non-paramatric component) by maximum likelihood. The generalized random effect assumption (or any other constraint that reduces the dimension of the regressors as conditioning variables) is restrictive, but it also turns out to have very powerful consequences for the incidental parameter bias.

We show that the rate at which the bias decreases with T depends on the smoothness of the true individual effect distribution, and that the bias can decrease at an arbitrary polynomial rate as long as the true individual effect distribution is suffi-

ciently smooth. The bias may therefore be much smaller than the one obtained from the existing methods in the large T panel literature. In particular in applications where T is modestly large and N is much larger than T, one may therefore be willing to impose the generalized random effect assumption together with a smoothness assumption on the individual effect distribution, in order to avoid (or substantially reduce) the incidental parameter problem.

The technical derivation of our result is done in two steps. Firstly, we derive the properties of the maximum likelihood estimator for the parameters of interest, assuming that some estimator for the individual effect distribution is given and is used to integrate out the individual effects from the likelihood function. We show that the resulting incidental parameter bias for the parameters of interest is bounded by an expression that involves the Hellinger distance between the true individual effect distribution and its estimator. The rate at which the incidental parameter bias vanishes as T increases therefore depends on the rate at which the individual effect distribution can be estimated. Secondly, we consider estimation of the individual effect distribution by maximum likelihood and show how in this case the convergence rate of the estimator in T depends on the smoothness properties of the true distribution. Our first result on the incidental parameter bias of the parameters of interest is also applicable to other estimators for the individual effect distribution, and it would clearly be interesting to explore alternative estimation approaches (beyond maximum likelihood) for this distribution in future research.

The purpose of this paper is not to replace the existing methods on bias correction in large T panel data, but to provide an interesting alternative with somewhat complementary properties. For example, the Jackknife bias correction methods developed in Hahn and Newey (2004) and Dhaene and Jochmans (2010) also allow for higher order bias correction (to order $1/T^2$, $1/T^3$, etc). Compared to this method, we have to impose a restriction on the correlation structure between the individual effects and the regressors, which is not required for the Jackknife. On the other hand, the Jackknife method needs to impose a stationarity assumption on all observed variables, and higher order Jackknife bias correction can significantly increase the standard error of the estimator, which is both not the case in our approach.

Our estimation approach is based on the integrated likelihood function (integrating over the individual effects) as opposed to the profile likelihood function (maximizing over the individual effects) that appears in the fixed effect estimation. Woutersen (2002) and Arellano and Bonhomme (2009) also use integration instead of profiling for the purpose of bias correction in large T panel data model, but

they only discuss how to reduce the bias to order $1/T^2$ or $o(1/T)$, respectively. For the most part these papers do not discuss consistent estimation of the individual effect distribution, which we show to be the key tool to achieve higher order bias correction. In the second part of Arellano and Bonhomme (2009), bias reduction to $o(1/T)$ is discussed for a parametric random effect model using joint maximum likelihood estimation of all parameters. The analysis in the present paper can be viewed as an extension of their results to the semi-parametric generalized random effect case and to higher order bias reduction. The econometric techniques used here are however quite different from their work, and in particular the higher order bias correction is crucial from an applied perspective.

Another related paper is Bester and Hansen (2007). They consider non-linear panel data models with "flexible correlated random effects" and discuss semiparamatric sieve estimation. The difference to the present paper is that they consider fixed T asymptotics, starting from the assumption that the model is identified at fixed T. Their paper is therefore complementary to our approach, just as the fixed T and large T panel data literature are complementary in general.

As mentioned above, we consider joint maximum likelihood estimation of the parameters of interest and the conditional distribution of the individual effects, treating the latter non-parametrically. There is a large literature on semi- and non-parametric estimation, including non-parametric density estimation (reviewed e.g. in Härdle and Linton (1994), Chen (2007), and Ichimura and Todd (2007)). We essentially employ a sieve approach, i.e. a different parameter set for the non-parametric component is chosen for different sample sizes in such a way that asymptotically a very large class of individual effect distributions can be approximated. Usually in sieve-estimation the parameter set is chosen sample size dependent for the purpose of keeping the sampling error in check. However, in our case the parameter set is chosen sample size dependent in order to keep the fixed T identification problem in check instead of controlling the sampling error, i.e. the role that is played by the large T asymptotics is somewhat non-standard. This is also illustrated by the fact that under the large T asymptotics the maximum likelihood estimator for the parameters of interest (parametric component) is consistent even if we plug in a fixed prior for the individual effect distribution (i.e. an inconsistent estimator for the non-parametric component). This is usually not the case in semi-parametric estimation problems, but is very intuitive in view of the existing large T panel literature (since profiling out and integrating out nuisance parameters should yield similar results for the parameters of interest). Our results are therefore more readily

interpreted from the perspective of this latter literature.

For our Monte Carlo simulations we consider a dynamic binary choice model, which is known not be identified at finite T, see e.g. Honoré and Tamer (2006). The simulations confirm our asymptotic results in that they show that the bias of our estimator for the parameters of interest is very small, as long as the true distribution for the individual effects is sufficiently smooth. The finite sample variance of our estimator is found to be very close to the the variance of the fixed effect maximum likelihood estimator, which is a very important result, since most large T bias correction techniques have a tendency to increase the finite sample variance of the estimator relative to the fixed effect MLE.

For future research it would be interesting to consider alternative estimation procedures for the individual effect distribution, both within the maximum likelihood framework, where one could explore the choice of alternative parameter sets for the non-parametric density estimation, but, as already mentioned above, also alternatives to maximum likelihood, e.g. the predictive recursion method that is considered recently in the statistics literature (Newton, Quintana and Zhang (1998), Newton (2002), Martin and Tokdar (2010)). Furthermore, it would be fascinating to explore more general ways to reduce the dimension of the conditioning variables in the individual effect distribution that go beyond our generalized random effect assumption, and that allow for a more general correlation structure between the individuals effects and the regressors. We formulated many of our results under high-level assumptions, instead of directly considering the generalized random effect case, exactly for the purpose of allowing these types of generalizations.

The paper is organized as follows. In Section 2 we introduce the model and some additional notation. Section 3 defines the estimators for the parameters of interest and the individual effect distribution, and provides a brief discussion of the main conceptual ideas and results of the paper. Section 4 derives the asymptotic distribution for the estimator of the parameters of interest under appropriate high-level assumptions, which is the main technical contribution of the paper. In Section 5 we apply these general results to the special case of generalized random effects. Monte Carlo simulations are presented in Section 6, and some concluding remarks are given in Section 7. The appendix contains figures and tables for the Monte Carlo simulations, and provides the regularity assumptions that are referred to in the theorems of the main text, as well as the proofs of these theorems.

2 Model

We consider a panel data model with N cross-sectional units and T time periods. A dependent variable Y_{it}, a vector of time-varying independent variables X_{it} and a vector of time-invariant independent variables Z_i are observed, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Let $Y_i = (Y_{i1}, \ldots, Y_{iT})$ and $X_i = (X_{i1}, \ldots, X_{iT}; Z_i)$, i.e. all independent variables are summarized by X_i. We assume that the unobserved heterogeneity in the distribution of Y_i conditional on X_i can be described by (a vector of) individual specific effects $\boldsymbol{\alpha}_i$. The random variables Y_i, X_i and $\boldsymbol{\alpha}_i$ take values in the sets $\mathcal{Y}_T$, $\mathcal{X}_T$ and $\mathcal{A}$, where $\mathcal{A} \subset \mathbb{R}^M$, and M is some finite positive integer. For elements of $\mathcal{Y}_T$, $\mathcal{X}_T$ and $\mathcal{A}$ we use the notation y_i, x_i and α_i, or simply y, x and α, i.e. we distinguish non-random from random objects by using lower case as opposed to capital letters and by not using bold face type for the fixed effects. We assume cross-sectional independence and that for each $i = 1, \ldots, N$ the distribution of Y_i conditional on X_i is given by

$$f_{Y|X}(y_i|x_i; \theta, \pi) \;=\; \int_{\mathcal{A}} f(y_i|x_i, \alpha; \theta)\, \pi(\alpha|x_i)\, d\alpha \;, \tag{2.1}$$

where $\theta \in \Theta \subset \mathbb{R}^K$ are the parameters of interest, $f(y_i|x_i, \alpha; \theta)$ is the distribution of Y_i conditional on X_i and $\boldsymbol{\alpha}_i$ for given θ, and $\pi(\alpha|x_i)$ is the distribution of the individual effects conditional on the regressors. Since the individual effects are unobserved they are integrated over in equation (2.1). In the following, when the distribution of one generic cross-sectional unit is considered, we will often drop the index i for notational convenience.

Equation (2.1) describes the distribution of Y given X as a mixture of distributions $f(y|x, \alpha; \theta)$ over the distribution of $\boldsymbol{\alpha}$. We impose a parametric model for $f(y|x, \alpha; \theta)$, i.e. we assume that $f(y|x, \alpha; \theta)$ is known up to the finite dimensional parameter θ. We furthermore assume that $f(y|x, \alpha; \theta)$ has the structure

$$f(y|x, \alpha; \theta) = \prod_{t=1}^{T} f_t(y_t|x, y^{(t-1)}, \alpha; \theta) \;, \tag{2.2}$$

with $y^{(t)} = (y_1, \ldots, y_t)$. Here, $f_t(y_t|x, y^{(t-1)}, \alpha; \theta)$ is the period likelihood function, which describes the distribution of Y_{it} conditional on X_i, lags of Y_{it} and individual effects $\boldsymbol{\alpha}_i$.

Some Further Notation

We use θ^0 and $\pi^0 = \pi^0(\alpha|x)$ to denote the true parameters of interest and the true conditional distribution of the individual effects, i.e. we assume that $f_{Y|X}(y|x; \theta^0, \pi^0)$

describes the actual distribution of Y_i conditional on the covariate value $X_i = x$.

Let $\Pi_T^{\mathcal{A}}$ be the set of all conditional probability densities $\pi(\alpha|x)$ with respect to the Lebesgue measure, over $\alpha \in \mathcal{A}$ and conditional on $x \in \mathcal{X}_T$. We only consider conditional distributions for the individual effects $\boldsymbol{\alpha}$ that are absolutely continuous with respect to the Lebesgue measure on $\mathcal{A}$, and we therefore use the terms distribution and density interchangeably, i.e. we often refer to $\pi(\alpha|x)$ as the distribution of $\boldsymbol{\alpha}$. The following subsets of $\Pi_T^{\mathcal{A}}$ impose a lower respectively upper bound on $\pi(\alpha|x)$

$$\Pi_T^{\text{low}} = \left\{ \pi \in \Pi_T^{\mathcal{A}} \,\middle|\, \forall x \in \mathcal{X}_T, \forall \alpha \in \mathcal{A} \,:\, \pi(\alpha|x) \geq \pi_T^{\text{low}}(\alpha|x) \right\},$$
$$\Pi_T^{\text{up}} = \left\{ \pi \in \Pi_T^{\mathcal{A}} \,\middle|\, \forall x \in \mathcal{X}_T, \forall \alpha \in \mathcal{A} \,:\, \pi(\alpha|x) \leq \pi_T^{\text{up}}(\alpha|x) \right\}. \tag{2.3}$$

Here, the bounds $\pi_T^{\text{low}}(\alpha|x)$ and $\pi_T^{\text{up}}(\alpha|x)$ do not integrate to one. We define the Hellinger distance between distributions $\pi, \pi^0 \in \Pi_T^{\mathcal{A}}$ and between the distributions of the outcome variable Y that are implied by π and π^0 (at the true θ^0) as follows

$$\mathcal{D}_{\text{H}}(\pi, \pi^0) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{A}} \left[\sqrt{\pi(\alpha|X_i)} - \sqrt{\pi^0(\alpha|X_i)} \right]^2 d\alpha \,},$$
$$\mathcal{D}_{\text{H}}(f_Y(\pi), f_Y(\pi^0)) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{Y}_T} \left[\sqrt{f_{Y|X}(y|X_i; \theta^0, \pi)} - \sqrt{f_{Y|X}(y|X_i; \theta^0, \pi^0)} \right]^2 dy}, \tag{2.4}$$

The Kullback Leibler divergence measures between $\pi, \pi^0 \in \Pi_T^{\mathcal{A}}$ and between their implied distributions for the outcome Y read

$$\mathcal{D}_{\text{KL}}(\pi || \pi^0) = \frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{A}} \log \left[\frac{\pi^0(\alpha|X_i)}{\pi(\alpha|X_i)} \right] \pi^0(\alpha|X_i) \, d\alpha \,,$$
$$\mathcal{D}_{\text{KL}}(f_Y(\pi) || f_Y(\pi^0)) = \frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{Y}_T} \log \left[\frac{f_{Y|X}(y|X_i; \theta^0, \pi^0)}{f_{Y|X}(y|X_i; \theta^0, \pi)} \right] f_{Y|X}(y|X_i; \theta^0, \pi^0) \, dy. \tag{2.5}$$

These two Hellinger distances and two Kullback Leibler divergences can all be viewed as distance measures between the distributions π and π^0. These distance measures are random variables since sample averages over functions of covariates appear in their definitions.

3 Description of Estimators and Main Results

The unknown parameters in the model (2.1) are the finite dimensional vector θ and the conditional distribution function $\pi = \pi(\alpha|x)$. The (log-) likelihood function

over these parameters is

$$L_{NT}(\theta, \pi) = \frac{1}{NT} \sum_{i=1}^{N} \log\ f_{Y|X}(Y_i|X_i; \theta, \pi). \tag{3.1}$$

For given π, the maximum likelihood estimator for θ and the corresponding profile likelihood function are

$$\hat{\theta}(\pi) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ L_{NT}(\theta, \pi), \qquad L_{NT}(\pi) = \max_{\theta \in \Theta} L_{NT}(\theta, \pi). \tag{3.2}$$

The maximum likelihood estimator for π is obtained by maximizing the profile likelihood $L_{NT}(\pi)$ over an appropriate set Π_T of individual effect distributions, i.e.

$$\hat{\pi} = \underset{\pi \in \Pi_T}{\operatorname{argmax}}\ L_{NT}(\pi), \qquad \hat{\theta} = \hat{\theta}(\hat{\pi}). \tag{3.3}$$

The choice of the parameter set Π_T crucially affects the properties of the joint maximum likelihood estimators $\hat{\pi}$ and $\hat{\theta}$. Clearly, we need $\Pi_T \subset \Pi_T^{\mathcal{A}}$ (the set of all conditional distributions that integrate to one), but the discussion below makes clear why it is important to constrain Π_T further. While a different set Π_T might be chosen for different values of N and T, it is the T-dependence of the parameter set that turns out to be decisive, which is why we make this dependence explicit in the subscript T.

It is well known that the fixed effect maximum likelihood estimator for θ has an incidental parameter bias of order $1/T$ (e.g. Hahn and Newey (2004)). Similarly, the estimator $\hat{\theta}$ in general also possesses an incidental parameter bias. The main objective of this paper is to show that the rate at which the bias of $\hat{\theta}$ vanishes with T is only restricted by the properties of the true distribution π^0, as long as the parameter set Π_T is chosen appropriately.

There are different types of restrictions that have to be imposed on π^0 and Π_T, which can be associated either with finite N sampling issues or finite T identification issues, and we discuss those separately in the following.

3.1 Sampling Issues (Generalized Random Effect Assumption)

The incidental parameter problem in panel data (Neyman and Scott (1948)) is very familiar when the individual effects $\boldsymbol{\alpha}_i$ are modeled as fixed effects, i.e. when a separate parameter α_i is introduced and estimated for each cross-sectional unit. The problem occurs because the number of parameters α_i grows with the sample

size, which results in an inconsistency of the estimator (e.g. the maximum likelihood estimator) for the parameters of interest under fixed T asymptotics.

Since we model the individual effects as correlated random effects, i.e. via their conditional distribution $\pi(\alpha|x)$, the problem is somewhat less obvious. However, the incidental parameter problem also arises in this approach whenever the support $\mathcal{X}_T$ of the regressors is "large" (i.e. either discrete with very many support points, or continuous and high-dimensional), which is to be expected in particular for large values of T. For example, if $\mathcal{X}_T$ is discrete with cardinality much larger than the sample size N, then we expect the realization of each X_i to be unique within the sample, so that there is only one available observation for the estimation of $\pi(\alpha|x)$ at $x = X_i$. In that case there is no difference between the unrestricted correlated random effects approach and the standard fixed effects approach.

For the correlated random effects approach this incidental parameter problem is resolved asymptotically for fixed T and $N \to \infty$ (e.g. for discrete $\mathcal{X}_T$ one eventually has many units with the same X_i under this asymptotic). However, this asymptotic consideration may not be relevant for the estimation problem at given finite sample size N, T, in particular if T is (moderately) large.

In order to estimate the conditional distribution $\pi(\alpha|x)$ consistently we thus either need the support $\mathcal{X}_T$ to be "small" in the first place (small number of continuous dimensions, or only few discrete support points, relative to N), or we have to make further assumptions to overcome this curse of dimensionality in the conditioning variables.

Various ways are conceivable to reduce the dimension of the conditioning variables. The restriction that we consider explicitly in this paper is what we call a "generalized random effect" assumption. Namely, we assume that there exists a partitioning of $\mathcal{X}_T$ into a finite number of groups such that the conditional densities $\pi(\alpha|x)$ and $\pi(\alpha|\tilde{x})$ are identical if x and $\tilde{x}$ belong to the same group, i.e. we impose a random effect assumption within each group. We assume that this partitioning is known. The number of groups G_T may increase with T, but not too rapidly. This generalized random effect assumption solves the incidental parameter problem, since we only need to estimate a different distribution $\pi(\alpha|x)$ for each group but not separately for each $x \in \mathcal{X}_T$.

The assumption that the true distribution π^0 satisifies this generalized random effect condition can also be written as

$$\boldsymbol{\alpha} \perp X \,\big|\, g, \tag{3.4}$$

i.e. the individual effects $\boldsymbol{\alpha}$ are independent of the regressors X once we condition on the group g. An important generalization of this conditional independence assumption is obtained by replacing the conditioning on the group g with a conditioning on some more general observed variable Z, i.e. $\boldsymbol{\alpha} \perp X \,\big|\, Z$. As long as the support of Z is not too large this more general condition also resolves the incidental parameter problem. The generalized random effect restriction which we consider in this paper is simply the special case where Z has discrete support. For future research it would clearly be interesting to consider the more general case as well. The existence for such a conditioning variable Z is also the basis for the control function approach, discussed e.g. in Imbens and Newey (2009). However, here we do not assume that the model is point identified at fixed T, even after the conditional independence assumption is imposed.

Example: The model which we consider explicitly in our Monte Carlo simulations below is the single index dynamic binary choice model, for which $Y_{it} \in \{0,1\}$, i.e. $\mathcal{Y}_T = \{0,1\}^T$, and

$$Y_{it} = 1\big\{\, \theta\, Y_{i,t-1} + \boldsymbol{\alpha}_i + \varepsilon_{it} \geq 0\big\}. \tag{3.5}$$

Here, $1\{.\}$ is the indicator function, ε_{it} is a random shock that is independent and identically distributed across i and t (with known distributions), and is independent of $\boldsymbol{\alpha}_i$. For simplicity we consider the case where no additional regressors X_{it} are present, but we assume that the initial period outcome variable Y_{i0} is observed (the total number of observed time-periods is thus $T+1$), and we treat this initial outcome as a conditioning variable, i.e. $Z_i = Y_{i0}$ and therefore $\mathcal{X}_T = \{0,1\}$ in the above notation. In this example one thus needs to estimate the parameter $\theta \in \mathbb{R}$ and the two densities $\pi(\alpha|Y_{i0} = 0)$ and $\pi(\alpha|Y_{i0} = 1)$. Since $\mathcal{X}_T$ is finite with a constant number of elements (independent of T), this model already satisfies the generalized random effects assumption, without imposing any further constraints.

3.2 Identification Issues (Smoothness Assumption on π)

The motivation for the large T asymptotics in the panel data literature is to resolve the incidental parameter estimation problem, to overcome the potential fixed T identification problem, and to do this in a way that is not model specific. We have argued that the generalized random effect assumption already overcomes the incidental parameter problem, but there may still remain an identification problem at finite T. For example, under the random effects approach there is no incidental

parameter problem in the dynamic binary choice model without additional regressors, which we just introduced. Nevertheless the model is not fixed T identified, as discussed in Honoré and Tamer (2006). It is this identification problem that motivates the use of large T asymptotics in the present paper.

If the particular model of interest is point identified at fixed T once the generalized random effect assumption is imposed, then one can use the sieve estimation approach of Bester and Hansen (2007), which allows for semi-parametric estimation at fixed T. However, to determine whether the model is point identified or not requires (a potentially complicated) model specific analysis. When T is sufficiently large we show that one can avoid this by considering the alternative asymptotics $N, T \to \infty$.

To discuss how the identification problem vanishes as $T \to \infty$ we consider the expected (log-) likelihood function

$$\overline{L}_{NT}(\theta, \pi) = \mathbb{E}\left[L_{NT}(\theta, \pi) \middle| X_1, \ldots, X_N\right]. \tag{3.6}$$

This expected likelihood is not quite the population likelihood function, since we still condition on the regressors X_i, i.e. $\overline{L}_{NT}(\theta, \pi)$ is still random, and for finite N not all of the possible variation in X_i is already accounted for in $\overline{L}_{NT}(\theta, \pi)$. For $\pi \in \Pi_T^{\mathcal{A}}$ we define[1]

$$\overline{\theta}(\pi) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \overline{L}_{NT}(\theta, \pi). \tag{3.7}$$

For sufficiently large values of N and T we assume that $\overline{\theta}(\pi^0) = \theta^0$ with probability one, i.e. if the true distribution for the individual effects is known, then θ is point-identified. Our analysis in the next section shows that if π satisfies a certain generalized Lipschitz condition with Lipschitz constant $\kappa_T = o(\sqrt{T})$, then under appropriate regularity conditions we have

$$\overline{\theta}(\pi) - \overline{\theta}(\pi^0) = \mathcal{O}_p\left(\frac{\kappa_T}{T}\right) \mathcal{D}_{\mathrm{H}}(\pi, \pi^0), \tag{3.8}$$

where $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0)$ is the Hellinger distance introduced above. This result has two important consequences:

Firstly, by choosing some fixed prior distribution π for each T we can achieve $\kappa_T = \mathcal{O}(1)$ and $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0) \leq \sqrt{2}$ (which always holds for the Hellinger distance). We thus obtain $\overline{\theta}(\pi) - \theta^0 = \mathcal{O}_p(1/T)$, i.e. the identification problem for θ vanishes as T becomes large, and the size of the identified set shrinks at least at the rate $1/T$.[2]

[1] Note that π depends on T, but we suppress this dependence.

[2] $\overline{\theta}(\pi)$ has a given value for each T and we have $\theta^0 = \overline{\theta}(\pi) + \mathcal{O}_p(1/T)$, which implies that the identified set for θ^0 has to shrink at the rate $1/T$.

