Essays in Panel Data Econometrics

Dissertation
zur Erlangung des akademischen Grades
Doctor Rerum Politicarum (Dr. rer. pol.)
im Fach Volkswirtschaftslehre
durch die Wirtschaftswissenschaftliche Fakultät
der Heinrich-Heine-Universität Düsseldorf

von:	Daniel Czarnowske geboren am 26.05.1986 in Salzgitter
Erstgutachter:	Prof. Dr. Florian Heiß
Zweitgutachter:	Prof. Dr. Joel Stiebale
Abgabedatum:	07.06.2021
Disputation:	05.10.2021

Acknowledgment

This thesis was written during my time as research assistant at the Chair of Statistics and Econometrics at the Heinrich-Heine-University Düsseldorf.

First of all, I want to thank my supervisor Florian Heiß. Your lectures sparked my interest in econometrics and programming. You encouraged me to pursue a PhD and always gave me the freedom to do research on topics that interested me. I want to thank my second supervisor Joel Stiebale for learning from your lecture on empirical industrial organization and your valuable feedback on my research projects. A lot of thanks to my colleagues. You always make work enjoyable. I loved our lively discussions during the breaks.

A big thank you also goes to my co-authors Philipp Berger and Amrei Stammann. It has been a pleasure to work and discuss with you, and to learn from each other. I am looking forward to more projects with you in the future. Many thanks to André Romahn, Thomas Zylkin, and Thomas Bauer for providing me with valuable feedback on one of my research projects.

I would also like to thank my family for supporting me and allowing me to study. Finally, I would like to give a special thank you to Amrei for her endless love and support. At times when I was stuck, you always put a smile on my face and made me feel better.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A major advantage of panel data is that they allow researchers to control for *unobserved heterogeneity* in their empirical analyses. Roughly speaking, researchers often choose between two models: *unobserved effects models*, which assume that the idiosyncratic error term can be decomposed into unobserved heterogeneity and a residual idiosyncratic error term, and *varying coefficients models*, which additionally allow slope parameter heterogeneity. The availability of comprehensive and especially long panel data, i. e. *large T-panels*, offers new opportunities to draw inference from both types of models but also poses new challenges.

In my thesis, I analyze existing inference methods and develop new inference methods for large T-panels. In Chapter 2 and 4, I examine how unbalancedness affects the asymptotic properties of two estimators for unobserved effects models. More precisely, I analyze the bias-corrected estimators of Fernández-Val and Weidner (2016), for fixed effects binary choice models, and the interactive fixed effects estimator of Bai (2009). The asymptotic properties for both estimators were derived for balanced panels. In Chapter 3, I extend the generic inference method of Chernozhukov, Fernández-Val, and Weidner (2020) for distribution regression models with unobserved effects. More specifically, I broaden its applicability to panel data applications with weakly exogenous regressors. In Chapter 5, I propose a novel estimation procedure based on the classifier-Lasso of Su, Shi, and Phillips (2016) to identify latent firm heterogeneity, i. e. slope parameter heterogeneity, in production functions. In the following, I provide a more detailed overview of each chapter.

Chapter 2 (joint work with Amrei Stammann) examines the finite sample properties of various bias-corrected estimators for fixed effects binary choice models with two unobserved effects proposed by Fernández-Val and Weidner (2016) in unbalanced panels. The consequences of unbalancedness have received little attention in the literature, although unbalanced panels are frequently used in practice. In simulation experiments, we find that unbalancedness is not harmless and that the wrong choice of an estimator can lead to severe biases. In particular, we find that split-panel jackknife estimators can be severely biased, while analytical bias corrections work well. We also present algorithms to accelerate the computation of bias-corrected estimators and provide an empirical illustration from labor economics using the *GSOEP*.

Chapter 3 (joint work with Philipp Berger and Amrei Stammann) presents a generic inference method for quantile functions and quantile effects in panel data applications with two unobserved effects and weakly exogenous regressors. More precisely, our inference method is an extension of Chernozhukov, Fernández-Val, and Weidner (2020), who assume strictly exogenous regressors. However, weak exogeneity is a more realistic assumption in panel data applications because it allows direct and indirect dynamic feedback between the regressors and the dependent variable. We confirm the good finite sample properties of our inference method in simulation experiments and find that Chernozhukov, Fernández-Val, and Weidner (2020)'s method is severely biased in the presence of weakly exogenous regressors. We also use our inference method to reassess parts of Arora, Belenzon, and Sheer (2021). In particular, we estimate quantile effects of knowledge spillovers and the use of research on firms' inventiveness.

Chapter 4 (joint work with Amrei Stammann) analyzes the finite sample properties of Bai (2009)'s interactive fixed effects (IFE) estimator in unbalanced panels. Compared to conventional fixed effects models, here unobserved heterogeneity is modeled as a low-rank factor structure and is thus allowed to vary over time. Since the estimator requires an additional data augmentation step in unbalanced panels, it is not clear whether the asymptotic theory derived for balanced panels nevertheless provides a reasonable approximation. Further, since the asymptotic theory is derived under the assumption that the rank of the factor structure is known, which is rarely the case, determining the rank is another practical challenge. Using simulation experiments, we analyze how different patterns and fractions of randomly missing data affect the performance of the IFE estimator and different estimators for the rank of the factor structure. We find that the finite sample properties of the IFE estimator are fairly well approximated by the asymptotic theory for balanced panels, while the accuracy of the estimators for the rank of the factor structure differs notably across different patterns and fractions of randomly missing data. We also present algorithms to accelerate the computation in unbalanced panels and reassess Acemoglu et al. (2019). Qualitatively, our results are similar to those of the authors, although they are of different magnitude.

Chapter 5 presents a new estimation procedure to identify latent firm heterogeneity in production functions. More specifically, I consider production functions that are heterogeneous across groups but homogeneous within groups, and where the group membership of firms is unknown. My estimation procedure embeds recent identification strategies from the production function literature into the classifier-Lasso (C-Lasso) of Su, Shi, and Phillips (2016). Simulation experiments demonstrate that firms are assigned to their correct latent group with probability approaching one. The asymptotic properties of my estimator are identical to an infeasible estimator that exploits the group membership of the firms. I also apply my estimation procedure to a panel of Chilean firms and find that firms from five industry sectors are classified into three latent groups.

References

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A. Robinson. 2019. “Democracy Does Cause Growth.” *Journal of Political Economy* 127 (1): 47–100.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021. “Knowledge Spillovers and Corporate Investment in Scientific Research.” *American Economic Review* 111 (3): 871–898.

Bai, Jushan. 2009. “Panel Data Models with Interactive Fixed Effects.” *Econometrica* 77 (4): 1229–1279.

Chernozhukov, Victor, Iván Fernández-Val, and Martin Weidner. 2020. “Network and panel quantile effects via distribution regression.” *Journal of Econometrics.*

Fernández-Val, Iván, and Martin Weidner. 2016. “Individual and time effects in nonlinear panel models with large N, T.” *Journal of Econometrics* 192 (1): 291–312.

Su, Liangjun, Zhentao Shi, and Peter C. B. Phillips. 2016. “Identifying Latent Structures in Panel Data.” *Econometrica* 84 (6): 2215–2264.

Chapter 2

Fixed Effects Binary Choice Models: Estimation and Inference with Long Panels

(joint work with Amrei Stammann)

2.1 Introduction

Empirical analyses explaining binary outcomes, such as labor force participation or exporting decisions, are quite common in economics. The increasing number and availability of large and long panel data sets offer several advantages to researchers compared to pure cross-sections or time series (see chapter 1.2 in Baltagi 2013 and Hsiao 2014 for a comprehensive list of advantages). Maybe the most important advantage is that they allow to control for different sources of unobserved heterogeneity. In panels it is natural to account for unobserved individual and time specific effects simultaneously, so-called two-way fixed effect models. The corresponding estimators treat the unobserved effects as additional parameters to be estimated and thus allow for unrestricted correlation patterns between the explanatory variables and the unobserved effects. As the researcher does not have to make any distributional assumptions about the unobserved heterogeneity, these models are very flexible and a natural candidate for many empirical applications.

In the early stage of panel data econometrics, panels consisted of relatively few observations per individual. Consequently, when deriving asymptotic properties of estimators, it is very often assumed that the number of individuals (N) grows and the number of time periods (T) is held fixed. Under this asymptotic framework, nonlinear fixed effects estimators are inconsistent, known as the incidental parameter problem (*IPP*) first mentioned by Neyman and Scott (1948). This strand of literature is therefore particularly interested in deriving fixed T consistent estimators. For instance, conditional logit estimators have been proposed for static and dynamic binary choice models with individual fixed effects (see Rasch 1960, Andersen 1970, Chamberlain 1980, and Honoré and Kyriazidou 2000). However, it is not possible to derive fixed T consistent fixed effects estimators for all kind of models, e. g. the probit model. Another drawback of all conditional logit estimators is that they preclude the estimation of partial effects, which are of great interest in economics (see Arellano and Hahn 2007, and Fernández-Val and Weidner 2018a).

For these reasons, among others, and further motivated by the seminal work of Phillips and Moon (1999) and the rising availability of comprehensive panel data, a growing literature now focuses on large N and T asymptotics. An appealing feature of this asymptotic framework is that *IPP* can be transformed into an asymptotic bias problem that can be corrected. Hahn and Kuersteiner (2002) were the first to exploit this asymptotic framework and to propose a bias-corrected estimator for dynamic linear panel models with individual fixed effects to address the inference problem induced by the Nickell (1981) bias. Hahn and Moon (2006) show that the same bias correction is also applicable if the model includes additional time fixed effects. In the meantime, several bias-corrected estimators for nonlinear models have been proposed (see among others Lancaster 2002, Woutersen 2002, Hahn and Newey 2004, Carro 2007, Fernández-Val 2009, Bester and Hansen 2009, Dhaene and Jochmans 2015, Fernández-Val and Weidner 2016, and Kim and Sun 2016). A remarkable difference compared to estimators for linear models is that the inclusion of time fixed effects leads to an additional bias as shown by Fernández-Val and Weidner (2016).[1]

Another apparent challenge that discourages researchers from using nonlinear fixed effects models is the computational burden associated with the estimation. This problem is especially severe when the model specification leads to high-dimensional fixed effects. If only one of the panel dimensions is large, the algorithm of Greene (2004) can significantly reduce the computational burden. If both dimensions are large algorithms like Guimarães and Portugal (2010) and Stammann (2018) can be used. From a practical point of view, however, it is not obvious how these algorithms can be combined with analytical bias corrections like the one

[1] We refer the interested reader to Arellano and Hahn (2007) and Fernández-Val and Weidner (2018a) for comprehensive overviews.

of Fernández-Val and Weidner (2016).

In this paper, we offer new insights that facilitate and validate the use of different (bias-corrected) estimators for binary choice models with two-way fixed effects. First, we show how to address the computational obstacles that often prevent the application of bias corrections. Second, we extend the simulation experiments of Fernández-Val and Weidner (2016) by several aspects to gain deeper insights into the statistical properties of different estimators. More specifically, we analyze additional analytical and split-panel jackknife bias-corrected estimators that were proposed but not studied by Fernández-Val and Weidner (2016). We also consider alternative estimators for average partial effects based on linear fixed effects models. These models are often used in empirical research to avoid the above mentioned pitfalls of nonlinear models. Because many real world panel data sets are unbalanced, our analysis also considers different patterns of randomly missing data. This aspect has so far received little attention in the literature, but is relevant for many empirical applications. Overall, we find that analytical bias corrections are preferable to split-panel jackknife approaches. In general, the latter show higher distortion, lower coverage, and are less robust to different patterns of randomly missing data. Third, we provide an illustrative example using an unbalanced panel from the German Socio-Economic Panel (see Wagner, Frick, and Schupp 2007) to investigate the inter-temporal labor force participation of 6,241 women between 1984 and 2013. Inspired by Hyslop (1999), we estimate a dynamic fixed effects probit model where bias corrections are required to address the inference problem. Finally, we offer the analytical bias correction of Fernández-Val and Weidner (2016) in our R package *alpaca* to encourage its application.[2]

The paper is organized as follows. In Section 2.2 we introduce the model, various bias corrections, and algorithms for handling panels with large N and T. In Section 2.3 we provide results of simulation experiments. In Section 2.4 we apply different bias-corrected estimators to an empirical example from labor economics. Finally, we give some concluding remarks in Section 2.5.

Throughout this paper, we follow conventional notation: scalars are represented in standard type, vectors and matrices in boldface, and all vectors are column vectors.

2.2 Bias Corrections for Fixed Effects Binary Choice Models

2.2.1 Model, Assumptions, and the Inference Problem

The fixed effects binary choice model considered in this paper can be derived from a latent variable model with two unobserved effects. Let

$$y_{it}^{*} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma_t + e_{it} \, ,$$

be the latent variable, where i and t are individual and time specific indexes.[3] However, instead of the latent variable, we only observe $y_{it} = 1$ if $y_{it}^{*} \geq 0$ and $y_{it} = 0$ otherwise. To allow for missing data, we define the following sets: $\mathcal{S}$ is a subset of $\{(i, t) \colon i \in \{1, \ldots, N\}, t \in \{1, \ldots, T\}\}$ containing all observed pairs of indexes and $\mathcal{S}_t = \{i \colon (i, t) \in \mathcal{S}\}$ and $\mathcal{S}_i = \{t \colon (i, t) \in \mathcal{S}\}$ are subsets of $\mathcal{S}$ containing all indexes of individuals observed in period t and points in time observed for an individual i, respectively. N and T are the number of individuals and time periods, and $n = |\mathcal{S}|$ is the sample size. Furthermore, $\mathbf{x}_{it}$ is the it-th row of the regressor matrix $\mathbf{X}$ and a J-dimensional vector of possibly predetermined explanatory variables, $\boldsymbol{\beta}$

[2] Until now, the analytical bias correction proposed by Fernández-Val and Weidner (2016) is only provided in a *Stata* routine by Cruz-Gonzalez, Fernández-Val, and Weidner (2017), which is not designed for long panels.

[3] Without loss of generality, i and t could also be indexes of a network, like trade flows between countries.

are the corresponding structural parameters, α_i and γ_t are incidental parameters that capture unobserved individual and time effects, and e_{it} is an idiosyncratic error term assumed to be mean zero and independent of $\mathbf{X}_i^t := (\mathbf{x}_{is})_{s \leq t \in \mathcal{S}_i}$ and $\boldsymbol{\pi} = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$. Note that the imposed exogeneity assumption is milder than the often assumed strict exogeneity condition, where e_{it} is independent of $\mathbf{X}_i^T := (\mathbf{x}_{is})_{s \in \mathcal{S}_i}$ instead of $\mathbf{X}_i^t$. The latter is often too restrictive in the context of panel data, because it is violated if regressors are affected by past outcomes, which, for instance, also precludes lagged dependent variables as regressors. In the presence of missing data, we have to additionally assume that conditional on $\mathbf{X}_i^t$ and $\boldsymbol{\pi}$ the observations are missing at random.

Assuming a certain distribution for e_{it} allows us to use maximum likelihood to derive a parametric estimator for fixed effects binary choice models. Let

$$l_{it}(\boldsymbol{\beta}, \alpha_i, \gamma_t) := y_{it} \log(F_{it}) + (1 - y_{it}) \log(1 - F_{it})$$

be the log-likelihood contribution of individual i at time t, where F_{it} is the cumulative distribution function of e_{it} evaluated at the linear index $\eta_{it} := \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma_t$. Common choices for F_{it} are the standard normal, the logistic, and the complementary log-log distribution. The corresponding maximum likelihood estimator is

$$\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) \in \underset{\boldsymbol{\beta} \in \mathbb{R}^J, \boldsymbol{\alpha} \in \mathbb{R}^N, \boldsymbol{\gamma} \in \mathbb{R}^T}{\arg\max} \sum_{(i,t) \in \mathcal{S}} l_{it}(\boldsymbol{\beta}, \alpha_i, \gamma_t) \, .$$

Although Fernández-Val and Weidner (2016) show consistency of $\hat{\boldsymbol{\beta}}$ under asymptotics where N and T grow at the same rate ($\text{plim}_{N,T \to \infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$), they also expose an asymptotic bias in the limiting distribution of the estimator with severe consequences for inference. To get a better understanding of this specific inference problem, we briefly summarize the key findings of Fernández-Val and Weidner (2016) for binary choice models and combine them with their conjecture about unbalanced panels stated in Fernández-Val and Weidner (2018a).

Under asymptotic sequences where $N, T \longrightarrow \infty$, $N/T \longrightarrow \kappa^2$, and $0 < \kappa < \infty$, an asymptotic approximation to the limiting distribution of $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} \overset{\text{a}}{\sim} \mathcal{N}(\boldsymbol{\beta} + \overline{T}^{-1}\mathbf{B}^{\beta} + \overline{N}^{-1}\mathbf{C}^{\beta}, \ \mathbf{V}^{\beta}) \, ,$$

where $\overline{N} = n/T$ and $\overline{T} = n/N$ are the average number of individuals and points in time, $\mathbf{B}^{\beta}$ and $\mathbf{C}^{\beta}$ are the leading terms of the asymptotic bias $\mathbf{b}^{\beta} := \overline{T}^{-1}\mathbf{B}^{\beta} + \overline{N}^{-1}\mathbf{C}^{\beta}$ stemming from the inclusion of individual and time fixed effects, and $\mathbf{V}^{\beta}$ is the asymptotic covariance matrix. Due to the asymptotic bias, the approximated limiting distribution is not correctly centered at $\boldsymbol{\beta}$, which means that, even if n is large, confidence intervals constructed around any $\hat{\boldsymbol{\beta}}$ might not cover the true value of the corresponding parameter with probability close to the desired nominal level. However, this inference problem can be corrected by forming suitable estimators for $\mathbf{b}^{\beta}$ that can be subtracted from $\hat{\boldsymbol{\beta}}$.

Most researchers are not directly interested in the structural parameters, but rather in average partial effects. Let

$$\boldsymbol{\Delta}_{itj} := \begin{cases} \beta_j \partial_\eta F_{it} & \text{(continous regressor)} \\ F_{it}|_{x_{itj}=1} - F_{it}|_{x_{itj}=0} & \text{(binary regressor)} \end{cases}$$

denote the partial effect of a change in x_{itj}, where x_{itj} is the j-th element in $\mathbf{x}_{it}$, $\partial_\eta F_{it}$ is the first-order partial derivative of F_{it} with respect to η_{it}, and $F_{it}|_{x_{itj}=k}$ indicates that x_{itj} in the linear index is replaced by k.[4] The average partial effects are then given by $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)$, where $\delta_j = n^{-1} \sum_{(i,t) \in \mathcal{S}} \Delta_{itj}$. An asymptotic approximation to the limiting distribution of the estimator of the average partial effects is given by

$$\hat{\boldsymbol{\delta}} \overset{\text{a}}{\sim} \mathcal{N}(\boldsymbol{\delta} + \overline{T}^{-1}\mathbf{B}^{\delta} + \overline{N}^{-1}\mathbf{C}^{\delta},\ \mathbf{V}^{\delta}) \, ,$$

where $\mathbf{V}^{\delta}$ is the asymptotic covariance matrix. Again, $\mathbf{B}^{\delta}$ and $\mathbf{C}^{\delta}$ are the leading terms of the asymptotic bias $\mathbf{b}^{\delta} := \overline{T}^{-1}\mathbf{B}^{\delta} + \overline{N}^{-1}\mathbf{C}^{\delta}$ stemming from the inclusion of individual and time fixed effects. Thus, as for $\hat{\boldsymbol{\beta}}$, there is an asymptotic bias problem.

In the next subsection, we review the different bias corrections proposed by Fernández-Val and Weidner (2016) with a slightly modified notation. We adapt their notation for two reasons: First, we consider panels that are potentially unbalanced, and second, we want to emphasize the link to recent advances in computational econometrics, which allow us to estimate (bias-corrected) binary choice models even when both panel dimensions are large.

2.2.2 Asymptotic Bias Corrections

Before we present the different bias corrections proposed by Fernández-Val and Weidner (2016), we have to introduce some additional notation. Let $\partial_{z^p} G_{it}$ denote the p-th order partial derivative of an arbitrary function G_{it} with respect to z_{it}. Further, we define $\partial_{\eta} l_{it} := H_{it}(y_{it} - F_{it})$, $\omega_{it} := H_{it} \partial_{\eta} F_{it}$, $H_{it} := \partial_{\eta} F_{it} / (F_{it}(1 - F_{it}))$, and $\nu_{it} := \partial_{\eta} l_{it} / \omega_{it}$. We use vector notation to indicate that we collect the different quantities for all observations, e. g. $\boldsymbol{\omega} := (\omega_{it})_{(i,t) \in \mathcal{S}}$. Finally we define the residual projection $\mathbb{M} := \mathbf{1}_n - \mathbb{P} := \mathbf{1}_n - \mathbf{D}(\mathbf{D}'\boldsymbol{\Omega}\mathbf{D})^{+}\mathbf{D}'\boldsymbol{\Omega}$, where $\mathbf{1}_n$ is an identity matrix of dimension $(n \times n)$, $\mathbf{D}$ is a sparse indicator matrix of dimension $(n \times N + T)$ arising from dummy encoding of individual and time identifiers, $(\cdot)^{+}$ refers to the Moore-Penrose inverse, and $\boldsymbol{\Omega}$ is a positive definite diagonal weighting matrix with $\text{diag}(\boldsymbol{\Omega}) := \boldsymbol{\omega}$. The corresponding sample analogues are indicated by a hat, i. e. we refer to $\hat{\eta}_{it} := \mathbf{x}_{it}'\hat{\boldsymbol{\beta}} + \hat{\alpha}_i + \hat{\gamma}_t$ as the sample analogue of η_{it}. Table 2.1 contains explicit expressions for distributions and the corresponding derivatives of frequently used binary choice models.