Secondly, the actual rate at which the identified set for θ^0 shrinks may be much faster, and depends on how fast the identified set for π^0 shrinks in terms of the Hellinger distance. This rate can be very high, depending on the smoothness assumptions that are imposed on the allowed conditional densities π. We are going to discuss this issue a little further now.

For simplicity, consider the case where θ^0 is known (the conclusions for the case where θ is estimated turn out to be equivalent), and define

$$\overline{\pi} = \underset{\pi \in \Pi_T}{\operatorname{argmax}} \, \overline{L}_{NT}(\theta^0, \pi), \tag{3.9}$$

for some appropriate parameter set Π_T. Here, the optimal $\overline{\pi}$ may not be unique, but we assume that it exists and that one of the optimal values is chosen if there are multiple ones. The rate at which $\mathcal{D}_{\mathrm{H}}(\overline{\pi}, \pi^0)$ vanishes as $T \to \infty$ provides an upper bound for the rate at which the identified set for π^0 shrinks with T.

Maximizing $\overline{L}_{NT}(\theta^0, \pi)$ is equivalent to minimizing $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))$, which is the Kullback Leibler divergence of the outcome variable distributions that are implied by π and π^0 (see definition above). Thus, the rate at which $\mathcal{D}_{\mathrm{H}}(\overline{\pi}, \pi^0)$ converges to zero as $T \to \infty$ depends on[3]

(i) How fast $\mathcal{D}_{\mathrm{KL}}(f_Y(\overline{\pi})||f_Y(\pi^0))$ converges to zero, i.e. how well the distribution of the outcome variable generated by an element in Π_T can approximate the true distribution of the data.

(ii) Whether a small value of $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))$ for $\pi \in \Pi_T$ also implies that $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0)$ is small.

Note that the first point demands Π_T to be large enough, while the second point requires it not to be too large. This is analogous to the problem one faces in non-parametric sieve estimation (see e.g. Chen (2007)), only that our condition (ii) is related to identification, while there it is about controlling the sample variation of the non-parametric estimator.

To discuss condition (i) one can e.g. use the inequality $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0)) \leq \mathcal{D}_{\mathrm{KL}}(\pi||\pi^0)$, which holds generally (chain rule for Kullback Leibler divergence). Satisfying condition (i) is therefore mainly an exercise in approximation theory, with the distance measure given by the Kullback Leibler divergence. There are many possibilities to approximate an unknown function, summarized e.g. in the review of Chen (2007). Since π is a probability density we also need to impose the constraints

[3] Assumption 4.3 below provides a formal statement of these conditions.

that the density is positive and integrates to one. The quality of the approximation of π^0 will strongly depend on how smooth π^0 is as a function of α.

To satisfy condition (ii) we are going to choose the set Π_T such that it only contains distributions that are sufficiently smooth in a well-defined sense, as will be discussed extensively in Section 5.

3.3 Main Results

In Section 4 below we derive the asymptotic properties of the estimator $\hat{\theta}$ under high-level assumptions on the parameter set Π_T and the true distribution π^0. In Section 5 we then discuss the generalized random effect assumption as one particular example how to satisfy these high-level assumptions. However, our asymptotic results for $\hat{\theta}$ in Section 4 are applicable more generally. The high-level assumptions we impose there reflect the restrictions discussed above, i.e. firstly that the dependence of $\boldsymbol{\alpha}$ and X needs to be restricted to avoid a curse of dimensionality problem when estimating $\pi(\alpha|x)$, secondly that the true distribution $\pi^0(\alpha|x)$ can be well-approximated by an element in Π_T, and thirdly that distributions in Π_T are sufficiently smooth to control for the fact that the model may not be identified at fixed T.

The method we use to control the asymptotic bias in $\hat{\theta}$ is to bound the difference between $\hat{\theta}$ and $\hat{\theta}(\pi^0)$, namely we show that

$$\hat{\theta} - \hat{\theta}(\pi^0) = \mathcal{O}_p\left(\frac{\kappa_T\,\mu_T}{T}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{3.10}$$

where κ_T is the generalized Lipschitz constant that describes the smoothness of the functions in Π_T, and μ_T is the rate at which the true distribution π^0 can be approximated by an element of Π_T in terms of Hellinger distance as $T \to \infty$. The result (3.10) is very powerful since the infeasible estimator $\hat{\theta}(\pi^0)$ has no asymptotic bias, i.e. the equation states that all terms that contribute to the asymptotic bias of $\hat{\theta}$ are of the order $\kappa_T\mu_T/T$, which can vanish very rapidly as T increases.

To describe the limiting distribution of $\hat{\theta}$ we note that under appropriate regularity conditions $\sqrt{NT}(\hat{\theta}(\pi^0) - \theta^0)$ is asymptotically normal with mean zero and variance $\mathcal{I}_0^{-1}$. Here, $\mathcal{I}_0$ is the large N, T limit of the appropriately scaled information matrix of the model for given π^0. Thus, as long as N is not growing too fast asymptotically, namely as long as $N = o\left(T/(\mu_T^2\kappa_T^2)\right)$, we can conclude that the right hand side of equation (3.10) is of order $o_p(1/\sqrt{NT})$, and therefore $\sqrt{NT}(\hat{\theta} - \theta^0) \to_d \mathcal{N}(0, \mathcal{I}_0^{-1})$ for $N, T \to \infty$.

We then apply these general results to the special case of generalized random effects, where the regressor domain $\mathcal{X}_T$ is decomposed into G_T groups and a random

effect assumption is imposed in each group. We discuss how Π_T can be chosen appropriately in that case. We then show that if $\pi^0(\alpha|x)$ is r times continuously differentiable in α with bounded derivatives and $r \geq 1$, then under appropriate regularity conditions we have

$$\frac{\kappa_T \, \mu_T}{T} = \frac{1}{T} \left(\frac{\log T}{T} \right)^{(r-1)/2}. \tag{3.11}$$

Thus, the rate at which the bias of $\hat{\theta}$ decreases in T is only restricted by the smoothness of π^0. This is a very powerful result: By imposing a generalized random effect assumption together with a smoothness assumption on the distribution π^0, we can obtain an estimator $\hat{\theta}$ whose asymptotic bias vanishes very rapidly. The rate at which the bias decreases in T can be substantially higher than the rates obtained from other bias correction techniques. Consequently, the estimator $\hat{\theta}$ can be asymptotically unbiased under a much larger range of asymptotics, allowing N to grow much faster than T. Our Monte Carlo simulations confirm this asymptotic result, since we find $\hat{\theta}$ to have very little bias also in scenarios where N is much larger than T, as long as π^0 can be well-approximated by an element of Π_T.

4 Asymptotic Analysis of the Estimators

In this section we analyze the large N, T asymptotic properties of the estimator $\hat{\theta}(\pi)$ that was introduced in (3.2), and of which the joint maximum likelihood estimator $\hat{\theta} = \hat{\theta}(\hat{\pi})$ is a special case. In subsection 4.1 we show uniform consistency of $\hat{\theta}(\pi)$ over all $\pi \in \Pi_T^{\text{low}}$, i.e. over all individual effect distributions $\pi = \pi(\alpha|x)$ that are appropriately bounded from below. Having established uniform consistency we then continue to analyze the local properties of the integrated likelihood function $L_{NT}(\theta, \pi)$ around θ^0. In subsection 4.2 we derive uniform bounds on the difference between the scores and the Hessians of $L_{NT}(\theta, \pi)$ between two different individual effect distributions. We then use these bounds for the score and Hessian to also bound the difference $\hat{\theta}(\hat{\pi}) - \hat{\theta}(\pi^0)$ in terms of the Hellinger distance between $\hat{\pi}$ and π^0. In subsection 4.3 we then derive the convergence rate of $\hat{\pi}$ to π^0 in terms of the Hellinger distance under appropriate assumptions. This also gives an upper bound for the convergence rate of $\hat{\theta}(\hat{\pi})$ to $\hat{\theta}(\pi^0)$. Since $\hat{\theta}(\pi^0)$ is asymptotically normal and unbiased we can use this result to characterize the asymptotic distribution of $\hat{\theta}(\hat{\pi})$.

4.1 Uniform Consistency of $\hat{\theta}(\pi)$

We are going to show uniform consistency of $\hat{\theta}(\pi)$ over a large class of distributions $\pi(\alpha|x)$ for the asymptotics $N, T \to \infty$. Our strategy for doing so is to relate the integrated likelihood that was defined in (3.1) to the profile likelihood, which reads

$$L^{\mathrm{p}}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N} \max_{\alpha\in\mathcal{A}} \log\, f(Y_i|X_i,\alpha;\theta) = \frac{1}{NT}\sum_{i=1}^{N} \log f(Y_i|X_i,\hat{\alpha}^{\mathrm{p}}_i(\theta);\theta), \quad (4.1)$$

where $\hat{\alpha}^{\mathrm{p}}_i(\theta) = \mathrm{argmax}_{\alpha\in\mathcal{A}}\, f(Y_i|X_i,\alpha;\theta)$. The profile likelihood is the one that is maximized in the fixed effect approach and the corresponding fixed effect estimator for the parameters of interest reads

$$\hat{\theta}^{\mathrm{p}} = \underset{\theta\in\Theta}{\mathrm{argmax}}\; L^{\mathrm{p}}_{NT}(\theta). \quad (4.2)$$

Consistency of $\hat{\theta}^{\mathrm{p}}$ is well-established in the literature on large T panel data. Our goal is therefore to show $\hat{\theta}(\pi) = \hat{\theta}^{\mathrm{p}} + o_p(1)$. In order for this to hold, the key condition on $\pi(\alpha|x)$ is the existence of a lower bound $\pi^{\mathrm{low}}_T(\alpha|x) > 0$, i.e. we impose the condition $\pi(\alpha|x) \geq \pi^{\mathrm{low}}_T(\alpha|x)$, or equivalently $\pi \in \Pi^{\mathrm{low}}_T$ in the notation introduced above. Further regularity conditions on the model and on $\pi^{\mathrm{low}}_T(\alpha|x)$ are presented in the appendix.

Theorem 4.1 (Consistency). *Let assumption B.1 be satisfied and let $N, T \to \infty$. Then we have*

$$\sup_{\pi\in\Pi^{\mathrm{low}}_T} \left\|\hat{\theta}(\pi) - \theta^0\right\| = o_p(1).$$

The proof of the theorem is based on relating the integrated likelihood function $L_{NT}(\theta,\pi)$ to the profile likelihood function $L^{\mathrm{p}}_{NT}(\theta)$. In general, as long as $\pi(\alpha|x)$ integrates to one, we have $L_{NT}(\theta,\pi) \leq L^{\mathrm{p}}_{NT}(\theta)$ (by the mean value theorem for integration). To prove the theorem we need to show that the opposite inequality also holds up to $o_p(1)$ in order to conclude that $L_{NT}(\theta,\pi) = L^{\mathrm{p}}_{NT}(\theta) + o_p(1)$, within a neighborhood of θ^0 and uniformly over $\pi \in \Pi^{\mathrm{low}}_T$. For this step, the lower bound on $\pi(\alpha|x)$ is required. Finally, under appropriate regularity conditions we know from the fixed effect panel literature that $L^{\mathrm{p}}_{NT}(\theta)$ has a well-seperated global maximum close to θ^0, so that the same must be true for $L_{NT}(\theta,\pi)$, which leads us to conclude $\hat{\theta}(\pi) = \theta^0 + o_p(1)$, uniformly over $\pi \in \Pi^{\mathrm{low}}_T$. For details we refer to the appendix.

Our assumptions allow the lower bound $\pi^{\mathrm{low}}_T(\alpha|x)$ to converge to zero at an arbitrary polynomial rate in T as $T \to \infty$, i.e. the constraint to have a lower bound for $\pi(\alpha|x)$ quickly becomes less and less restrictive as T increases.

It is crucial that the consistency result for $\hat{\theta}(\pi)$ holds uniformly over the set Π_T^{low}, since ultimately we want to consider $\hat{\theta}(\hat{\pi})$, where $\hat{\pi}$ is an estimator for π^0, i.e. is random. The uniform consistency results then shows that $\hat{\theta}(\hat{\pi})$ is consistent as long as $\hat{\pi}$ takes values in Π_T^{low}.

4.2 Score and Hessian of the Integrated Likelihood

Having established uniform consistency of $\hat{\theta}(\pi)$, we are now going to study the local properties of the integrated likelihood function $L_{NT}(\theta, \pi)$ as a function of θ around the true parameter θ^0, in particular the score and the Hessian of $L_{NT}(\theta, \pi)$ at θ^0. This local analysis of $L_{NT}(\theta, \pi)$ can then be combined with the uniform consistency result to derive the asymptotic properties of $\hat{\theta}(\pi)$.

While it is sufficient for consistency of $\hat{\theta}(\pi)$ to impose a lower bound on $\pi(\alpha|x)$ it now becomes crucial to also impose a smoothness condition on the density $\pi(\alpha|x)$ in the form of a generalized Lipschitz condition. To formulate the Lipschitz condition we require a distance measure in the space of individual effects $\mathcal{A}$. For each $x \in \mathcal{X}_T$ let $d_x : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ be a measurable non-negative function, such that $d_x(\alpha, \alpha) = 0$ for all $\alpha \in \mathcal{A}$. In many applications one might simply set $d_x(\alpha, \beta) = \|\alpha - \beta\|$. However, in general $d_x(\alpha, \beta)$ need neither be symmetric nor satisfy a triangle inequality. We define the following parameter set for π

$$\Pi_{T,\kappa}^{\text{lip}} = \left\{ \pi \in \Pi_T^{\mathcal{A}} \,\middle|\, \forall x \in \mathcal{X}_T,\, \forall \alpha, \beta \in \mathcal{A} \,:\, |\pi(\alpha|x) - \pi(\beta|x)| \leq \kappa_T \, \pi(\alpha|x) \, d_x(\alpha, \beta) \right\}. \tag{4.3}$$

This is the set of all conditional distributions that satisfy a Lipschitz type condition with generalized Lipschitz constant κ_T. In contrast to a standard Lipschitz condition, there also appears $\pi(\alpha|x)$ on the right hand side of the condition. This is natural here, since it is important to control the magnitude of relative changes of $\pi(\alpha|x)$ across α, as opposed to absolute changes. Locally the condition is simply a standard Lipschitz condition on $\log \pi(\alpha|x)$.

Theorem 4.2. *Consider the limit $N, T \to \infty$ and let assumption B.2(iii) and (vi) be satisfied. Let $\kappa_T > 0$ be a series such that $\kappa_T^{-1} = \mathcal{O}(1)$ and $\kappa_T = \mathcal{O}(T)$. Then*

$$\sup_{\pi_1, \pi_2 \in \Pi_{T,\kappa}^{\text{lip}}} \left\| \frac{\partial L_{NT}(\theta^0, \pi_1)}{\partial \theta} - \frac{\partial L_{NT}(\theta^0, \pi_2)}{\partial \theta} \right\| = \mathcal{O}_p\left(\frac{\kappa_T}{T}\right).$$

$$\sup_{\pi_1, \pi_2 \in \Pi_{T,\kappa}^{\text{lip}}} \left\| \frac{\partial^2 L_{NT}(\theta^0, \pi_1)}{\partial \theta \partial \theta'} - \frac{\partial^2 L_{NT}(\theta^0, \pi_2)}{\partial \theta \partial \theta'} \right\| = \mathcal{O}_p\left(\frac{\kappa_T}{\sqrt{T}}\right).$$

Again, it is crucial that the result of Theorem 4.2 holds uniformly over π, since we are ultimately interested in applications where π is replaced by the estimator $\hat{\pi}$. We focus on the result for the score. The theorem implies that all scores $\partial L_{NT}(\theta^0, \pi)/\partial\theta$ for different $\pi(\alpha|x)$ become asymptotically close to each other as long $\pi(\alpha|x)$ satisfies a Lipschitz type condition with $\kappa_T = o(T)$. The most interesting application of the theorem is obtained by setting $\pi_1 = \pi^0$. The score $\partial L_{NT}(\theta^0, \pi)/\partial\theta$ is unbiased for $\pi = \pi^0$. Assuming that $\pi^0 \in \Pi^{\text{lip}}_{T,\kappa}$, we can thus conclude from the theorem that the bias of the score of the integrated likelihood is at most of order κ_T/T for all $\pi \in \Pi^{\text{lip}}_{T,\kappa}$.

The bound on the bias of the score provided by Theorem 4.2 is independent of π. However, we clearly expect the bias of the score to be small whenever π is close to the true distribution π^0, and in the following we formalize this intuition. To do so, we decompose the score of the integrated likelihood as follows

$$\frac{\partial L_{NT}(\theta^0, \pi)}{\partial\theta} = \frac{\partial L_{NT}(\theta^0, \pi^0)}{\partial\theta} + \frac{\partial \overline{L}_{NT}(\theta^0, \pi)}{\partial\theta} + \frac{\nu_{NT}(\pi)}{\sqrt{NT}}, \tag{4.4}$$

where $\overline{L}_{NT}(\theta, \pi)$ is the expected likelihood conditional on the regressors $X_1, \ldots, X_N$, which was introduced in equation (3.6), and $\nu_{NT}(\pi)$ is given by

$$\begin{aligned}
\nu_{NT}(\pi) &= \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \nu_{NT,i}(\pi) \\
\nu_{NT,i}(\pi) &= \frac{\partial \log\, f_{Y|X}(Y_i|X_i; \theta^0, \pi)}{\partial\theta} - \frac{\partial \log\, f_{Y|X}(Y_i|X_i; \theta^0, \pi^0)}{\partial\theta} \\
&\qquad - \mathbb{E}\left[\frac{\partial \log\, f_{Y|X}(Y_i|X_i; \theta^0, \pi)}{\partial\theta} \middle| X_i \right].
\end{aligned} \tag{4.5}$$

In equation (4.4) the first term of the decomposition is the score at π^0, which is unbiased, i.e. the bias of the integrated likelihood score originates from the remaining two terms. In the following we provide bounds on these two terms.

Theorem 4.3. *Let $\kappa_T > 0$ be a series such that $\kappa_T^{-1} = \mathcal{O}(1)$ and $\kappa_T = \mathcal{O}(T)$, and assume that there exist an upper bound $\pi_T^{\text{up}}(\alpha|x)$ such that $\pi^0(\alpha|x) \leq \pi_T^{\text{up}}(\alpha|x)$ for all $\alpha \in \mathcal{A}$ and $x \in \mathcal{X}_T$. Let assumption B.2(iv) be satisfied and consider the limit $N, T \to \infty$. Then there exists a series of random variables $C_T > 0$ with $C_T = \mathcal{O}_p(1)$ such that for all $\pi \in \Pi^{\text{lip}}_{T,\kappa} \cap \Pi^{\text{up}}_T$*

$$\left\| \frac{\partial \overline{L}_{NT}(\theta^0, \pi)}{\partial\theta} \right\| \leq C_T \, \frac{\kappa_T}{T} \, \mathcal{D}_{\text{H}}(\pi, \pi^0).$$

Note that C_T is independent of π, so that the theorem can also be applied when an estimator $\hat{\pi}$ is plugged in. The notation Π_T^{up} for the set of all conditional probability densities that are bounded from above by $\pi_T^{\text{up}}(\alpha|x)$ was introduced earlier.

At the true parameters we know that $\partial \overline{L}_{NT}(\theta^0, \pi^0)/\partial\theta = 0$, and one expects intuitively that the expected score $\partial \overline{L}_{NT}(\theta^0, \pi)/\partial\theta$ should be close to zero whenever π is close to π^0. Theorem 4.3 formalizes this intuition and shows that an appropriate distance measure for the individual effect distributions is the Hellinger distance.

Next, we discuss the asymptotic properties of $\nu_{NT}(\pi)$, which is the sum over the difference of individual score functions minus the mean of this difference (the score at π^0 is mean zero, so this term does not appear explicitly in equation (4.5)).

Lemma 4.4. *Let assumption B.2(i) and (ii) be satisfied, consider the limit $N, T \to \infty$, and let $\kappa_T > 0$ be a series such that $\kappa_T^{-1} = \mathcal{O}(1)$. For all series π_T with $\pi_T \in \Pi_{T,\kappa}^{\text{lip}}$ we have $\mathbb{E}\left[(\nu_{NT}(\pi_T))^2\right] = \mathcal{O}\left(\kappa_T^2/T\right)$, and therefore $\nu_{NT}(\pi_T) = \mathcal{O}_p\left(\kappa_T/\sqrt{T}\right)$.*

From this lemma we conclude that $\nu_{NT}(\pi) = o_p(1)$ for those π that satisfy a generalized Lipschitz condition with $\kappa_T = o(\sqrt{T})$.[4] However, the lemma does not make a uniform statement over all π that satisfy this condition. Therefore we cannot apply the lemma when π is replaced by the estimator $\hat{\pi}$. Another way of seeing that there is a complication here is to realize that the result $\nu_{NT}(\pi) = \mathcal{O}_p\left(\kappa_T/\sqrt{T}\right)$ is derived in the lemma by evaluating the second moment of $\nu_{NT}(\pi)$, using the fact that the $\nu_{NT,i}(\pi)$ are mean zero and independent across i conditional on the regressors. This, however, only holds for non-stochastic π and the argument breaks down when the estimator $\hat{\pi}$ is plugged in. The problem of generalizing a pointwise convergence or boundedness result to the corresponding uniform result is well-known in the literature on empirical processes (see e.g. Andrews (1994) for a review). The key in going from the pointwise to the uniform result is to impose a condition (e.g. a stochastic equicontinuity condition) that guarantees that the parameter set (in our case Π_T) is sufficiently "small" in an appropriate sense. This problem is therefore directly related to our discussion in Section 3.1, where we have argued that the space of conditional densities $\pi(\alpha|x)$ is too "large" for our purposes and needs to be restricted. Instead of making such a restriction explicit here, we impose the following high-level assumption, which has the advantage that our results in this section can be applied under various restrictions on Π_T that satisfy this assumption.

Assumption 4.1. $\sup_{\pi \in \Pi_T} \|\nu_{NT}(\pi)\| = o_p(1)$.

[4] In Lemma 4.4 we make the T-dependence of the conditional densities $\pi(\alpha|x)$ explicit (note that $x \in \mathcal{X}_T$, i.e. the dimension of x changes with T), while we usually suppress this T-dependence.

Using this assumption and applying the preceding results of this section, we obtain the following corollary.

Corollary 4.5. *Let $\kappa_T > 0$ be a series such that $\kappa_T^{-1} = \mathcal{O}(1)$ and $\kappa_T = o(\sqrt{T})$, and assume that an upper bound $\pi_T^{\mathrm{up}}(\alpha|x)$ exists such that $\pi^0(\alpha|x) \leq \pi_T^{\mathrm{up}}(\alpha|x)$ for all $\alpha \in \mathcal{A}$ and $x \in \mathcal{X}_T$. Let assumption B.2 and 4.1 be satisfied and consider the limit $N, T \to \infty$. Let $\hat{\pi}$ be an estimator that takes values in $\Pi_T \subset \Pi_{T,\kappa}^{\mathrm{lip}} \cap \Pi_T^{\mathrm{up}}$ for all T. Then we have*

$$\hat{\theta}(\hat{\pi}) - \hat{\theta}(\pi^0) = \mathcal{O}_p\left(\frac{\kappa_T}{T}\right)\mathcal{D}_{\mathrm{H}}(\hat{\pi}, \pi^0) + o_p\left(\frac{1}{\sqrt{NT}}\right).$$

This result holds independently of whether $\hat{\pi}$ is the joint maximum likelihood estimator introduced in equation (3.3), i.e. it can also be applied for alternative estimation procedures for the individual effect distribution. It is interesting to compare Corollary 4.5 to Theorem 4 in Arellano and Bonhomme (2009), which makes a very similar statement. They use the L^2 Kullback Leibler loss instead of the Hellinger distance and they assume a parametric specification for π and impose a random effect assumption. The key difference, however, is that they need to impose that N/T converges to a constant, while we impose no restriction on how N and T go infinity. This means that our result can be applied to discuss higher order bias correction of $\hat{\theta}(\hat{\pi})$.

Note that $\hat{\theta}(\pi^0)$ is asymptotically unbiased and that the term $o_p\left(1/\sqrt{NT}\right)$ can be ignored in the limiting distribution of $\hat{\theta}(\hat{\pi})$ since it is small compared to the asymptotic standard error of $\hat{\theta}(\hat{\pi})$. Thus, according to corollary 4.5 the asymptotic bias of $\hat{\theta}(\hat{\pi})$ crucially depends on the asymptotics of $\mathcal{D}_{\mathrm{H}}(\hat{\pi}, \pi^0)$, i.e. how fast $\hat{\pi}$ converges to the true distribution in terms of the Hellinger distance. This is what we are going to study next.

4.3 Joint Maximum Likelihood Estimation of θ and π

We now want to study the properties of the joint maximum likelihood estimators $\hat{\theta}$ and $\hat{\pi}$ that were introduced in (3.3). The profile likelihood $L_{NT}(\pi)$ was defined in (3.2). A Taylor expansion in θ gives

$$\begin{aligned} L_{NT}(\pi) = L_{NT}(\theta^0, \pi) &+ (\hat{\theta}(\pi) - \theta^0)' \frac{\partial L_{NT}(\theta^0, \pi)}{\partial \theta} \\ &+ \frac{1}{2}(\hat{\theta}(\pi) - \theta^0)' \frac{\partial L_{NT}(\theta^0, \pi)}{\partial \theta \partial \theta'}(\hat{\theta}(\pi) - \theta^0) + \mathcal{O}_p\left(\|\hat{\theta}(\pi) - \theta^0\|^3\right). \quad (4.6) \end{aligned}$$

Using our results on $\hat{\theta}(\pi)$ as well as the bounds on the score and Hessian from the last subsection it is easy to provide appropriate bounds on the terms in $L_{NT}(\pi)$ that involve $\hat{\theta}(\pi)$. The new task here is to handle the term $L_{NT}(\theta^0, \pi)$. We have

$$L_{NT}(\theta^0, \pi) = L_{NT}(\theta^0, \pi^0) - \frac{1}{T}\,\mathcal{D}_{\mathrm{KL}}\left(f_Y(\pi)||f_Y(\pi^0)\right) + \frac{1}{T\sqrt{N}}\psi(\pi), \tag{4.7}$$

where we have defined

$$\begin{aligned}\psi_{NT}(\pi) &= \sqrt{N}T\left[L_{NT}(\theta^0, \pi^0) - L_{NT}(\theta^0, \pi) - \overline{L}_{NT}(\theta^0, \pi^0) + \overline{L}_{NT}(\theta^0, \pi)\right] \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\log\frac{f_{Y|X}(Y_i|X_i;\theta^0,\pi^0)}{f_{Y|X}(Y_i|X_i;\theta^0,\pi)} - \mathbb{E}\left[\log\frac{f_{Y|X}(Y_i|X_i;\theta^0,\pi^0)}{f_{Y|X}(Y_i|X_i;\theta^0,\pi)}\middle|X_i\right]\right\}.\end{aligned} \tag{4.8}$$

Note that $\psi_{NT}(\pi)$ and $\nu_{NT}(\pi)$ are closely related, namely $\nu_{NT}(\pi)$ is obtained from $\psi_{NT}(\pi)$ by taking the derivative with respect to θ^0 and multiplying with minus $T^{-1/2}$. Both are zero mean empirical processes with index π. Analogously to $\nu_{NT}(\pi)$ we can bound $\psi_{NT}(\pi)$ pointwise by evaluating the second moment, as stated in the following lemma.

Lemma 4.6. *Let assumption B.3 be satisfied, let $\pi^0 \in \Pi_T^{\mathrm{up}}$ and let $\pi_T \in \Pi_T^{\mathrm{low}}$. Then we have $\psi_{NT}(\pi_T) = \mathcal{O}_p(1)\sqrt{\mathcal{D}_{\mathrm{KL}}(f_Y(\pi_T)||f_Y(\pi^0)}$.*

However, just as we discussed for $\nu_{NT}(\pi)$ above the pointwise bound is not sufficient for our purposes and we impose a high-level assumption to guarantee the uniform bound. Satisfying this assumption again requires to constraint the parameter set Π_T appropriately.

Assumption 4.2. *Let there exist $\kappa_T > 0$ with $\kappa_T^{-1} = \mathcal{O}(1)$ and $\kappa_T = o(\sqrt{T})$, and a series of random numbers $c_T = o_p(1)$ such that $\Pi_T \subset \Pi_{T,\kappa}^{\mathrm{lip}}$ and $\forall \pi \in \Pi_T$:*

$$|\psi_{NT}(\pi)| \leq c_T\,\frac{\sqrt{T}}{\kappa_T}\sqrt{\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))}.$$

Apart from $\psi_{NT}(\pi)$, the other term on the right hand side of equation (4.7) which depends on π is $\mathcal{D}_{\mathrm{KL}}\left(f_Y(\pi)||f_Y(\pi^0)\right)$. If we consider the population level in the cross-sectional dimension and assume θ^0 is known, then maximizing the expected likelihood over π is equivalent to minimizing the Kullback Leibler divergence $\mathcal{D}_{\mathrm{KL}}\left(f_Y(\pi)||f_Y(\pi^0)\right)$. However, the model may not be identified, i.e. minimizing $\mathcal{D}_{\mathrm{KL}}\left(f_Y(\pi)||f_Y(\pi^0)\right)$ may not necessarily give $\pi = \pi^0$. It is, however, crucial for our discussion that minimizing $\mathcal{D}_{\mathrm{KL}}\left(f_Y(\pi)||f_Y(\pi^0)\right)$ over $\pi \in \Pi_T$ yields a small value of $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0)$ at least for sufficiently large values of T, because in the last section we

have shown that the convergence rate of $\mathcal{D}_{\mathrm{H}}(\hat{\pi}, \pi^0)$ determines the rate at which the bias of $\hat{\theta}$ converges to zero as $T \to \infty$. This motivates the following assumption.

Assumption 4.3. *Let there exists $\mu_T = o(1)$, and a series of random numbers $C_T = \mathcal{O}_p(1)$, such that*

(i) $\inf_{\pi \in \Pi_T} \mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0)) = \mathcal{O}_p(\mu_T^2)$.

(ii) $\forall \pi \in \Pi_T$: $\quad \mathcal{D}_{\mathrm{H}}(\pi, \pi^0) \leq C_T \left[\sqrt{\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))} + \mu_T \right]$.

Assumption 4.3(i) demands that the true distribution of the dependent variable Y can be approximated better and better (as $T \to \infty$) in terms of the Kullback Leibler divergence by a distribution of Y that is implied by some $\pi \in \Pi_T$. This assumption is trivially satisfied if $\pi^0 \in \Pi_T$, but unless we are willing to assume a parametric form for π it is not reasonable to expect $\pi^0 \in \Pi_T$. For the semiparametric approach we have to choose Π_T such that every distribution π^0 that satisfies certain regularity conditions can be approximated well by an element in Π_T. The rate of convergence of this approximation of π^0 is given by μ_T. This rate crucially depends on the (smoothness) properties of π^0.

In this sense Assumption 4.3(i) requires Π_T to be sufficiently "large". In contrast, assumption 4.3(ii) demands Π_T to be sufficiently "small", namely it should not contain ill-behaved distributions for which $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))$ is small, while $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0)$ is not. Note that μ_T appears on the right hand side of Assumption 4.3(ii), i.e. the assumption allows for the possibility that $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))$ is (close to) zero but $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0)$ is not, which is important since the model may not be identified at finite T. However, as T becomes large the assumption requires this identification problem to vanish at the rate μ_T.

We have $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0) \leq \sqrt{\mathcal{D}_{\mathrm{KL}}(\pi||\pi^0)}$, which is a general relation between Hellinger distance and Kullback Leibler divergence. Thus, a slightly stronger form of Assumption 4.3(ii) is obtained by replacing $\mathcal{D}_{\mathrm{H}}(\pi, \pi^0)$ with $\sqrt{\mathcal{D}_{\mathrm{KL}}(\pi||\pi^0)}$ on the left hand side of the assumption. This is interesting, because by the "chain rule for the Kullback Leibler divergence" we have $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0)) \leq \mathcal{D}_{\mathrm{KL}}(\pi||\pi^0)$. Thus, Assumption 4.3($ii$) essentially requires that this general inequality can be reverted for $\pi \in \Pi_T$, allowing for some random pre-factor C_T and some "slackness" μ_T.

Theorem 4.7. *Let assumption 4.1, 4.2, 4.3 and B.2 be satisfied. Then we have*

(i) $\quad \mathcal{D}_{\mathrm{H}}(\hat{\pi}, \pi^0) = \mathcal{O}_p(\mu_T) + o_p\left(\frac{1}{\kappa_T}\sqrt{\frac{T}{N}}\right)$,

(ii) $\quad \hat{\theta} - \hat{\theta}(\pi^0) = \mathcal{O}_p\left(\frac{\kappa_T\,\mu_T}{T}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).$

We want to give some intuition on the proof of this theorem. Consider $\overline{\pi} = \operatorname{argmax}_{\pi\in\Pi_T} \overline{L}_{NT}(\theta^0, \pi)$, which is an infeasible "estimator" for π^0 based on the expected likelihood and on knowledge of θ^0. It is easy to see that Assumption 4.3 implies $\mathcal{D}_{\mathrm{H}}(\overline{\pi}, \pi^0) = \mathcal{O}_p(\mu_T)$. Part (i) of Theorem 4.7 generalizes this result to the feasible estimator $\hat{\pi}$. The fact that $\hat{\pi}$ is obtained from $L_{NT}(\theta, \pi)$ instead of the expected likelihood $\overline{L}_{NT}(\theta, \pi)$ can be controlled by the decomposition (4.7) and assumption 4.2, and results in the additional term $o_p\left(\frac{1}{\kappa_T}\sqrt{\frac{T}{N}}\right)$. This terms also accounts for the fact that θ^0 is not known and $\hat{\theta}(\pi)$ is used instead — the results of the previous subsection are crucial to show this. The second part of Theorem 4.7 then follows directly from the first part and Corollary 4.5 above. For the details we refer to the appendix.

Part two of Theorem 4.7 provides a bound on the difference between the feasible maximum likelihood estimator $\hat{\theta}$ and the infeasible "miracle" maximum likelihood estimator $\hat{\theta}(\pi^0)$ that could be obtained if the distribution of the individual effects were known. Under appropriate regularity conditions one can show that $\hat{\theta}(\pi^0)$ is asymptotically normal and unbiased. Unbiasedness stems from the fact that the score $\partial L_{NT}(\theta^0, \pi)/\partial\theta$ is unbiased for $\pi = \pi^0$. Asymptotic normality can, for example, be shown by relating $\hat{\theta}(\pi^0)$ to the fixed effect estimator $\hat{\theta}^p$, whose asymptotic theory is well-studied in the large T literature. For the sake of generality, we formulate this as an assumption, which can be justified in different ways.

Assumption 4.4. *As $N, T \to \infty$ we have $\sqrt{NT}(\hat{\theta}(\pi^0) - \theta^0) \;\to_d\; \mathcal{N}\left(0,\, \mathcal{I}_0^{-1}\right)$, for a positive definite $K \times K$ matrix $\mathcal{I}_0$.*

From Theorem 4.7(ii) we then obtain the following corollary.

Corollary 4.8. *Let the assumptions of Theorem 4.1 and 4.7 as well as Assumption 4.4 be satisfied, and let $N = o\left(T/(\mu_T^2\kappa_T^2)\right)$ for $N, T \to \infty$. Then we have*

$$\sqrt{NT}(\hat{\theta} - \theta^0) \;\underset{d}{\to}\; \mathcal{N}\left(0,\, \mathcal{I}_0^{-1}\right).$$

Thus, as long as N does not grow too fast relative to T as $N, T \to \infty$, we find that $\hat{\theta}$ is asymptotically unbiased. How rapidly N is allowed to go to infinity relative to T is specified by μ_T and κ_T, the details of which depend on the particular low-level assumptions that are made to justify the high-level assumption in this section. One concrete example for this is discussed in the following.

5 Generalized Random Effects

To apply the results of the last section one needs to satisfy the high-level assumptions 4.1, 4.2 and 4.3. Assumptions 4.1 and 4.2 demand specific uniform bounds on the empirical processes $\nu_{NT}(\pi)$ and $\psi_{NT}(\pi)$. We gave conditions under which these bounds are satisfied pointwise in Lemma 4.4 and 4.6, but in order to show the uniform result over $\pi \in \Pi_T$ one needs to impose restrictions on the parameter set of allowed individual effect distributions Π_T. Since assumption 4.3(i) demands the true distribution $\pi^0(\alpha|x)$ to be well approximated by an element in Π_T,[5] the restrictions on Π_T require us to also impose restrictions on π^0. In particular, the unrestricted correlated random effect model, where apart from basic regularity conditions no restriction is imposed on $\pi(\alpha|x)$, is ruled out by these high-level assumptions, unless the regressor domain $\mathcal{X}_T$ is discrete and only contains a small number of elements relative to N.[6] If $\mathcal{X}_T$ contains too many elements, then the set of conditional distributions $\pi(\alpha|x)$ is too rich, so that the uniform bounds in assumption 4.1 and 4.2 cannot be satisfied. In other words, we are facing a curse of dimensionality problem in estimating the conditional distribution $\pi(\alpha|x)$ if the domain of the conditioning variables is too large.

In order to overcome this problem one needs to restrict the correlation structure between the individual effects and the regressors. Different ways of doing this are conceivable. Here, we want to explore the generalized random effect restriction, which assumes a partitioning of $\mathcal{X}_T$ into groups and imposes a random effect assumption within each group. Let $\mathcal{G}_T$ be a known partition of $\mathcal{X}_T$ into G_T groups. We assume that the distributions $\pi(\alpha|x)$ and $\pi(\alpha|\tilde{x})$ are identical if x and $\tilde{x}$ belong to the same group, i.e. if $x, \tilde{x} \in g$ with $g \in \mathcal{G}_T$. The set of distributions for which this constraint holds is given by

$$\Pi_T^{\mathcal{G}} = \left\{ \pi \in \Pi_T^{\mathcal{A}} \,\middle|\, \forall g \in \mathcal{G}_T, \forall x, \tilde{x} \in g, \forall \alpha \in \mathcal{A} : \ \pi(\alpha|x) = \pi(\alpha|\tilde{x}) \right\}. \tag{5.1}$$

We assume $\pi^0 \in \Pi_T^{\mathcal{G}}$, i.e. the true distribution satisfied the generalized random effect assumption, and we restrict the parameter set Π_T to be a subset of $\Pi_T^{\mathcal{G}}$. The generalized random effect assumption reduces the dimension of Π_T dramatically, compared to the unrestricted correlated random effect case. Once this assumption

[5] Actually, assumption 4.3(i) only requires the outcome variable distributions implied by π^0 and the one implied by some $\pi \in \Pi_T$ to be close to each other, but the easiest way to satisfy this assumption is to show that π^0 itself can be approximated well by $\pi \in \Pi_T$ in terms of Kullback Leibler distance.

[6] A small dimensional continuous support $\mathcal{X}_T$ can also be admissible, as long as a smoothness assumption of $\pi(\alpha|x)$ as a function of x is imposed.

and some further regularity conditions on Π_T are imposed one can use methods from empirical process theory to show that the uniform bounds in Assumption 4.1 and 4.2 are satisfied as long as the number of groups G_T increases sufficiently slowly as $N, T \to \infty$.

5.1 Imposing an Appropriate Smoothness Assumption

We now want to discuss how to choose $\Pi_T \subset \Pi_T^{\mathcal{G}}$ such that Assumption 4.3 is satisfied, and which rates μ_T and κ_T can be obtained. Assumption 4.3(ii) is an approximate identification condition for π within Π_T. For given θ^0,[7] the assumption demands π to be close to π^0 (in terms of Hellinger distance) whenever the distributions for the outcome variable Y that are implied by π and π^0 are close to each other (in terms of Kullback Leibler divergence). However, this identification of π from the distribution of Y is only required approximately, since μ_T appears as a slackness parameter in the assumption. Only as T becomes large the slackness μ_T converges to zero, so that the individual effect distribution is identified in the limit $T \to \infty$.

Determining the distribution $\pi(\alpha|x)$ from the distribution of Y is an ill-posed inverse problem for fixed T, as discussed e.g. in Bonhomme (2010). One way to understand why this problem is solved asymptotically as T becomes large is to ask how one could estimate $\pi(\alpha|x)$ within the fixed effect approach. The fixed effect approach starts from the realizations α_i^0 for the individual effects of each cross-sectional unit, which are distributed according to $\pi(\alpha|X_i)$. If the realizations α_i^0 would be observed, then, once the generalized random effect assumption is imposed, one could estimate $\pi(\alpha|x)$ consistently for fixed T as $N \to \infty$ by using a standard kernel density estimation within each group $g \in \mathcal{G}_T$. In reality the α_i^0 are not observed, but the fixed effect approach provides estimators $\hat{\alpha}_i^{\mathrm{p}}$, which are obtained from maximizing the profile likelihood jointly with the parameters of interest. As T becomes large, we have $\hat{\alpha}_i^{\mathrm{p}} = \alpha_i^0 + \mathcal{O}_p(T^{-1/2})$. Thus, a Kernel density estimator over $\hat{\alpha}_i^{\mathrm{p}}$ within each group provides an estimator for $\pi(\alpha|x)$ which is consistent as both N and T become large.

We do not consider the Kernel density estimator over $\hat{\alpha}_i^{\mathrm{p}}$ further, because it does not allow for higher order bias correction (it gives some fixed convergence rate of $\mathcal{D}_{\mathrm{H}}(\hat{\pi}, \pi^0)$ in T, which is not optimal). However, apart from the fact that $\pi(\alpha|x)$ can be consistently estimated as $N, T \to \infty$, this estimator illustrates another very

[7] Note that θ^0 enters in the definition of $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0)$.

important point: For a given value of T, the model intrinsically determines the "resolution" of the individual effects. In the fixed effect approach, this "resolution" is described by the variance of $\hat{\alpha}_i^{\mathrm{p}}$. For large values of T the variance of $\hat{\alpha}_i^{\mathrm{p}}$ is approximated by $T^{-1}\mathcal{I}^{-1}(\alpha_i^0, \theta^0, X_i)$, where $\mathcal{I}(\alpha, \theta, x)$ is the information matrix of the model with respect to the individual effects, namely

$$\begin{aligned}\mathcal{I}(\alpha, \theta, x) &= -\,\mathbb{E}\left[\frac{1}{T}\,\frac{\partial^2 \log f(Y_i|X_i, \alpha; \theta)}{\partial\alpha\partial\alpha'}\bigg| X_i = x,\, \boldsymbol{\alpha} = \alpha,\, \theta\right] \\ &= -\,\frac{1}{T}\int_{y\in\mathcal{Y}_T} \frac{\partial^2 \log f(y|x, \alpha; \theta)}{\partial\alpha\partial\alpha'}\, f(y|x, \alpha; \theta)\, dy\,. \qquad (5.2)\end{aligned}$$

This information matrix also plays an important role for our maximum likelihood estimator of $\pi(\alpha|x)$, and for the question how to choose Π_T such that assumption 4.3(ii) is satisfied. To understand this, consider a quadratic expansion of $\log f(Y_i|X_i, \alpha; \theta)$ in α around its maximum value $\hat{\alpha}_i^{\mathrm{p}} = \hat{\alpha}_i^{\mathrm{p}}(\theta) = \operatorname{argmax}_\alpha f(Y_i|X_i, \alpha; \theta)$. For large T we have

$$\begin{aligned}f(Y_i|X_i, \alpha; \theta) &\approx f(Y_i|X_i, \hat{\alpha}_i^{\mathrm{p}}; \theta)\ \exp\left[\frac{1}{2}\,(\alpha - \hat{\alpha}_i^{\mathrm{p}})'\,\frac{\partial^2 \log f(Y_i|X_i, \hat{\alpha}_i^{\mathrm{p}}; \theta)}{\partial\alpha\partial\alpha'}\,(\alpha - \hat{\alpha}_i^{\mathrm{p}})\right] \\ &\approx f(Y_i|X_i, \hat{\alpha}_i^{\mathrm{p}}; \theta)\ \exp\left[-\frac{T}{2}\,(\alpha - \hat{\alpha}_i^{\mathrm{p}})'\,\mathcal{I}(\hat{\alpha}_i^{\mathrm{p}}, \theta, X_i)\,(\alpha - \hat{\alpha}_i^{\mathrm{p}})\right].\end{aligned}$$

In the last line we used that under appropriate regularity conditions the Hessian converges to its expected value as T becomes large. Thus, the functional form of $f(Y_i|X_i, \alpha; \theta)$ in α approximates a non-normalized multivariate normal pdf with mean $\hat{\alpha}_i^{\mathrm{p}}$ and variance $\mathcal{I}^{-1}(\hat{\alpha}_i^{\mathrm{p}}, \theta, X_i)/T$. This variance therefore describes how fast $f(Y_i|X_i, \alpha; \theta)$ is varying as a function of α, which is important, since according to equation (2.1) the distribution of Y_i given X_i only depends on $\pi(\alpha|X_i)$ via integration over α, with $f(Y_i|X_i, \alpha; \theta)$ also appearing in the integrand. Therefore, variations in $\pi(\alpha|X_i)$ over α that are more rapid than the variations in $f(Y_i|X_i, \alpha; \theta)$ over α will tend to be "washed out" by the integration, i.e. very rapid fluctuations in $\pi(\alpha|X_i)$ have very little effect on the distribution of Y_i given X_i.

Thus, if we do not impose an appropriate smoothness assumption on $\pi(\alpha|x)$, then $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))$ can be close to zero, while π and π^0 are quite different. How smooth $\pi(\alpha|x)$ needs to be at a particular value of α is determined by $\mathcal{I}^{-1}(\alpha, \theta^0, x)/T$.