Table 2.1: *Distributions and their Derivatives*

	Logit	Probit	Complementary Log-Log
F_{it}	$(1 + \exp(-\eta_{it}))^{-1}$	$\Phi(\eta_{it})$	$1 - \exp(-\exp(\eta_{it}))$
$\partial_{\eta} F_{it}$	$F_{it}(1 - F_{it})$	$\phi(\eta_{it})$	$\exp(\eta_{it} - \exp(\eta_{it}))$
$\partial_{\eta^2} F_{it}$	$\partial_{\eta} F_{it}(1 - 2F_{it})$	$-\eta_{it}\phi(\eta_{it})$	$\partial_{\eta} F_{it}(1 - \exp(\eta_{it}))$
$\partial_{\eta^3} F_{it}$	$\partial_{\eta} F_{it}((1 - 2F_{it})^2 - 2\partial_{\eta} F_{it})$	$(\eta_{it}^2 - 1)\phi(\eta_{it})$	$\partial_{\eta^2} F_{it}(2 - \exp(\eta_{it})) - \partial_{\eta} F_{it}$

Note: $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution and probability density function of the standard normal distribution.

Fernández-Val and Weidner (2016) distinguish between two types of bias corrections: analytical and split-panel jackknife. The latter exploits the relation between sample size and bias to form a non-parametric estimator of the asymptotic bias and is an extension of Dhaene and Jochmans (2015), whereas the former

[4] For simplicity, we ignore the possibility of more complex functional forms in the linear index, e. g. polynomials.

relies on explicit expressions derived from asymptotic expansions.[5] A bias-corrected estimator for $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}} - \hat{\mathbf{b}}^{\beta} \,,$$

where $\hat{\mathbf{b}}^{\beta}$ is an estimator of the asymptotic bias such that $\tilde{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \mathbf{V}^{\beta})$.

We start with an analytical bias correction at the level of the estimator. An explicit expression for an estimator of the asymptotic bias is

$$\hat{\mathbf{b}}^{\beta}_{\text{abc}} := \widehat{\mathbf{W}}^{-1}(\widehat{\mathbf{B}}^{\beta} + \widehat{\mathbf{C}}^{\beta}) \,,$$

where

$$\begin{aligned}
\widehat{\mathbf{B}}^{\beta} &:= -\frac{1}{2}\sum_{i=1}^{N} \frac{\sum_{t\in\mathcal{S}_i} \widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}(\widehat{\mathbb{M}}\mathbf{X})_{it} + 2\sum_{l=1}^{L}\tau_i(l)\sum_{t>l\in\mathcal{S}_i}\partial_{\eta}\hat{l}_{it-l}\hat{\omega}_{it}(\widehat{\mathbb{M}}\mathbf{X})_{it}}{\sum_{t\in\mathcal{S}_i}\hat{\omega}_{it}} \,, \\
\widehat{\mathbf{C}}^{\beta} &:= -\frac{1}{2}\sum_{t=1}^{T} \frac{\sum_{i\in\mathcal{S}_t} \widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}(\widehat{\mathbb{M}}\mathbf{X})_{it}}{\sum_{i\in\mathcal{S}_t}\hat{\omega}_{it}} \,, \\
\widehat{\mathbf{W}} &:= \sum_{(i,t)\in\mathcal{S}} \hat{\omega}_{it}(\widehat{\mathbb{M}}\mathbf{X})_{it}(\widehat{\mathbb{M}}\mathbf{X})'_{it} \,,
\end{aligned}$$

L is the bandwidth parameter of the truncated spectral density estimator suggested by Hahn and Kuersteiner (2007), and $\tau_i(l) := |\mathcal{S}_i|/(|\mathcal{S}_i| - l)$ is a finite sample adjustment proposed by Fernández-Val and Weidner (2016). The corresponding estimator of the asymptotic covariance is $\widehat{\mathbf{V}}^{\beta} := \widehat{\mathbf{W}}^{-1}$. By making the stronger strict exogeneity assumption, we can set $L = 0$ and drop the second term in $\widehat{\mathbf{B}}^{\beta}$, so that the expressions for $\widehat{\mathbf{B}}^{\beta}$ and $\widehat{\mathbf{C}}^{\beta}$ become identical except for the indexes.[6] However, this stronger assumption is difficult to motivate in practice. Therefore, Fernández-Val and Weidner (2016, 2018a) recommend to check the sensitivity of the estimates using different values of $L \in \{0, \ldots, 4\}$.

To possibly further improve its finite sample properties, the analytical bias correction can be further iterated. More precisely, we start with an initial $\tilde{\boldsymbol{\beta}}$, afterwards we recompute $\hat{\mathbf{b}}^{\text{abc}}$, update $\tilde{\boldsymbol{\beta}}$, and repeat this procedure a finite number of times. Arellano and Hahn (2007) refer to this approach as infinitely repeated analytical bias correction.

The idea of the split-panel jackknife is to split the panel into sub panels and use them to construct an estimator of the asymptotic bias from different sub panel estimators. We consider two different splitting strategies: the first strategy (*SPJ1*) is described in Fernández-Val and Weidner (2016) and the second one (*SPJ2*) in Cruz-Gonzalez, Fernández-Val, and Weidner (2017). Let

$$\hat{\mathbf{b}}^{\beta}_{\text{spj1}} := \hat{\boldsymbol{\beta}}^{N} + \hat{\boldsymbol{\beta}}^{T} - 2\hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\mathbf{b}}^{\beta}_{\text{spj2}} := \hat{\boldsymbol{\beta}}^{NT} - \hat{\boldsymbol{\beta}}$$

[5] The idea to reduce bias using jackknife techniques originates from Quenouille (1949, 1956). In the context of dynamic models they were first mentioned by Hu (2002).

[6] The second term in $\widehat{\mathbf{B}}^{\beta}$ can be interpreted as an estimator for a Nickell (1981)-type bias.

be estimators of the asymptotic bias, where

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{N} &:= \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_{\{i \leq \lceil N/2 \rceil\}} + \hat{\boldsymbol{\beta}}_{\{i \geq \lfloor N/2+1 \rfloor\}}\right) , \\
\hat{\boldsymbol{\beta}}^{T} &:= \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_{\{t \leq \lceil T/2 \rceil\}} + \hat{\boldsymbol{\beta}}_{\{t \geq \lfloor T/2+1 \rfloor\}}\right) , \\
\hat{\boldsymbol{\beta}}^{NT} &:= \frac{1}{4}\left(\hat{\boldsymbol{\beta}}_{\{i \leq \lceil N/2 \rceil \wedge t \leq \lceil T/2 \rceil\}} + \hat{\boldsymbol{\beta}}_{\{i \leq \lceil N/2 \rceil \wedge t \geq \lfloor T/2+1 \rfloor\}} + \right. \\
&\qquad \left. \hat{\boldsymbol{\beta}}_{\{i \geq \lfloor N/2+1 \rfloor \wedge t \leq \lceil T/2 \rceil\}} + \hat{\boldsymbol{\beta}}_{\{i \geq \lfloor N/2+1 \rfloor \wedge t \geq \lfloor T/2+1 \rfloor\}}\right) ,
\end{aligned}$$

$\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are floor and ceiling functions, and the subscript in curly brackets indicates the condition to construct the corresponding sub panel. For instance, $\{i \leq \lceil N/2 \rceil \wedge t \leq \lceil T/2 \rceil\}$ means that the corresponding sub panel only contains the first half of all individuals in the first half of the observation period. In the presence of missing data, we follow the suggestion of Fernández-Val and Weidner (2018a) and ignore the attrition process. Note that, contrary to analytical bias corrections, the split-panel jackknife requires an additional unconditional homogeneity assumption (see assumption 4.3 in Fernández-Val and Weidner 2016). For instance, this condition rules out trends or structural breaks in the explanatory variables. Note further that both splitting strategies can lead to overlapping sub panels that introduce an additional variance inflation as pointed out by Dhaene and Jochmans (2015).[7]

The analytical bias corrections can also be applied at the level of score. The corresponding bias-corrected estimator is the solution to the following system of equations:

$$(\widehat{\mathbb{M}}\mathbf{X}(\boldsymbol{\beta}))'\widehat{\boldsymbol{\Omega}}(\boldsymbol{\beta})\hat{\nu}(\boldsymbol{\beta}) = \widehat{\mathbf{B}}^{\beta} + \widehat{\mathbf{C}}^{\beta} .$$

Fernández-Val and Weidner (2016) also suggest a continuously updated score correction where $\widehat{\mathbf{B}}^{\beta}$ and $\widehat{\mathbf{C}}^{\beta}$ are replaced by $\widehat{\mathbf{B}}^{\beta}(\boldsymbol{\beta})$ and $\widehat{\mathbf{C}}^{\beta}(\boldsymbol{\beta})$ such that $\boldsymbol{\beta}$ and the asymptotic bias are estimated simultaneously.

Finally, we describe the bias corrections for the average partial effects. A bias-corrected estimator for $\boldsymbol{\delta}$ is

$$\tilde{\boldsymbol{\delta}} := \hat{\boldsymbol{\delta}} - \hat{\mathbf{b}}^{\delta} ,$$

where $\hat{\mathbf{b}}^{\delta}$ is an estimator of the asymptotic bias such that $\tilde{\boldsymbol{\delta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\delta}, \mathbf{V}^{\delta})$. Again, we can either use explicit expressions or the split-panel jackknife described earlier to construct an estimator of the asymptotic bias. Because the application of the split-panel jackknife is generic and already known from the estimation of $\boldsymbol{\beta}$, we omit it for brevity

In the following we assume that $\hat{\boldsymbol{\Delta}}_{it}$ and $\hat{\boldsymbol{\delta}}$ are constructed from $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\pi}}$, where

$$\tilde{\boldsymbol{\pi}} := (\tilde{\alpha}, \tilde{\gamma}) \in \underset{\alpha \in \mathbb{R}^{N}, \gamma \in \mathbb{R}^{T}}{\arg\max} \sum_{(i,t) \in \mathcal{S}} l_{it}(\tilde{\boldsymbol{\beta}}, \alpha_i, \gamma_t) .$$

An analytical estimator of the asymptotic bias is

$$\hat{\mathbf{b}}_{\text{abc}}^{\delta} := n^{-1}(\widehat{\mathbf{B}}^{\delta} + \widehat{\mathbf{C}}^{\delta}) ,$$

[7] Dhaene and Jochmans (2015) show how to construct non-overlapping sub panels.

where

$$\begin{aligned}\widehat{\mathbf{B}}^{\delta} &:= \frac{1}{2}\sum_{i=1}^{N} \frac{\sum_{t\in\mathcal{S}_i} \widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}(\widehat{\mathbb{P}}\widehat{\mathbf{\Psi}})_{it} + \partial_{\eta^2}\hat{\mathbf{\Delta}}_{it} - 2\sum_{l=1}^{L}\tau_i(l)\sum_{t>l\in\mathcal{S}_i}\partial_{\eta}\hat{l}_{it-l}\hat{\omega}_{it}(\widehat{\mathbb{M}}\widehat{\mathbf{\Psi}})_{it}}{\sum_{t\in\mathcal{S}_i}\hat{\omega}_{it}}, \\ \widehat{\mathbf{C}}^{\delta} &:= \frac{1}{2}\sum_{t=1}^{T} \frac{\sum_{i\in\mathcal{S}_t} \widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}(\widehat{\mathbb{P}}\widehat{\mathbf{\Psi}})_{it} + \partial_{\eta^2}\hat{\mathbf{\Delta}}_{it}}{\sum_{i\in\mathcal{S}_t}\hat{\omega}_{it}},\end{aligned}$$

and $\widehat{\mathbf{\Psi}}_{it} := -\partial_{\eta}\hat{\mathbf{\Delta}}_{it}/\hat{\omega}_{it}$. Again, by making the stronger strict exogeneity assumption, we can set $L = 0$ and drop the last term in $\widehat{\mathbf{B}}^{\delta}$. Let $\bar{\hat{\mathbf{\Delta}}}_{it} := \hat{\mathbf{\Delta}}_{it} - \hat{\boldsymbol{\delta}}$, the corresponding estimator of the asymptotic covariance is

$$\widehat{\mathbf{V}}^{\delta} = n^{-2}\Big(\Big(\sum_{(i,t)\in\mathcal{S}}\bar{\hat{\mathbf{\Delta}}}_{it}\Big)\Big(\sum_{(i,t)\in\mathcal{S}}\bar{\hat{\mathbf{\Delta}}}_{it}\Big)' + \sum_{(i,t)\in\mathcal{S}}\widehat{\mathbf{\Gamma}}_{it}\widehat{\mathbf{\Gamma}}_{it}' + 2\sum_{i=1}^{N}\sum_{s>t\in\mathcal{S}_i}\bar{\hat{\mathbf{\Delta}}}_{it}\widehat{\mathbf{\Gamma}}_{is}'\Big),$$

where

$$\widehat{\mathbf{\Gamma}}_{it} := \Big(\sum_{(i,t)\in\mathcal{S}}\partial_{\beta}\hat{\mathbf{\Delta}}_{it} - (\widehat{\mathbb{P}}\mathbf{X})_{it}\partial_{\eta}\hat{\mathbf{\Delta}}_{it}\Big)'\widehat{\mathbf{W}}^{-1}(\widehat{\mathbb{M}}\mathbf{X})_{it}\partial_{\eta}\hat{l}_{it} - (\widehat{\mathbb{P}}\widehat{\mathbf{\Psi}})_{it}\partial_{\eta}\hat{l}_{it}\,.$$

The first and second term measure the uncertainty caused by the substitution of population by sample means and by the estimation of the structural parameters. The last term is a covariance between both sources of uncertainty that can be dropped if we make the stronger strict exogeneity assumption.[8]

So far we learned how to mitigate the inference problem. In the next subsection we show how the computational costs associated with the estimation and the application of bias corrections can be reduced substantially.

2.2.3 Feasible Estimation with Long Panel Data

In applications where both panel dimensions are large, estimation with standard software quickly becomes very time-consuming or even infeasible. However, recent advances in computational econometrics embed special solvers in the optimization algorithm of the maximum likelihood estimator to address this problem (see Guimarães and Portugal 2010 and Stammann 2018). We first summarize the basic idea of Stammann (2018), who proposed an algorithm that can be interpreted as a generalization of Greene (2004) to more than one fixed effect, and then present an extension that can be used to estimate $\boldsymbol{\pi}$ for a given $\tilde{\boldsymbol{\beta}}$.[9] This extension is necessary for the bias corrections of the average partial effects. Details about the derivation, the algorithms, and an example code are included in the Appendix 2.A.

We start with the estimation of $\boldsymbol{\beta}$. Each step of the optimization algorithm involves solving a weighted least squares problem. More precisely, in iteration r we set

$$\boldsymbol{\theta}^{\langle r+1\rangle} = (\boldsymbol{\beta}^{\langle r+1\rangle}, \boldsymbol{\pi}^{\langle r+1\rangle}) = (\mathbf{Z}'\mathbf{\Omega}^{\langle r\rangle}\mathbf{Z})^{+}\mathbf{Z}'\mathbf{\Omega}^{\langle r\rangle}\mathbf{w}^{\langle r\rangle},$$

[8] The estimator of the asymptotic covariance can be adjusted to take into account additional sampling assumptions with respect to the unobserved effects. For instance, if $\{\alpha_i\}_N$ and $\{\gamma_t\}_T$ are assumed to be sequences of independent random variables, where $\alpha_i \perp \gamma_t \,\forall\, i,t$, the estimator of the asymptotic covariance becomes

$$\widehat{\mathbf{V}}^{\delta} = n^{-2}\sum_{i=1}^{N}\Big(\sum_{(t,s)\in\mathcal{S}_i}\bar{\hat{\mathbf{\Delta}}}_{it}\bar{\hat{\mathbf{\Delta}}}_{is}' + \sum_{(i',t)\in\mathcal{S}}\bar{\hat{\mathbf{\Delta}}}_{it}\bar{\hat{\mathbf{\Delta}}}_{i't}' + \sum_{t\in\mathcal{S}_i}\widehat{\mathbf{\Gamma}}_{it}\widehat{\mathbf{\Gamma}}_{it}' + 2\sum_{s>t\in\mathcal{S}_i}\bar{\hat{\mathbf{\Delta}}}_{it}\widehat{\mathbf{\Gamma}}_{is}'\Big).$$

[9] To be more specific, Stammann, Heiss, and McFadden (2016) show that the algorithm of Greene (2004) can also be derived using the Frisch-Waugh-Lovell theorem. Stammann (2018) combines the resulting projections with Halperin (1962)'s method of alternating projections, leading to a very efficient algorithm for any number of fixed effects.

where $\mathbf{Z} := (\mathbf{X}, \mathbf{D})$ and $\mathbf{w}^{\langle r \rangle} := \boldsymbol{\nu}^{\langle r \rangle} + \boldsymbol{\eta}^{\langle r \rangle}$. Remember, $\boldsymbol{\Omega}^{\langle r \rangle}$ is a diagonal matrix with $\boldsymbol{\omega}^{\langle r \rangle} := \operatorname{diag}(\boldsymbol{\Omega}^{\langle r \rangle})$ and $\boldsymbol{\eta}^{\langle r \rangle}$ is the collection of linear indexes. Because the rank of $\mathbf{D}$ increases with the sample size, solving the optimization problem quickly becomes infeasible. However we can formulate an alternative weighted least squares problem based on "demeaned" variables so that

$$\boldsymbol{\beta}^{\langle r+1 \rangle} = ((\mathbb{M}\mathbf{X}^{\langle r \rangle})' \boldsymbol{\Omega}^{\langle r \rangle} \mathbb{M}\mathbf{X}^{\langle r \rangle})^{-1} (\mathbb{M}\mathbf{X}^{\langle r \rangle})' \boldsymbol{\Omega}^{\langle r \rangle} \mathbb{M}\mathbf{w}^{\langle r \rangle} .$$

Afterwards we can use

$$\boldsymbol{\eta}^{\langle r+1 \rangle} = \mathbf{w}^{\langle r \rangle} - (\mathbb{M}\mathbf{w}^{\langle r \rangle} - \mathbb{M}\mathbf{X}^{\langle r \rangle} \boldsymbol{\beta}^{\langle r+1 \rangle})$$

to update $\boldsymbol{\omega}$ and $\mathbf{w}$ for the subsequent iteration. Consequently, $\boldsymbol{\beta}$ are the coefficients obtained by a regression of a weighted two-way demeaned $\mathbf{w}$ on a weighted two-way demeaned $\mathbf{X}$ using $\boldsymbol{\omega}$ as weights. $\boldsymbol{\eta}$ are the corresponding fitted values. Thus we can update $\boldsymbol{\beta}$, $\mathbf{w}$, and $\boldsymbol{\omega}$ in each iteration without explicitly updating the incidental parameters. The practical advantage is that $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ can be very efficiently computed with any software routine developed for weighted least squares problems with high-dimensional fixed effects.[10] Note that we can also use the same software routines to compute all terms that depend on $\mathbb{M}$ and $\mathbb{P}$, as in some expressions of the bias corrections.

Next, we modify the algorithm of Stammann (2018) to estimate $\mathbf{D}\boldsymbol{\pi}$ given a fixed $\tilde{\boldsymbol{\beta}}$. Note that we estimate $\mathbf{D}\boldsymbol{\pi}$ instead of $\boldsymbol{\pi}$ as this is simpler and fully sufficient to compute the linear index needed for F_{it} and its derivatives. Again, we start with the weighted least squares problem in iteration r:

$$\boldsymbol{\pi}^{\langle r+1 \rangle} = (\mathbf{D}' \boldsymbol{\Omega}^{\langle r \rangle} \mathbf{D})^{+} \mathbf{D}' \boldsymbol{\Omega}^{\langle r \rangle} \tilde{\mathbf{w}}^{\langle r \rangle} ,$$

where $\tilde{\mathbf{w}}^{\langle r \rangle} := \mathbf{w}^{\langle r \rangle} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\eta}^{\langle r \rangle} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{D}\boldsymbol{\pi}^{[r]}$. Left multiplying by $\mathbf{D}$ yields

$$\mathbf{D}\boldsymbol{\pi}^{\langle r+1 \rangle} = \mathbb{P}\,\tilde{\mathbf{w}}^{\langle r \rangle} = \tilde{\mathbf{w}}^{\langle r \rangle} - \mathbb{M}\,\tilde{\mathbf{w}}^{\langle r \rangle} ,$$

which reveals that we only need to demean $\tilde{\mathbf{w}}^{\langle r \rangle}$ to update the linear index.

Finally, we give a short impression about the capabilities of the algorithms presented in this section. Therefore, we estimate the specification from our empirical illustration using a standard routine and the alternative algorithm we suggest. The former requires more than twelve hours whereas our suggestion only needs three seconds.

2.3 Simulation Experiments

In this section, we extend the analysis of Fernández-Val and Weidner (2016) by two aspects. First, we compare the various analytical and split-panel jackknife bias corrections discussed earlier in this paper with respect to their finite sample properties in balanced panels. Second, we analyze whether the improved inference of bias corrections, which is well studied in balanced panels, also shows up in unbalanced panels. For brevity, we restrict ourselves to the analysis of dynamic designs, because Fernández-Val and Weidner (2016) have already shown that the results with respect to exogenous regressors are very similar for static and dynamic designs in balanced panels.

[10] Some examples available in popular statistical software are *reghdfe* by Correia (2016) for *Stata*, *lfe* by Gaure (2013a) for *R*, *FixedEffectModels* by Matthieu Gomez for *Julia*, and *pyhdfe* by Jeff Gortmaker for *Python*.

Besides the asymptotically biased maximum likelihood estimator (*MLE*), we consider four different analytical bias corrections for the structural parameters. Two of them are applied at the level of the estimator, whereas the others are the solution of modified score equations. *ABC1* is the analytical bias correction analyzed in Fernández-Val and Weidner (2016). *ABC2* is essentially *ABC1*, but additionally iterated until convergence. *ABC3* and *ABC4* are estimators at the level of the score, whereas the latter recomputes the asymptotic bias in each iteration of the nonlinear solver. The analytical bias-corrected estimators of the average partial effects are labeled analogously. Further we analyze the two different splitting strategies (*SPJ1–2*) for the split-panel jackknife bias correction and an alternative estimator for the average partial effects (*LPM*) based on the bias-corrected ordinary least squares estimator proposed by Hahn and Kuersteiner (2002) and the truncated spectral density estimator of Hahn and Kuersteiner (2007).[11]

We adapt the dynamic design of Fernández-Val and Weidner (2016) to allow for the possibility of missing data:

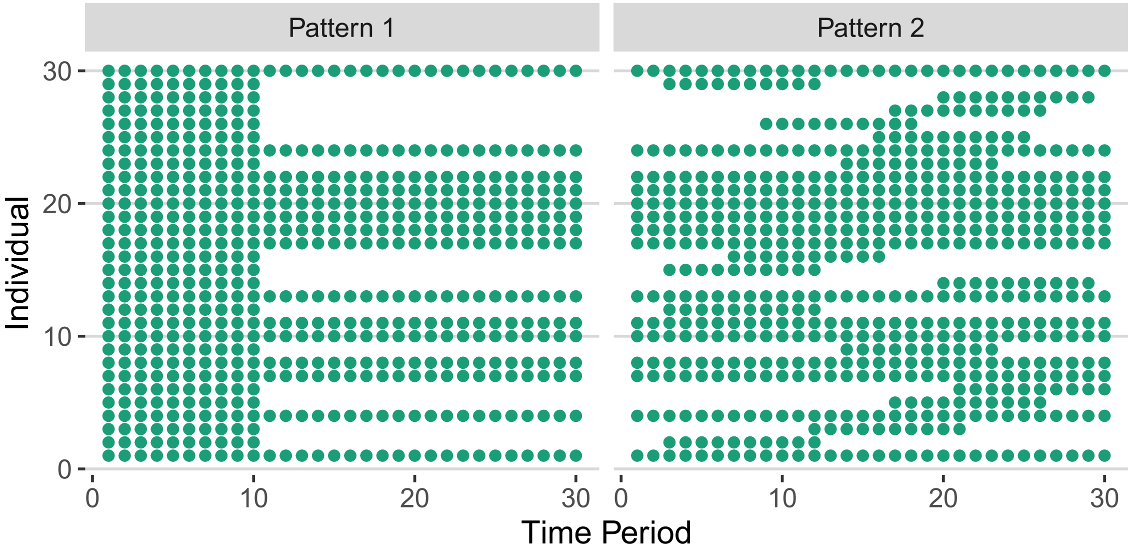
$$\begin{aligned}
y_{it} &= \mathbf{1}\{\rho y_{it-1} + \beta x_{it} + \alpha_i + \gamma_t \geq \epsilon_{it}\}\,, \\
y_{i0} &= \mathbf{1}\{\beta x_{i0} + \alpha_i + \gamma_0 \geq \epsilon_{i0}\}\,, \\
x_{it} &= 0.5 x_{it-1} + \alpha_i + \gamma_t + \nu_{it}\,, \\
(i &= 1, \ldots, N,\ t = 1 \leq s_i, \ldots, T_i \leq T)\,,
\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function, $\alpha_i, \gamma_t \sim$ iid. $\mathcal{N}(0, 1/16)$, $\epsilon_{it} \sim$ iid. $\mathcal{N}(0,1)$, $\nu_{it} \sim$ iid. $\mathcal{N}(0, 0.5)$, and $x_{i0} \sim$ iid. $\mathcal{N}(0,1)$. The corresponding parameters are $\rho = 0.5$ and $\beta = 1$. We consider balanced and unbalanced panels with sample sizes that reflect commonly used panel data sets ($N \gg T$). More specifically, we consider two different patterns of randomly missing data that mimic common situations where some people drop out of a survey and are replaced if necessary. To describe the different patterns of missing data, we distinguish between two types of individuals: *type 1* and *type 2*. The former drop out, whereas the latter are observed over the entire time horizon. To be more precise, let N_1 and N_2 be the number of *type 1* and *type 2* individuals in the unbalanced panel such that $N = N_1 + N_2$. Likewise T_1 and T_2 denote the number of consecutive points in time such that $T_1 < T_2$. Both patterns of missing data differ only in the starting point (s_i) of the time series of each *type 1* individual. We set $s_i = 1$ in *Pattern 1* and sample s_i with equal probability from $\{1, \ldots, T_2 - T_1 + 1\}$ in *Pattern 2*. For clarification, we set $s_i = 1$ for all *type 2* individuals irrespective of the pattern. Figure 2.1 provides a graphical illustration of both patterns. We generate balanced panel data sets with $N = 200$ and $T_i = T \in \{15, 20, 25\}$ and unbalanced panel data sets with $(N_1, N_2) \in \{(300, 100), (150, 150), (60, 180)\}$, $T_1 = 10$, and $T_2 = 30$. The different pairs (N_1, N_2) are chosen such that the average number of individuals ($\overline{N}$) and points in time ($\overline{T}$) are $\overline{N} = 200$ and $\overline{T} \in \{15, 20, 25\}$.

To analyze the finite sample performance of the different estimators, we focus on biases relative to the truth and empirical coverage probabilities of 95% confidence intervals. The latter statistic is especially important, because even if the relative bias is quite small, it might still be large compared to the dispersion of the estimator, with severe consequences for inference. The *MLE*, *ABC1–4*, and *SPJ1–2* standard errors of the average partial effects are computed using the expression in footnote 8, which takes the independent sampling of the unobserved effects into account. The *LPM* standard errors are based on the cluster-robust covariance estimator of Cameron, Gelbach, and Miller (2011) to deal with within-individual and within-time

[11] Hahn and Moon (2006) show that the bias correction of Hahn and Kuersteiner (2002) can be applied to dynamic linear models with individual and time specific effects.

Figure 2.1: *Patterns of Randomly Missing Observations*



correlation of the error terms induced by the probit data generating process. Because there is no obvious way to choose an optimal bandwidth for the estimation of the spectral expectations, we try different choices from a set of values and then report only the results of the choice with the best finite sample performance. We choose L from $\{1, \ldots, 4\}$ for *ABC1–4*, as suggested by Fernández-Val and Weidner (2016, 2018a), and from $\{1, \ldots, \overline{T} - 1\}$ for *LPM*. All results are based on 1,000 replications.[12]

We start by comparing the different analytical bias corrections in a balanced panel. Table 2.2 reports the relative biases and coverage probabilities of the estimators for the structural parameters. As expected, all corrections reduce a larger fraction of the bias and improve coverage as T increases. The difference between the various estimators is in most cases negligible small. Interestingly, *ABC2* does not perform better than *ABC1*, which is remarkable, because often, as for instance in Arellano and Hahn (2007) and Fernández-Val and Weidner (2018a), it is noted, that a further iteration of *ABC1* could improve the finite sample performance

Table 2.2: *Analytical Bias Corrections: Coefficients*

	$\hat{\rho}$				$\hat{\beta}$			
	ABC1	ABC2	ABC3	ABC4	ABC1	ABC2	ABC3	ABC4
				Panel A: Relative Bias				
$T = 15$	-3.886	-6.580	-4.673	-5.204	1.004	2.666	1.710	1.761
$T = 20$	-2.575	-4.227	-2.979	-3.426	0.606	1.575	0.979	1.035
$T = 25$	-1.815	-2.945	-2.078	-2.432	0.319	0.969	0.562	0.614
				Panel B: Coverage Probability				
$T = 15$	0.932	0.916	0.929	0.930	0.943	0.919	0.939	0.941
$T = 20$	0.947	0.931	0.944	0.939	0.941	0.930	0.938	0.941
$T = 25$	0.945	0.942	0.944	0.941	0.943	0.935	0.941	0.941

Note: The biases are in percentage of the truth; *ABC1–ABC4* refer to the analytically bias-corrected estimators with bandwidth $L = 2$; results based on 1,000 repetitions.

[12] We use the *lfe* package of Gaure (2013a) for the estimation of linear fixed effects models, the nonlinear equations solver (*nleqslv*) of Hasselman (2018) for analytical bias corrections at the level of the scores, and R version 4.0.2 R Core Team 2021.

of the estimator.[13] *ABC3* and *ABC4* perform equally well. Our results for the average partial effects are very similar and provided in Table 2.B.1 of the Appendix 2.B.

Next, we compare the two different split-panel jackknife estimators for the structural parameters in a balanced panel. Table 2.3 reports the relative biases and coverage probabilities. Similar to the analytical bias corrections, we find improved finite sample properties as T increases. One exception is the bias of the estimators in $T = 20$. However, although the bias increases, coverage does not decrease. Overall, we find almost identical properties of both estimators, which is remarkable, because we would expect *SPJ2* to have higher dispersion, as significantly smaller sub panels are used to estimate the asymptotic bias. Results for the average partial effects are similar and reported in Table 2.B.2 of the Appendix 2.B.

Table 2.3: *Split-Panel Jackknife Bias Corrections: Coefficients*

	Relative Bias				Coverage Rate			
	$\hat{\rho}$		$\hat{\beta}$		$\hat{\rho}$		$\hat{\beta}$	
	SPJ1	SPJ2	SPJ1	SPJ2	SPJ1	SPJ2	SPJ1	SPJ2
$T = 15$	-0.030	0.540	-0.744	-1.620	0.907	0.907	0.895	0.884
$T = 20$	3.171	3.564	-1.897	-2.544	0.916	0.917	0.900	0.875
$T = 25$	-0.379	-0.158	-0.600	-0.986	0.930	0.931	0.904	0.900

Note: The biases are in percentage of the truth; *SPJ1–2* refer to the different split-panel jackknife bias-corrected estimators; results based on 1,000 repetitions.

Now we analyze whether the finite sample properties of the different estimators are affected by the two patterns of randomly missing data. Because we already found that the finite sample properties of the different analytical and split-panel jackknife bias corrections barely differ among themselves, we restrict our final analysis to *MLE*, *ABC1*, *SPJ1*, and *LPM*. Table 2.4 and 2.5 report the relative biases and coverage probabilities of the estimators for the structural parameters and average partial effects in balanced and unbalanced panels. We start with the analysis of the finite sample properties in a balanced panel, as these will serve as a benchmark for the properties in unbalanced panels. First, we find that the estimators for the effects of y_{it-1} are worse than those for x_{it}. For instance, we observe larger biases and coverage probabilities that are further away from their nominal level. The distortions in the coefficients are also apparent in the estimates of the average partial effects, which is in contrast to the negligible small biases in the average partial effects of x_{it}.[14] In general, the bias corrections perform well in reducing the relative biases and improving coverage. As in Fernández-Val and Weidner (2016), the properties of *SPJ1* are worse than those of *ABC1*. *LPM* as an alternative estimator for the average partial effects works well too.[15] Interestingly, *LPM* tends to overestimate the average partial effects of y_{it-1}, while the other estimators underestimate them. Further, the optimal bandwidths of *LPM* for the estimation of the spectral expectations are much larger than for *ABC1* and increase rapidly with T. This indicates that the temporal dependence induced by the probit data generating process can be very strong, which in turn makes the choice of appropriate bandwidths for *LPM* difficult in practice. Next we compare the finite sample properties of the different estimators in unbalanced panels with our benchmark. First, by comparing the relative biases of *MLE*, we can confirm the conjecture of Fernández-Val and Weidner (2018a) that the

[13] Juodis (2015) analyzed an iterated analytical bias correction, similar to *ABC2*, for a static design with only individual unobserved effects with similar findings.

[14] Hahn and Newey (2004), Fernández-Val (2009), and Fernández-Val and Weidner (2016) have similar findings for average partial effects of an exogenous regressor in balanced panels.

[15] Fernández-Val (2009) has similar findings for linear probability models with only individual fixed effects.

Table 2.4: *Finite Sample Properties: Lagged Dependent Variable*

		Relative Bias			Coverage Rate		
		Bal	P1	P2	Bal	P1	P2
		Panel A: Coefficient					
$T = \overline{T} = 15$	MLE	-41.872	-40.832	-40.192	0.142	0.009	0.014
	ABC1	-3.886	-5.248	-4.787	0.932	0.911	0.936
	SPJ1	-0.030	-32.175	-19.733	0.907	0.115	0.441
$T = \overline{T} = 20$	MLE	-31.360	-30.098	-29.931	0.232	0.104	0.111
	ABC1	-2.575	-2.998	-2.894	0.947	0.951	0.940
	SPJ1	3.171	-13.741	-8.846	0.916	0.668	0.822
$T = \overline{T} = 25$	MLE	-24.997	-24.405	-23.860	0.306	0.263	0.278
	ABC1	-1.815	-2.057	-1.554	0.945	0.949	0.942
	SPJ1	-0.379	-4.595	-2.377	0.930	0.912	0.925
		Panel B: Average Partial Effect					
$T = \overline{T} = 15$	MLE	-49.162	-48.230	-47.611	0.030	0.001	0.002
	ABC1 (2)	-5.212	-6.960	-6.419	0.910	0.873	0.894
	SPJ1	-10.657	-39.059	-26.846	0.825	0.022	0.183
	LPM (4)	5.269	1.800	2.153	0.913	0.961	0.930
$T = \overline{T} = 20$	MLE	-37.768	-36.550	-36.496	0.083	0.017	0.018
	ABC1 (2)	-3.200	-3.816	-3.849	0.911	0.918	0.910
	SPJ1	-2.929	-18.853	-13.854	0.903	0.488	0.666
	LPM (7)	4.398	-0.209	-0.087	0.910	0.954	0.936
$T = \overline{T} = 25$	MLE	-30.634	-30.027	-29.551	0.147	0.105	0.110
	ABC1 (2)	-2.156	-2.479	-2.014	0.924	0.930	0.927
	SPJ1	-3.944	-8.102	-5.813	0.892	0.832	0.873
	LPM (13)	1.620	0.676	1.213	0.922	0.937	0.924

Note: The biases are in percentage of the truth; *Bal*, *P1*, and *P2* refer to balanced panel, *Pattern 1*, and *Pattern 2*; *MLE*, *ABC1*, *SPJ1*, and *LPM* denote the (bias-corrected) estimators; values in parentheses indicate "optimal" bandwidth choices for the spectral density estimator; results based on 1,000 repetitions.

Table 2.5: *Finite Sample Properties: Exogenous Regressor*

		Relative Bias			Coverage Rate		
		Bal	P1	P2	Bal	P1	P2
		Panel A: Coefficient					
$T = \overline{T} = 15$	MLE	14.551	13.635	13.508	0.226	0.051	0.034
	ABC1	1.004	0.849	0.761	0.943	0.944	0.952
	SPJ1	-0.744	8.995	5.790	0.895	0.343	0.640
$T = \overline{T} = 20$	MLE	10.657	9.838	9.962	0.329	0.205	0.205
	ABC1	0.606	0.301	0.400	0.941	0.943	0.948
	SPJ1	-1.897	3.292	2.130	0.900	0.814	0.886
$T = \overline{T} = 25$	MLE	8.401	8.126	8.113	0.408	0.346	0.357
	ABC1	0.319	0.277	0.261	0.943	0.951	0.955
	SPJ1	-0.600	0.621	0.109	0.904	0.934	0.930
		Panel B: Average Partial Effect					
$T = \overline{T} = 15$	MLE	2.845	2.176	2.134	0.878	0.882	0.862
	ABC1 (2)	-0.096	-0.433	-0.458	0.943	0.932	0.932
	SPJ1	2.617	1.744	2.797	0.870	0.819	0.812
	LPM (4)	0.206	0.162	0.237	0.915	0.914	0.908
$T = \overline{T} = 20$	MLE	2.285	1.731	1.693	0.881	0.874	0.890
	ABC1 (2)	0.044	-0.271	-0.316	0.942	0.924	0.925
	SPJ1	1.364	0.946	1.494	0.909	0.896	0.895
	LPM (7)	0.166	0.312	0.225	0.923	0.912	0.924
$T = \overline{T} = 25$	MLE	1.848	1.675	1.614	0.888	0.903	0.890
	ABC1 (2)	0.015	-0.048	-0.112	0.932	0.941	0.933
	SPJ1	0.889	0.735	0.836	0.917	0.924	0.908
	LPM (13)	0.402	0.395	0.276	0.916	0.927	0.912

Note: The biases are in percentage of the truth; *Bal*, *P1*, and *P2* refer to balanced panel, *Pattern 1*, and *Pattern 2*; *MLE*, *ABC1*, *SPJ1*, and *LPM* denote the (bias-corrected) estimators; values in parentheses indicate “optimal” bandwidth choices for the spectral density estimator; results based on 1,000 repetitions.

magnitude of the asymptotic bias in the limiting distribution depends on $\overline{N}$ and $\overline{T}$. However, the coverage is worse because the sample size of an unbalanced panel is larger than that of a balanced panel (with equal $\overline{N}$ and $\overline{T}$), which in turn implies a smaller standard deviation of the estimators and thus distortions that are larger relative to the variance of the estimator. Second, compared to the benchmark we notice some substantial differences in the performance of *SPJ1*. Whereas the properties of *ABC1* and *LPM* are unaffected by the different patterns of missing data, they are partially significantly worse for *SPJ1*, especially for $T = \overline{T} = 15$. *Pattern 1* stands out in particular, because it clearly shows that the reduction of bias and improvement of coverage are deteriorating. An intuitive explanation is that the splitting strategy leads to sub panels of widely differing sizes. This issue is not so severe in *Pattern 2*, but the performance is still worse than the benchmark.

Finally, we briefly summarize the key findings. We find no differences in the finite sample performance when we compare the various analytically bias-corrected and the different split-panel jackknife estimators among themselves. Although the latter have the advantage that they are relatively easy to implement, this convenience is associated with some performance losses. More precisely, split-panel jackknife estimators have higher distortion and react sensitive to different patterns of randomly missing data, whereas the performance of the analytical bias corrections is unaffected. An alternative estimator for the average partial effects based on the bias-corrected ordinary least squares estimator works well too, but to find an appropriate bandwidth for the required spectral density estimator might be challenging in practice.

2.4 Empirical Illustration

We analyze the inter-temporal labor force participation of women using longitudinal micro data (1984–2013) from the German Socio Economic Panel (*GSOEP*). More specifically, we investigate how fertility decisions and non-labor income jointly affect women's labor force participation using an unbalanced panel data set of 6,241 women observed consecutively for at least ten years. Further details about the sample are provided in the Appendix 2.C.

In spirit of Hyslop (1999), we estimate the following model specification:

$$\begin{aligned} y_{it} &= \mathbf{1}\{\rho y_{it-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma_t + e_{it} \geq 0\} \, , \\ (i &= 1, \ldots, N, \; t = 1 \leq s_i, \ldots, T_i \leq T) \, , \end{aligned}$$

where y_{it} is an indicator equal to one if woman i participates in the labor market at time t, $\mathbf{x}_{it}$ is a vector of explanatory variables, $\boldsymbol{\beta}$ are the corresponding parameters, α_i and γ_t are unobserved effects that capture individual specific taste for labor and permanent income as well as control for the business cycle and other time specific shifts in preferences, and e_{it} is an idiosyncratic error term independent of $\mathbf{X}_i^t$ and $\boldsymbol{\pi}$ with mean zero. As our data set is unbalanced, we have to additionally assume that conditional on $\mathbf{X}_i^t$ and $\boldsymbol{\pi}$ the attrition process is random. We consider the following explanatory variables: the number of children in different age groups, various non-labor income classes, and an indicator that is equal to one if a birth occurs in the next year. Further control variables are age squared, marital status, a regional identifier for Eastern Germany, the number of children between zero and one in the previous year, and the number of other household members.

Table 2.6 shows some descriptive statistics of our sample. The average participation rate is 73% in the full sample and 67% for the group of movers who change their labor force participation decision at least once. The group of women who never participate is the smallest and most different from the other groups. On average, this group is older, more likely to be married, and lives in Western Germany. Contrary, women

Table 2.6: *Descriptive Statistics*

	Full	Always	Never	Movers
Participation	0.73 (0.44)	1.00 (0.00)	0.00 (0.00)	0.67 (0.47)
Age	40.48 (11.02)	42.69 (9.80)	48.02 (10.65)	38.21 (11.06)
Married	0.70 (0.46)	0.67 (0.47)	0.93 (0.26)	0.69 (0.46)
Middle Class	0.44 (0.50)	0.42 (0.49)	0.48 (0.50)	0.44 (0.50)
Upper Class	0.01 (0.08)	0.01 (0.08)	0.01 (0.10)	0.01 (0.08)
East	0.22 (0.41)	0.28 (0.45)	0.05 (0.22)	0.20 (0.40)
#Children 0-1	0.05 (0.21)	0.02 (0.13)	0.04 (0.21)	0.06 (0.25)
#Children 2-4	0.12 (0.36)	0.05 (0.23)	0.12 (0.36)	0.16 (0.41)
#Children 5-18	0.71 (0.95)	0.53 (0.80)	0.81 (1.14)	0.79 (0.99)
#HH older	2.25 (0.84)	2.18 (0.80)	2.61 (0.94)	2.25 (0.84)
Birth$_{t+1}$	0.03 (0.18)	0.01 (0.11)	0.03 (0.16)	0.04 (0.21)
#Observations	97,465	33,574	7,175	56,716
#Individuals (N)	6,241	2,306	477	3,458
#Years (T)	28	28	28	28
Avg. #Individuals ($\overline{N}$)	3,481	1,199	256	2,026
Avg. #Years ($\overline{T}$)	16	15	15	16

Note: Means and standard deviations in parentheses.
Source: GSOEP 1984–2013.

who always participate have less children and live in smaller households. The full sample comprises 97,465 observations and consists of 6,241 women observed for a maximum of 28 years. As some households drop out of the *GSOEP* and are replaced by new ones, this leads to a pattern of missing data similar to *Pattern 2* in the simulation study. On average, we observe each woman for roughly 16 years.