To specify an appropriate parameter set Π_T that accounts for this smoothness requirement, we first need to introduce some further notation. Let $\phi(\alpha; \beta, \Omega)$ be the multivariate normal pdf with mean β and variance matrix Ω. For $x \in \mathcal{X}_T$ and

$\alpha, \beta \in \mathcal{A}$ we define the kernel function

$$\mathcal{K}_T^{\Omega}(\alpha, \beta; x) = \frac{\phi(\alpha; \beta, \Omega_T(\beta, x))}{\int_{\mathcal{A}} \phi(\gamma; \beta, \Omega_T(\beta, x))\, d\gamma}, \tag{5.3}$$

where $\Omega_T(\beta, x)$ is a positive definite $M \times M$ matrix for all values of T, β and x. In order to be compatible with the generalized random effect assumption we impose the restriction $\Omega_T(\beta, x) = \Omega_T(\beta, \tilde{x})$ for all x, $\tilde{x}$ in the same group. When the kernel $\mathcal{K}_T^{\Omega}$ is applied to $\pi \in \Pi_T^{\mathcal{G}}$ one obtains another conditional distribution in $\Pi_T^{\mathcal{G}}$, namely

$$\left[\mathcal{K}_T^{\Omega}\, \pi\right](\alpha|x) = \int_{\mathcal{A}} \mathcal{K}_T^{\Omega}(\alpha, \beta; x)\, \pi(\beta|x)\, d\beta. \tag{5.4}$$

Using this kernel function we now define the parameter set of individual effects distributions as follows

$$\Pi_T = \mathcal{K}_T^{\Omega} \left(\Pi_T^{\mathcal{G}} \cap \Pi_T^{\text{up}} \cap \Pi_T^{\text{low}}\right). \tag{5.5}$$

This is the set of all distributions that can be generated by applying the Kernel $\mathcal{K}_T^{\Omega}$ to an element of $\Pi_T^{\mathcal{G}} \cap \Pi_T^{\text{up}} \cap \Pi_T^{\text{low}}$ (the set of distribution that satisfy the generalized random effect assumption as well as some appropriate upper and lower bound). We assume $\pi^0 \in \Pi_T^{\text{up}} \cap \Pi_T^{\text{low}}$. The parameter set satisfies $\Pi_T \subset \Pi_T^{\mathcal{G}}$. Here we impose upper and lower bound restrictions for technical reasons. In our Monte Carlo simulations below we do not impose any upper or lower bound restriction on $\pi(\alpha|x)$, i.e. we simply use $\Pi_T = \mathcal{K}_T^{\Omega}\, \Pi_T^{\mathcal{G}}$, which turns out not to affect the good performance of the estimators.

Since applying $\mathcal{K}_T^{\Omega}$ is a Gaussian kernel smoothing with variance $\Omega_T(\alpha, x)$, the smoothness of the distributions in Π_T depends on the choice of variance matrix $\Omega_T(\alpha, x)$. The larger $\Omega_T(\alpha, x)$, the smoother are the distributions in Π_T. Motivated by the above discussion, we choose

$$\Omega_T(\alpha, x) = \frac{\rho_T}{T} \left[\frac{1}{N_{g(x)}} \sum_{\left\{i \in \{1,\ldots,N\} \,:\, X_i \in g(x)\right\}} \mathcal{I}(\alpha, \tilde{\theta}, X_i)\right]^{-1}, \tag{5.6}$$

where $\tilde{\theta}$ is some preliminary consistent estimator for θ^0, e.g. the fixed effect maximum likelihood estimator, $\rho_T > 0$ is a scalar bandwidth parameter, $g(x) \in \mathcal{G}_T$ denotes to the group to which $x \in \mathcal{X}_T$ belongs (i.e. $x \in g(x)$), and $N_{g(x)}$ is the number of observed individuals in this group, which is also the set of individuals that are summed over in (5.6). For the bandwidth we require

$$\frac{\rho_T}{\log T} \to \text{const.} > 0, \qquad \text{as } N, T \to \infty. \tag{5.7}$$

We assume that $\pi^0(\alpha|x)$ satisfied the generalized random effect assumption and is r times continuously differentiable in α with bounded derivatives, where $r \geq 1$. Under appropriate further regularity conditions one can then show that Assumption 4.3 is satisfied for the above choice of Π_T and Ω_T with

$$\kappa_T = \sqrt{\frac{T}{\log T}}\,, \qquad \mu_T = \left(\frac{\log T}{T}\right)^{r/2}. \tag{5.8}$$

By theorem 4.7 the bias of $\hat{\theta}$ converges at the rate $\kappa_T \mu_T / T$, which now equals $T^{-1}\,[(\log T)/T]^{(r-1)/2}$. Thus, the smoother π^0 is, the faster the bias of $\hat{\theta}$ converges to zero.

Alternative choices for the parameter set Π_T are conceivable. The advantage of defining Π_T as the image of a Gaussian kernel smoothing operator is that the smoothness properties of $\pi(\alpha|x)$ can be controlled separately for each value of α and within each group in terms of the variance matrix $\Omega_T(\alpha, x)$, which is related to the information matrix. The individual effect distributions in Π_T are simply infinite mixtures of normal distributions with different means and specified variances. Further technical details and motivation for this construction of Π_T are discussed in the appendix.

5.2 Computation

In contrast to standard choices for a non-parametric sieve space, the parameter set Π_T defined in (5.5) is still infinite dimensional. However, the numerical implementation of the estimator requires to discretize Π_T. A convenient method for doing this is to discretize the support $\mathcal{A}$ of the individual effects. This discretization is to be chosen such that it does not affect the properties of the estimator. As explained above, one can interpret the standard error of the fixed effect estimator for the individual effects (which for large T is given by the square root of the diagonal elements of $T^{-1}\mathcal{I}^{-1}(\alpha, \theta^0, x)$) as the "resolution" in the space $\mathcal{A}$ that is implied by the model. The numerical error due to the discretization of $\mathcal{A}$ will be small as long as the step-size for the discretization of $\mathcal{A}$ is chosen sufficiently small relative to this standard error. In particular, it is natural here to choose a different discretization step-size for different values of α and different groups $g \in \mathcal{G}_T$. For example, in our Monte Carlo simulations below the set $\mathcal{A}$ is one-dimensional and we choose the discretization step-size for given α and $x \in g$ equal to 1/6 of the approximate standard error $\sqrt{T^{-1}\mathcal{I}^{-1}(\alpha, \tilde{\theta}, x)}$, using some preliminary consistent estimator $\tilde{\theta}$ of θ^0. We verified that the choice of 1/6 did not affect the performance of the estimator in

that case, i.e. a smaller discretization yields essentially the same estimator for the parameters of interest.[8] If the set $\mathcal{A}$ is unbounded, then the discretization requires to impose some bounds, which can however be chosen very broadly.

The discretization of $\mathcal{A}$ and the calculation of the Kernel variance $\Omega_T(\alpha, x)$ both require evaluation of the information matrix $\mathcal{I}(\alpha, \theta, x)$, which involves integration over $y \in \mathcal{Y}_T$. Unless the model allows for an analytical evaluation of $\mathcal{I}(\alpha, \theta, x)$, it will often be easiest to perform a Monte Carlo integration to determine $\mathcal{I}(\alpha, \theta, x)$, i.e. to draw many y from the distribution $f(y|x, \alpha; \theta)$ and to use minus $1/T$ times the sample average of the corresponding Hessians as an approximation of $\mathcal{I}(\alpha, \theta, x)$. It is sufficient for our procedure to get a reasonable approximation of $\mathcal{I}(\alpha, \theta, x)$, i.e. a small sampling error in $\mathcal{I}(\alpha, \theta, x)$ does not affect the performance of the estimation.

What can affect the performance, however, is the choice of bandwidth ρ_T that enters into $\Omega_T(\alpha, x)$. In view of the resolution discussion in the space $\mathcal{A}$ it is natural to choose $\rho_T \geq 1$. We want to leave a rigorous treatment of the data dependent bandwidth choice for future research. In our Monte Carlo simulations we choose $\rho_T = 4$ everywhere, which seemed a reasonable compromise between being able to approximate many density functions with Π_T (Assumption 4.3(i)) and making sure that the model is approximately identified (Assumption 4.3(ii)).

After discretization of $\mathcal{A}$ and choice of ρ_T the parametrization of Π_T is determined and is finite dimensional. For ease of notation we consider in the rest of this section the pure random effect case where the number of groups $G_T = 1$, and we assume that $\mathcal{A}$ is one-dimensional. The generalization to $G_T > 1$ and higher-dimensional $\mathcal{A}$ is straightforward. Let α_q^*, $q = 1, \ldots, Q$, be the chosen discretization of $\mathcal{A}$, and let $\mathcal{K}_{qr}^{\Omega}$, $q, r = 1, \ldots, Q$ be the corresponding discretization of $\mathcal{K}_T^{\Omega}(\alpha, \beta, x)$,[9] which in the pure random effect case does not depend on x. The discretized version

[8] Note that we only discretized $\mathcal{A}$ for the estimation procedure, but not for data generating process of the Monte Carlo simulation, i.e. the realizations α_i^0 were chosen from an unrestricted continuous distributions, only restricted by the computer precision.

[9] One can, for example, choose

$$\mathcal{K}_{qr}^{\Omega} = \Phi\left(\frac{\alpha_{q+1}^* + \alpha_q^*}{2}; \alpha_r^*, \Omega_T(\alpha_r^*)\right) - \Phi\left(\frac{\alpha_q^* + \alpha_{q-1}^*}{2}; \alpha_r^*, \Omega_T(\alpha_r^*)\right),$$

for $q = 2, \ldots, Q-1$, where $\Phi(\alpha; \beta, \Omega)$ is the cdf of a normal distribution with mean β and variance Ω. For $q = 1$ the second term is omitted, and for $q = Q$ the first term is set to 1. This definition guarantees $\sum_q \mathcal{K}_{qr}^{\Omega} = 1$.

of the integrated likelihood function, defined in (3.1), then reads

$$L_{NT}^{\text{approx}}(\theta, \pi^{\text{disc}}) = \frac{1}{NT} \sum_{i=1}^{N} \log \left[\sum_{q,r=1}^{Q} f(Y_i|X_i, \alpha_q^*; \theta)\, \mathcal{K}_{qr}^{\Omega}\, \pi_r^{\text{disc}} \right], \tag{5.9}$$

where π_r^{disc}, $r = 1, \ldots, Q$, describes the distribution to which the Kernel smoothing is applied in order to parameterize Π_T, and the superscript "disc" refers to discretization. The constraints $\sum_r \pi_r^{\text{disc}} = 1$ and $\pi_r^{\text{disc}} \geq 0$, for all $r = 1, \ldots, Q$, need to be imposed, and further upper and lower bounds on π_r^{disc} can also be imposed without difficulty.

The estimators $\hat{\theta}$ and $\hat{\pi}^{\text{disc}}$ are obtained by joint maximization of $L_{NT}^{\text{approx}}(\theta, \pi^{\text{disc}})$ over θ and π^{disc}. The number of parameter π^{disc} may be quite large, but this is numerically unproblematic, since for given θ the objective function is smooth and concave over π^{disc} and the constraints on π^{disc} are linear, i.e. the optimization problem over π^{disc} is very well-behaved, and the corresponding gradient and Hessian can be easily calculated analytically.

The structure of $L_{NT}^{\text{approx}}(\theta, \pi^{\text{disc}})$ as a function of both θ and π^{disc} may, however, be more complicated. In principle, multiple local maxima could exist, and it may therefore be necessary to repeat the joint maximization over $L_{NT}(\theta, \pi)$ with multiple starting values, or to perform an initial grid search over $\theta \in \Theta$.

An interesting alternative option — which we also make use of in our Monte Carlo simulations — is to perform the optimization sequentially. Starting with some consistent preliminary estimator for θ, one first optimizes over π^{disc} for given θ, then takes the optimal value π^{disc} and optimizes over θ with π^{disc} fixed, and so on. Asymptotically, one can show that already after a finite number of repetitions this sequential approach yields an estimator for θ whose asymptotic bias decreases at the same rate as the joint maximum likelihood estimator. In our Monte Carlo setup this sequential approach turned out to converge quite rapidly. Note that once the estimator $\hat{\pi}^{\text{disc}}$ is found, one can obtain the actual estimator for the distribution π by applying the Kernel function $\mathcal{K}_{rq}^{\Omega}$.

6 Monte Carlo Simulations

In our simulation study we consider the dynamic binary choice model without additional regressors, as introduced in equation (3.5). In this model only the binary outcome variable Y_{it} is observed for time periods $t = 0, \ldots, T$ and cross-sectional units $i = 1, \ldots, N$. The parameter of interest $\theta \in \Theta = \mathbb{R}$ and the individual effect

$\alpha_i \in \mathcal{A} = \mathbb{R}$ are both scalars. The initial period outcome Y_{i0} is used as a conditioning variable. We consider a probit model, i.e. shocks ε_{it} are standard normally distributed, *iid* across both i and t. The (log-) likelihood function of the model thus reads

$$L_{NT}(\theta, \pi) = \frac{1}{NT} \sum_{i=1}^{N} \log \int_{\mathbb{R}} \left\{ \prod_{t=1}^{T} [\Phi(\theta\, Y_{i,t-1} + \alpha)]^{Y_{it}} [1 - \Phi(\theta\, Y_{i,t-1} + \alpha)]^{1-Y_{it}} \right\} \pi(\alpha|Y_{i0}) d\alpha, \tag{6.1}$$

where $\Phi(.)$ is the cdf of the standard normal distribution. The unknown parameters that enter the likelihood function are θ and the two densities $\pi(\alpha|Y_{i0} = 0)$ and $\pi(\alpha|Y_{i0} = 1)$.

For the data generating process we choose the true parameter of interest $\theta^0 = 1$. The conditional distribution $\boldsymbol{\alpha}|Y_{i0} = 1$ is chosen to be a t-distribution with 5 degrees of freedom, centered at $\alpha = 0$ and rescaled such that its standard error is σ_π. The conditional distribution $\boldsymbol{\alpha}|Y_{i0} = 0$ is chosen to be an equal mixture of two t-distributions with 5 degrees of freedom, one centered at $\alpha = 1$ and one centered at $\alpha = -1$, and both rescaled such that each has a standard error σ_π. Thus, σ_π parameterizes the smoothness of the conditional density $\pi(\alpha|Y_{i0})$, and we consider different values for σ_π in the simulations below. Finally, we let $Y_{i0} = 0$ and $Y_{i0} = 1$ both occur with probability 0.5 in the data generating process.

To estimate the model, we discretize the set $\mathcal{A} = \mathbb{R}$ by choosing a lower bound of $\alpha = -9$, an upper bound of $\alpha = 9$, and a discretization step-size that is sufficiently small relative to the variance of the fixed effect estimator for α, as described in the computation section above.

In Figure 1 we plot $\sqrt{T^{-1}\mathcal{I}^{-1}(\alpha, \theta^0, Y_{i0})}$ as a function of α for $Y_{i0} = 0$ and 1, and for $T = 12$ and 24. We have argued that this quantity, which approximated the standard error of the fixed effect estimator for α, can be viewed as the "resolution scale" that the model provides for the estimation of the individual effect distribution. The figure shows that we cannot expect to resolve much structure in $\pi(\alpha|Y_{i0})$ for say $\alpha < -2.5$ and $\alpha > 2$, since $\sqrt{T^{-1}\mathcal{I}^{-1}(\alpha, \theta^0, Y_{i0})}$ then becomes quickly very large. The figure also shows that we can expect to resolve somewhat finer structures for $T = 24$ than for $T = 12$. Note also that $\sqrt{T^{-1}\mathcal{I}^{-1}(\alpha, \theta^0, Y_{i0})}$ is not symmetric around $\alpha = 0$ (it would only be for $\theta = 0$) and that it is slightly different for $Y_{i0} = 0$ and $Y_{i0} = 1$.

We choose a bandwidth $\rho_T = 4$ for all simulations. According to our Kernel

construction we approximate $\pi(\alpha|Y_{i0})$ as a mixture of normal distributions with arbitrary means α, but given variances $\Omega_T(\alpha, Y_{i0}) = \rho_T T^{-1}\mathcal{I}^{-1}(\alpha, \tilde{\theta}, Y_{i0})$. We plot a few of these normal distributions that are used as "basis functions" in Figure 2 for $T = 12$ and in Figure 3 for $T = 24$, for $\tilde{\theta} = \theta^0$ (the dependence on $\tilde{\theta}$ is not very strong, i.e. using an estimator $\tilde{\theta}$ — as we do in the actual estimation procedure — does not change these plots much). As was expected from Figure 1, these basis functions become more and more wide, corresponding to less and less resolution power, as the absolute value of α becomes larger. One can also see that the basis functions for $T = 24$ are more narrow, i.e. are able to resolve more structure in $\pi(\alpha|Y_{i0})$. By choosing a smaller value for the bandwidth ρ_T one could resolve finer structures in the individual effect distribution. However, smaller values of ρ_T also mean that the identification of $\pi(\alpha|Y_{i0})$ from the distribution of Y_{it} becomes more problematic, and one needs to compromise between these two competing goals.

Once the bandwidth ρ_T and thus the basis functions are chosen, the key question is whether the true distribution $\pi^0(\alpha|Y_{i0})$ can be well-approximated by these basis functions (i.e. by an element in the parameter set $\Pi_T = \mathcal{K}_T^{\Omega}\Pi_T^{\mathcal{A}}$). This will crucially depend on how smooth $\pi^0(\alpha|Y_{i0})$ is, which in our setup is regulated by the parameter σ_π. Note that we can only expect a good performance of our joint maximum likelihood estimator for θ and π, if the true distribution π^0 can be well-approximated by an element in Π_T.

Figure 4 shows the maximum likelihood estimator for $\pi(\alpha|Y_{i0})$ obtained from a sample of size $T = 12$ and $N = 10000$ (taking $\theta = \theta^0$ as given) for different values of σ_π. Here, we have used $N = 10000$, which is larger than the sample sizes in the actual Monte Carlo simulations below, in order to keep the sampling error small, so that we can focus on the question of whether π^0 can be well-approximated at $T = 12$ for our bandwidth choice. The figure shows that the approximation of π^0 is relatively good at $\sigma_\pi = 1.4$, but rather bad at $\sigma_\pi = 0.7$. Thus, we expect the joint maximum likelihood estimator $\hat{\theta}$ to have little bias in the case $\sigma_\pi = 1.4$ but potentially large bias in the case $\sigma_\pi = 0.7$. It turns out that the approximation of π^0 in the intermediate case $\sigma_\pi = 1$ is still sufficiently good to obtain little bias of $\hat{\theta}$, but from Figure 4 alone it would probably not be clear what to expect in that case.

Table 1 contains our actual Monte Carlo results for various estimators of the parameters of interest at $T = 12$, and for different values of N and σ_π. We performed 1000 simulation repetitions for $N = 100$ and 500, and 500 repetitions for $N = 2500$. The fixed effect estimators (based on the profile likelihood) for θ that we consider are the fixed effect maximum likelihood estimator (FE-MLE), the first order split panel

Jackknife estimator (FE-JACK-1), which eliminates the asymptotic bias of order $1/T$, and the second order split panel Jackknife estimator (FE-JACK-2), which in addition eliminates the asymptotic bias of order $1/T^2$. These Jackknife estimators are obtained by estimating two sub-panels of sample size $T/2$ (and also three sub-panels of sample size $T/3$ for FE-JACK-2), and then appropriately forming linear combinations of the estimators at different sample size, as described in Dhaene and Jochmans (2010), and originally proposed by Hahn and Newey (2004) for panels without time-correlation. Here, we use the Jackknife method, since to our knowledge it is the only bias correction method in the literature so far that in principle allows for arbitrary higher order bias correction, which makes it a natural object of comparison for our random effect method. The random effect estimators (based on the integrated likelihood) that we calculate are the random effect miracle estimator $\hat{\theta}(\pi^0)$ (RE-MIR), which is infeasible since it assumes knowledge of π^0, the random effect estimator with fixed prior distribution $\hat{\theta}(\pi^{\text{prior}})$ (RE-PRIOR), using a normal prior distribution $\pi^{\text{prior}}(\alpha|Y_{i0})$ with mean zero and standard error equal to 4 for both values of Y_{i0}, and finally our joint maximum likelihood estimator (RE-MLE) that is obtained by maximizing the integrated likelihood over $\theta \in \mathbb{R}$ and $\pi \in \Pi_T$.

Table 1 shows that, as expected, for $T = 12$ the FE-MLE is severely biased due to the incidental parameter problem. The first order Jackknife bias correction eliminates around 90% of this bias, and the second order Jackknife correction reduces the bias even further. However, the bias correction also increases the standard error of the estimator, by around 20% for the first order correction and by around 100% for the second order correction. Both Jackknife estimators are known to have the same asymptotic variance as the MLE, but the phenomenon of finite sample variance inflation is also well-known. In terms of root mean square error the second order Jackknife estimator performs worse than the first order Jackknife estimator in all our simulations, due to the much larger standard error. In the following we therefore concentrate on the comparison between the first order Jackknife estimator and the RE-MLE.

The RE-MLE performs very well in the $T = 12$ simulations for $\sigma_\pi = 1.4$ and $\sigma_\pi = 1$. In those cases the RE-MLE is essentially unbiased at $N = 500$ and $N = 2500$ (the bias is below 5% significance given the number of simulation repetitions), and it has a bias at $N = 100$ which is still very small relative to the standard error at $N = 100$. Furthermore, it has a standard error that is almost identical to the standard error of the FE-MLE, and which is therefore smaller than the standard error of the FE-JACK-1. Note that the bias of the FE-JACK-1 is essentially independent of N,

while its standard error decreases like $N^{-1/2}$. In our simulation design at $T = 12$ we find that at $N = 2500$ the bias and the standard error of the FE-JACK-1 are almost identical, i.e. for all values of N larger than 2500 the bias will dominate the standard error of the FE-JACK-1. Even at $N = 500$ the bias is about half the size of the standard error for the FE-JACK-1, which would be very problematic for testing purposes. Thus, in particular for large values of N the RE-MLE therefore performs much better than the FE-JACK-1 for $\sigma_\pi = 1.4$ and 1.

However, as already anticipated from Figure 4, this is not true for $\sigma_\pi = 0.7$. In that case we find the bias of the RE-MLE to be around twice the bias of the FE-JACK-1, which also results in a larger root mean square error for large values of N. The properties of the FE-JACK-1 are essentially independent from σ_π, since in the fixed effect approach the properties of the individual effect distribution are not important. In contrast, for our random effect approach it makes a big difference whether $\sigma_\pi = 1$ or $\sigma_\pi = 0.7$, since in one case the true individual effect distribution can be reasonably approximated, while in the other case it cannot. These results very clearly show the tradeoff one faces between using the RE-MLE and using the FE-JACK-1.