We consider the following estimators for the structural parameters and average partial effects: *MLE*, *ABC1*, *SPJ1*, and *LPM*. The estimators are labeled and bandwidths are chosen as in the simulation study. Table 2.7 reports the corresponding estimates. All results are in line with the theoretical model of Hyslop (1999). We find positive state dependence and negative effects of transitory non-labor income, number of children, and expectations about future fertility. All effects are significant at the 5% level. Remarkably, most probit estimates of the average partial effects are very close to each other. Exceptions are those with respect to lagged participation which range from 0.23 up to 0.31. *LPM* estimates are also quite close except for lagged participation and number of children between zero and one. Overall we find evidence for strong state dependence. A woman who has currently a job increases her probability to participate in the future by 23–54 percentage points. Further we find that women respond heterogeneously to changes in transitory non-labor income. Being in the middle class reduces the participation probability by roughly one percentage point compared to a woman in the lower income class. The reduction associated with belonging to the upper income class is significantly stronger with three up to five percentage points. Finally we find that the number of children reduces the likelihood of participation substantially. As expected, the effect is declining with the age of children. Each additional child between zero and one reduces the probability to participate 20 up to 30 percentage points. For children older than four, the reduction is only one percentage point. The results are largely consistent with the empirical findings of Hyslop (1999). However, contrary to Hyslop (1999), we find that future birth always negatively affects current participation decision irrespective of the chosen estimator. This might support the author's perfect foresight assumption with respect to life-cycle fertility decisions.

Table 2.7: *Labor Force Participation of Women*

	Coefficient			Average Partial Effect			
	MLE	ABC1	SPJ1	MLE	ABC1	SPJ1	LPM
Participation$_{t-1}$	1.469	1.657	1.712	0.233	0.300	0.307	0.540
	(0.017)	(0.017)	(0.017)	(0.019)	(0.023)	(0.006)	(0.015)
Middle Class	-0.122	-0.112	-0.114	-0.013	-0.013	-0.014	-0.014
	(0.023)	(0.023)	(0.023)	(0.003)	(0.003)	(0.002)	(0.003)
Upper Class	-0.365	-0.330	-0.416	-0.042	-0.041	-0.051	-0.034
	(0.130)	(0.132)	(0.132)	(0.017)	(0.016)	(0.016)	(0.016)
#Children 0-1	-1.822	-1.603	-1.704	-0.198	-0.190	-0.206	-0.304
	(0.036)	(0.035)	(0.036)	(0.017)	(0.015)	(0.006)	(0.010)
#Children 2-4	-0.427	-0.304	-0.335	-0.046	-0.036	-0.042	-0.039
	(0.026)	(0.025)	(0.026)	(0.005)	(0.004)	(0.003)	(0.005)
#Children 5-18	-0.159	-0.113	-0.112	-0.017	-0.013	-0.014	-0.014
	(0.012)	(0.012)	(0.013)	(0.002)	(0.002)	(0.001)	(0.002)
Birth$_{t+1}$	-0.560	-0.514	-0.572	-0.065	-0.066	-0.074	-0.080
	(0.038)	(0.038)	(0.038)	(0.007)	(0.007)	(0.005)	(0.007)

Note: MLE, *ABC1*, *SPJ1*, and *LPM* denote the (bias-corrected) estimators; bandwidths are 2 and 4 for *ABC1* and *LPM*; standard errors in parentheses; *LPM* standard errors are robust to heteroskedasticity and clustered by woman and year; estimates relative to a low income woman in the west.
Further control variables: squared age, married, east, lag of number of children between zero and one, and number of household members above 18.
Source: GSOEP 1984–2013.

Finally, we check the sensitivity of *ABC1* and *LPM* to different bandwidth choices and conduct a simulation study calibrated to our empirical illustration. The results are reported in Table 2.C.1 and 2.C.2 of the Appendix 2.C. With respect to the different bandwidth choices, we find that especially the *ABC1* estimates are very robust. We find the largest variation with respect to the effect of lagged participation. The other effects are almost indistinguishable. Most *LPM* estimates are robust as well. Exceptions are the effects of lagged participation, being in the upper class, and number of children between zero and one. The results of the calibrated simulation study confirm that the finite sample properties of *MLE* can be improved. However, the improvement is less good compared to the simulation experiments with artificial data. The performance of *SPJ1* and *LPM* is in some cases significantly worse than that of *ABC1*.

2.5 Concluding Remarks

In this paper, we offer new relief and guidance for empirical researchers by showing how popular binary choice estimators benefit from recent advances in econometrics. Especially the analytically bias-corrected estimator of Fernández-Val and Weidner (2016) convinced by its good performance in all cases.

Although we have focused on panel data binary choice models, we would like to point out that the bias corrections derived by Fernández-Val and Weidner (2016) are far more general. First, they can also be used to reduce the asymptotic bias of other popular nonlinear maximum likelihood estimators e. g. poisson and tobit. The algorithms described in this paper can be easily adapted to these problems. Second, the bias corrections can also be applied if we observe cross-sections of networked activities instead of panels e. g. a cross-section of bilateral trade flows. For instance, Cruz-Gonzalez, Fernández-Val, and Weidner (2017) use the bias corrections of Fernández-Val and Weidner (2016) to mitigate the asymptotic bias problem in a

Helpman, Melitz, and Rubinstein (2008)-type model to determine the extensive margin of trade.

Recently, there are also some extensions of Fernández-Val and Weidner (2016). For instance, Weidner and Zylkin (2020) and Hinz, Stammann, and Wanner (2020) have extended these bias corrections to a special three-way error component that is particularly relevant for the estimation of the intensive and extensive margin of trade in a panel of bilateral trade flows. Further, Chernozhukov, Fernández-Val, and Weidner (2020) use a sequence of binary choice regressions to estimate distribution functions of non-binary outcomes conditional on strictly exogenous regressors and two unobserved effects. The resulting inference problem is addressed by, among other things, the sequential application of analytic bias corrections. These extensions can also benefit from the findings of this paper.

Finally, future research could investigate whether bootstrap procedures can further improve inference. In an additional simulation study we find that bootstrapping a bias-corrected estimator in spirit of Kaffo (2014) slightly improves the coverage of the estimators for the structural parameters but not for the average partial effects.

References

Andersen, Erling Bernhard. 1970. "Asymptotic Properties of Conditional Maximum-Likelihood Estimators." *Journal of the Royal Statistical Society: Series B (Methodological)* 32 (2): 283–301.

Arellano, Manuel, and Jinyong Hahn. 2007. "Understanding Bias in Nonlinear Panel Models: Some Recent Developments." In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress,* edited by Richard Blundell, Whitney Newey, and Torsten Persson, 3:381–409. Econometric Society Monographs. Cambridge University Press.

Baltagi, Badi H. 2013. *Econometric Analysis of Panel Data.* 5th ed. John Wiley & Sons.

Bester, C. Alan, and Christian Hansen. 2009. "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects." *Journal of Business & Economic Statistics* 27 (2): 131–148.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference With Multiway Clustering." *Journal of Business & Economic Statistics* 29 (2): 238–249.

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

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *The Review of Economic Studies* 47 (1): 225–238.

Chernozhukov, Victor, Iván Fernández-Val, and Martin Weidner. 2020. "Network and panel quantile effects via distribution regression." *Journal of Econometrics.*

Correia, Sergio. 2016. "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator." *Working Paper.*

Cruz-Gonzalez, Mario, Iván Fernández-Val, and Martin Weidner. 2017. "Bias corrections for probit and logit models with two-way fixed effects." *The Stata Journal* 17 (3): 517–545.

Dhaene, Geert, and Koen Jochmans. 2015. "Split-panel Jackknife Estimation of Fixed-effect Models." *The Review of Economic Studies* 82 (3): 991–1030.

Fernández-Val, Iván. 2009. "Fixed effects estimation of structural parameters and marginal effects in panel probit models." *Journal of Econometrics* 150 (1): 71–85.

Fernández-Val, Iván, and Martin Weidner. 2016. "Individual and time effects in nonlinear panel models with large N, T." *Journal of Econometrics* 192 (1): 291–312.

———. 2018a. "Fixed Effects Estimation of Large-T Panel Data Models." *Annual Review of Economics* 10 (1): 109–138.

Gaure, Simen. 2013a. "lfe: Linear group fixed effects." *The R Journal* 5 (2): 104–117.

Greene, William H. 2004. "The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects." *The Econometrics Journal* 7 (1): 98–119.

Guimarães, Paulo, and Pedro Portugal. 2010. "A simple feasible procedure to fit models with high-dimensional fixed effects." *Stata Journal* 10 (4): 628–649.

Hahn, Jinyong, and Guido Kuersteiner. 2002. “Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when Both n and T Are Large.” *Econometrica* 70 (4): 1639–1657.

———. 2007. “Bandwidth Choice for Bias Estimators in Dynamic Nonlinear Panel Models.” *Working Paper.*

Hahn, Jinyong, and Hyungsik Roger Moon. 2006. “Reducing Bias of MLE in a Dynamic Panel Model.” *Econometric Theory* 22 (3): 499–512.

Hahn, Jinyong, and Whitney Newey. 2004. “Jackknife and Analytical Bias Reduction for Nonlinear Panel Models.” *Econometrica* 72 (4): 1295–1319.

Halperin, Israel. 1962. “The product of projection operators.” *Acta Sci. Math. (Szeged)* 23:96–99.

Hasselman, Berend. 2018. *nleqslv: Solve Systems of Nonlinear Equations.* R package version 3.3.2. `https://CRAN.R-project.org/package=nleqslv`.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein. 2008. “Estimating trade flows: Trading partners and trading volumes.” *The Quarterly Journal of Economics* 123 (2): 441–487.

Hinz, Julian, Amrei Stammann, and Joschka Wanner. 2020. “State Dependence and Unobserved Heterogeneity in the Extensive Margin of Trade.” *arXiv preprint arXiv:2004.12655.*

Honoré, Bo E., and Ekaterini Kyriazidou. 2000. “Panel Data Discrete Choice Models with Lagged Dependent Variables.” *Econometrica* 68 (4): 839–874.

Hsiao, Cheng. 2014. *Analysis of Panel Data.* 3rd ed. Econometric Society Monographs. Cambridge University Press.

Hu, Luojia. 2002. “Estimation of a Censored Dynamic Panel Data Model.” *Econometrica* 70 (6): 2499–2517.

Hyslop, Dean R. 1999. “State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women.” *Econometrica* 67 (6): 1255–1294.

Juodis, Artūras. 2015. “Iterative Bias Correction Procedures Revisited: A Small Scale Monte Carlo Study.” *Working Paper.*

Kaffo, Maximilien. 2014. “Bootstrap inference for nonlinear dynamic panel data models with individual fixed effects.” *Working Paper.*

Kim, Min Seong, and Yixiao Sun. 2016. “BOOTSTRAP AND k-STEP BOOTSTRAP BIAS CORRECTIONS FOR THE FIXED EFFECTS ESTIMATOR IN NONLINEAR PANEL DATA MODELS.” *Econometric Theory* 32 (6): 1523–1568.

Lancaster, Tony. 2002. “Orthogonal Parameters and Panel Data.” *The Review of Economic Studies* 69, no. 3 (July): 647–666.

Neyman, Jerzy, and Elizabeth L. Scott. 1948. “Consistent Estimates Based on Partially Consistent Observations.” *Econometrica* 16 (1): 1–32.

Nickell, Stephen. 1981. “Biases in Dynamic Models with Fixed Effects.” *Econometrica* 49 (6): 1417–1426.

Phillips, Peter C. B., and Hyungsik R. Moon. 1999. “Linear Regression Limit Theory for Nonstationary Panel Data.” *Econometrica* 67 (5): 1057–1111.

Quenouille, M. H. 1949. “Approximate Tests of Correlation in Time-Series.” *Journal of the Royal Statistical Society. Series B (Methodological)* 11 (1): 68–84.

———. 1956. “Notes on Bias in Estimation.” *Biometrika* 43 (3/4): 353–360.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Rasch, George. 1960. “Probabilistic models for some intelligence and attainment tests: Danish institute for Educational Research.” *Denmark Paedogiska, Copenhagen.*

Stammann, Amrei. 2018. “Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects.” *arXiv preprint arXiv:1707.01815.*

Stammann, Amrei, Florian Heiss, and Daniel McFadden. 2016. “Estimating Fixed Effects Logit Models with Large Panel Data.” *Working Paper.*

Wagner, Gert G., Joachim R. Frick, and Jürgen Schupp. 2007. “The German Socio-Economic Panel study (SOEP)-evolution, scope and enhancements.”

Weidner, Martin, and Thomas Zylkin. 2020. “Bias and Consistency in Three-Way Gravity Models.” *arXiv preprint arXiv:1909.01327.*

Woutersen, Tiemen. 2002. “Robustness against incidental parameters.” *Working Paper.*

Appendix

2.A Feasible Estimation for Long Panels

2.A.1 Derivation of the Algorithms

We briefly review the derivation of the algorithm proposed by Stammann (2018) for binary choice models with two unobserved effects.

We start with the weighted least squares problem in iteration r:

$$\boldsymbol{\theta}^{\langle r+1\rangle} = (\boldsymbol{\beta}^{\langle r+1\rangle}, \boldsymbol{\pi}^{\langle r+1\rangle}) = (\mathbf{Z}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{Z})^{+}\mathbf{Z}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{w}^{\langle r\rangle}\,, \tag{2.1}$$

where $\boldsymbol{\pi}^{\langle r\rangle} = (\boldsymbol{\alpha}^{\langle r\rangle}, \boldsymbol{\gamma}^{\langle r\rangle})$, $\mathbf{Z} = (\mathbf{X}, \mathbf{D})$, $\boldsymbol{\Omega}^{\langle r\rangle}$ is a diagonal weighting matrix with $\operatorname{diag}(\boldsymbol{\Omega}^{\langle r\rangle}) = (\omega_{it}^{\langle r\rangle})_{(i,t)\in\mathcal{S}}$, $\mathbf{w}^{\langle r\rangle} = \boldsymbol{\nu}^{\langle r\rangle} + \boldsymbol{\eta}^{\langle r\rangle}$, $\boldsymbol{\nu}^{\langle r\rangle} = (\nu_{it}^{\langle r\rangle})_{(i,t)\in\mathcal{S}}$, and $\boldsymbol{\eta}^{\langle r\rangle} = (\eta_{it}^{\langle r\rangle})_{(i,t)\in\mathcal{S}}$. The corresponding expressions for ω_{it} and ν_{it} are given in section 2.2. (2.1) implies the following normal equations:

$$\begin{aligned} \mathbf{X}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{X}\boldsymbol{\beta}^{\langle r+1\rangle} + \mathbf{X}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{D}\boldsymbol{\pi}^{\langle r+1\rangle} &= \mathbf{X}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{w}^{\langle r\rangle}\,, & (2.2)\\ \mathbf{D}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{X}\boldsymbol{\beta}^{\langle r+1\rangle} + \mathbf{D}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{D}\boldsymbol{\pi}^{\langle r+1\rangle} &= \mathbf{D}'\boldsymbol{\Omega}^{\langle r\rangle}\mathbf{w}^{\langle r\rangle}\,. & (2.3)\end{aligned}$$

Re-arranging (2.3) yields

$$\mathbf{D}\boldsymbol{\pi}^{\langle r+1\rangle} = \mathbb{P}\,\mathbf{w}^{\langle r\rangle} - \mathbb{P}\,\mathbf{X}^{\langle r\rangle}\boldsymbol{\beta}^{\langle r+1\rangle}\,. \tag{2.4}$$

Substituting (2.4) into (2.2) results in an alternative weighted least squares problem

$$\boldsymbol{\beta}^{\langle r+1\rangle} = ((\mathbb{M}\,\mathbf{X}^{\langle r\rangle})'\boldsymbol{\Omega}^{\langle r\rangle}\,\mathbb{M}\,\mathbf{X}^{\langle r\rangle})^{-1}(\mathbb{M}\,\mathbf{X}^{\langle r\rangle})'\boldsymbol{\Omega}^{\langle r\rangle}\,\mathbb{M}\,\mathbf{w}^{\langle r\rangle}\,. \tag{2.5}$$

based on transformed variables that allows us to update $\boldsymbol{\beta}$ without explicitly updating the incidental parameters $\boldsymbol{\pi}$. This transformation can be interpreted as a weighted within-transformation ("demeaning") as known from linear fixed effects models. Consequently, $\boldsymbol{\beta}$ can be obtained by a regression of the transformed $\mathbf{w}$ on the transformed $\mathbf{X}$ using $\boldsymbol{\omega}$ as weights. Meanwhile, this type of optimization problem can be solved very efficiently thanks to special software routines. Some examples available in popular statistical software are *reghdfe* by Correia (2016) for *Stata*, *lfe* by Gaure (2013a) for *R*, *FixedEffectModels* by Matthieu Gomez for *Julia*, and *pyhdfe* by Jeff Gortmaker for *Python*. To complete the optimization algorithm we need to find a way to update ω_{it} and ν_{it}. Fortunately, it is quite easy to update η_{it} from already computed quantities. We can exploit the fact that the residuals of (2.1) and (2.5) are equal, as shown by Gaure (2013b). Therefore

$$\boldsymbol{\eta}^{\langle r+1\rangle} = \mathbf{w}^{\langle r\rangle} - (\mathbb{M}\,\mathbf{w}^{\langle r\rangle} - \mathbb{M}\,\mathbf{X}^{\langle r\rangle}\boldsymbol{\beta}^{\langle r+1\rangle}) \tag{2.6}$$

are simply the fitted values of (2.5). We can sketch the entire algorithm as follows:

Algorithm 1. IWLS Algorithm

Set an initial $\boldsymbol{\eta}$, e. g. $\boldsymbol{\eta} = \mathbf{0}_n$, and repeat the following steps.

Step 1. Compute $\boldsymbol{\omega}$ and $\mathbf{w}$.

Step 2. Solve (2.5), where $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are the coefficients and fitted values.

Step 3. Check convergence.

Finally, we outline the algorithm for updating the linear index η_{it} given a fixed $\tilde{\boldsymbol{\beta}}$. For instance, this algorithm is required for the bias corrections of the average partial effects. Remember,

$$\mathbf{D}\boldsymbol{\pi}^{\langle r+1 \rangle} = \mathbb{P}\,\tilde{\mathbf{w}}^{\langle r \rangle} = \tilde{\mathbf{w}}^{\langle r \rangle} - \mathbb{M}\,\tilde{\mathbf{w}}^{\langle r \rangle}$$

with $\tilde{\mathbf{w}}^{\langle r \rangle} = \mathbf{w}^{\langle r \rangle} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ is sufficient to update $\boldsymbol{\eta}^{\langle r \rangle} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{D}\boldsymbol{\pi}^{\langle r \rangle}$. We can sketch the entire algorithm as follows:

Algorithm 2. IWLS Offset Algorithm

Set an initial $\boldsymbol{\eta}$, e. g. $\boldsymbol{\eta} = \mathbf{0}_n$, and repeat the following steps.

Step 1. Compute $\boldsymbol{\omega}$ and $\mathbf{w}^*$.

Step 2. Compute $\mathbf{D}\boldsymbol{\pi}$ and update $\boldsymbol{\eta}$.

Step 3. Check convergence.

In the next subsection we present an example code in which we combine the algorithms described here with the analytical bias correction of Fernández-Val and Weidner (2016).