Given our bandwidth choice we thus found that we cannot properly resolve the distribution π^0 for $\sigma_\pi = 0.7$ at $T = 12$. This problem is, however, automatically resolved if T becomes larger. Figure 5 shows that at $T = 24$ one can already approximate the true individual effects distribution for $\sigma_\pi = 0.7$ relatively well, and Table 2 shows the corresponding Monte Carlo results for the parameters of interest. In that case, the bias of the RE-MLE is again very small relative to its standard error, and for $N = 500$ and $N = 2500$ the RE-MLE therefore again performs much better than the FE-JACK-1.[10]

[10] It is also interesting that in all our simulations the standard errors of the FE-MLE, the RE-PRIOR and the RE-MLE are very similar, while the infeasible RE-MIR has a somewhat smaller variance. Asymptotically all these standard errors will converge, but in finite sample knowing the true π^0 not only results in an insignificant bias of the RE-MIR estimator, but also in an increased efficiency in terms of standard error. Finally, we want to point out that the performance of the FE-MLE and the RE-PRIOR estimator is very similar not only in terms of standard error but also in terms of bias. Both estimators have a bias of similar order that decreases at the rate of $1/T$.

7 Conclusions

This paper presents an alternative approach to higher order bias correction in non-linear panel data model with large N and T. Instead of profiling out the individual effects (fixed effect approach) we propose to integrate out the individual effects from the likelihood function, and use the resulting integrated likelihood to estimate the parameters of interest. We show that if a consistent estimator for the individual effect distribution is used to integrate out the individual effects, then the rate at which the bias of the estimator for the parameters of interest decreases with T is proportional to the rate at which the estimator of the individual effect distribution approaches the true distributions (in terms of Hellinger distance). Compared to the fixed effect maximum likelihood estimator, which has a bias of order $1/T$, we can thus obtain a significant improvement in the convergence rate of the bias, as long as a good estimator for the individual effects distribution is available. The bias that results from our estimation approach can also be significantly lower than the one obtained from existing bias correction techniques. This result on the bias correction for the parameters of interest is applicable to all estimators of the individual effect distribution that satisfy some weak regularity conditions.

The estimator for the distribution that we consider explicitly in this paper is the joint maximum likelihood estimator, which maximizes the likelihood function jointly with the estimator for the individual effects. The properties of this estimator are crucially determined by the choice of parameter set of distributions over which the estimation is performed. To allow for non-parametric estimation of the individual effect distribution this parameter set needs to be chosen sample size dependent, analogously to a semi-parametric sieve estimation approach. Under appropriate high-level assumptions on this parameter set and on the true distribution of the individual effects we then derive the convergence rates of the estimator for the distribution (in terms of Hellinger distance) and thus also obtain (an upper bound on) the convergence rate of the incidental parameter bias of the estimator for the parameters of interest.

The high-level assumptions that are employed to derive these general results require a restriction on the correlation structure between the regressors and the individual effects. As a concrete example of such a restriction we consider the case of generalized random effects, which demands that individuals can be partitioned into groups and imposes a random effect assumption in each group. No further parametric assumptions are made on the distribution of the individual effects. We discuss

how to choose an appropriate parameter set for the individual effect distribution in this case, and show that the convergence rate of the incidental parameter bias only depends on the smoothness properties of the true individual effect distribution. As long as this distribution is sufficiently smooth, the bias can decrease at an arbitrary polynomial rate in T.

For future work it would be fascinating to discuss alternative choices for the estimator of the individual effect distribution. Furthermore, it would be interesting to discuss alternative restrictions on the correlation structure between the regressors and the individual effects, which go beyond the generalized random effect assumption discussed explicitly in the present paper. An important extension would also be the estimation of policy parameters like marginal effects. Finally, it is also important to develop a data dependent selection method for the bandwidth ρ_T that enters in the non-parametric estimation of the individual effect distribution.

A Figures and Tables

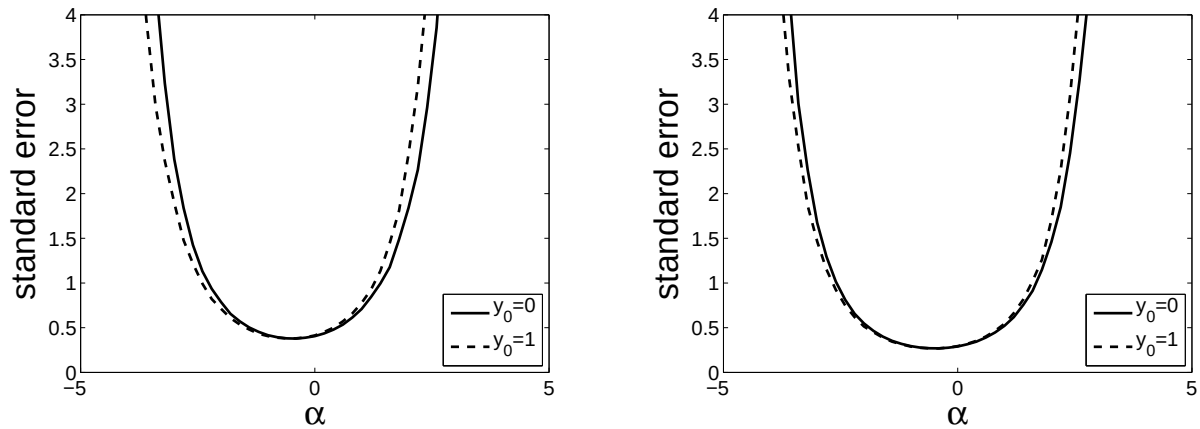


Figure 1: For $T = 12$ (left diagram) and $T = 24$ (right diagram) we plot $\sqrt{T^{-1}\mathcal{I}^{-1}(\alpha, \theta^0, y_{i0})}$ for $\theta^0 = 1$ and $y_{i0} = 0$ and 1. For large T this quantity approximates the standard error of the fixed effect estimator for the individual effects α.

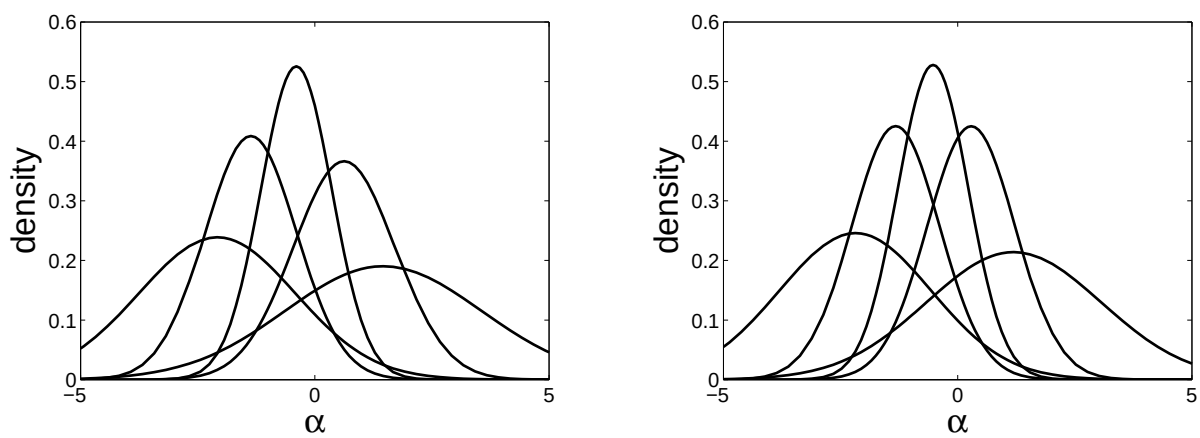


Figure 2: For $T = 12$ some examples of the "basis functions" that are used to approximate the true distributions are plotted for $y_{i0} = 0$ (left) and $y_{i0} = 1$ (right).

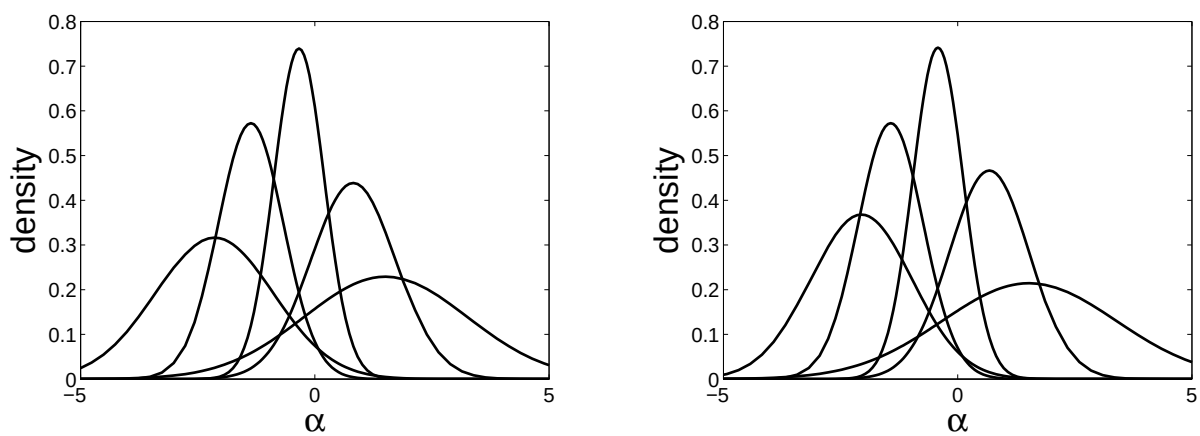
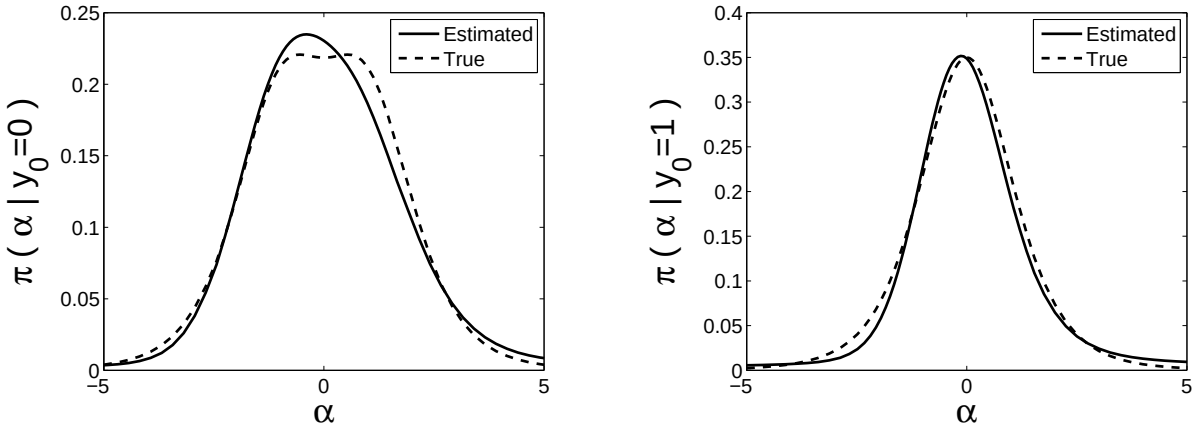
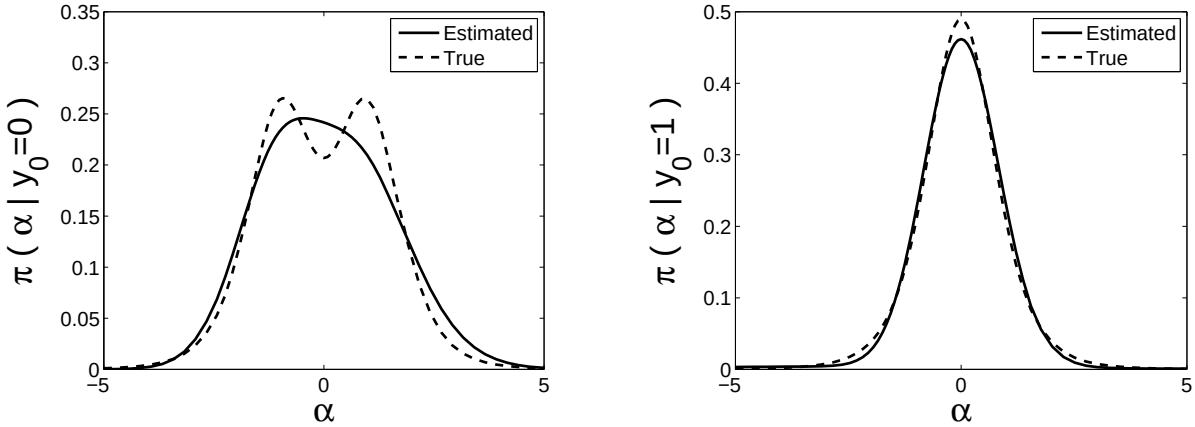


Figure 3: Same as Figure 2, but for $T = 24$. Note the smaller width of "basis functions" compared to the $T = 12$ case.

• $\sigma_\pi = 1.4$:



• $\sigma_\pi = 1$:



• $\sigma_\pi = 0.7$:

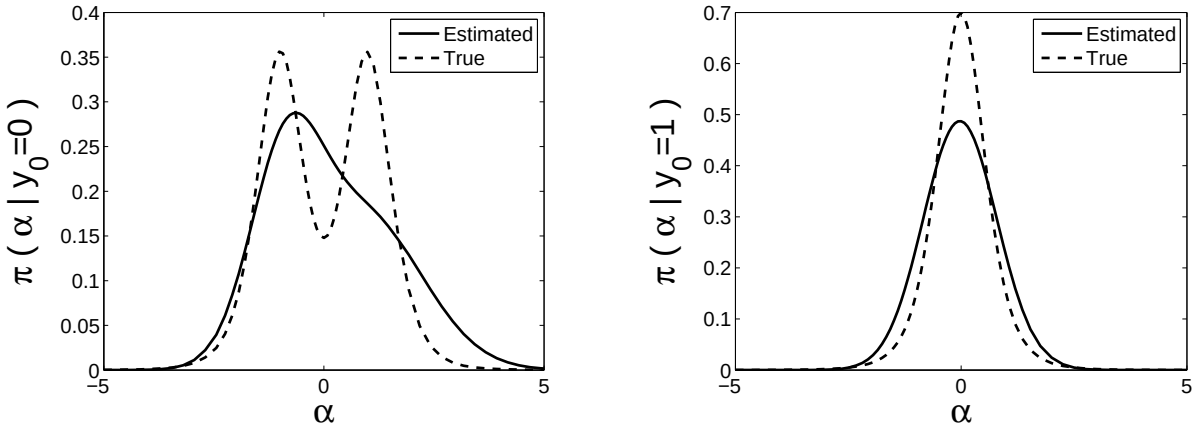


Figure 4: For different values of σ_π the true conditional distribution $\pi^0(\alpha|y_{i0})$ is plotted as a dotted line for both $y_{i0} = 0$ (left diagram) and $y_{i0} = 1$ (right diagram). The solid lines are the corresponding maximum likelihood estimators for $\pi(\alpha|y_{i0})$ obtained from a sample with $T = 12$ and $N = 10000$. The structure of the true distribution cannot be resolved very well for the case $\sigma_\pi = 0.7$, given our particular bandwidth choice $\rho_T = 4$.

	$T = 12$, $\sigma_\pi = 1.4$								
	$N = 100$			$N = 500$			$N = 2500$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse
FE-MLE	-0.313	0.124	0.336	-0.308	0.056	0.313	-0.3115	0.0243	0.3125
FE-JACK-1	0.029	0.151	0.153	0.037	0.067	0.077	0.0308	0.0299	0.0429
FE-JACK-2	-0.014	0.252	0.253	0.003	0.115	0.115	-0.0044	0.0516	0.0517
RE-MIR	0.003	0.109	0.109	0.003	0.050	0.050	-0.0003	0.0214	0.0214
RE-PRIOR	-0.171	0.124	0.212	-0.169	0.056	0.178	-0.1728	0.0242	0.1745
RE-MLE	-0.006	0.129	0.129	0.003	0.059	0.059	-0.0002	0.0257	0.0257

	$T = 12$, $\sigma_\pi = 1$								
	$N = 100$			$N = 500$			$N = 2500$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse
FE-MLE	-0.320	0.112	0.339	-0.314	0.052	0.319	-0.3160	0.0253	0.3169
FE-JACK-1	0.025	0.138	0.141	0.029	0.063	0.069	0.0278	0.0290	0.0402
FE-JACK-2	-0.008	0.244	0.244	-0.004	0.106	0.106	-0.0059	0.0493	0.0497
RE-MIR	0.003	0.095	0.095	0.002	0.045	0.045	0.0012	0.0204	0.0204
RE-PRIOR	-0.196	0.113	0.227	-0.194	0.052	0.201	-0.1948	0.0240	0.1963
RE-MLE	-0.016	0.116	0.117	-0.002	0.055	0.055	-0.0005	0.0249	0.0249

	$T = 12$, $\sigma_\pi = 0.7$								
	$N = 100$			$N = 500$			$N = 2500$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse
FE-MLE	-0.332	0.112	0.350	-0.323	0.050	0.327	-0.3237	0.0221	0.3244
FE-JACK-1	0.017	0.135	0.136	0.023	0.060	0.064	0.0225	0.0266	0.0349
FE-JACK-2	-0.017	0.230	0.230	-0.011	0.102	0.102	-0.0083	0.0449	0.0457
RE-MIR	-0.005	0.090	0.090	0.002	0.040	0.040	0.0004	0.0179	0.0179
RE-PRIOR	-0.223	0.112	0.250	-0.214	0.051	0.220	-0.2149	0.0223	0.2161
RE-MLE	-0.049	0.113	0.124	-0.037	0.051	0.063	-0.0424	0.0223	0.0479

Table 1: Results for the bias, standard error (std) and root mean square error (rmse) of different estimators for the parameter of interest θ in simulations at $T = 12$ and for different values of N and σ_π. The results are based on 1000 repetitions for $N = 100$ and $N = 500$, and on 500 repetitions for $N = 2500$. The estimators are the fixed effect MLE (FE-MLE), first and second order split panel Jackknife (FE-JACK-1 and 2), the infeasible random effect miracle estimator (RE-MIR), a random effect estimator with fixed prior distribution (RE-PRIOR), and the random effect joint MLE over θ and π (RE-MLE).

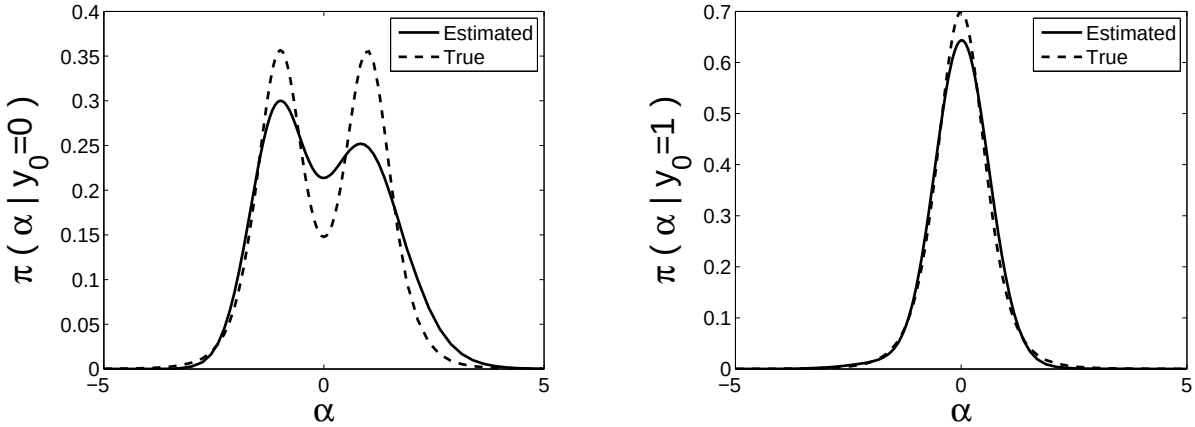


Figure 5: Same as Figure 4, but with $T = 24$ and only for $\sigma_\pi = 0.7$.

	$T = 24$, $\sigma_\pi = 0.7$								
	$N = 100$			$N = 500$			$N = 2500$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse
FE-MLE	-0.166	0.078	0.183	-0.1662	0.0334	0.1695	-0.1664	0.0164	0.1672
FE-JACK-1	0.008	0.087	0.087	0.0068	0.0373	0.0379	0.0062	0.0176	0.0187
FE-JACK-2	0.007	0.115	0.116	0.0061	0.0517	0.0520	0.0048	0.0226	0.0231
RE-MIR	0.003	0.066	0.066	-0.0003	0.0288	0.0288	0.0004	0.0143	0.0143
RE-PRIOR	-0.116	0.079	0.140	-0.1178	0.0336	0.1226	-0.1180	0.0166	0.1192
RE-MLE	-0.007	0.080	0.081	-0.0028	0.0349	0.0350	-0.0003	0.0170	0.0170

Table 2: Same as Table 1, but with $T = 24$ and only for $\sigma_\pi = 0.7$.

B Assumptions

B.1 Assumptions for Consistency

To state our assumption we first define

$$\mathcal{I}_i = -\frac{1}{T}\frac{\partial^2 \log f(Y_i|X_i, \hat{\alpha}_i^{\mathrm{p}}(\theta^0); \theta^0)}{\partial\alpha\partial\alpha'},$$
$$\mathcal{B}_{c,i} = \left\{\alpha \in \mathbb{R}^M \;\middle|\; \left[\alpha - \hat{\alpha}_i^{\mathrm{p}}(\theta^0)\right]' \mathcal{I}_i \left[\alpha - \hat{\alpha}_i^{\mathrm{p}}(\theta^0)\right] \leq \frac{c}{T}\right\}. \tag{B.1}$$

Assumption B.1. *There exist $c_1, c_2, c_3, c_4, c_5 > 0$ such that wpa1*

(i) $\hat{\theta}^{\mathrm{p}} = \theta^0 + o_p(1), \qquad L_{NT}^{\mathrm{p}}(\hat{\theta}^{\mathrm{p}}) = L_{NT}^{\mathrm{p}}(\theta^0) + o_p(1).$

(ii) $\forall\theta \in \Theta: \quad L_{NT}^{\mathrm{p}}(\theta) \leq L_{NT}^{\mathrm{p}}(\hat{\theta}^{\mathrm{p}}) - \min(c_1,\, c_2\|\theta - \hat{\theta}^{\mathrm{p}}\|^2).$

(iii) $\forall i \in \{1, \ldots, N\},\ \forall\alpha \in \mathcal{B}_{c_3,i} \cap \mathcal{A}:$

$$\frac{1}{T}\log f(Y_i|X_i, \alpha; \theta^0) \geq \frac{1}{T}\log f(Y_i|X_i, \hat{\alpha}_i^{\mathrm{p}}; \theta^0) - \frac{c_4}{2}\left(\alpha - \hat{\alpha}_i^{\mathrm{p}}\right)' \mathcal{I}_i \left(\alpha - \hat{\alpha}_i^{\mathrm{p}}\right),$$

where $\hat{\alpha}_i^{\mathrm{p}} = \hat{\alpha}_i^{\mathrm{p}}(\theta^0)$.

(iv) $\forall i \in \{1, \ldots, N\}: \quad \mathrm{vol}(\mathcal{B}_{c_3,i} \cap \mathcal{A}) \geq c_5 \mathrm{vol}(\mathcal{B}_{c_3,i}).$

(v) $\frac{1}{N}\sum_{i=1}^{N}\sqrt{\|\mathcal{I}_i^{-1}\|} = o_p(T^{3/2}).$

(vi) The logarithm of $\pi_T^{\mathrm{low}}(\alpha|x)$ is Lipschitz continuous in α with a Lipschitz constant that is uniformly bounded over $x \in \mathcal{X}_T$. Furthermore, $\pi_T^{\mathrm{low}}(\alpha|x)$ satisfies

$$\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sqrt{\det \mathcal{I}_i}}{\pi_T^{\mathrm{low}}(\hat{\alpha}_i^{\mathrm{p}}(\theta^0)|X_i)} \leq o_p(T).$$

Part (i) of this assumption demands consistency of the fixed effect effect estimator and some continuity of the profile likelihood around θ^0. Part (ii) requires that $L_{NT}^{\mathrm{p}}(\theta)$ has a properly isolated maximum at $\hat{\theta}^{\mathrm{p}}$ with a non-degenerate Hessian. Part (iii) is a similar assumption on the maximum of $\log f(Y_i|X_i, \alpha; \theta^0)$ in α. Part (iv) demands that the boundary of $\mathcal{A}$ is well-behaved, where "vol" refers to the volume of a set. Part (v) requires a lower bound on the eigenvalues of $\mathcal{I}_i$, here $\|.\|$ is the operator norm. Part (vi) is a regularity condition on the lower bound $\pi_T^{\mathrm{low}}(\alpha|x)$, which still allows this lower bound to decrease in T at any polynomial rate.

B.2 Further Regularity Conditions on the Model

Define

$$f_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta,\pi) = \frac{f(y|x,\alpha;\theta)\,\pi(\alpha|x)}{f_{Y|X}(y|x;\theta,\pi)}\,, \tag{B.2}$$

This is the posterior distribution of $\boldsymbol{\alpha}$ for given $Y = y$ under the prior $\pi(\alpha|x)$, for given values of x and θ. Similarly, the posterior distribution of $\boldsymbol{\alpha}$ under a uniform prior reads

$$f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta) = \frac{f(y|x,\alpha;\theta)}{\int_{\mathcal{A}} f(y|x,\beta;\theta)d\beta}. \tag{B.3}$$

It is convenient to introduce the following notation.

$$\begin{aligned}
J^{(1)}(y,x) &= \frac{1}{\sqrt{T}}\int_{\mathcal{A}} \frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta}\, f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0)\,d\alpha,\\
J^{(2)}(y,x) &= \frac{1}{T}\int_{\mathcal{A}} \frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta}\,\frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta'}\, f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0)\,d\alpha,\\
H_{k_1k_2}(y,x) &= \frac{1}{T^2}\int_{\mathcal{A}} \left(\frac{\partial^2 \log f(y|x,\alpha;\theta^0)}{\partial\theta_{k_1}\partial\theta'_{k_2}} + \frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta_{k_1}}\,\frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta_{k_2}}\right)^2\\
&\qquad\qquad f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0)\,d\alpha,\\
D^{(q)}(y,x) &= \int_{\mathcal{A}}\int_{\mathcal{A}} \left[\sqrt{T}\,d_x(\alpha,\beta)\right]^q f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0)\,d\alpha\, f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\beta|y,x;\theta^0)\,d\beta,
\end{aligned} \tag{B.4}$$

where $q = 2, 4$. Note that in the definition of $J^{(1)}(y,x)$ the factor $\frac{1}{\sqrt{T}}$ is the appropriate normalization for the score function $\frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta_k}$, since the score at the true parameters has zero mean, and since $f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0)$ will be centered around the realized value α_i^0 if evaluated at $y = Y_i$ and $x = X_i$. Similarly, for $H_{k_1k_2}(y,x)$ the expression $\frac{\partial^2 \log f(y|x,\alpha;\theta^0)}{\partial\theta_{k_1}\partial\theta'_{k_2}} + \frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta_{k_1}}\,\frac{\partial \log f(y|x,\alpha;\theta^0)}{\partial\theta_{k_2}}$ is mean zero at α_i^0, so that $1/T$ is the appropriate normalization for the square of this expression. Also, in the definition of $D^{(q)}(y,x)$ it is natural to rescale $d_x(\alpha,\beta)$ by $\sqrt{T}$, since the standard deviation of the distribution $f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta)$ is of order $1/\sqrt{T}$. We make the following high-level assumptions.

Assumption B.2. *We assume that*

(i) Y_i *and* X_i *are independently and identically distributed across* i.

(ii) $\mathbb{E}\,J^{(2)}(Y_i,X_i) = \mathcal{O}(1)$, *and* $\mathbb{E}\,D^{(q)}(Y_i,X_i) = \mathcal{O}(1)$, *for* $q = 2, 4$.

(iii) $\frac{1}{N}\sum_{i=1}^N \left[J_k^{(1)}(Y_i,X_i)\right]^2 = \mathcal{O}_p(1)$, $\frac{1}{N}\sum_{i=1}^N \left[J_{k_1k_2}^{(2)}(Y_i,X_i)\right]^2 = \mathcal{O}_p(1)$, $\frac{1}{N}\sum_{i=1}^N H(Y_i,X_i) = \mathcal{O}_p(1)$, *and* $\frac{1}{N}\sum_{i=1}^N \left[D^{(q)}(Y_i,X_i)\right]^2 = \mathcal{O}_p(1)$, *for* $q = 2, 4$.

(iv) $\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}\left(J^{(1)}(Y,X)\middle|X=X_i,\,\boldsymbol{\alpha}=\alpha\right)\right]^2\pi_T^{\mathrm{up}}(\alpha|X)\,d\alpha=\mathcal{O}_p(1/T)$,

$\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}\left(J^{(2)}(Y,X)\middle|X=X_i,\,\boldsymbol{\alpha}=\alpha\right)\right]^2\pi_T^{\mathrm{up}}(\alpha|X)\,d\alpha=\mathcal{O}_p(1)$,

$\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}\left(D^{(q)}(Y,X)\middle|X=X_i,\,\boldsymbol{\alpha}=\alpha\right)\right]^2\pi_T^{\mathrm{up}}(\alpha|X)\,d\alpha=\mathcal{O}_p(1)$, $q=2,4$.

(v) $\sqrt{NT}\,\frac{\partial L_{NT}(\theta^0,\pi^0)}{\partial\theta}=\mathcal{O}_p(1)$, *and* $\exists c>0$ *such that* $\frac{\partial^2 L_{NT}(\theta^0,\pi^0)}{\partial\theta\partial\theta'}>c$, *wpa1.*

(vi) $\frac{\partial^3 L_{NT}(\theta,\pi)}{\partial\theta_{k_1}\partial\theta_{k_2}\partial\theta_{k_3}}=\mathcal{O}_p(1)$, *uniformly in a neighborhood of* θ^0 *and over* $\pi\in\Pi_{T,\kappa}^{\mathrm{lip}}$ *with* $\kappa_T=\sqrt{T}$.

These regularity assumptions look complicated. However, a key advantage of our analysis of the integrated likelihood is that it does not involve a Laplace approximation and therefore allows the distributions $\pi(\alpha|x)$ to be e.g. non-differentiable in α — only a Lipschitz condition is imposed.

Assumption B.3. $\pi_T^{\mathrm{up}}(\alpha|x)/\pi_T^{\mathrm{low}}(\alpha|x)$ *is uniformly bounded over* $\alpha\in\mathcal{A}$, $x\in\mathcal{X}_T$ *and* T.

Assumption B.3 is a convenient technical condition, but can probably relaxed without affecting the validity of our conclusions. For the moment, we leave this generalization for future work.

C Proofs

C.1 Proofs for Section 4.1

Proof of Theorem 4.1. By the mean value theorem for integration there exist $\tilde{\alpha}_i(\theta,\pi,Y_i,X_i)\in\mathcal{A}$ such that

$$L_{NT}(\theta,\pi)=\frac{1}{NT}\sum_{i=1}^{N}\log f(Y_i|X_i,\tilde{\alpha}_i(\theta,\pi,Y_i,X_i);\theta), \tag{C.1}$$

and therefore $L_{NT}(\theta,\pi)\leq L_{NT}^{\mathrm{p}}(\theta)$. We have thus obtained an upper bound on $L_{NT}(\theta,\pi)$. Next, we derive a lower bound on $L_{NT}(\theta^0,\pi)$. Let $\hat{\alpha}_i^{\mathrm{p}}=\hat{\alpha}_i^{\mathrm{p}}(\theta^0)$. We

have wpa1 that

$$
\begin{aligned}
&L_{NT}(\theta^0,\pi)\\
&= \frac{1}{NT}\sum_{i=1}^{N}\log\int_{\mathcal{A}} f(Y_i|X_i,\alpha;\theta^0)\,\pi(\alpha|X_i)\,d\alpha\\
&\geq \frac{1}{NT}\sum_{i=1}^{N}\log\int_{\mathcal{B}^{c_3,i}\cap\mathcal{A}} f(Y_i|X_i,\hat{\alpha}_i^{\mathrm{p}};\theta^0)\exp\left[-\frac{c_4\,T}{2}\,(\alpha-\hat{\alpha}_i^{\mathrm{p}})'\,\mathcal{I}_i\,(\alpha-\hat{\alpha}_i^{\mathrm{p}})\right]\pi_T^{\mathrm{low}}(\alpha|X_i)d\alpha\\
&\geq L_{NT}^{\mathrm{p}}(\theta^0)+\frac{1}{NT}\sum_{i=1}^{N}\log\left[\exp\left[-\frac{c_3c_4}{2}\right]\inf_{\alpha\in\mathcal{B}^{c_3,i}\cap\mathcal{A}}\pi_T^{\mathrm{low}}(\alpha|X_i)\left(\int_{\mathcal{B}^{c_3,i}\cap\mathcal{A}}d\alpha\right)\right]\\
&= L_{NT}^{\mathrm{p}}(\theta^0)-\frac{c_3c_4}{2T}+\frac{1}{NT}\sum_{i=1}^{N}\inf_{\alpha\in\mathcal{B}^{c_3,i}\cap\mathcal{A}}\log\pi_T^{\mathrm{low}}(\alpha|X_i)+\frac{1}{NT}\sum_{i=1}^{N}\log\mathrm{vol}(\mathcal{B}^{c_3,i}\cap\mathcal{A})\\
&\geq L_{NT}^{\mathrm{p}}(\theta^0)+\frac{1}{NT}\sum_{i=1}^{N}\log\pi_T^{\mathrm{low}}(\hat{\alpha}_i^{\mathrm{p}}|X_i)-\frac{b}{NT}\sum_{i=1}^{N}\sqrt{\frac{c_3\,\|\mathcal{I}_i^{-1}\|}{T}}+\frac{1}{NT}\sum_{i=1}^{N}\log\left[c_5\,\mathrm{vol}(\mathcal{B}^{c_3,i})\right]+o_p(1)\\
&= L_{NT}^{\mathrm{p}}(\theta^0)+\frac{1}{NT}\sum_{i=1}^{N}\log\frac{\pi_T^{\mathrm{low}}(\hat{\alpha}_i^{\mathrm{p}}|X_i)}{T^{M/2}\sqrt{\det\mathcal{I}_i}}\\
&\geq L_{NT}^{\mathrm{p}}(\theta^0)+o_p(1),
\end{aligned}
\tag{C.2}
$$

uniformly over $\pi\in\Pi_T^{\mathrm{low}}$. Here, $b>0$ is the Lipschitz constant of $\log\pi_T^{\mathrm{low}}$. Then we have uniformly over $\pi\in\Pi_T^{\mathrm{low}}$

$$
L_{NT}^{\mathrm{p}}(\hat{\theta}(\pi))\geq L_{NT}(\hat{\theta}(\pi),\pi)\geq L_{NT}(\theta^0,\pi)\geq L_{NT}^{\mathrm{p}}(\theta^0)+o_p(1)=L_{NT}^{\mathrm{p}}(\hat{\theta}^{\mathrm{p}})+o_p(1). \tag{C.3}
$$

Applying our assumption on the shape of $L_{NT}^{\mathrm{p}}(\theta)$ we thus obtain

$$
c_2\left\|\hat{\theta}(\pi)-\hat{\theta}^{\mathrm{p}}\right\|^2\leq L_{NT}^{\mathrm{p}}(\hat{\theta}^{\mathrm{p}})-L_{NT}^{\mathrm{p}}(\hat{\theta}(\pi))=o_p(1), \tag{C.4}
$$

which implies $\|\hat{\theta}(\pi)-\hat{\theta}^{\mathrm{p}}\|=o_p(1)$, and therefore $\|\hat{\theta}(\pi)-\theta^0\|=o_p(1)$, uniformly over $\pi\in\Pi_T^{\mathrm{low}}$. ■

C.2 Proofs for Section 4.2

Lemma C.1. *For all $\kappa_T>0$, $y\in\mathcal{Y}_T$ and $x\in\mathcal{X}_T$ we have*

(i)
$$
\sup_{\pi_1,\pi_2\in\Pi_\kappa^{\mathrm{lip}}}\left(\frac{\partial\log\,f_{Y|X}(y|x;\theta^0,\pi_1)}{\partial\theta_k}-\frac{\partial\log\,f_{Y|X}(y|x;\theta^0,\pi_2)}{\partial\theta_k}\right)^2 \leq 8\,\kappa_T^2\,J_{kk}^{(2)}(y,x)\left(D^{(2)}(y,x)+\frac{\kappa_T^2}{2T}D^{(4)}(y,x)\right).
$$

$$(ii)\quad \sup_{\pi\in\Pi_\kappa^{\text{lip}}}\left(\frac{\partial\log\, f_{Y|X}(y|x;\theta^0,\pi)}{\partial\theta_k}-\int_{\mathcal{A}}\frac{\partial\log f(y|x,\alpha;\theta^0)}{\partial\theta_k}f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0)d\alpha\right)^2$$
$$\leq 4\,\kappa_T^2\,J_{kk}^{(2)}(y,x)\,\left(D^{(2)}(y,x)+\frac{\kappa_T^2}{2T}D^{(4)}(y,x)\right).$$

In addition, either let $\tilde{\mathbb{E}}$ be a (conditional) expected value over the random variable $\tilde{Y}=Y$ and $\tilde{X}=X$, or let $\tilde{\mathbb{E}}=\frac{1}{N}\sum_{i=1}^N$ be a sample average over the sample $\tilde{Y}=Y_i$ and $\tilde{X}=X_i$. Then we have

$$(iii)\quad \sup_{\pi_1,\pi_2\in\Pi_\kappa^{\text{lip}}}\left[\tilde{\mathbb{E}}\left(\frac{\partial\log\, f_{Y|X}(\tilde{Y}|\tilde{X};\theta^0,\pi_1)}{\partial\theta_k}-\frac{\partial\log\, f_{Y|X}(\tilde{Y}|\tilde{X};\theta^0,\pi_2)}{\partial\theta_k}\right)\right]^2$$
$$\leq 8\,\kappa_T^2\,\left(\tilde{\mathbb{E}}\,J_{kk}^{(2)}(\tilde{Y},\tilde{X})\right)\,\left[\left(\tilde{\mathbb{E}}\,D^{(2)}(\tilde{Y},\tilde{X})\right)+\frac{\kappa_T^2}{2T}\left(\tilde{\mathbb{E}}\,D^{(4)}(\tilde{Y},\tilde{X})\right)\right].$$

$$(iv)\quad \sup_{\pi_1,\pi_2\in\Pi_\kappa^{\text{lip}}}\left[\tilde{\mathbb{E}}\left(\frac{\partial\log\, f_{Y|X}(\tilde{Y}|\tilde{X};\theta^0,\pi_1)}{\partial\theta_k}-\int_{\mathcal{A}}\frac{\partial\log f(\tilde{Y}|\tilde{X},\alpha;\theta^0)}{\partial\theta_k}f^{\text{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|\tilde{Y},\tilde{X};\theta^0)d\alpha\right)\right]^2$$
$$\leq 4\,\kappa_T^2\,\left(\tilde{\mathbb{E}}\,J_{kk}^{(2)}(\tilde{Y},\tilde{X})\right)\,\left[\left(\tilde{\mathbb{E}}\,D^{(2)}(\tilde{Y},\tilde{X})\right)+\frac{\kappa_T^2}{2T}\left(\tilde{\mathbb{E}}\,D^{(4)}(\tilde{Y},\tilde{X})\right)\right].$$

Proof. Part (i): Applying Chebyshev's inequality one gets

$$\begin{aligned}
&\left(\frac{\partial\log\, f_{Y|X}(y|x;\theta^0,\pi_1)}{\partial\theta_k}-\frac{\partial\log\, f_{Y|X}(y|x;\theta^0,\pi_2)}{\partial\theta_k}\right)^2\\
&=\left(\int_{\mathcal{A}}\frac{\partial\log f(y|x,\alpha;\theta^0)}{\partial\theta_k}f(y|x,\alpha;\theta^0)\left[\frac{\pi_1(\alpha|x)}{f_{Y|X}(y|x;\theta^0,\pi_1)}-\frac{\pi_2(\alpha|x)}{f_{Y|X}(y|x;\theta^0,\pi_2)}\right]\,d\alpha\right)^2\\
&\leq\underbrace{\int_{\mathcal{A}}\left[\frac{\partial\log f(y|x,\alpha;\theta^0)}{\partial\theta_k}\right]^2\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}\,d\alpha}_{=T\,J^{(2)}(y,x)}\\
&\quad\underbrace{\int_{\mathcal{A}}\left[\frac{\pi_1(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_1)}-\frac{\pi_2(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_2)}\right]^2\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}\,d\alpha}_{\equiv b(y,x)}.
\end{aligned}\tag{C.5}$$

For the integrand in the second term we have

$$\begin{aligned}
&\left[\frac{\pi_1(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_1)}-\frac{\pi_2(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_2)}\right]^2\\
&=\left[\left(\frac{\pi_1(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_1)}-1\right)-\left(\frac{\pi_2(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_2)}-1\right)\right]^2\\
&\leq 2\left(\frac{\pi_1(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_1)}-1\right)^2+2\left(\frac{\pi_2(\alpha|x)\int_{\mathcal{A}}f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_2)}-1\right)^2.
\end{aligned}\tag{C.6}$$

Furthermore

$$\begin{aligned}\left|\frac{\pi_1(\alpha|x)\int_{\mathcal{A}} f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_1)}-1\right| &= \left|\int_{\mathcal{A}} \frac{f(y|x,\beta;\theta^0)\left[\pi_1(\alpha|x)-\pi_1(\beta|x)\right]}{f_{Y|X}(y|x;\theta^0,\pi_1)}d\beta\right| \\ &\leq \left|\int_{\mathcal{A}} \frac{f(y|x,\beta;\theta^0)\left|\pi_1(\alpha|x)-\pi_1(\beta|x)\right|}{f_{Y|X}(y|x;\theta^0,\pi_1)}d\beta\right| \\ &\leq \kappa_T \int_{\mathcal{A}} \frac{f(y|x,\beta;\theta^0)\pi_1(\beta|x)d_x(\beta,\alpha)}{f_{Y|X}(y|x;\theta^0,\pi_1)}d\beta \\ &= \kappa_T \int_{\mathcal{A}} d_x(\beta,\alpha) f_{\boldsymbol{\alpha}|Y,X}(\beta|y,x;\theta^0,\pi_1)d\beta. \qquad \text{(C.7)}\end{aligned}$$

Therefore, also applying Jensen's inequality (namely $[\mathbb{E}Z]^2 \leq \mathbb{E}[Z^2]$), we obtain

$$\begin{aligned}A_1 &\equiv \int_{\mathcal{A}} \left(\frac{\pi_1(\alpha|x)\int_{\mathcal{A}} f(y|x,\beta;\theta^0)d\beta}{f_{Y|X}(y|x;\theta^0,\pi_1)}-1\right)^2 \frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\beta;\theta^0)d\beta}\,d\alpha \\ &\leq \kappa_T^2 \int_{\mathcal{A}}\int_{\mathcal{A}} d_x^2(\beta,\alpha) f_{\boldsymbol{\alpha}|Y,X}(\beta|y,x;\theta^0,\pi_1)\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\,d\beta d\alpha \\ &\leq \kappa_T^2 \int_{\mathcal{A}}\int_{\mathcal{A}} d_x^2(\beta,\alpha)\frac{f(y|x,\beta;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\,d\beta d\alpha \\ &\quad + \kappa_T^2 \int_{\mathcal{A}}\int_{\mathcal{A}} d_x^2(\beta,\alpha)\left|f_{\boldsymbol{\alpha}|Y,X}(\beta|y,x;\theta^0,\pi_1)-\frac{f(y|x,\beta;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\right|\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\,d\beta d\alpha \\ &\leq \kappa_T^2 \int_{\mathcal{A}}\int_{\mathcal{A}} d_x^2(\beta,\alpha)\frac{f(y|x,\beta;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\,d\beta d\alpha \\ &\quad + \kappa_T^2\sqrt{A_1\int_{\mathcal{A}}\int_{\mathcal{A}} d_x^4(\beta,\alpha)\frac{f(y|x,\beta;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\,d\beta d\alpha}, \qquad \text{(C.8)}\end{aligned}$$

where in the last step we applied Chebyshev's inequality. This implies

$$\begin{aligned}A_1 &\leq 2\kappa_T^2 \int_{\mathcal{A}}\int_{\mathcal{A}} d_x^2(\beta,\alpha)\frac{f(y|x,\beta;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\,d\beta d\alpha \\ &\quad + \kappa_T^4 \int_{\mathcal{A}}\int_{\mathcal{A}} d_x^4(\beta,\alpha)\frac{f(y|x,\beta;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\frac{f(y|x,\alpha;\theta^0)}{\int_{\mathcal{A}} f(y|x,\gamma;\theta^0)d\gamma}\,d\beta d\alpha \\ &= \frac{2\kappa_T^2}{T}D^{(2)}(y,x)+\frac{\kappa_T^4}{T^2}D^{(4)}(y,x). \qquad \text{(C.9)}\end{aligned}$$

By symmetry we obtain the same results for π_2, and we denote the corresponding term by A_2. Combining the above inequalities we find

$$b(y,x) \leq 2A_1 + 2A_2 \leq \frac{8\kappa_T^2}{T}D^{(2)}(y,x)+\frac{4\kappa_T^4}{T^2}D^{(4)}(y,x). \qquad \text{(C.10)}$$

Combining the above results gives part (i) of the lemma.

Part (ii) of the lemma is obtained analogously, but in that case there is no A_2 term, so that the bound is a factor two smaller.