2.A.2 Example Code

```
########################################################
# Example R-Code for Logit MLE
########################################################


## Load required packages and data set
require(bife)
require(lfe)
data <- psid

## Helper functions

# Log-Likelihood function
logl <- function(y, eta) sum(log(plogis((2.0 * y - 1.0) * eta)))

# Derivatives of CDF
partial2F <- function(eta) dlogis(eta) * (1.0 - 2.0 * plogis(eta))
partial3F <- function(eta) {
dlogis(eta) * ((1.0 - 2.0 * plogis(eta))^2 - 2.0 * dlogis(eta))
}

## Estimate structural parameters

# Set initial \beta, \eta, and termination criteria
n <- nrow(data)
fval <- - 1.0e100
eta <- numeric(n)
maxiter <- 100L
tol <- 1.0e-05

# Find optimal \beta and \eta
for (iter in 1:maxiter) {
# Store previous function value
fvalold <- fval

# Update weights and working response
omega <- dlogis(eta)
data$w <- (data$LFP - plogis(eta)) / omega + eta

# Update \beta and \eta
reg <- felm(w ~ KID1 + KID2 + KID3 + log(INCH) | ID + TIME,
data, weights = omega)
beta <- coef(reg)
eta <- as.vector(fitted.values(reg))
fval <- logl(data$LFP, eta)

# Check convergence
if (abs(fval - fvalold) / (abs(fvalold) + 0.1) < tol) break
}

# Final estimates
betahat <- beta
etahat <- eta

## Debias maximum likelihood estimates of the structural parameters

# Compute required derivatives and demeaned regressor matrix
d1lhat <- data$LFP - plogis(etahat)
omegahat <- dlogis(etahat)
d2Fhat <- partial2F(etahat)
reg <- felm(KID1 + KID2 + KID3 + log(INCH) ~ 0 | ID + TIME,
data, weights = omegahat)
MhatX <- residuals(reg)

# Estimate asymptotic bias
Tempmat <- MhatX * d2Fhat
Bhatnum <- aggregate(Tempmat, list(data$ID), sum)[, - 1L]
Bhatdenom <- aggregate(omegahat, list(data$ID), sum)[, - 1L]
Bhat <- - colSums(Bhatnum / Bhatdenom) / 2.0
Chatnum <- aggregate(Tempmat, list(data$TIME), sum)[, - 1L]
Chatdenom <- aggregate(omegahat, list(data$TIME), sum)[, - 1L]
Chat <- - colSums(Chatnum / Chatdenom) / 2.0
What <- crossprod(MhatX * sqrt(omegahat))
bhat <- solve(What, Bhat + Chat)

# Estimate asymptotic covariance matrix
Vhat <- solve(What)

# Debias estimates and compute standard errors
betatilde <- betahat - bhat
sebetatilde <- sqrt(diag(Vhat))

## Debias maximum likelihood estimates of the APEs
```

```
# Compute fixed part of \eta
X <- model.matrix(LFP ~ KID1 + KID2 + KID3 + log(INCH) + 0, data)
Xbetatilde <- as.vector(X %*% betatilde)

# Set initial \eta and termination criteria
fval <- - 1.0e100
eta <- numeric(n)

# Update \eta given debiased \beta
for (iter in 1:maxiter) {
# Store previous function value
fvalold <- fval

# Update weights and working response
omega <- dlogis(eta)
data$wast <- (data$LFP - plogis(eta)) / omega + eta - Xbetatilde

# Update \eta
reg <- felm(wast ~ 0 | ID + TIME, data, weights = omega)
Dpi <- as.vector(fitted.values(reg))
eta <- Xbetatilde + Dpi
fval <- logl(data$LFP, eta)

# Check convergence
if (abs(fval - fvalold) / (abs(fvalold) + 0.1) < tol) break
}

# Final estimates
etatilde <- eta

# Compute APEs and derivatives evaluated at debiased coefficients
Deltahat <- sapply(betatilde, function(x) x * dlogis(etatilde))
d1Deltahat <- sapply(betatilde, function(x) x * partial2F(etatilde))
d2Deltahat <- sapply(betatilde, function(x) x * partial3F(etatilde))
deltahat <- colMeans(Deltahat)
Psihat <- - d1Deltahat / omegahat
reg <- felm(Psihat ~ 0 | ID + TIME, data, weights = omegahat)
PhatPsihat <- fitted.values(reg)

# Estimate asymptotic bias
Tempmat <- PhatPsihat * d2Fhat + d2Deltahat
```

```
Bhatnum <- aggregate(Tempmat, list(data$ID), sum)[, - 1L]
Bhatdenom <- aggregate(omegahat, list(data$ID), sum)[, - 1L]
Bhat <- colSums(Bhatnum / Bhatdenom) / 2.0
Chatnum <- aggregate(Tempmat, list(data$TIME), sum)[, - 1L]
Chatdenom <- aggregate(omegahat, list(data$TIME), sum)[, - 1L]
Chat <- colSums(Chatnum / Chatdenom) / 2.0
bhat <- (Bhat + Chat) / n

# Estimate asymptotic covariance matrix
V1 <- tcrossprod(colSums(Deltahat - deltahat))
J <- crossprod(MhatX, d1Deltahat)
diag(J) <- diag(J) + sum(dlogis(etatilde))
JV <- t(J) %*% solve(What)
Gammahat <- tcrossprod(MhatX * d1lhat, JV) - PhatPsihat * d1lhat
V2 <- crossprod(Gammahat)
Vhat <- (V1 + V2) / n^2

# Debias estimates and compute standard errors
deltatilde <- deltahat - bhat
sedeltatilde <- sqrt(diag(Vhat))
```

2.B Additional Simulation Results

Table 2.B.1: *Analytical Bias Corrections: Average Partial Effects*

	$\hat{\delta}_y$				$\hat{\delta}_x$			
	ABC1	ABC2	ABC3	ABC4	ABC1	ABC2	ABC3	ABC4
	Panel A: Relative Bias							
$T = 15$	-5.212	-8.492	-6.248	-6.805	-0.096	1.050	0.379	0.433
$T = 20$	-3.200	-5.223	-3.744	-4.221	0.044	0.725	0.297	0.352
$T = 25$	-2.156	-3.543	-2.512	-2.892	0.015	0.477	0.181	0.230
	Panel B: Coverage Probability							
$T = 15$	0.910	0.874	0.897	0.895	0.943	0.930	0.941	0.943
$T = 20$	0.911	0.900	0.908	0.910	0.942	0.933	0.941	0.939
$T = 25$	0.924	0.920	0.923	0.925	0.932	0.927	0.928	0.928

Note: The biases are in percentage of the truth; *ABC1–ABC4* refer to the analytically bias-corrected estimators with bandwidth $L = 2$; results based on 1,000 repetitions.

Table 2.B.2: *Split-Panel Jackknife Bias Corrections: Average Partial Effects*

	Relative Bias				Coverage Rate			
	$\hat{\delta}_y$		$\hat{\delta}_x$		$\hat{\delta}_y$		$\hat{\delta}_x$	
	SPJ1	SPJ2	SPJ1	SPJ2	SPJ1	SPJ2	SPJ1	SPJ2
$T = 15$	-10.657	-10.609	2.617	2.460	0.825	0.825	0.870	0.874
$T = 20$	-2.929	-2.890	1.364	1.241	0.903	0.898	0.909	0.913
$T = 25$	-3.944	-3.927	0.889	0.817	0.892	0.890	0.917	0.914

Note: The biases are in percentage of the truth; *SPJ1–2* refer to the different split-panel jackknife bias-corrected estimators; results based on 1,000 repetitions.

2.C Empirical Illustration

2.C.1 Data Preparation

We use the *Cross-National Equivalent File* ($PEQUV-File version 30) of the *GSOEP* and restrict the sample to women between 16 and 65 that are observed consecutively for at least ten years and do not receive any retirement income. A woman is assumed to participate in labor-force if she has positive income from individual labor and works at least 52 hours a year. A proxy for transitory non-labor income is constructed from post-government household income minus woman's individual labor earnings. All income variables are converted to constant 2010 EURO using the consumer price index. We additionally correct the labor income by a household specific tax rate. To make income comparable between different household sizes, we use an equivalence scale proposed by Buhmann et al. (1988) and divide the transitory non-labor income by the square root of household members. To allow for heterogeneous effects of transitory non-labor income on participation decisions, we define the three income classes: lower, middle, and upper. A woman belongs to the lower class if she has a non-labor income of less than 11,278 EURO at her disposal. Contrary a woman is in the upper income class if she has more than 56,391 EURO available. Women in between this interval belong to the middle class. Those numbers are equal to 60% and 300% of the annual median equivalence income.[16] The class distinction is taken from the *Armuts- und Reichtumsbericht* of the federal government.[17] Finally, we group the federal states into Eastern and Western Germany to control for regional differences that still exists after reunification. More precisely, Schleswig-Holstein, Hamburg, Lower-Saxony, Bremen, Hessen, Baden-Wuerttemberg, Bavaria, North-Rhine-Westfalia, Rheinland-Pfalz, and Saarland are classified as Western Germany and Berlin, Brandenburg, Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt, and Thueringia as Eastern Germany.

[16] https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Lebensbedingungen-Armutsgefaehrdung/Tabellen/einkommensverteilung-silc.html.

[17] https://www.armuts-und-reichtumsbericht.de/.

2.C.2 Additional Results

Table 2.C.1: *Sensitivity to Different Bandwidth Choices*

	Coefficient	Average Partial Effect	
	ABC1	ABC1	LPM
Participation$_{t-1}$	[1.580, 1.674]	[0.284, 0.303]	[0.453, 0.540]
Middle Class	[-0.116, -0.107]	[-0.014, -0.013]	[-0.014, -0.012]
Upper Class	[-0.347, -0.321]	[-0.044, -0.040]	[-0.043, -0.031]
#Children 0-1	[-1.610, -1.594]	[-0.191, -0.190]	[-0.313, -0.303]
#Children 2-4	[-0.318, -0.302]	[-0.038, -0.036]	[-0.054, -0.039]
#Children 5-18	[-0.120, -0.113]	[-0.014, -0.013]	[-0.019, -0.013]
Birth$_{t+1}$	[-0.520, -0.501]	[-0.066, -0.064]	[-0.082, -0.080]

Note: The intervals show the range of the estimates; *ABC1* and *LPM* denote the bias-corrected estimators; bandwidths are chosen from $\{1, \ldots, 4\}$ and $\{1, \ldots, 15\}$ for *ABC1* and *LPM*.
Further control variables: squared age, married, east, lag of number of children between zero and one, and number of household members above 18.
Source: GSOEP 1984–2013.

Table 2.C.2: *Results of the Calibrated Simulation Study*

	Coefficient			Average Partial Effect			
	MLE	ABC1	SPJ1	MLE	ABC1	SPJ1	LPM
				Panel A: Relative Bias			
Participation$_{t-1}$	-21.681	-7.173	-12.310	-38.464	-16.452	-24.779	95.404
Middle Class	27.591	8.583	19.596	12.289	4.626	12.369	13.353
Upper Class	24.191	5.734	31.748	9.089	1.641	23.081	-10.789
#Children 0-1	20.875	4.879	11.889	6.402	1.083	5.341	52.210
#Children 2-4	49.423	10.962	24.895	31.531	6.946	19.140	27.993
#Children 5-18	46.165	12.428	23.739	28.661	8.357	17.626	-5.823
Birth$_{t+1}$	10.080	0.656	5.118	-4.889	-4.539	-1.693	23.556
				Panel B: Coverage Probability			
Participation$_{t-1}$	0.000	0.000	0.000	0.000	0.339	0.000	0.000
Middle Class	0.713	0.934	0.807	0.948	0.969	0.834	0.916
Upper Class	0.884	0.954	0.483	0.928	0.933	0.492	0.930
#Children 0-1	0.000	0.431	0.001	1.000	1.000	0.680	0.000
#Children 2-4	0.000	0.647	0.071	0.348	0.997	0.195	0.173
#Children 5-18	0.008	0.731	0.326	0.613	0.983	0.473	0.939
Birth$_{t+1}$	0.714	0.956	0.877	0.985	0.983	0.906	0.333

Note: The biases are in percentage of the truth; *MLE*, *ABC1*, *SPJ1*, and *LPM* denote the (bias-corrected) estimators; bandwidths are 2 and 4 for *ABC1* and *LPM*; *LPM* standard errors are robust to heteroskedasticity and clustered by woman and year; results based on 1,000 repetitions.
Further control variables: squared age, married, east, lag of number of children between zero and one, and number of household members above 18.
Source: GSOEP 1984–2013.

Chapter 3

Distribution Regression with Fixed Effects and Weakly Exogenous Regressors

(joint work with Philipp Berger and Amrei Stammann)

3.1 Introduction

Policy interventions can often be inadequately evaluated using average effects. For instance, if a policy is intended to increase the income of low-income households, i. e. the left tail of the income distribution, a positive average effect does not necessarily imply that the policy intervention was successful. Quantile effects provide a convenient way to evaluate the policy intervention along the entire outcome distribution. A popular method to obtain these effects is the quantile regression originating from Koenker and Bassett (1978). A drawback of quantile regression is that the inference made for quantile effects is invalid if the outcome variable is discrete or mixed.

Chernozhukov, Fernández-Val, and Melly (2013) introduce distribution regression as a flexible alternative to quantile regression that can naturally handle continuous, discrete, or mixed outcomes. Unlike quantile regression, distribution regression does not directly estimate the conditional quantile functions. Instead, an indirect route is taken by first estimating the conditional distribution and then exploiting the relationship between distribution and quantile function, where the latter is the inverse of the former. Quantile effects are then computed as differences between two quantile functions. Recently, Chernozhukov et al. (2020) propose a generic inference method to obtain confidence bands for quantile functions and quantile effects via distribution regression that relies entirely on valid uniform confidence bands for the corresponding distribution functions. The idea is to invert uniform confidence bands for two distribution functions to obtain uniform bands for the two quantile functions. Afterwards, the Minkowski difference is used to obtain uniform confidence bands for quantile effects. Chernozhukov, Fernández-Val, and Weidner (2020) extend the generic inference method presented in Chernozhukov et al. (2020) to network and panel data applications with strictly exogenous regressors. More specifically, they consider distribution regression models with two fixed effects for classical panels (individual and time) or networks (sender and receiver). Since the fixed-effects estimator for distribution regression models is essentially a sequence of fixed effects binary choice estimators, it also suffers from the well-known incidental parameter problem (Neyman and Scott 1948). Chernozhukov, Fernández-Val, and Weidner (2020) address the inference problem by extending previous results on bias corrections from Fernández-Val and Weidner (2016), which were proposed in the context of nonlinear fixed effects models with two unobserved effects.

In this paper, we extend the inference method of Chernozhukov, Fernández-Val, and Weidner (2020) for the case of weakly exogenous regressors, which is particularly relevant in panel applications where the strict exogeneity assumption is often too restrictive. For instance, in our empirical illustration, we analyze how knowledge spillovers and investments in research affect firms' inventiveness. But if firms' investment decisions are influenced by the success of past inventions, then the strict exogeneity assumption is violated. Likewise, research questions based on dynamic models cannot be properly analyzed. Our contribution is meant to be complementary to Chernozhukov, Fernández-Val, and Weidner (2020) and ensures a wide applicability of distribution regression with two unobserved effects in the context of classical panels and networks. Like Chernozhukov, Fernández-Val, and Weidner (2020), we extend the bias correction developed by Fernández-Val and Weidner (2016), but additionally allow the inclusion of weakly exogenous regressors.[1] We demonstrate the importance and effectiveness of our inference method in simulation experiments.[2] Finally, we provide an empirical illustration from innovation economics in which we reassess parts of Arora, Belenzon,

[1] Fernández-Val and Weidner (2016) consider the cases of strictly and weakly exogenous regressors and propose bias-corrected estimators for both.

[2] A rigorous proof for our inference method is work in progress and will be added later.

and Sheer (2021). More specifically, we analyze quantile effects of knowledge outflow and the use of internal and external research on firms' inventiveness as measured by the number of patents. Qualitatively, we can confirm their results, but also find evidence of heterogeneous effects along the distribution of patents.

The paper is organized as follows. We present the model and address the inference problem in Section 3.2. We demonstrate the performance of our inference method in Section 3.3. We provide an empirical illustration in Section 3.4. Finally, we give some concluding remarks in Section 3.5.

3.2 Distribution Regression Model, Assumptions, and the Inference Problem

We consider the distribution regression model of Chernozhukov, Fernández-Val, and Weidner (2020) for panel applications with two unobserved effects and weakly exogenous regressors. We model the conditional distribution of y_{it} given $(x_{it}, \alpha_i, \gamma_t)$ as

$$F_{y_{it}}(y \mid x_{it}, \alpha_i, \gamma_t) = \Lambda(\pi_{it}(y)) \, , \quad \pi_{it}(y) := x_{it}'\beta(y) + \alpha_i v(y) + \gamma_t \varphi(y) \, , \tag{3.1}$$

where $i = \{1, \dots, N\}$ and $t = \{1, \dots, T\}$ are individual- and time-specific indexes, y_{it} is the outcome variable with region of interest $\mathcal{Y}$, $\pi_{it}(y)$ is the linear index, x_{it} is a vector of regressors, α_i and γ_t are unobserved individual and time effects, $\beta(y)$, $v(y)$, and $\varphi(y)$ are unknown model coefficients that are allowed to vary with $y \in \mathcal{Y}$ to account for heterogeneity in the distribution, and $\Lambda(\pi) := 1/(1 + \exp(-\pi))$ is the logistic cumulative distribution function (CDF). We assume a potentially unbalanced panel data set of N individuals observed consecutively for at most T time periods. Formally, we observe panel data $\{(y_{it}, x_{it}) \colon (i,t) \in \mathcal{D}\}$, where $\mathcal{D} \subseteq \{(i,t) \colon i \in \{1, \dots, N\}, t \in \{1, \dots, T\}\}$ is a subset of all observed index pairs and $n := |\mathcal{D}|$ is the sample size. Further, we define $\mathcal{D}_i := \{t_i, \dots, T_i\} = \{t \colon (i,t) \in \mathcal{D}\}$ and $\mathcal{D}_t := \{i \colon (i,t) \in \mathcal{D}\}$, where t_i and T_i denote the beginning and end of the time series of individual i. The model coefficients for each y can be estimated by a logistic regression of $\mathbf{1}\{y_{it} \leq y\}$ on the regressors and fixed effects.[3]

Chernozhukov, Fernández-Val, and Weidner (2020) use distribution regression primarily to construct counterfactual distribution functions, which in turn are used to obtain quantile effect functions. Let $x_{it,k}$ denote a counterfactual outcome of x_{it}. For instance, $x_{it,k}$ might be the outcome after we increase a continuous treatment variable by one unit and hold the other regressors and unobserved effects fixed. The corresponding counterfactual distribution and quantile functions are

$$F_k(y) := \frac{1}{n} \sum_{i=1}^{N} \sum_{t=t_i}^{T_i} \Lambda(\pi_{it,k}(y)) \, , \quad \pi_{it,k}(y) := x_{it,k}'\beta(y) + \alpha_i v(y) + \gamma_t \varphi(y)$$

and

$$Q_k(\tau) := F_k^{\leftarrow}(\tau) := \min(\inf\{y \in \mathcal{Y} \colon F_k(y) \geq \tau\}, \sup\{y \in \mathcal{Y}\}) \, ,$$

where $\tau \in (0,1)$ are quantile indexes and $\inf\{\varnothing\} = \infty$. The quantile functions for two different counterfactual outcomes $k \in \{0,1\}$ can then be used to construct quantile effect functions $\Delta(\tau) = Q_1(\tau) - Q_0(\tau)$. For instance, $x_{it,1}$ might be the counterfactual outcome if we increase the treatment variable by one unit and $x_{it,0} = x_{it}$ might be the observed outcome. A distinctive feature of the inference method of Chernozhukov, Fernández-Val, and Weidner (2020) is that it relies entirely on valid uniform confidence bands for the corresponding counterfactual distribution functions. Uniform confidence bands for the quantile effect

[3] In practice, y is an element of a finite subset of $\mathcal{Y}$, e. g. a grid point on an equidistant grid covering $\mathcal{Y}$.

functions can then be constructed from the inverted uniform confidence bands for both counterfactual distribution functions and the Minkowski difference of these two bands.

To make valid inference in the presence of weakly exogenous regressors, we need to extend the inference method of Chernozhukov, Fernández-Val, and Weidner (2020), which already corrects for the incidental parameter problem. In particular, we have to extend their bias correction to deal with the additional Nickell (1981)-type bias induced by the violation of the strict exogeneity assumption, similar to the bias correction of Fernández-Val and Weidner (2016).

We make the following assumptions, with assumptions (iii), (iv), (vi), and (vii) taken directly from Chernozhukov, Fernández-Val, and Weidner (2020):

Assumption 1. Sampling and Panel Model Conditions.

(i) Sampling: Conditional on the unobserved effects, $\{(y_i^T, x_i^T) \colon i \in \{1, \ldots, N\}\}$ is independent over i, where $y_i^T := \{y_{it} \colon t \in \mathcal{D}_i\}$ and $x_i^T := \{x_{it} \colon t \in \mathcal{D}_i\}$. For each $i \in \{1, \ldots, N\}$, $\{(y_{it}, x_{it}) \colon t \in \mathcal{D}_i\}$ is a strong mixing array with mixing coefficients satisfying $\sup_{i \in \{1,\ldots,N\}} \; a_i(h) = O(h^{-\mu})$ with $\nu > 0$ and $\mu > 4(8+\nu)/\nu$ as $h \to \infty$, where

$$a_i(h) := \sup_{s \in \{t_i, \ldots, T_i\}} \; \sup_{A \in \mathcal{A}_s^i, B \in \mathcal{B}_{s+h}^i} |P(A \cap B) - P(A)P(B)|$$

with $\mathcal{A}_t^i := \sigma((y_{is}, x_{is}) \colon s \in \{t_i, \ldots, t\})$ and $\mathcal{B}_t^i := \sigma((y_{is}, x_{is}) \colon s \in \{t, \ldots, T_i\})$.

(ii) Model: For all $y \in \mathcal{Y}$ and for $x_i^t := \{x_{is} \colon s \in \{t_i, \ldots, t\}\}$,

$$F_{y_{it}}(y \mid x_{it}, \alpha_i, \gamma_t) = P(y_{it} \leq y \mid x_{it}, \alpha_i, \gamma_t) = P(\epsilon_{it} < \pi_{it}(y) \mid x_{it}, \alpha_i, \gamma_t) = \Lambda(\pi_{it}(y)) \, ,$$

$\theta(y) := (\beta(y), \upsilon(y), \varphi(y))$, $\epsilon_{it} \mid x_i^t, \alpha, \gamma \sim \Lambda$, and $y \mapsto \theta(y)$ is a measurable vector-valued function.

(iii) Compactness: x_{it} has compact support χ. $\upsilon(y)$ and $\varphi(y)$ are uniformly bounded over i, t, N, T and $\mathcal{Y}$.

(iv) Compactness and smoothness: $\mathcal{Y}$ is either a discrete finite set or a bounded interval on $\mathbb{R}$. For the bounded interval case, we additionally assume that the conditional density function $f_{y_{it}}(y \mid x_{it}, \alpha_i, \gamma_t)$ exists, is uniformly bounded above and away from zero, and is uniformly continuous in y on the interior of $\mathcal{Y}$, uniformly over the support of $(x_{it}, \alpha_i, \gamma_t)$.

(v) Missing data: For each $i \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$, there is a fixed number of missing observations such that $\max_{i \in \{1,\ldots,N\}} \; (T - |\mathcal{D}_i|) \leq c_2$ and $\max_{t \in \{1,\ldots,T\}} \; (N - |\mathcal{D}_t|) \leq c_2$, where c_2 is a finite constant independent of the sample size. Conditional on (x_i^t, α, γ), the outcome y_{it} is independent of the attrition process.