Part (iii) and (iv) are also obtained by following the same arguments, but with $\tilde{E}$ taken into account whenever Chebyshev's inequality is applied. ■

Proof of Theorem 4.2. # Part I (Score): Applying part (*iii*) of Lemma C.1 yields

$$\sup_{\pi_1,\pi_2\in\Pi_\kappa^{\text{lip}}}\left(\frac{\partial L_{NT}(\theta^0,\pi_1)}{\partial\theta}-\frac{\partial L_{NT}(\theta^0,\pi_2)}{\partial\theta_k}\right)^2$$
$$=\sup_{\pi_1,\pi_2\in\Pi_\kappa^{\text{lip}}}\left[\frac{1}{NT}\sum_{i=1}^N\left(\frac{\partial\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi_1)}{\partial\theta_k}-\frac{\partial\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi_2)}{\partial\theta_k}\right)\right]^2$$
$$\leq\frac{8\,\kappa_T^2}{T^2}\left(\frac{1}{N}\sum_{i=1}^N J_{kk}^{(2)}(Y_i,X_i)\right)\left[\left(\frac{1}{N}\sum_{i=1}^N D^{(2)}(Y_i,X_i)\right)+\frac{\kappa_T^2}{2T}\left(\frac{1}{N}\sum_{i=1}^N D^{(4)}(Y_i,X_i)\right)\right]. \tag{C.11}$$

Together with the assumptions, this shows the result.

Part II (Hessian): We have

$$\frac{\partial^2 L_{NT}(\theta^0,\pi)}{\partial\theta\partial\theta'}=\frac{1}{NT}\sum_{i=1}^N\frac{\partial^2\,\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi)}{\partial\theta\partial\theta'}$$
$$=\frac{1}{NT}\sum_{i=1}^N\Bigg\{\int_{\mathcal{A}}\left[\frac{\partial^2\log f(Y_i|X_i,\alpha;\theta^0)}{\partial\theta\partial\theta'}+\frac{\partial\log f(Y_i|X_i,\alpha;\theta^0)}{\partial\theta}\frac{\partial\log f(Y_i|X_i,\alpha;\theta^0)}{\partial\theta'}\right]$$
$$f_{\boldsymbol{\alpha}|Y,X}(\alpha|Y_i,X_i;\theta^0,\pi)\,d\alpha-\frac{\partial\,\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi)}{\partial\theta}\frac{\partial\,\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi)}{\partial\theta'}\Bigg\} \tag{C.12}$$

Thus, we have

$$\frac{\partial^2 L_{NT}(\theta^0,\pi_1)}{\partial\theta\partial\theta'}-\frac{\partial^2 L_{NT}(\theta^0,\pi_2)}{\partial\theta\partial\theta'}=A_1-A_2, \tag{C.13}$$

where

$$A_1=\frac{1}{NT}\sum_{i=1}^N\int_{\mathcal{A}}\left[\frac{\partial^2\log f(Y_i|X_i,\alpha;\theta^0)}{\partial\theta\partial\theta'}+\frac{\partial\log f(Y_i|X_i,\alpha;\theta^0)}{\partial\theta}\frac{\partial\log f(Y_i|X_i,\alpha;\theta^0)}{\partial\theta'}\right]$$
$$f_{\boldsymbol{\alpha}|Y,X}(\alpha|Y_i,X_i;\theta^0,\pi_1)-f_{\boldsymbol{\alpha}|Y,X}(\alpha|Y_i,X_i;\theta^0,\pi_2)\,d\alpha, \tag{C.14}$$

and

$$A_2=\frac{1}{NT}\sum_{i=1}^N\Bigg[\frac{\partial\,\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi_1)}{\partial\theta}\frac{\partial\,\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi_1)}{\partial\theta'}$$
$$-\frac{\partial\,\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi_2)}{\partial\theta}\frac{\partial\,\log\, f_{Y|X}(Y_i|X_i;\theta^0,\pi_2)}{\partial\theta'}\Bigg] \tag{C.15}$$

Analogous to the proof of Lemma C.1 that was used in Part I we obtain the following bound for A_1:

$$A_{1,k_1k_2}^2\leq\frac{8\,\kappa_T^2}{T}\left(\frac{1}{N}\sum_{i=1}^N H_{k_1k_2}(Y_i,X_i)\right)\left[\left(\frac{1}{N}\sum_{i=1}^N D^{(2)}(Y_i,X_i)\right)+\frac{\kappa_T^2}{2T}\left(\frac{1}{N}\sum_{i=1}^N D^{(4)}(Y_i,X_i)\right)\right]. \tag{C.16}$$

Therefore, $A_1 = \mathcal{O}_p(\kappa_T/\sqrt{T})$ under our assumptions, uniformly over π_1 and π_2. Applying part (ii) of Lemma C.1 we obtain

$$
\begin{aligned}
|A_{1,k_1k_2}| \leq \frac{2}{N\sqrt{T}} \sum_{i=1}^{N} \Bigg[& J^{(1)}_{k_1}(Y_i, X_i) \sqrt{4\,\kappa_T^2\, J^{(2)}_{k_2k_2}(Y_i, X_i)\, \left(D^{(2)}(Y_i, X_i) + \frac{\kappa_T^2}{2T} D^{(4)}(Y_i, X_i) \right)} \\
& + \text{same term with } k_1 \leftrightarrow k_2 \\
& + 4\,\kappa_T^2\, J^{(2)}_{k_2k_2}(Y_i, X_i) \left(D^{(2)}(Y_i, X_i) + \frac{\kappa_T^2}{2T} D^{(4)}(Y_i, X_i) \right) \Bigg] \\
\leq \frac{8\,\kappa_T^2}{\sqrt{T}} & \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[J^{(1)}_{k_1}(Y_i, X_i) \right]^2} \sqrt{\frac{1}{N} \sum_{i=1}^{N} J^{(2)}_{k_2k_2}(Y_i, X_i) \left(D^{(2)}(Y_i, X_i) + \frac{\kappa_T^2}{2T} D^{(4)}(Y_i, X_i) \right)} \\
& + \text{same term with } k_1 \leftrightarrow k_2 \\
& + \frac{8\,\kappa_T^2}{\sqrt{T}} \frac{1}{N} \sum_{i=1}^{N} J^{(2)}_{k_2k_2}(Y_i, X_i) \left(D^{(2)}(Y_i, X_i) + \frac{\kappa_T^2}{2T} D^{(4)}(Y_i, X_i) \right) \Bigg].
\end{aligned}
\tag{C.17}
$$

Furthermore, we have

$$
\begin{aligned}
& \frac{1}{N} \sum_{i=1}^{N} J^{(2)}_{kk}(Y_i, X_i) \left(D^{(2)}(Y_i, X_i) + \frac{\kappa_T^2}{2T} D^{(4)}(Y_i, X_i) \right) \\
& \leq \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[J^{(2)}_{kk}(Y_i, X_i) \right]^2} \left(\sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[D^{(2)}(Y_i, X_i) \right]^2} + \frac{\kappa_T^2}{2T} \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[D^{(4)}(Y_i, X_i) \right]^2} \right).
\end{aligned}
\tag{C.18}
$$

Thus, our assumptions also guarantee $A_2 = \mathcal{O}_p(\kappa_T/\sqrt{T})$, uniformly over π_1 and π_2, so that the same holds for the difference of the Hessians. ■

Proof of Theorem 4.3. We have

$$
\begin{aligned}
\frac{\partial \overline{L}_{NT}(\theta^0, \pi)}{\partial \theta} &= \frac{1}{NT} \sum_{i=1}^{N} \left\{ \int_{\mathcal{A}} \mathbb{E}_{Y|X_i,\alpha} \left[\frac{\partial \log\, f_{Y|X}(Y|X_i; \theta, \pi)}{\partial \theta} \right] \pi^0(\alpha|X_i)\, d\alpha \right\} \\
&= \frac{1}{NT} \sum_{i=1}^{N} \left\{ \int_{\mathcal{A}} \mathbb{E}_{Y|X_i,\alpha} \left[\frac{\partial \log\, f_{Y|X}(Y|X_i; \theta, \pi)}{\partial \theta} \right] \left[\pi^0(\alpha|X_i) - \pi(\alpha|X_i) \right]\, d\alpha \right\} \\
&= \frac{1}{NT} \sum_{i=1}^{N} \Bigg\{ \int_{\mathcal{A}} \mathbb{E}_{Y|X_i,\alpha} \left[\frac{\partial \log\, f_{Y|X}(Y|X_i; \theta, \pi)}{\partial \theta} \right] \left[\sqrt{\pi^0(\alpha|X_i)} + \sqrt{\pi(\alpha|X_i)} \right] \\
&\qquad \left[\sqrt{\pi^0(\alpha|X_i)} - \sqrt{\pi(\alpha|X_i)} \right]\, d\alpha \Bigg\}.
\end{aligned}
\tag{C.19}
$$

Applying Chebyshev's inequality we find

$$
\begin{aligned}
&\left|\frac{\partial \overline{L}_{NT}(\theta^0,\pi)}{\partial \theta_k}\right| \\
&\leq \frac{1}{T}\,\mathcal{D}_{\mathrm{H}}(\pi,\pi^0)\sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}\frac{\partial\,\log\,f_{Y|X}(Y|X_i;\theta^0,\pi)}{\partial\theta_k}\right]^2\left[\sqrt{\pi^0(\alpha|X_i)}+\sqrt{\pi(\alpha|X_i)}\right]^2 d\alpha} \\
&\leq \frac{\sqrt{2}}{T}\,\mathcal{D}_{\mathrm{H}}(\pi,\pi^0)\sqrt{\underbrace{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}\frac{\partial\,\log\,f_{Y|X}(Y|X_i;\theta^0,\pi)}{\partial\theta_k}\right]^2\left[\pi^0(\alpha|X_i)+\pi(\alpha|X_i)\right]d\alpha}_{=B(\pi)}}.
\end{aligned}
\tag{C.20}
$$

Using the upper bound on $\pi^0(\alpha|X_i)$ and $\pi(\alpha|X_i)$ we find

$$
B(\pi)\leq\frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}\frac{\partial\,\log\,f_{Y|X}(Y|X_i;\theta^0,\pi)}{\partial\theta_k}\right]^2\pi_T^{\mathrm{up}}(\alpha|X_i)\,d\alpha \tag{C.21}
$$

For the integrand in the last expression we have

$$
\begin{aligned}
&\left[\mathbb{E}_{Y|X_i,\alpha}\frac{\partial\,\log\,f_{Y|X}(Y|X_i;\theta^0,\pi)}{\partial\theta_k}\right]^2 \\
&\leq\Bigg[\underbrace{\mathbb{E}_{Y|X_i,\alpha}\int_{\mathcal{A}}\frac{\partial\log f(Y|X_i,\alpha;\theta^0)}{\partial\theta_k}f^{\mathrm{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|Y,X_i;\theta^0)d\alpha}_{=T\mathbb{E}_{Y|X_i,\alpha}J_k^{(1)}(Y,X_i)} \\
&\qquad+\mathbb{E}_{Y|X_i,\alpha}\left(\frac{\partial\,\log\,f_{Y|X}(Y|X_i;\theta^0,\pi)}{\partial\theta_k}-\int_{\mathcal{A}}\frac{\partial\log f(Y|X_i,\alpha;\theta^0)}{\partial\theta_k}f^{\mathrm{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|Y,X_i;\theta^0)d\alpha\right)\Bigg]^2 \\
&\leq 2T\left[\mathbb{E}_{Y|X_i,\alpha}J_k^{(1)}(Y,X_i)\right]^2 \\
&\qquad+2\left[\mathbb{E}_{Y|X_i,\alpha}\left(\frac{\partial\,\log\,f_{Y|X}(Y|X_i;\theta^0,\pi)}{\partial\theta_k}-\int_{\mathcal{A}}\frac{\partial\log f(Y|X_i,\alpha;\theta^0)}{\partial\theta_k}f^{\mathrm{unif}}_{\boldsymbol{\alpha}|Y,X}(\alpha|Y,X_i;\theta^0)d\alpha\right)\right]^2 \\
&\leq 2T\left[\mathbb{E}_{Y|X_i,\alpha}J_k^{(1)}(Y,X_i)\right]^2 \\
&+4\,\kappa_T^2\,\left(\mathbb{E}_{Y|X_i,\alpha}\,J_{kk}^{(2)}(Y,X_i)\right)\left[\left(\mathbb{E}_{Y|X_i,\alpha}\,D^{(2)}(Y,X_i)\right)+\frac{\kappa_T^2}{2T}\left(\mathbb{E}_{Y|X_i,\alpha}\,D^{(4)}(Y,X_i)\right)\right],
\end{aligned}
\tag{C.22}
$$

where we applied part (iv) or Lemma C.1. By Chebyshev's inequality and applying

the assumptions, we thus obtain

$$
\begin{aligned}
B(\pi) \leq\ & 4T\,\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}J_k^{(1)}(Y,X_i)\right]^2\pi_T^{\mathrm{up}}(\alpha|X_i)d\alpha \\
& + 8\kappa_T^2\sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}\,J_{kk}^{(2)}(Y,X_i)\right]^2\pi_T^{\mathrm{up}}(\alpha|X_i)d\alpha} \\
& \qquad \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}\,D^{(2)}(Y,X_i)\right]^2\pi_T^{\mathrm{up}}(\alpha|X_i)d\alpha} \\
& + \frac{4\kappa_T^4}{T}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}\,J_{kk}^{(2)}(Y,X_i)\right]^2\pi_T^{\mathrm{up}}(\alpha|X_i)d\alpha} \\
& \qquad \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\mathbb{E}_{Y|X_i,\alpha}\,D^{(4)}(Y,X_i)\right]^2\pi_T^{\mathrm{up}}(\alpha|X_i)d\alpha} \\
=\ & \mathcal{O}_p(1)+\mathcal{O}_p(\kappa_T^2) \qquad \text{(C.23)}
\end{aligned}
$$

Combining the above results gives the statement in the theorem. ■

Proof of Lemma 4.4. Applying part (iii) of Lemma C.1 we find

$$
\begin{aligned}
\mathbb{E}\nu_{NT}^2(\pi_T) &= \frac{1}{NT}\sum_{i=1}^{N}\mathbb{E}\nu_{NT,i}^2(\pi_T) \\
&\leq \frac{1}{T}\mathbb{E}\left[\frac{\partial\log\,f_{Y|X}(Y|X;\theta^0,\pi)}{\partial\theta}-\frac{\partial\log\,f_{Y|X}(Y|X;\theta^0,\pi^0)}{\partial\theta}\right]^2 \\
&\leq \frac{8\kappa_T^2}{T}\left(\mathbb{E}\,J_{kk}^{(2)}(Y,X)\right)\left[\left(\mathbb{E}\,D^{(2)}(Y,X)\right)+\frac{\kappa_T^2}{2T}\left(\mathbb{E}\,D^{(4)}(Y,X)\right)\right] \\
&\leq \mathcal{O}\left(\frac{\kappa_T^2}{T}\right), \qquad \text{(C.24)}
\end{aligned}
$$

and therefore $\nu_{NT}(\pi_T)=\mathcal{O}_p(\kappa_T/\sqrt{T})$, uniformly over $\pi\in\Pi_{T,\kappa}^{\mathrm{lip}}$. ■

Proof of Corollary 4.5. Consistency of $\hat{\theta}(\pi^0)$ and $\hat{\theta}(\hat{\pi})$ together with Assumption B.2(v) and (vi) and Theorem 4.2 imply that

$$
\begin{aligned}
\left(\hat{\theta}(\pi^0)-\theta^0\right) &= \left(\frac{\partial^2 L_{NT}(\theta^0,\pi^0)}{\partial\theta\partial\theta'}\right)^{-1}\frac{\partial L_{NT}(\theta^0,\pi^0)}{\partial\theta}+o_p\left((NT)^{-1/2}\right), \\
\left(\hat{\theta}(\hat{\pi})-\theta^0\right) &= \left(\frac{\partial^2 L_{NT}(\theta^0,\hat{\pi})}{\partial\theta\partial\theta'}\right)^{-1}\frac{\partial L_{NT}(\theta^0,\hat{\pi})}{\partial\theta}+o_p\left((NT)^{-1/2}\right). \qquad \text{(C.25)}
\end{aligned}
$$

Therefore

$$\hat{\theta}(\hat{\pi}) - \hat{\theta}(\pi^0) = \left(\frac{\partial^2 L_{NT}(\theta^0, \pi^0)}{\partial\theta\partial\theta'}\right)^{-1} \left[\frac{\partial L_{NT}(\theta^0, \hat{\pi})}{\partial\theta} - \frac{\partial L_{NT}(\theta^0, \pi^0)}{\partial\theta}\right]$$
$$+ \left[\left(\frac{\partial^2 L_{NT}(\theta^0, \hat{\pi})}{\partial\theta\partial\theta'}\right)^{-1} - \left(\frac{\partial^2 L_{NT}(\theta^0, \pi^0)}{\partial\theta\partial\theta'}\right)^{-1}\right] \frac{\partial L_{NT}(\theta^0, \hat{\pi})}{\partial\theta}$$
$$+ o_p\left((NT)^{-1/2}\right) \tag{C.26}$$

By Assumption B.2(v) and Theorem 4.3 we find the first term on the right hand side of (C.26) to be of order $\mathcal{O}_p(\kappa_T/T)\mathcal{D}_{\mathrm{H}}(\hat{\pi}, \pi^0)$. By Theorem 4.2 and again Assumption B.2(v) we find the second term on the right hand side to be of order $o_p\left((NT)^{-1/2}\right)$. For this last step we need to bound the difference between the inverse of two matrices, which can e.g. by done by using the general matrix relation $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, which implies $\|A^{-1} - B^{-1}\| \leq \|A^{-1}\|\|B - A\|\|B^{-1}\|$, where the norm is the operator norm. The statement of the corollary thus follows from (C.26). ■

C.3 Proofs for Section 4.3

Proof of Lemma 4.6. Cross-sectional independence implies that

$$\mathbb{E}(\psi_{NT}^2(\pi_T)|X_1, \ldots, X_N) \leq \left[\mathcal{D}_{\mathrm{KL}}^{(2)}(f_Y(\pi_T)||f_Y(\pi^0))\right]^2, \tag{C.27}$$

where

$$\mathcal{D}_{\mathrm{KL}}^{(2)}(f_Y(\pi)||f_Y(\pi^0)) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{Y}_{\mathcal{T}}}\left(\log\left[\frac{f_{Y|X}(y|X_i;\theta^0,\pi^0)}{f_{Y|X}(y|X_i;\theta^0,\pi)}\right]\right)^2 f_{Y|X}(y|X_i;\theta^0,\pi^0)\,dy}. \tag{C.28}$$

Therefore

$$\psi_{NT}(\pi_T) = \mathcal{O}_p(1)\,\mathcal{D}_{\mathrm{KL}}^{(2)}(f_Y(\pi_T)||f_Y(\pi^0). \tag{C.29}$$

By assumption $\pi^0(\alpha|x)/\pi_T(\alpha|x) \leq \pi_T^{\mathrm{up}}(\alpha|x)/\pi_T^{\mathrm{low}}(\alpha|x)$ is bounded. This also implies that $\frac{f_{Y|X}(y|x;\theta^0,\pi^0)}{f_{Y|X}(y|x;\theta^0,\pi)}$ is bounded. Note that for all $0 < z \leq b$ we have $(\log z)^2 \leq d^2(\log z + 1/z - 1)$, with $d^2 = b^2/(b-1)$. Thus, there exists a constants d such that

$$\mathcal{D}_{\mathrm{KL}}^{(2)}(f_Y(\pi)||f_Y(\pi^0)) \leq d\sqrt{\mathcal{D}_{\mathrm{KL}}(f_Y(\pi)||f_Y(\pi^0))}\,. \tag{C.30}$$

This proofs the lemma. ■

Proof of Theorem 4.7. Assumption 4.3(i) guarantees that there exists $\tilde{\pi}_T \in \Pi_T$ such that $\mathcal{D}_{\mathrm{KL}}(f_Y(\tilde{\pi})||f_Y(\pi^0)) = \mathcal{O}_p(T^{-2\mu})$. For such a $\tilde{\pi}_T$ we have

$$\begin{aligned} T\left[L_{NT}(\tilde{\pi}_T) - L_{NT}(\theta^0, \pi^0)\right] &\geq T\left[L_{NT}(\theta^0, \tilde{\pi}_T) - L_{NT}(\theta^0, \pi^0)\right] \\ &= \mathcal{O}_p(T^{-2\mu}) + o_p\left(\sqrt{\frac{T}{N}}\frac{T^{-\mu}}{\kappa_T}\right). \end{aligned} \tag{C.31}$$

Here, we have also used assumption 4.2. The optimal $\hat{\pi}$ needs to satisfy $L_{NT}(\hat{\pi}) \geq L_{NT}(\theta^0, \tilde{\pi}_T)$. Therefore

$$T\left[L_{NT}(\hat{\pi}) - L_{NT}(\theta^0, \pi^0)\right] \geq \mathcal{O}_p(T^{-2\mu}) + o_p\left(\sqrt{\frac{T}{N}}\frac{T^{-\mu}}{\kappa_T}\right). \tag{C.32}$$

Using the expansion (4.6) and our results from on the score and Hessian and on $\theta(\pi)$, the last inequality yields

$$\begin{aligned} &\mathcal{D}_{\mathrm{KL}}\left(f_Y(\hat{\pi})||f_Y(\pi^0)\right) \\ &\leq \mathcal{O}_p(T^{-2\mu}) + o_p\left(\sqrt{\frac{T}{N}}\frac{T^{-\mu}}{\kappa_T}\right) + o_p\left(\frac{1}{\kappa_T}\sqrt{\frac{T}{N}}\right)\sqrt{\mathcal{D}_{\mathrm{KL}}\left(f_Y(\hat{\pi})||f_Y(\pi^0)\right)} \\ &\qquad + T\left[\mathcal{O}_p\left(\frac{1}{\sqrt{NT}}\right) + \mathcal{O}_p\left(\frac{\kappa_T}{T}\right)\mathcal{D}_{\mathrm{H}}(\hat{\pi}, \pi^0)\right]^2 \\ &\leq \mathcal{O}_p(T^{-2\mu}) + o_p\left(\sqrt{\frac{T}{N}}\frac{T^{-\mu}}{\kappa_T}\right) + o_p\left(\frac{1}{\kappa_T}\sqrt{\frac{T}{N}}\right)\sqrt{\mathcal{D}_{\mathrm{KL}}\left(f_Y(\hat{\pi})||f_Y(\pi^0)\right)} \\ &\quad + \mathcal{O}_p\left(\frac{1}{N}\right) + \mathcal{O}_p\left(\frac{\kappa_T}{\sqrt{NT}}\right)\left[\sqrt{\mathcal{D}_{\mathrm{KL}}(f(\hat{\pi})||f(\pi^0))} + T^{-\mu}\right] \\ &\qquad + \mathcal{O}_p\left(\frac{\kappa_T^2}{T}\right)\left[\sqrt{\mathcal{D}_{\mathrm{KL}}(f(\hat{\pi})||f(\pi^0))} + T^{-\mu}\right]^2. \end{aligned} \tag{C.33}$$

From this we can conclude that

$$\begin{aligned} \sqrt{\mathcal{D}_{\mathrm{KL}}(f(\hat{\pi})||f(\pi^0))} &= \mathcal{O}_p(T^{-\mu}) + o_p\left(\frac{1}{\kappa_T}\sqrt{\frac{T}{N}}\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N}}\right) + o_p\left(\sqrt{\sqrt{\frac{T}{N}}\frac{T^{-\mu}}{\kappa_T}}\right) \\ &= \mathcal{O}_p(T^{-\mu}) + o_p\left(\frac{1}{\kappa_T}\sqrt{\frac{T}{N}}\right). \end{aligned} \tag{C.34}$$

By assumption 4.3(ii) this implies part (i) of the theorem. Part (ii) of the theorem follows from part (i) by applying Corollary 4.5. ■

D Further Discussions for Section 5

We now present the technical justification for the choice of parameter set Π_T in equation (5.5).