(vi) Non-collinearity: There exists a constant $c_3 > 0$, independent of the sample size, such that

$$\min_{\{\delta \in \mathbb{R}^{\dim(x)} \colon \, \|\delta\|=1\}} \; \min_{(a,b) \in \mathbb{R}^{N+T}} \left[\frac{1}{n} \sum_{i=1}^{N} \sum_{t=t_i}^{T_i} (x_{it}'\delta - a_i - b_t)^2 \right] \geq c_3 \, ,$$

i. e. x_{it} has sufficient variation after projecting out individual and time fixed effects.

(vii) Asymptotics: We consider asymptotic sequences where $N_n, T_n \to \infty$ with $N_n/T_n \to c$ and $0 < c < \infty$ as $n \to \infty$.

Remark 1. (Differences in Model Assumptions). To emphasize that our extension is primarily relevant to panel data applications, we use the usual i and t indexes to refer to individual and time periods. We additionally assume that each individual time series is consecutive. (i) We allow for weak temporal dependence instead of conditional independence over i and t. As in Fernández-Val and Weidner (2016), we do so by imposing a strong mixing condition to deal with the weak temporal dependence and achieve proper bounds. (ii) We only require $\epsilon_{it} \mid x_i^t, \alpha, \gamma \sim \Lambda$ instead of $\epsilon_{it} \mid x_i^T, \alpha, \gamma \sim \Lambda$. Thus, we relax the strong exogeneity assumption and allow for weakly exogenous regressors, like predetermined outcome variables. (v) We add the conditional missing at random assumption.

We show how the limiting distributions of the estimators for the model coefficients and the marginal distribution functions, as defined by Theorem 1 in Chernozhukov, Fernández-Val, and Weidner (2020), are affected by weakly exogenous regressors. For this we need to introduce some additional notation. We denote $\Lambda_{it}^{(q)}(y)$ and $\Lambda_{it,k}^{(q)}(y)$ as the q-th order partial derivatives of the logistic CDF evaluated at $\pi_{it}(y)$ and $\pi_{it,k}(y)$, respectively. Further, we define the projections

$$\mathbb{P}\, v \in \underset{p \in \{a_i + b_t :\, a \in \mathbb{R}^N,\, b \in \mathbb{R}^T\}}{\arg\min} \sum_{i=1}^{N} \sum_{t=t_i}^{T_i} \Lambda_{it}^{(1)} (v_{it} - p_{it})^2$$

and $\mathbb{M}\, v := v - \mathbb{P}\, v$, where v is an arbitrary n-dimensional vector. For matrices, we apply the projections column-wise. We refer to $\tilde{x}_{it}, \tilde{x}_{it,k}, \Psi_{it,k}$, and $\widetilde{\Psi}_{it,k}$ as the it-th row of $\mathbb{M}\, X$, $\mathbb{M}\, X_k$, $\mathbb{P}\, f_k$, and $\mathbb{M}\, f_k$, respectively, where $X := (x_{it})_{(i,t) \in \mathcal{D}}$, $X_k := (x_{it,k})_{(i,t) \in \mathcal{D}}$, and $f_k := (\Lambda_{it,k}^{(1)} / \Lambda_{it}^{(1)})_{(i,t) \in \mathcal{D}}$. The theorem itself remains unchanged, except for the model assumptions, but we need to adjust the definition of the bias terms that result from the inclusion of individual fixed effects. In particular, we redefine $B^{(\beta)}(y) := B_1^{(\beta)}(y) + B_2^{(\beta)}(y)$ and $B^{(\Lambda)}(y) := B_1^{(\Lambda)}(y) + B_2^{(\Lambda)}(y)$, where

$$\begin{aligned}
W(y) &:= \frac{1}{n} \sum_{i=1}^{N} \sum_{t=t_i}^{T_i} \Lambda_{it}^{(1)} \tilde{x}_{it}(y) \tilde{x}_{it}'(y)\,, \\
B_1^{(\beta)}(y) &:= -\frac{1}{2N} W^{-1}(y) \sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i} \Lambda_{it}^{(2)}(y) \tilde{x}_{it}(y)}{\sum_{t=t_i}^{T_i} \Lambda_{it}^{(1)}(y)}\,, \\
B_2^{(\beta)}(y) &:= -\frac{1}{N} W^{-1}(y) \sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i - 1} \sum_{s=t+1}^{T_i} (1\{y_{it} \le y\} - \Lambda_{it}(y)) \Lambda_{is}^{(1)}(y) \tilde{x}_{is}(y)}{\sum_{t=t_i}^{T_i} \Lambda_{it}^{(1)}(y)}\,, \\
B_1^{(\Lambda)}(y) &:= \frac{1}{2N} \sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i} \Lambda_{it,k}^{(2)}(y) - \Lambda_{it}^{(2)}(y) \Psi_{it,k}(y)}{\sum_{t=t_i}^{T_i} \Lambda_{it}^{(1)}(y)}\,, \quad \text{and} \\
B_2^{(\Lambda)}(y) &:= \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i - 1} \sum_{s=t+1}^{T_i} (1\{y_{it} \le y\} - \Lambda_{it}(y)) \Lambda_{is}^{(1)}(y) \widetilde{\Psi}_{is,k}(y)}{\sum_{t=t_i}^{T_i} \Lambda_{it}^{(1)}(y)}\,.
\end{aligned}$$

The expressions $W(y)$, $B_1^{(\beta)}(y)$, and $B_1^{(\Lambda)}(y)$ have been derived by Chernozhukov, Fernández-Val, and Weidner (2020), with different notation. The other expressions $B_2^{(\beta)}(y)$ and $B_2^{(\lambda)}(y)$ are Nickell (1981)-type biases, similar to those derived by Fernández-Val and Weidner (2016).

We construct plug-in estimators for the bias terms $B^{(\beta)}(y)$ and $B^{(\Lambda)}(y)$ from the corresponding expressions above. For clarification, by plug-in estimator we mean the corresponding quantity evaluated at the fixed effects estimators instead of the true parameter values. For instance, we denote $\hat{\pi}_{it}(y) := x_{it}' \hat{\beta}(y) + \widehat{\alpha_i \upsilon}(y) + \widehat{\gamma_t \varphi}(y)$ as the plug-in estimator for $\pi_{it}(y)$. The estimators for the bias terms are $\widehat{B}^{(\beta)}(y) := \widehat{B}_1^{(\beta)}(y) + \widehat{B}_2^{(\beta)}(y)$ and

$\widehat{B}^{(\Lambda)}(y) := \widehat{B}_1^{(\Lambda)}(y) + \widehat{B}_2^{(\Lambda)}(y)$, where

$$\begin{aligned}
\widehat{W}(y) &:= \frac{1}{n}\sum_{i=1}^{N}\sum_{t=t_i}^{T_i} \widehat{\Lambda}_{it}^{(1)} \hat{\hat{x}}_{it}(y)\hat{\hat{x}}'_{it}(y)\,, \\
\widehat{B}_1^{(\beta)}(y) &:= -\frac{1}{2N}\widehat{W}^{-1}(y)\sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i} \widehat{\Lambda}_{it}^{(2)}(y)\hat{\hat{x}}_{it}(y)}{\sum_{t=t_i}^{T_i} \widehat{\Lambda}_{it}^{(1)}(y)}\,, \\
\widehat{B}_2^{(\beta)}(y) &:= -\frac{1}{N}\widehat{W}^{-1}(y)\sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i-1}\sum_{s=t+1}^{\min(t+L,T_i)} w_i(t,s)(1\{y_{it} \leq y\} - \widehat{\Lambda}_{it}(y))\widehat{\Lambda}_{is}^{(1)}(y)\hat{\hat{x}}_{is}(y)}{\sum_{t=t_i}^{T_i} \widehat{\Lambda}_{it}^{(1)}(y)} \\
\widehat{B}_1^{(\Lambda)}(y) &:= \frac{1}{2N}\sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i} \widehat{\Lambda}_{it,k}^{(2)}(y) - \widehat{\Lambda}_{it}^{(2)}(y)\widehat{\Psi}_{it,k}(y)}{\sum_{t=t_i}^{T_i} \widehat{\Lambda}_{it}^{(1)}(y)}\,, \\
\widehat{B}_2^{(\Lambda)}(y) &:= \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{t=t_i}^{T_i-1}\sum_{s=t+1}^{\min(t+L,T_i)} w_i(t,s)(1\{y_{it} \leq y\} - \widehat{\Lambda}_{it}(y))\widehat{\Lambda}_{is}^{(1)}(y)\widehat{\widehat{\Psi}}_{is,k}(y)}{\sum_{t=t_i}^{T_i} \widehat{\Lambda}_{it}^{(1)}(y)}\,,
\end{aligned}$$

L is a bandwidth parameter for the truncation kernel (Hahn and Kuersteiner 2007) in $\widehat{B}_2^{(\beta)}(y)$ and $\widehat{B}_2^{(\Lambda)}(y)$, satisfying $L \longrightarrow \infty$ and $\max_{i\in\{1,\ldots,N\}} L/|\mathcal{D}_i| \longrightarrow 0$, and $w_i(t,s) := |\mathcal{D}_i|/(|\mathcal{D}_i| - s + t)$ is a finite sample adjustment suggested by Fernández-Val and Weidner (2016). Note that if $L = 0$, the estimators are the same as those in Chernozhukov, Fernández-Val, and Weidner (2020).

After removing the additional bias from the asymptotic distribution, uniform confidence bands can be constructed using the multiplier bootstrap procedure presented in Chernozhukov, Fernández-Val, and Weidner (2020).[4][5] Details are omitted for brevity.

3.3 Simulation Experiments

We analyze the finite sample performance of our inference method for weakly exogenous regressors through simulation experiments. Our data generating process is inspired by the dynamic process of Fernández-Val and Weidner (2016).[6] However, contrary to them, we violate the strict exogeneity assumption by a more general form of feedback from past realizations of the outcome variables to the regressors. The design is meant to mimic our application in section 3.4, where we do not want to make the strict exogeneity assumption as we expect some of the regressors to be at least partly affected by past outcomes.

[4] Intuitively, Chernozhukov, Fernández-Val, and Weidner (2020) use their bootstrap procedure to obtain multiple testing adjusted critical values for the confidence bands.

[5] The multiplier bootstrap in Chernozhukov, Fernández-Val, and Weidner (2020) is equal to the “score-based” wild bootstrap described in Kline and Santos (2012).

[6] Fernández-Val and Weidner (2016) use a similar data generation process for a dynamic probit model with individual and time fixed effects, where the strict exogeneity assumption is violated by the presence of a predetermined outcome variable as a regressor. Their additional regressor is continuous and strictly exogeneous.

We generate the outcome variable by the following left- and right-censored data generating process:

$$\begin{aligned}
y_{it} &= \min\{\max\{x_{it} - d_{it} + \alpha_i + \gamma_t + \epsilon_{it},\ y_{\text{lower}}\},\ y_{\text{upper}}\}\,, \\
x_{it} &= 0.5\,(y_{it-1} - x_{it-1}) + \alpha_i + \gamma_t + \mu_{it}\,, \\
d_{it} &= \mathbf{1}\{0.5\,(d_{it-1} - y_{it-1}) + \alpha_i + \gamma_t + \nu_{it} > 0\}\,, \\
y_{i0} &= \min\{\max\{x_{i0} - d_{i0} + \alpha_i + \gamma_0 + \epsilon_{i0},\ y_{\text{lower}}\},\ y_{\text{upper}}\}\,, \\
x_{i0} &= \mu_{i0}\,,\ d_{i0} = \mathbf{1}\{\nu_{i0} > 0\}\,, \\
(i &= 1, \dots, N,\ t = 1, \dots, T)\,,
\end{aligned}$$

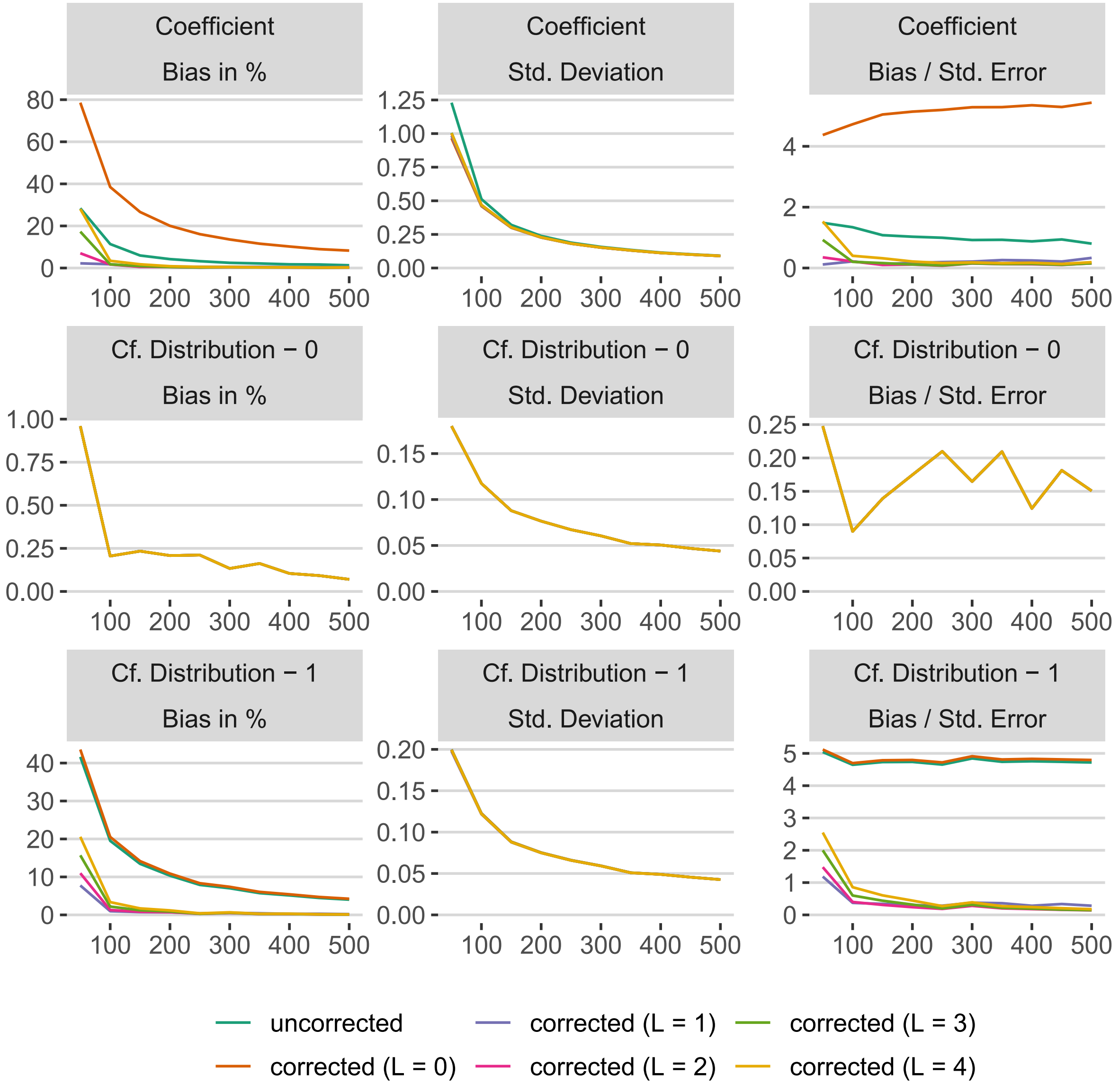
where y_{lower} and y_{upper} are lower and upper censoring thresholds, $\epsilon_{it} := \log(u_{it}/(1 - u_{it}))$, $u_{it} \sim$ iid. $\mathcal{U}(0, 1)$, $\alpha_i, \gamma_t \sim$ iid. $\mathcal{N}(0, 1/16)$, and $\mu_{it}, \nu_{it} \sim$ iid. $\mathcal{N}(0, 0.5)$. We set $y_{\text{lower}} = -4$ and $y_{\text{upper}} = 2$, which roughly corresponds to the 10% and 90% percentiles of the uncensored outcome. We consider panels with more individuals than time periods and let N and T grow at a fixed rate, i. e. $N(T) := 5T$ with $T \in \{10, 20, \dots, 90, 100\}$. For all sizes, we use a fixed grid of 120 equidistant points $\mathcal{Y} = \{-4, -3.95, \dots, 1.9, 1.95\}$ and report results for bias-corrected and uncorrected estimators. For the bias-corrected estimators, we consider different bandwidth parameters $L \in \{0, \dots, 4\}$ for the truncation kernel, where $L = 0$ corresponds to the bias correction of Chernozhukov, Fernández-Val, and Weidner (2020). We construct uniform confidence bands for the model coefficients $\beta(y)$ and counterfactual distributions $F_k(y)$ by the multiplier bootstrap with 1,000 replications and standard normal weights. All results are based on 1,000 repetitions.[7]

Figures 3.3.1 and 3.3.2 report absolute biases relative to the truth in percent, standard deviations, and ratios of absolute biases to standard errors for different estimators of the model coefficients and counterfactual distributions. The results are pointwise in the sense that we compute all statistics separately for each $y \in \mathcal{Y}$ and then integrate over $\mathcal{Y}$.[8] Given the asymptotic theory, we expect the biases to decrease with increasing sample size, and to do so without increasing the variance of the estimators. The ratio of biases to standard errors is particularly relevant from a practical point of view and is intended to show that even if the bias appears to be small, it is still relatively large compared to the dispersion of the estimator. Figure 3.3.1 reports the results for the continuous regressor x_{it}. The treatment levels for the counterfactual distributions are $x_{it,0} = x_{it}$ and $x_{it,1} = x_{it} + 1$. We find that the biases decrease with the sample size. In general, the biases of the estimators for the model coefficient are more severe than for the counterfactual distributions. Note that the counterfactual distribution $F_0(y)$ for the continuous regressor corresponds to the empirical distribution, which is asymptotically unbiased. All bias-corrected estimators with $L > 0$ significantly reduce the bias while maintaining the variance of the corresponding uncorrected estimator. For instance, for $N(T) = 200$, the biases of the estimators for $F_1(y)$ are close to zero, while the bias of the uncorrected estimator is still about 10%. Interestingly, the bias-corrected estimators with $L = 0$ perform worse than the uncorrected ones. We find that the biases relative to the dispersion of the estimators remain substantial and roughly constant as the sample size increases. Again, all bias-corrected estimators with $L > 0$ significantly reduce the bias. However, this statistic illustrates that even in large sample sizes where the relative bias is quite small, the bias is still apparent and invalidates inference. For instance, for $N(T) = 500$, the bias relative to the dispersion of the uncorrected estimator for $F_1(y)$ is about 20 times that of the bias-corrected estimators with $L > 0$, although

[7] We use the *alpaca* package of Stammann (2018) for the estimation of the fixed effects logit models. All computations were carried out using R version 4.0.3 (R Core Team 2021).

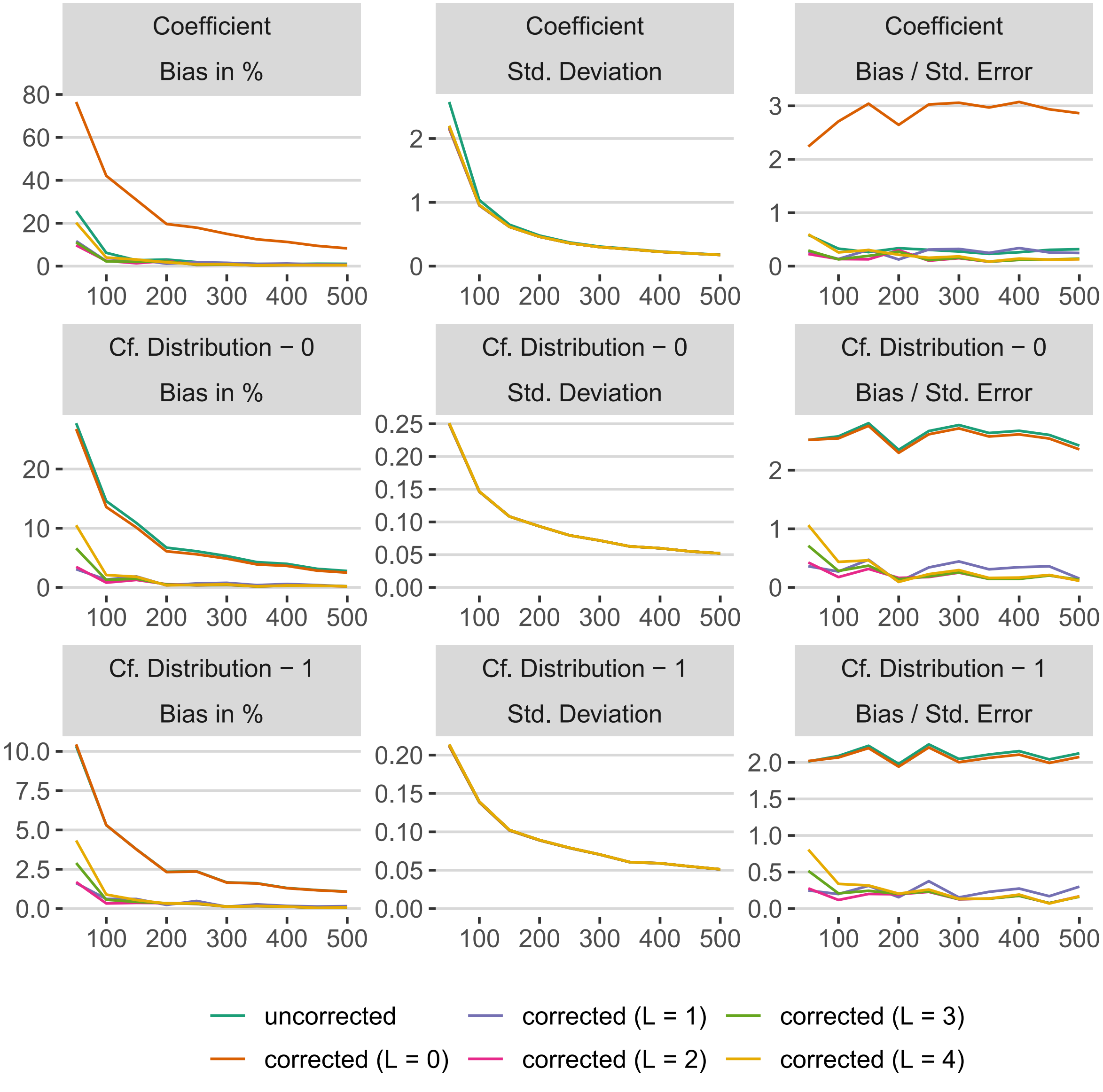
[8] We approximate the integrals using the trapezoidal rule.