D.1 Approximating Unknown Distributions

For simplicity we consider the case where $\mathcal{A} = \mathbb{R}$, i.e. the number of dimensions of the incidental parameter space is $M = 1$, and there are no additional restrictions on $\mathcal{A}$. In that case we have $\mathcal{K}_T^{\Omega}(\alpha, \beta; x) = \phi(\alpha; \beta, \Omega_T(\beta, x)$, which is a standard normal pdf with mean β and variance $\Omega_T(\alpha, x) = \frac{\rho_T}{T}\,\Lambda_T(\alpha, x)$, where we denote the inverse that appears in equation (5.6) by $\Lambda_T(\alpha, x)$. In the rest of this subsection we drop the dependence on the regressor value x for notational convenience.

We have to show that an unknown density $\pi^0(\alpha)$ can be approximated well by $\pi^{\text{approx}}(\alpha) = \int_{\mathbb{R}} \phi(\alpha; \beta, \Omega_T(\beta))\pi(\beta)d\beta$ for some appropriate choice of density $\pi(\beta)$. First we note that if both $\pi^0(\alpha)$ and $\Omega_T(\alpha)$ are arbitrarily often differentiable, then we can achieve $\pi^0 = \pi^{\text{approx}}$ by choosing

$$\pi(\alpha) = \frac{1}{\sum_{q=0}^{\infty} \frac{1}{2^q\, q!} \frac{d^{2q}}{d\alpha^{2q}} \left[\Omega_T(\alpha)\right]^q}\, \pi^0(\alpha). \tag{D.1}$$

This expression has to be understood as a formal power expansion, which can be rewritten as

$$\begin{aligned}
\pi(\alpha) &= \frac{1}{1 + \sum_{q=1}^{\infty} \frac{1}{2^q\, q!} \frac{d^{2q}}{d\alpha^{2q}} \left[\Omega_T(\alpha)\right]^q}\, \pi^0(\alpha) = \sum_{r=0}^{\infty} \left[-\sum_{q=1}^{\infty} \frac{1}{2^q\, q!} \frac{d^{2q}}{d\alpha^{2q}} [\Omega_T(\alpha)]^q \right]^r \pi^0(\alpha) \\
&= \pi^0(\alpha) - \frac{1}{2}\frac{d^2}{d\alpha^2}\left[\Omega_T(\alpha)\pi^0(\alpha)\right] + \frac{1}{4}\frac{d^2}{d\alpha^2}\left[\Omega_T(\alpha)\frac{d^2}{d\alpha^2}\left[\Omega_T(\alpha)\pi^0(\alpha)\right]\right] + \ldots \\
&= \sum_{q=0}^{\infty} \left(\frac{\rho_T}{T}\right)^q A_q(\alpha)\,,
\end{aligned} \tag{D.2}$$

where the first few expansion coefficients $A_q(\alpha)$ read

$$\begin{aligned}
A_0(\alpha) &= \pi^0(\alpha)\,, \\
A_1(\alpha) &= -\frac{1}{2}\frac{d^2}{d\alpha^2}\left[\Lambda_T(\alpha)\pi^0(\alpha)\right]\,, \\
A_2(\alpha) &= \frac{1}{4}\frac{d^2}{d\alpha^2}\left[\Lambda_T(\alpha)\frac{d^2}{d\alpha^2}\left[\Lambda_T(\alpha)\pi^0(\alpha)\right]\right] - \frac{1}{8}\frac{d^4}{d\alpha^4}\left[[\Lambda_T(\alpha)]^2\pi^0(\alpha)\right],
\end{aligned} \tag{D.3}$$

and the expression for all higher $A_q(\alpha)$ can be obtained by expanding the first line of equation (D.2) and sorting terms by powers of $\Omega_T(\alpha)$.[11] Under appropriate regularity conditions we have $\int_{\mathbb{R}} A_q(\alpha)d\alpha = 0$ for all $q \geq 1$. For example, we have $\int_{\mathbb{R}} A_1(\alpha)d\alpha = \lim_{\alpha\to-\infty} \frac{1}{2}\frac{d}{d\alpha}\left[\Lambda_T(\alpha)\pi^0(\alpha)\right] - \lim_{\alpha\to+\infty} \frac{1}{2}\frac{d}{d\alpha}\left[\Lambda_T(\alpha)\pi^0(\alpha)\right]$, and we assume that these limits are both zero.

[11] In the special case where $\Lambda_T(\alpha)$ does not depend on α one obtains the simple general formula $A_q(\alpha) = (-2^q q!)^{-1}\,\Lambda_T^q\, d^{2q}/d\alpha^{2q}\pi^0(\alpha)$.

The highest derivatives of $\pi^0(\alpha)$ and $\Lambda_T(\alpha)$ that appear in $A_q(\alpha)$ are of order $2q$. Thus, if $\pi^0(\alpha)$ is r times differentiable, and assuming that $\Lambda_T(\alpha)$ is also sufficiently often differentiable, we can choose

$$\pi(\alpha) = \sum_{q=0}^{\lfloor r/2 \rfloor} \left(\frac{\rho_T}{T}\right)^q A_q(\alpha), \tag{D.4}$$

where $\lfloor r/2 \rfloor$ denotes the largest integer smaller or equal to $r/2$. For large T this choice of $\pi(\alpha)$ is close to $\pi^0(\alpha)$, so that $\pi(\alpha) \geq 0$ is satisfied asymptotically. For this choice of $\pi(\alpha)$ one can show that under appropriate regularity conditions $\mathcal{D}_{\mathrm{KL}}(\pi^{\mathrm{approx}}, \pi^0) = \mathcal{O}_p[(\rho_T/T)^r]$. Since $\mathcal{D}_{\mathrm{KL}}(f_Y(\pi^{\mathrm{approx}})||f_Y(\pi^0)) \leq \mathcal{D}_{\mathrm{KL}}(\pi^{\mathrm{approx}}||\pi^0)$ this implies that Assumption 4.3(i) is satisfied with $\mu_T = (\rho_T/T)^{r/2}$.

D.2 Approximate Identification of $\pi(\alpha|x)$

Assumption 4.3(ii) is an approximate identification condition of $\pi = \pi(\alpha|x)$ within the set Π_T. The identification is approximate since the "slackness" μ_T appears on the right hand side of the inequality in the assumption. We are going to define an infeasible parameter set that satisfies Assumption 4.3(ii) with $\mu_T = 0$ for algebraic reasons and then show that the kernel construction of Section 5.1 provides a feasible parameter set that approximates the infeasible one sufficiently well. We define

$$\begin{aligned} f_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta,\pi) &= \frac{f(y|x,\alpha;\theta)\,\pi(\alpha|x)}{f_{Y|X}(y|x;\theta,\pi)}\,, \\ \mathcal{K}_T^0(\alpha,\beta;x) &= \int_{\mathcal{Y}_T} f_{\boldsymbol{\alpha}|Y,X}\left(\alpha|y,x;\theta^0,\pi^0\right) f(y|x,\beta;\theta^0)dy. \end{aligned} \tag{D.5}$$

This is not a Bayesian paper, but $f_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta,\pi)$ clearly has a Bayesian interpretation, namely it is the posterior distribution of $\boldsymbol{\alpha}$ for given $Y = y$ under the prior $\pi(\alpha|x)$, for given values of x and θ. In $\mathcal{K}_T^0(\alpha,\beta;x)$ this posterior distribution is integrated over the true distribution of Y, i.e. $\mathcal{K}_T^0(\alpha,\beta;x)$ is the expected posterior distribution under the prior π^0 conditional on the individual effect equal to β. $\mathcal{K}_T^0(\alpha,\beta;x)$ is a kernel function that defines an endomorphism of distributions over $\mathcal{A}$ for each $x \in \mathcal{X}_T$. Namely, for $\pi \in \Pi_T^{\mathcal{A}}$ we have

$$\left[\mathcal{K}_T^0\,\pi\right](\alpha|x) = \int_{\mathcal{A}} \mathcal{K}_T^0(\alpha,\beta;x)\pi(\beta|x)d\beta. \tag{D.6}$$

For given $q \in \mathbb{N}$ an infeasible parameter set is defined by

$$\Pi_T^{(q)} = (\mathcal{K}_T^0)^q \left(\Pi_T^{\mathrm{up}} \cap \Pi_T^{\mathrm{low}}\right). \tag{D.7}$$

This is the set of all distributions that can be generated by q consecutive applications of the kernel $\mathcal{K}_T^0$ to an element of $\Pi_T^{\text{up}} \cap \Pi_T^{\text{low}}$ (the set of distribution that satisfy some appropriate upper and lower bound). The parameter set $\Pi_T^{(q)}$ is infeasible, since π^0 and θ^0 enter into the definition of $\mathcal{K}_T^0$. We assume $\pi^0 \in \Pi_T^{\text{up}} \cap \Pi_T^{\text{low}}$. Then we have $\pi^0 \in \Pi_T^{(q)}$ for all $q \in \mathbb{N}$ because π^0 is a fix point of $\mathcal{K}_T^0$.

The main motivation for defining $\Pi_T^{(q)}$ is the following algebraic result: for every q there exists a constant c_q such that for all $\pi \in \Pi_T^{(q)}$

$$\mathcal{D}_{\text{H}}(\pi, \pi^0) \leq c_q \ \left[\mathcal{D}_{\text{H}}(f(\pi), f(\pi^0))\right]^{\left(1-\frac{1}{2q+1}\right)}. \tag{D.8}$$

The proof is given below. Since $\mathcal{D}_{\text{H}}(f(\pi), f(\pi^0)) \leq \sqrt{\mathcal{D}_{\text{KL}}(f(\pi)||f(\pi^0))}$ this means that if q becomes large the set $\Pi_T^{(q)}$ approximately satisfies Assumption 4.3(ii).

In order to approximate this infeasible parameter set by a feasible one, we note that $\mathcal{K}_T^0(\alpha, \beta; x)$ has generic properties as T becomes large. Namely, under some regularity conditions on π^0 one can apply a Laplace approximation argument to show that the distribution of α whose probability density function is given by $\mathcal{K}_T^0(\alpha, \beta; x)$ becomes a Gaussian distribution with mean β and variance $2\mathcal{I}^{-1}(\beta, \theta^0, x)/T$ as $T \to \infty$, where the variance contains the inverse of the information matrix.

Thus, applying $\mathcal{K}_T^0$ to a distribution becomes asymptotically equivalent to applying a Gaussian kernel smoothing with variance $2\mathcal{I}^{-1}(\beta, \theta^0, x)$. This suggests to define a feasible parameter set by replacing $(\mathcal{K}_T^0)^q$ with a Gaussian kernel of appropriate variance, which is exactly the construction of section 5.1, with $q \to \infty$ as $N, T \to \infty$.

We now want to prove inequality (D.8). Note that for $\pi \in \Pi_T^{(1)}$ there exist $\tilde{f}(y|x)$ and $\tilde{\pi}(\alpha|x)$ such that

$$\tilde{f}(y|x) = \int_{\mathcal{A}} f(y|x, \alpha; \theta^0)\, \tilde{\pi}(\alpha|x) d\alpha, \quad \pi(\alpha|x) = \int_{\mathcal{Y}_T} \frac{f(y|x, \alpha; \theta)\, \pi^0(\alpha|x)}{f_{Y|X}(y|x; \theta, \pi^0)}\, \tilde{f}(y|x) dy. \tag{D.9}$$

Therefore

$$
\begin{aligned}
\mathcal{D}_{\mathrm{H}}^2(\pi,\pi^0) &= \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[\sqrt{\pi(\alpha|X_i)}-\sqrt{\pi^0(\alpha|X_i)}\right]^2 d\alpha \\
&= 2-\frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\sqrt{\frac{\pi^0(\alpha|X_i)}{\pi(\alpha|X_i)}}\,\pi(\alpha|X_i)d\alpha \\
&= 2-\frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\int_{\mathcal{Y}_T}\frac{1}{\sqrt{\frac{\pi(\alpha|X_i)}{\pi^0(\alpha|X_i)}}}\,f_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0,\pi^0)\,\tilde{f}(y|x)dyd\alpha \\
&\le 2-\frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{Y}_T}\frac{1}{\sqrt{\int_{\mathcal{A}}\frac{\pi(\alpha|X_i)}{\pi^0(\alpha|X_i)}f_{\boldsymbol{\alpha}|Y,X}(\alpha|y,x;\theta^0,\pi^0)d\alpha}}\,\tilde{f}(y|x)dy \\
&= 2-\frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{Y}_T}\sqrt{\frac{f_{Y|X}(y|x;\theta^0,\pi^0)}{f_{Y|X}(y|x;\theta^0,\pi)}}\,\tilde{f}(y|x)dy \\
&= \mathcal{D}_{\mathrm{H}}^2(f_{Y|X}(\pi),f_{Y|X}(\pi^0)) \\
&\quad+\frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{Y}_T}\left[1-\sqrt{\frac{f_{Y|X}(y|x;\theta^0,\pi^0)}{f_{Y|X}(y|x;\theta^0,\pi)}}\right]\left[\tilde{f}(y|x)-f_{Y|X}(y|x;\theta^0,\pi)\right]dy \\
&= \mathcal{D}_{\mathrm{H}}^2(f_{Y|X}(\pi),f_{Y|X}(\pi^0)) \\
&\quad-\frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{Y}_T}\left[1-\sqrt{\frac{f_{Y|X}(y|x;\theta^0,\pi^0)}{f_{Y|X}(y|x;\theta^0,\pi)}}\right]\left[1-\sqrt{\frac{\tilde{f}(y|x)}{f_{Y|X}(y|x;\theta^0,\pi)}}\right] \\
&\qquad\qquad\left[1+\sqrt{\frac{\tilde{f}(y|x)}{f_{Y|X}(y|x;\theta^0,\pi)}}\right]f_{Y|X}(y|x;\theta^0,\pi)dy \\
&\le \mathcal{D}_{\mathrm{H}}^2(f_{Y|X}(\pi),f_{Y|X}(\pi^0)) \\
&\quad+2\sqrt{2}\,\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi),f_{Y|X}(\pi^0))\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi),\tilde{f})\left[1+\sup_{y,x}\sqrt{\frac{\tilde{f}(y|x)}{f_{Y|X}(y|x;\theta^0,\pi)}}\right]
\end{aligned}
\tag{D.10}
$$

Here we have used Jensen's inequality in the fourth line and Chebychev's inequality in the last step. Note that

$$
\sup_{y,x}\frac{\tilde{f}(y|x)}{f_{Y|X}(y|x;\theta^0,\pi)}=\sup_{y,x}\frac{f_{Y|X}(y|x;\theta^0,\tilde{\pi})}{f_{Y|X}(y|x;\theta^0,\pi)}\le\sup_{\alpha,x}\frac{\tilde{\pi}(\alpha|x)}{\pi(\alpha|x)}. \tag{D.11}
$$

Applying the triangle inequality $\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi),\tilde{f})\le\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi),f_{Y|X}(\pi^0))+\mathcal{D}_{\mathrm{H}}(\tilde{f},f_{Y|X}(\pi^0))$ we thus obtain

$$
\mathcal{D}_{\mathrm{H}}^2(\pi,\pi^0)\le a_1\,\mathcal{D}_{\mathrm{H}}^2(f_{Y|X}(\pi),f_{Y|X}(\pi^0))+a_2\,\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi),f_{Y|X}(\pi^0))\mathcal{D}_{\mathrm{H}}(\tilde{f},f_{Y|X}(\pi^0)), \tag{D.12}
$$

for suitable constants a_1 and a_2. Next, we use $\tilde{f}(y|x) = f_{Y|X}(y|x;\theta^0,\tilde{\pi})$ to obtain

$$\begin{aligned}
\mathcal{D}_{\mathrm{H}}^2(\tilde{f}, f_{Y|X}(\pi^0)) &= 2 - \frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{Y}_{\mathcal{T}}}\sqrt{\frac{f_{Y|X}(y|x;\theta^0,\pi^0)}{\tilde{f}(y|x)}}f_{Y|X}(y|x;\theta^0,\tilde{\pi})dy \\
&= 2 - \frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{Y}_{\mathcal{T}}}\int_{\mathcal{A}}\sqrt{\frac{f_{Y|X}(y|x;\theta^0,\pi^0)}{\tilde{f}(y|x)}}f(y|x,\alpha;\theta^0)\,\tilde{\pi}(\alpha|x)d\alpha dy \\
&\leq 2 - \frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\frac{1}{\sqrt{\int_{\mathcal{Y}_{\mathcal{T}}}\frac{\tilde{f}(y|x)}{f_{Y|X}(y|x;\theta^0,\pi^0)}f(y|x,\alpha;\theta^0)dy}}\,\tilde{\pi}(\alpha|x)d\alpha \\
&= 2 - \frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\sqrt{\frac{\pi^0(\alpha|x)}{\pi(\alpha|x)}}\,\tilde{\pi}(\alpha|x)d\alpha \\
&= \mathcal{D}_{\mathrm{H}}^2(\pi,\pi^0) + \frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{A}}\left[1-\sqrt{\frac{\pi^0(\alpha|x)}{\pi(\alpha|x)}}\right]\left[\tilde{\pi}(\alpha|x)-\pi(\alpha|x)\right]d\alpha \\
&= \mathcal{D}_{\mathrm{H}}^2(\pi,\pi^0) + 2\sqrt{2}\,\mathcal{D}_{\mathrm{H}}(\pi,\pi^0)\,\mathcal{D}_{\mathrm{H}}(\tilde{\pi},\pi)\,\left[1+\sup_{\alpha,x}\sqrt{\frac{\tilde{\pi}(\alpha|x)}{\pi(\alpha|x)}}\right]
\end{aligned} \tag{D.13}$$

Again, using the triangle inequality for the Hellinger distance, we thus obtain

$$\mathcal{D}_{\mathrm{H}}^2(\tilde{f}, f_{Y|X}(\pi^0)) \leq a_3\mathcal{D}_{\mathrm{H}}^2(\pi,\pi^0) + a_4\mathcal{D}_{\mathrm{H}}(\pi,\pi^0)\mathcal{D}_{\mathrm{H}}(\tilde{\pi},\pi^0)\ , \tag{D.14}$$

for suitable constants a_3 and a_4. Combining this with the above result gives

$$\begin{aligned}
\mathcal{D}_{\mathrm{H}}^2(\pi,\pi^0) \leq a_1\,\mathcal{D}_{\mathrm{H}}^2(f_{Y|X}(\pi), f_{Y|X}(\pi^0)) + a_2a_3\,\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi), f_{Y|X}(\pi^0))\mathcal{D}_{\mathrm{H}}(\pi,\pi^0) \\
+ a_2a_4\,\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi), f_{Y|X}(\pi^0))\sqrt{\mathcal{D}_{\mathrm{H}}(\pi,\pi^0)\mathcal{D}_{\mathrm{H}}(\tilde{\pi},\pi^0)}.
\end{aligned} \tag{D.15}$$

From this, we can conclude

$$\mathcal{D}_{\mathrm{H}}(\pi,\pi^0) \leq c_1[\mathcal{D}_{\mathrm{H}}(f_{Y|X}(\pi), f_{Y|X}(\pi^0))]^{2/3}\ , \tag{D.16}$$

for a suitable constant c_1. By iterating the above proof for $\pi \in \Pi_T^{(1)}$ one obtains the result for general $\Pi_T^{(q)}$.

References

Andrews, D. (1994). Empirical process methods in econometrics. Handbook of Econometrics, Volume IV, eds. RF Engle and DL McFadden.

Arellano, M. and Bonhomme, S. (2009). Robust priors in nonlinear panel data models. *Econometrica*, 77(2):489–536.

Arellano, M. and Hahn, J. (2007). Understanding bias in nonlinear panel models: Some recent developments. *Econometric Society Monographs*, 43:381.

Arellano, M. and Honoré, B. (2001). Panel data models: some recent developments. *Handbook of econometrics*, 5:3229–3296.

Bester, A. and Hansen, C. (2007). Flexible Correlated Random Effects Estimation in Panel Models with Unobserved Heterogeneity. Technical report, mimeo.

Bester, C. and Hansen, C. (2009). A penalty function approach to bias reduction in nonlinear panel models with fixed effects. *Journal of Business and Economic Statistics*, 27(2):131–148.

Bonhomme, S. (2010). Functional Differencing. Technical report, Mimeo.

Carro, J. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics*, 140(2):503–528.

Chamberlain, G. (1984). Panel Data. *Griliches and M. Intrilligator, eds., Handbook of Econometrics, Chapter 22*, pages 1247–1318.

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. *Handbook of Econometrics*, 6:5549–5632.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2009a). Identification and Estimation of Marginal Effects in Nonlinear Panel Models.

Chernozhukov, V., Fernandez-Val, I., and Newey, W. (2009b). Quantile and average effects in nonseparable panel models. *CeMMAP working papers*.

Chernozhukov, V., Hahn, J., and Newey, W. (2005). Bound analysis in panel models with correlated random effects. Technical report, Technical report, MIT, UCLA and MIT.

Dhaene, G. and Jochmans, K. (2010). Split-panel jackknife estimation of fixed-effect models.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150:71–85.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica*, 70(4):1639–1657.

Hahn, J. and Kuersteiner, G. (2004). Bias reduction for dynamic nonlinear panel models with fixed effects. *Unpublished manuscript*.

Hahn, J. and Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319.

Härdle, W. and Linton, O. (1994). Applied nonparametric methods. *Handbook of Econometrics*, 4:2295–2339.

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

Ichimura, H. and Todd, P. (2007). Implementing nonparametric and semiparametric estimators. *Handbook of Econometrics*, 6:5369–5468.

Imbens, G. and Newey, W. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413.

Lancaster, T. (2002). Orthogonal parameters and panel data. *The Review of Economic Studies*, 69(3):647–666.

Martin, R. and Tokdar, S. (2010). Semiparametric inference in mixture models with predictive recursion. Manuscript.

Newton, M. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, 64(2):306–322.

Newton, M., Quintana, F., and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. *Practical nonparametric and semiparametric Bayesian statistics*, 133:45–61.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Phillips, P. C. B. and Moon, H. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica*, 67(5):1057–1111.

Woutersen, T. (2002). Robustness against incidental parameters. *Unpublished manuscript*.