Figure 3.3.1: *Bias, Standard Deviation, and Bias / Standard Error – Continuous Regressor*



Note: $N(T) = \{50, 100, \ldots, 500\}$ on the x-axis; *Coefficient* refers to the model coefficient and *Cf. Distribution - 0/1* denote the counterfactual distributions $F_0(y)$ and $F_1(y)$, where $x_{it,0} = x_{it}$ and $x_{it,1} = x_{it} + 1$; *Bias in %* refers to absolute bias relative to the truth in percent, *Std. Deviation* is the Monte-Carlo standard deviation, and *Bias / Std. Error* denotes absolute bias relative to average standard errors; *uncorrected* refers to the uncorrected estimator, *corrected (L = 0)* is the bias-corrected estimator of Chernozhukov, Fernández-Val, and Weidner (2020) for strictly exogenous regressors, and *corrected (L > 0)* are bias-corrected estimators with different bandwidths for the truncation kernel; all results are integrated over $\mathcal{Y}$ and based on 1,000 repetitions.

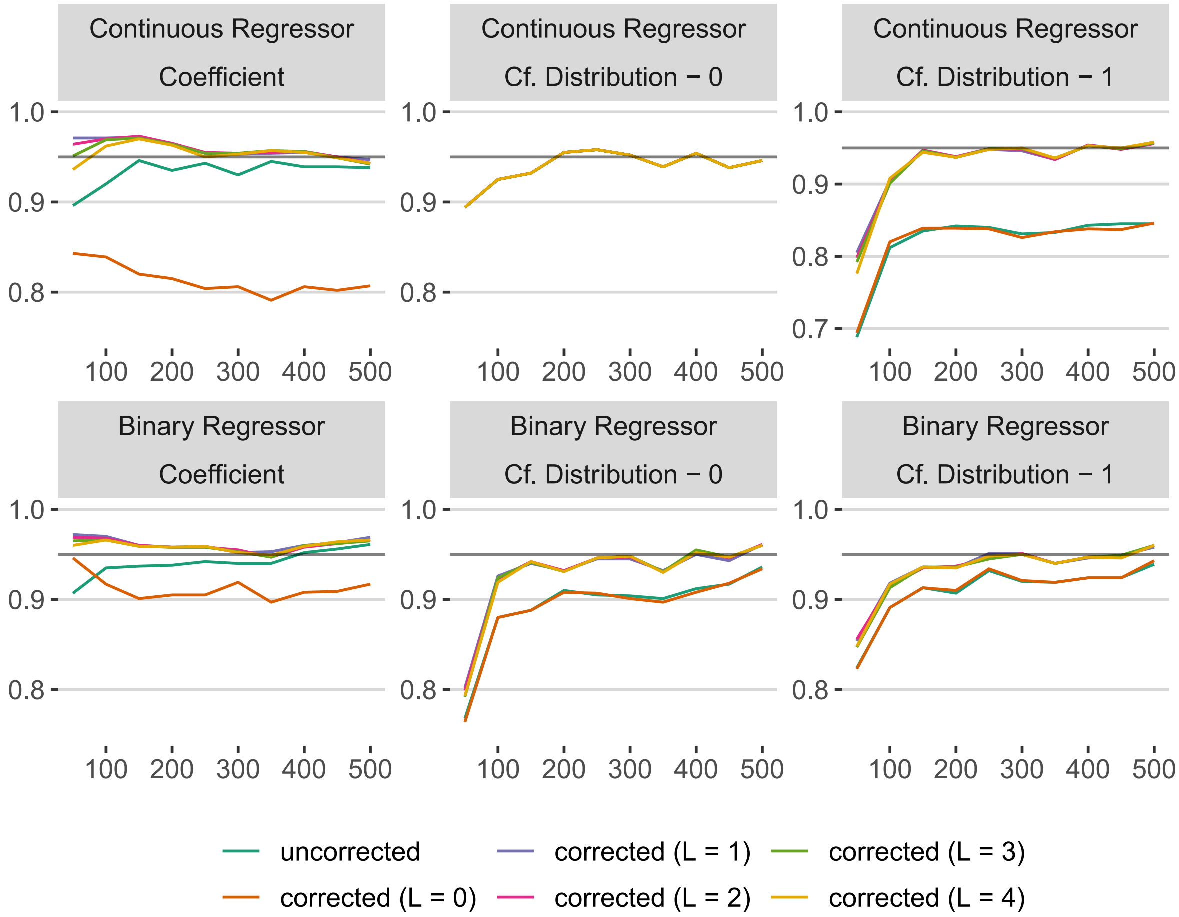
Figure 3.3.2: *Bias, Standard Deviation, and Bias / Standard Error – Binary Regressor*



Note: $N(T) = \{50, 100, \ldots, 500\}$ on the x-axis; *Coefficient* refers to the model coefficient and *Cf. Distribution - 0/1* denote the counterfactual distributions $F_0(y)$ and $F_1(y)$, where $d_{it,0} = 0$ and $d_{it,1} = 1$; *Bias in %* refers to absolute bias relative to the truth in percent, *Std. Deviation* is the Monte-Carlo standard deviation, and *Bias / Std. Error* denotes absolute bias relative to average standard errors; *uncorrected* refers to the uncorrected estimator, *corrected (L = 0)* is the bias-corrected estimator of Chernozhukov, Fernández-Val, and Weidner (2020) for strictly exogenous regressors, and *corrected (L > 0)* are bias-corrected estimators with different bandwidths for the truncation kernel; all results are integrated over $\mathcal{Y}$ and based on 1,000 repetitions.

the relative bias is less than 5%. Figure 3.3.2 reports the results for the binary regressor d_{it}. The treatment levels for the counterfactual distributions are $d_{it,0} = 0$ and $d_{it,1} = 1$. Note that unlike for the continuous regressor, both counterfactual distributions exhibit an asymptotic bias. Overall, the results are qualitatively comparable to the continuous regressor, although the biases are of a different magnitude.

Figure 3.3.3: *Coverage of 95% Uniform Confidence Bands*



Note: $N(T) = \{50, 100, \ldots, 500\}$ on the x-axis; *Coefficient* refers to the model coefficients and *Cf. Distribution - 0/1* denote the counterfactual distributions $F_0(y)$ and $F_1(y)$, where $x_{it,0} = x_{it}$ and $x_{it,1} = x_{it} + 1$ for the continuous regressor and $d_{it,0} = 0$ and $d_{it,1} = 1$ for the binary regressor; *uncorrected* refers to the uncorrected estimator, *corrected (L = 0)* is the bias-corrected estimator of Chernozhukov, Fernández-Val, and Weidner (2020) for strictly exogenous regressors, and *corrected (L > 0)* are bias-corrected estimators with different bandwidths for the truncation kernel; uniform confidence bands with 0.95 nominal level are constructed by the multiplier bootstrap with 1,000 replications and standard normal weights; all results are based on 1,000 repetitions.

Figure 3.3.3 reports coverage probabilities of uniform confidence bands with 95% nominal level for the model coefficients and the counterfactual distributions. All bias-corrected estimators with $L > 0$ have coverage probabilities close to their nominal level for $N(T) \geq 100$ and outperform the uncorrected and bias-corrected estimator with $L = 0$. One exception are the uniform confidence bands for the model coefficients. Here, the uncorrected estimators perform as well as the bias-corrected estimators with $L > 0$. Again, we find that the bias-corrected estimators with $L = 0$ often perform worse than the uncorrected ones.

Overall, we find that the bias-corrected estimators for weakly exogenous regressors perform as expected. They reduce bias substantially without affecting the variance of the estimator, and they improve inference significantly.

3.4 Empirical Illustration

We analyze how knowledge spillovers and investment in research affect firms' inventiveness. This question, among others, was recently studied by Arora, Belenzon, and Sheer (2021). Unlike other studies, they use publicly available data on publications and patents, which in turn allows us to use them for our illustration and to reassess some of their findings.[9]

Table 3.4.1 reports descriptive statistics of the model variables in the data. The inventiveness of firms' is measured by the number of patents. The variables of interest are internal use, spillout, and citations to rivals. Internal use and citations to rivals refer to the use of research, either internal or by rivals, and spillout denotes the knowledge outflow to rivals. Internal use and citations to rivals are measured as the number of citations of patents in internal and rivals' publications, respectively, and spillout is the number of citations of publications in rivals' patents.[10] The data consists of 53,110 observations of 3,807 publicly traded US manufacturing firms observed continuously for 3 to 36 years between 1980 and 2015. On average we observe 1,475 firms for 14 years. All model variables are right-skewed and, with the exception of the R&D stock, have a substantial portion of zero-valued observations. In particular, the strong skewness requires a special treatment in the estimation to ensure resistance against the influence of observations with unusually large values.[11] Fortunately, the distribution regression model provides the required resistance in a natural and convenient way.[12]

Table 3.4.1: *Descriptive Statistics of Model Variables*

	Mean	Std. Dev	Share Zeros	Percentiles			
				50%	90%	95%	99%
Patents	24.22	138.40	0.33	2.00	36.00	89.00	446.00
Publications stock	72.34	477.56	0.34	1.48	66.30	183.69	1,691.08
Internal use	1.81	23.17	0.91	0.00	0.00	3.00	38.00
Citations to rivals	0.16	0.67	0.81	0.00	0.37	0.96	2.92
Spillout	1.44	16.08	0.87	0.00	0.18	1.99	24.90
R&D stock	390.97	2,352.50	0.05	20.94	431.97	1,160.11	7,637.64

Note: Mean, standard deviation, share of zero-valued observations, and percentiles of the model variables; 53,110 consecutive firm-year observations; 3,807 firms observed for 3 to 36 years; average numbers of firms and years are 1,475 and 14, respectively.
Source: Arora, Belenzon, and Sheer (2021).

Arora, Belenzon, and Sheer (2021) predicted that the use of research, both internally or externally by rivals, should have a positive effect on firms' inventiveness. Knowledge outflows, in turn, should have no

[9] The data are part of the replication package provided by the authors.

[10] Further details on the construction of the model variables are provided in Arora, Belenzon, and Sheer (2021).

[11] Methods for this type of problem are often subsumed under "robust regressio", e. g. Huber (1973).

[12] Another common way to deal with this kind of skewness is to use logarithms. However, since we have a significant proportion of zero-valued observations that would be dropped if we used the logarithm, this strategy is not practical for our purpose.

effect. To test their predictions, we specify the following conditional distribution of Patents$_{it}$:

$$\begin{aligned}
F_{\text{patents}_{it}}(y \mid \mathbf{x}_{it}, \alpha_i, \gamma_t) &= \Lambda(\pi_{it}(y)), \quad y \in \mathcal{Y}, \\
\pi_{it}(y) &= \beta_1(y)\,\mathbf{1}\{\text{Internal use}_{it-1} > 0\} + \beta_2(y)\,\mathbf{1}\{\text{Spillout}_{it-1} > 0\} + \\
&\quad \beta_3(y)\,\mathbf{1}\{\text{Citations to rivals}_{it-1} > 0\} + \beta_4(y)\,\text{arsinh}(\text{Publications stock}_{it-1}) + \\
&\quad \beta_5(y)\,\text{arsinh}(\text{R\&D stock}_{it-1}) + \alpha_i \upsilon(y) + \gamma_t \varphi(y), \\
(i &= 1, \ldots, N, t = t_i, \ldots, T_i),
\end{aligned}$$

where i and t are firm- and year-specific indexes, $\Lambda(\cdot)$ is the logistic CDF, $\{\beta_j(y)\}_{j=1}^5$, $\upsilon(y)$, and $\varphi(y)$ are model coefficients, α_i and γ_t are unobserved firm and year effects, and $\text{arsinh}(x) := \log(x + \sqrt{x^2 + 1})$ denotes the inverse hyperbolic sine.[13] Since the model coefficients and unobserved effects can be different at the support points, we allow for the possibility that research or knowledge outflows affect only certain parts of the outcome distribution. In contrast, parametric models, such as the Poisson model, restrict research or knowledge outflows to homogeneously affect all parts of the distribution of outcomes. For instance, if research increases the inventiveness of less innovative firms, then research also increases the inventiveness of very innovative firms.

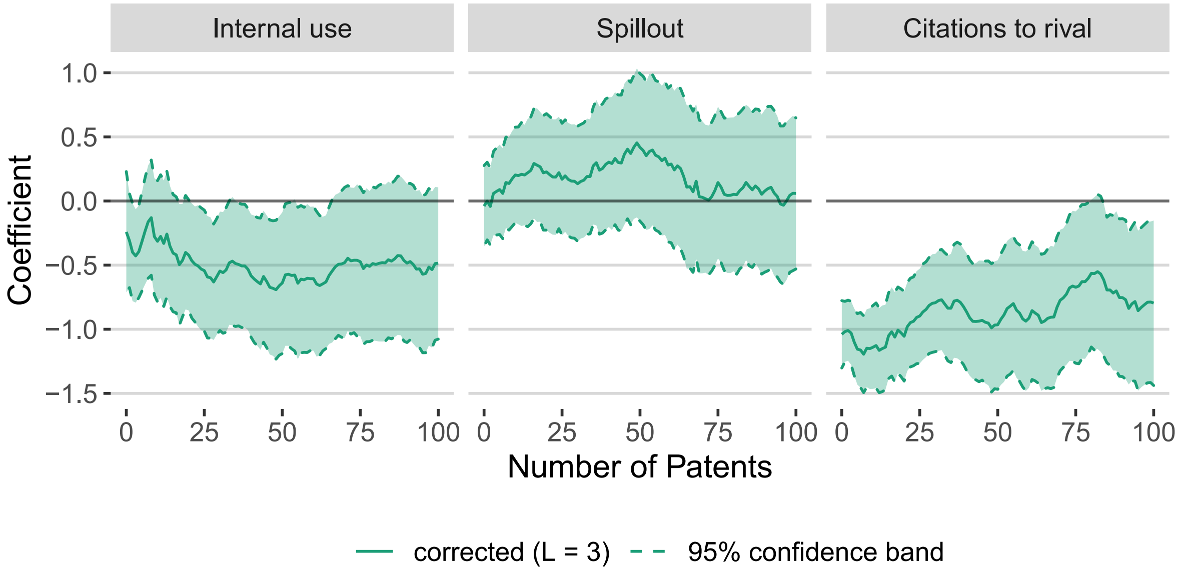
To test the predictions, it suffices to restrict our attention to $\{\beta_j(y)\}_{j=1}^3$ and $\{(F_{1,j}(y), F_{0,j}(y))\}_{j=1}^3$, where the latter are counterfactual distributions for patents at different levels of the binary treatment variables. For instance, $F_{1,1}(y)$ and $F_{0,1}(y)$ denote the counterfactual distributions for hypothetical situations in which all and no firms use internal research, respectively, and all else remains equal. The corresponding quantile effects are constructed from the counterfactual distributions. We use a grid of all possible outcomes between 0 and the 99% quantile of patents, i. e. $\mathcal{Y} \in \{0, 1, \ldots, 445, 446\}$, and allow for the possibility that some of the regressors are only weakly exogenous. For instance, the number of patents that can potentially be cited by internal and external research might depend on the number of past innovations. Further, the R&D budget could be partially predetermined by the success of past innovations. We report bias-corrected estimates of the model coefficients and quantile effects of interest. We choose $L = 3$ as bandwidth for the truncation kernel and construct uniform confidence bands by the multiplier bootstrap with 1,000 replications, clustering on firms, and standard normal weights.

Figure 3.4.1 shows estimates of the model coefficients and 95% uniform confidence bands for $y \in \{0, 1, \ldots, 99, 100\}$. Remember that although the model coefficients have no meaningful direct interpretation, they are informative about changes in the conditional quantile function due to a change in the corresponding treatment variable. Thus, they provide a first ground to test the predictions of Arora, Belenzon, and Sheer (2021). Our results support the predictions. We find that the use of research, in general, has a significant positive effect on inventiveness, while knowledge outflows have no significant effect. Interestingly, internal use of research is significant only for some parts of the region of interest, while external use by rivals is significant for almost the entire region.

Figure 3.4.2 shows quantile treatment effects of the internal and external use of research and knowledge outflows for quantiles with indexes $[0.2, 0.95]$. For comparison, we additionally report poisson estimates of the quantile treatment effects and correct the bias induced by the violation of the strict exogeneity assumption with the split-panel jackknife of Dhaene and Jochmans (2015) as suggested by Fernández-Val and Weidner

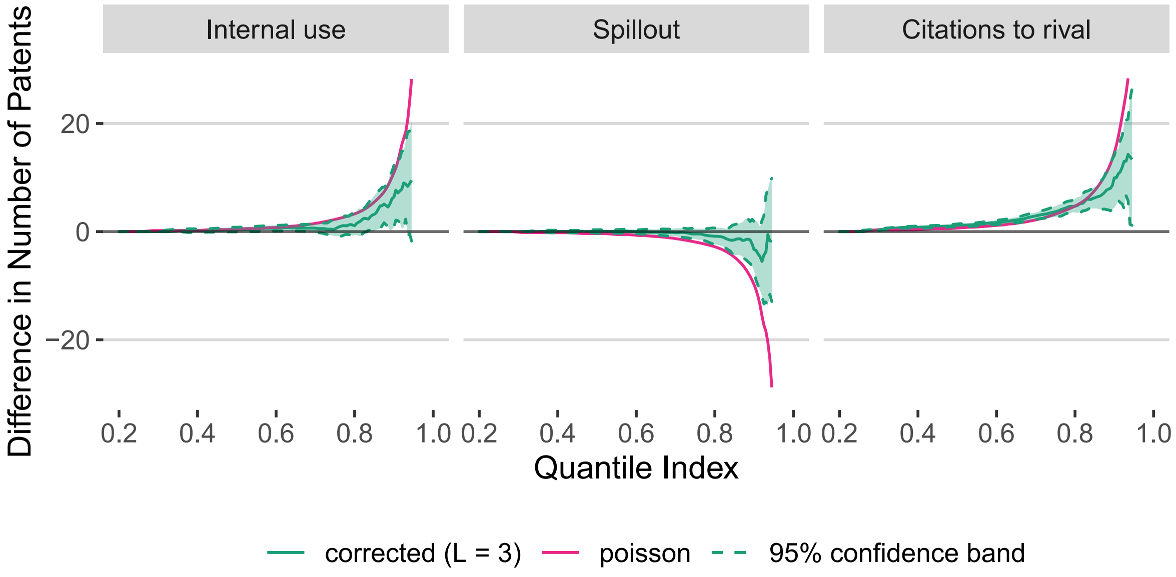
[13] Unlike the logarithm, the inverse hyperbolic sine transformation can be applied to zero-valued observations. Since the inverse hyperbolic sine behaves like $\log(2)\log(x)$ for sufficiently large values, this transformation allows us to take advantage of the logarithm without dropping observations.

Figure 3.4.1: *Estimates and 95% Uniform Confidence Bands for $\beta_1(y)$, $\beta_2(y)$, and $\beta_3(y)$*



Note: $\beta_1(y)$, $\beta_2(y)$, and $\beta_3(y)$ are the model coefficients of $\mathbf{1}\{\text{Internal use}_{it-1} > 0\}$, $\mathbf{1}\{\text{Spillout}_{it-1} > 0\}$, and $\mathbf{1}\{\text{Citations to rivals}_{it-1} > 0\}$, respectively; y on the x-axis; bias-corrected estimates with $L = 3$ and uniform confidence bands constructed by the multiplier bootstrap with 1,000 replications, clustering on firms, and standard normal weights on the y-axis.

Figure 3.4.2: *Estimates and 95% Uniform Confidence Bands for the Quantile Treatment Effects on Inventiveness*



Note: Quantile treatment effects computed from counterfactual distributions for patents at different levels of $\mathbf{1}\{\text{Internal use}_{it-1} > 0\}$, $\mathbf{1}\{\text{Spillout}_{it-1} > 0\}$, and $\mathbf{1}\{\text{Citations to rivals}_{it-1} > 0\}$, respectively; quantile indexes on the x-axis; bias-corrected estimates with $L = 3$, uniform confidence bands constructed by the multiplier bootstrap with 1,000 replications and standard normal weights, and split-panel jackknife bias-corrected poisson estimates on the y-axis.

(2016). Again, our results support the predictions of Arora, Belenzon, and Sheer (2021). We find that research has a positive and increasing effect on inventiveness. Internal use of research only affects the upper tail of the distribution, while external use by rivals affects almost the entire distribution. Further, the external use of research has a much stronger positive effect. The effect of knowledge outflows, on the other hand, is close to zero and insignificant along the entire distribution. In general, we find that firms at the upper end of the distribution benefit disproportionately more from the use of research. The poisson estimates for the quantile treatment effects of research have some substantial overlap with the uniform confidence bands. However, in particular, the estimates for the upper end of the distribution seem to overstate the effect of research. In contrast, the poisson estimates for the quantile treatment effects of knowledge outflows and those of distribution regression differ much more.

3.5 Concluding Remarks

This paper complements the work of Chernozhukov, Fernández-Val, and Weidner (2020) by allowing weakly exogenous regressors, thus widening the applicability of distribution regression for panel data applications. It would be interesting to extend the distribution regression framework to models with a panel network structure and three unobserved effects as it is standard in the literature on international trade.

References

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021. “Knowledge Spillovers and Corporate Investment in Scientific Research.” *American Economic Review* 111 (3): 871–898.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. “Inference on Counterfactual Distributions.” *Econometrica* 81 (6): 2205–2268.

Chernozhukov, Victor, Iván Fernández-Val, Blaise Melly, and Kaspar Wüthrich. 2020. “Generic inference on quantile and quantile effect functions for discrete outcomes.” *Journal of the American Statistical Association* 115 (529): 123–137.

Chernozhukov, Victor, Iván Fernández-Val, and Martin Weidner. 2020. “Network and panel quantile effects via distribution regression.” *Journal of Econometrics.*

Dhaene, Geert, and Koen Jochmans. 2015. “Split-panel Jackknife Estimation of Fixed-effect Models.” *The Review of Economic Studies* 82 (3): 991–1030.

Fernández-Val, Iván, and Martin Weidner. 2016. “Individual and time effects in nonlinear panel models with large N, T.” *Journal of Econometrics* 192 (1): 291–312.

Hahn, Jinyong, and Guido Kuersteiner. 2007. “Bandwidth Choice for Bias Estimators in Dynamic Nonlinear Panel Models.” *Working Paper.*

Huber, Peter J. 1973. “Robust Regression: Asymptotics, Conjectures and Monte Carlo.” *The Annals of Statistics* 1 (5): 799–821.

Kline, Patrick, and Andres Santos. 2012. “A Score Based Approach to Wild Bootstrap Inference.” *Journal of Econometric Methods* 1 (1): 23–41.

Koenker, Roger, and Gilbert Bassett. 1978. “Regression Quantiles.” *Econometrica* 46 (1): 33–50.

Neyman, Jerzy, and Elizabeth L. Scott. 1948. “Consistent Estimates Based on Partially Consistent Observations.” *Econometrica* 16 (1): 1–32.

Nickell, Stephen. 1981. “Biases in Dynamic Models with Fixed Effects.” *Econometrica* 49 (6): 1417–1426.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Stammann, Amrei. 2018. “Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects.” *arXiv preprint arXiv:1707.01815.*

Chapter 4

Inference in Unbalanced Panel Data Models with Interactive Fixed Effects

(joint work with Amrei Stammann)

4.1 Introduction

Economists are often concerned that unobserved heterogeneity is correlated with some variables of interest and thus leads to inconsistent estimates of the corresponding common parameters $\boldsymbol{\beta}$. If panel data is available, fixed effects models are frequently used to address this issue. One critical assumption of these models is that the unobserved heterogeneity has to be additively separable in both panel dimensions. For instance, if a panel consists of N individuals observed for T time periods, the researcher has to assume that the unobserved individual and/or time effects enter the model additively. If this is not the case, for instance because both effects are multiplicatively interacted, fixed effects models are not suitable to solve the endogeneity problem. This concern motivates so-called interactive fixed effects (IFE) estimators that model the unobserved heterogeneity as a low rank factor structure $\boldsymbol{\lambda}_i'\mathbf{f}_t$, where $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ are individual- and time-specific effects, respectively (see among others Holtz-Eakin, Newey, and Rosen 1988, Pesaran 2006, and Bai 2009).[1] Throughout this article, we refer to $\boldsymbol{\lambda}_i$ as factor loadings and $\mathbf{f}_t$ as common factors.

Inspired by Anderson and Hsiao (1982), Holtz-Eakin, Newey, and Rosen (1988) propose a quasi-differencing approach for panels with large N but small T. First they eliminate the factor loadings from the estimation equation and then estimate the remaining common factors and parameters using lagged regressors as instrumental variables. Although this estimator is consistent under fixed T asymptotics, its is well known that for large T, the number of instruments and parameters leads to biased estimates (see Newey and Smith 2004). Recently, the literature considers estimators that require N and T to be sufficiently large. Pesaran (2006) suggests a common correlated effects (CEE) estimator in the spirit of Mundlak (1978) and Chamberlain (1982, 1984), which uses cross-sectional averages of the dependent variable and the regressors to control for the unobserved common factors. Pesaran (2006)'s estimator is at least $\sqrt{N}$ consistent without the need to know the true rank of the factor structure or to impose strong factor assumptions as in Bai (2009) and Moon and Weidner (2015, 2017). However, in order to use cross-sectional averages as proxy variables for the unobserved common factors, additional parametric assumptions on the joint probability distributions of the dependent variable and the regressors are required. Bai (2009) suggests a different estimator that treats the common factors and factor loadings as additional parameters to be estimated.[2] His estimator is closely related to Bai (2003)'s principal components estimator for pure factor models and has the advantage that we do not need to make distributional assumptions about unobserved heterogeneity. Under the assumption that the true number of factors is known, Bai (2009) shows $\sqrt{NT}$ consistency irrespective of cross-sectional and/or time-serial dependence in the idiosyncratic error term. However, the presence of cross-sectional and/or time-serial dependence leads to an asymptotic bias in the limiting distribution of the estimator that can be corrected (see Bai 2009). Moon and Weidner (2017) derived an additional correction for the Nickell (1981)-type bias stemming from the inclusion of weakly exogenous regressors. Because the true number of factors is usually unknown, Moon and Weidner (2015) show that under certain conditions, and as long as the number of factors used to estimate $\boldsymbol{\beta}$ is larger than the true number, the estimator may have the same limiting distribution as Bai (2009)'s estimator but remains at least $\sqrt{\min(N,T)}$ consistent. Intuitively, there may therefore be a loss of efficiency due to the inclusion of too many irrelevant factors. However, given a consistent estimator for $\boldsymbol{\beta}$, the number of factors can be estimated using estimators for pure factor models (see

[1] Bonhomme and Manresa (2015) suggest a related but different approach. Instead of imposing rank restrictions on the time-varying unobserved heterogeneity, they use a clustering approach to assign each cross-sectional unit to a specific group where the corresponding group-specific heterogeneity is allowed to vary over time.

[2] For a detailed discussion of the different interactive fixed effects estimators we refer the reader to Bai (2009) and Moon and Weidner (2015, 2017).

among others Buja and Eyuboglu 1992, Bai and Ng 2002, Hallin and Liška 2007, Alessi, Barigozzi, and Capasso 2010, Onatski 2010, Ahn and Horenstein 2013, and Dobriban and Owen 2019). A recent comparison of some popular estimators is given in Choi and Jeong (2019). For the IFE estimator such a comparison does not exist so far.

In applied work, it is often the case that some of the observations are missing. One frequent reason is attrition. For instance, some individuals drop out of a panel because they move or leave the participating household. In some of these cases, those individuals are replaced by new survey participants. In macroeconomic panels, it also occurs that some countries are divided into several independent countries. Further, some survey designs replace individuals because of non-response. All these cases lead to very different patterns of missing data that usually, in the absence sample selection problems, do not affect the properties of the estimators (see Fernández-Val and Weidner 2018b). In the presence of missing data, the principal component estimator of Bai (2009) requires an additional data augmentation step based on the EM algorithm of Stock and Watson (1998, 2002) (see Appendix of Bai 2009 and Bai, Liao, and Yang 2015). Bai, Liao, and Yang (2015) show consistency of the EM-type estimator in simulation studies, but do not further investigate the limiting behavior of the estimator.

We make the following contributions. First, we extend the work of Bai, Liao, and Yang (2015) and analyze the limiting behavior of their suggested estimator in simulation experiments. We find that the limiting behavior of their estimator can be approximated fairly well by the inference theory of Bai (2009) and Moon and Weidner (2017), although the goodness depends on the fraction and pattern of missing data. Second, we present some algorithms that reduce the computational costs in unbalanced panels. Third, because the limiting theory of Bai (2009) and Moon and Weidner (2017) assumes that the true number of factors is known, we additionally investigate the finite sample performance of some frequently used estimators for the number of factors, i. e. Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013), and Dobriban and Owen (2019). Although we find that all estimators perform well for balanced data and different configurations of the idiosyncratic error term, their accuracy varies substantially with different fractions and patterns of missing data. Fourth, we reassess the baseline analysis of Acemoglu et al. (2019) using Bai (2009)'s IFE estimator. We qualitatively confirm Acemoglu et al. (2019)'s main results and find significant effects of democratization on growth. In their preferred specification, we estimate a long-run effect of 18%, which is pretty close to the 20% reported by the authors, but the instantaneous effect of democratization almost halves to 0.6%.

The paper is organized as follows. We introduce the model and the estimator in Section 4.2. We briefly review some estimators for the number of factors in Section 4.3. We provide results of simulation experiments in Section 4.4. We reassess Acemoglu et al. (2019) using the IFE estimator in Section 4.5. We briefly discuss the handling of endogenous regressors as in Moon and Weidner (2017) and Moon, Shum, and Weidner (2018) and consider an alternative estimator suggested by Moon and Weidner (2019) in Section 4.6. Finally, we conclude in Section 4.7.

Throughout this article, we follow conventional notation: scalars are represented in standard type, vectors and matrices in boldface, and all vectors are column vectors. Let $\mathbf{A}$ be a $N \times N$ matrix and $\mathbf{B}$ be a $N \times T$ matrix. We refer to $[\mathbf{B}]_{ij}$ as the ij-th element of $\mathbf{B}$, where i is a row index and j is a column index, and we define $\mathbb{M}_{\mathbf{B}} := \mathbf{1}_N - \mathbb{P}_{\mathbf{B}}$, where $\mathbf{1}_N$ is a $N \times N$ identity matrix, $\mathbb{P}_{\mathbf{B}} := \mathbf{B}(\mathbf{B}'\mathbf{B})^{\dagger}\mathbf{B}'$, and $(\cdot)^{\dagger}$ is the Moore-Penrose inverse.

4.2 Model, Estimation, and Inference

4.2.1 Model, Estimator, and Asymptotic Distribution

We analyze the following unobserved effects model:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\mathbf{f}_t + e_{it} \,, \tag{4.1}$$

where i and t are individual and time specific indexes, $\mathbf{x}_{it} := (x_{1,it}, \dots, x_{K,it})'$ is a vector of K regressors, $\boldsymbol{\beta}$ is the corresponding vector of common parameters, and e_{it} is an idiosyncratic error term. To allow for missing data, we introduce $\mathcal{D}$ as a subset of observed pairs of indexes, i. e. $\mathcal{D} \subseteq \{(i,t) \colon i \in \{1, \dots, N\}, t \in \{1, \dots, T\}\}$, where $n = |\mathcal{D}|$ is the sample size and N and T are the number of individuals and time periods, respectively. We assume that all observations are conditionally missing at random. The unobserved effects are expressed as a factor structure of rank $R \ll \min(N,T)$, where $\boldsymbol{\lambda}_i := (\lambda_{i1}, \dots, \lambda_{iR})'$ is a vector of factor loadings and $\mathbf{f}_t := (f_{t1}, \dots, f_{tR})'$ is a vector of common factors. Note that (4.1) collapses to the conventional fixed effects model if $\boldsymbol{\lambda}_i = (\alpha_i, 1)'$ and $\mathbf{f}_t = (1, \delta_t)'$. In contrast to the fixed effects model, the factor structure allows to capture more general patterns of heterogeneity. For instance, unobserved temporal shocks triggered by financial crises may affect each country's output differently (see Bai (2009) for some additional motivating examples).

Following Bai (2009) and Moon and Weidner (2017), we treat $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ as parameters to be estimated, and allow the common factors and loadings to be related to the regressors. To stress the similarity with the conventional fixed effects model, we refer to (4.1) as interactive fixed effects model. Given R, Moon and Weidner (2015, 2017) suggest the following IFE estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\min}\; Q(\boldsymbol{\beta}) \,,$$

where

$$Q(\boldsymbol{\beta}) := \min_{\boldsymbol{\Lambda}, \mathbf{F}} \frac{1}{n} \sum_{(i,t) \in \mathcal{D}} \left(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \boldsymbol{\lambda}_i'\mathbf{f}_t\right)^2 \tag{4.2}$$

is the profile objective function, $\boldsymbol{\Lambda} := (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N)'$ is a $N \times R$ matrix of factor loadings, and $\mathbf{F} := (\mathbf{f}_1, \dots, \mathbf{f}_T)'$ is a $T \times R$ matrix of common factors. Note that the minimizing common factors $\hat{\mathbf{f}}_t(\boldsymbol{\beta})$ and loadings $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ are not uniquely determined without imposing further restrictions. [3] Further, the objective function is globally non-convex due to the rank constraint imposed on the factor structure as pointed out by Moon and Weidner (2019).

Moon and Weidner (2017) show consistency of the IFE estimator using an asymptotic framework where $N, T \longrightarrow \infty$. In particular, their assumptions require that the true number of factors is known, weak exogeneity of the regressors, i. e. $\mathbb{E}[e_{it} \mid (\mathbf{x}_{is})_{s \leq t \in \mathcal{D}_i}, \boldsymbol{\Lambda}, \mathbf{F}] = 0$, and some additional assumptions about "low-rank" regressors that ensure that these regressors are not fully absorbed by the factor structure. Intuitively, this is comparable to the conventional fixed effects model where effects of time-invariant regressors are not identified.[4] We briefly describe the asymptotic distribution of the IFE estimator derived by Moon and Weidner

[3] However, their products are uniquely determined. We will return to this issue in the next subsection.

[4] Bai (2009) also shows consistency of the IFE estimator using different asymptotics that requires strict exogeneity of all regressors, i. e. $\mathbb{E}[e_{it} \mid (\mathbf{x}_{is})_{s \in \mathcal{D}_i}, \boldsymbol{\Lambda}, \mathbf{F}] = 0$. However, his framework is not suitable for our purposes because we consider lagged dependent variables as regressors in our empirical illustration.

(2017) and adapt their notation to unbalanced panels, following Fernández-Val and Weidner (2018b).[5] Under asymptotic sequences where $N/T \longrightarrow \kappa^2$ and $0 < \kappa < \infty$, the IFE estimator has the following limiting distribution:

$$\hat{\boldsymbol{\beta}} \overset{\text{a}}{\sim} \mathcal{N}(\boldsymbol{\beta} - \overline{N}^{-1}\mathbf{B} - \overline{T}^{-1}\mathbf{C}_1 - \overline{T}^{-1}\mathbf{C}_2, \mathbf{V}), \tag{4.3}$$

where $\overline{N} = n/T$, $\overline{T} = n/N$, $\mathbf{B}$, $\mathbf{C}_1$, and $\mathbf{C}_2$ are bias terms, and $\mathbf{V}$ is a covariance matrix. $\mathbf{C}_1$ stems from the inclusion of weakly exogenous regressors, like lagged outcome variables, and is a generalization of the Nickell (1981) bias. $\mathbf{B}$ and $\mathbf{C}_2$ arise if the idiosyncratic error term is heteroskedastic or correlated across individuals and time periods, respectively.

4.2.2 Estimation Algorithm

Moon and Weidner (2017) show that (4.2), for balanced panels, can be reformulated as

$$Q(\boldsymbol{\beta}) = \frac{1}{NT} \sum_{r=R+1}^{\min(N,T)} \mu_r\left(\mathbf{W}(\boldsymbol{\beta})'\mathbf{W}(\boldsymbol{\beta})\right), \tag{4.4}$$

where $\mathbf{W}(\boldsymbol{\beta})$ is a $N \times T$ matrix, $[\mathbf{W}(\boldsymbol{\beta})]_{it} := y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}$, and $\mu_r(\cdot)$ denotes the r-th largest eigenvalue. However, in unbalanced panels, some of the entries in $\mathbf{W}(\boldsymbol{\beta})$ are missing and we cannot simply apply the eigenvalue decomposition as in balanced panels.

We follow the suggestion of Bai (2009) and Bai, Liao, and Yang (2015) and combine (4.4) with an EM algorithm proposed by Stock and Watson (1998, 2002), i. e. we augment the missing data in the E-step and apply the eigenvalue decomposition to the complete data in the M-step. Following Bai, Liao, and Yang (2015), we can rearrange (4.1) to

$$y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} = [\mathbf{W}(\boldsymbol{\beta})]_{it} = \boldsymbol{\lambda}_i'\mathbf{f}_t + e_{it},$$

which means that, for a known $\boldsymbol{\beta}$, we can augment missing observations in $\mathbf{W}(\boldsymbol{\beta})$ with estimates of $\boldsymbol{\Lambda}$ and $\mathbf{F}$. As Bai (2009), we impose the following normalizing restrictions to uniquely determine the common factors and loadings, i. e. $\mathbf{F}'\mathbf{F}\ /T = \mathbf{1}_R$ and $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\ = \text{diag}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} \in \mathbb{R}^R$.[6] Given these restrictions, $\widehat{\mathbf{F}}(\boldsymbol{\beta})$ is equal to the first R eigenvectors of $\mathbf{W}(\boldsymbol{\beta})'\mathbf{W}(\boldsymbol{\beta})$ multiplied by $\sqrt{T}$ and $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}) = \mathbf{W}(\boldsymbol{\beta})\widehat{\mathbf{F}}(\boldsymbol{\beta})/T$.[7]

Before we describe the entire EM algorithm, we introduce some additional notation. Let

$$[\mathfrak{P}_{\mathcal{D}}(\mathbf{A})]_{it} := \begin{cases} [\mathbf{A}]_{it} & \text{if } (i,t) \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

be a projection operator that replaces missing observations of any $N \times T$ matrix $\mathbf{A}$ with zeros. Likewise, $\mathfrak{P}_{\mathcal{D}}^{\perp}(\mathbf{A})$ is the complementary operator that replaces non-missing observations with zeros.

Algorithm 1. EM algorithm

Given $\boldsymbol{\beta}$ and R, initialize $\mathbf{W}^{\perp} = \mathbf{0}_{N \times T}$ and repeat the following steps until convergence

Step 1. Set $\widetilde{\mathbf{W}}(\boldsymbol{\beta}) = \mathfrak{P}_{\mathcal{D}}(\mathbf{W}(\boldsymbol{\beta})) + \mathfrak{P}_{\mathcal{D}}^{\perp}(\mathbf{W}^{\perp})$

[5] Chen, Fernández-Val, and Weidner (2019) use the same notation for nonlinear models with interactive fixed effects.

[6] Other valid normalizing restrictions are discussed in Bai and Ng (2013).

[7] If $T > N$, it is computationally more efficient to impose $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\ /N = \mathbf{1}_R$ and $\mathbf{F}'\mathbf{F}\ = \text{diag}(\boldsymbol{\alpha})$ and estimate $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta})$ as the first R eigenvectors of $\mathbf{W}(\boldsymbol{\beta})\mathbf{W}(\boldsymbol{\beta})'$ multiplied by $\sqrt{N}$ and $\widehat{\mathbf{F}}(\boldsymbol{\beta}) = \mathbf{W}(\boldsymbol{\beta})'\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta})/N$.

Step 2. Compute $\widehat{\mathbf{F}}(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta})$ by decomposing $\widetilde{\mathbf{W}}(\boldsymbol{\beta})$

Step 3. Update $\mathbf{W}^{\perp} = \widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta})\widehat{\mathbf{F}}(\boldsymbol{\beta})'$

For a given $\boldsymbol{\beta}$ and R, we start by replacing missing observations in $\mathbf{W}(\boldsymbol{\beta})$ with zeros. We denote this augmented matrix as $\widetilde{\mathbf{W}}(\boldsymbol{\beta})$. Afterwards we estimate $\widehat{\mathbf{F}}(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta})$ from $\widetilde{\mathbf{W}}(\boldsymbol{\beta})$, replace the missing observations in $\widetilde{\mathbf{W}}(\boldsymbol{\beta})$ with $\hat{\boldsymbol{\lambda}}_i'(\boldsymbol{\beta})\,\hat{\mathbf{f}}_t(\boldsymbol{\beta})$, and repeat these two steps until convergence.

Let $\widetilde{\mathbf{W}}(\boldsymbol{\beta})$ denote the augmented matrix after convergence, a general IFE objective function in spirit of Moon and Weidner (2015, 2017) is then given by

$$\widetilde{Q}\,(\boldsymbol{\beta}) = \frac{1}{n} \sum_{r=R+1}^{\min(N,T)} \mu_r\left(\widetilde{\mathbf{W}}(\boldsymbol{\beta})'\widetilde{\mathbf{W}}(\boldsymbol{\beta})\right),$$

where in case of balanced data $\widetilde{\mathbf{W}}(\boldsymbol{\beta})$ is simply $\mathbf{W}(\boldsymbol{\beta})$. Thus, in case of missing data, we can complement the IFE estimator suggested by Moon and Weidner (2015, 2017) with an additional data augmentation step.[8]

4.2.3 Bias and Covariance Estimators

Before we describe the estimators for the biases and the covariances, we need to introduce some additional notation. Let $\mathbf{A} := (\mathbf{a}_1, \ldots, \mathbf{a}_T)'$, $\mathbf{B} := (\mathbf{b}_1, \ldots, \mathbf{b}_N)'$, and $\mathbf{z} := (z_{it})_{(i,t)\in\mathcal{D}}$ be an arbitrary n-dimensional vector. We denote $\check{\mathbf{z}}$, $\grave{\mathbf{z}}$, and $\acute{\mathbf{z}}$ as the residuals of the following least squares problems:

$$\begin{aligned} \check{z}_{it} &:= z_{it} - \hat{\boldsymbol{\lambda}}_i'\hat{\mathbf{a}}_t - \hat{\mathbf{f}}_t'\hat{\mathbf{b}}_i\,, \text{ where} \\ (\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) &\in \underset{\mathbf{A}\in\mathbb{R}^{T\times R},\,\mathbf{B}\in\mathbb{R}^{N\times R}}{\arg\min} \sum_{(i,t)\in\mathcal{D}} (z_{it} - \hat{\boldsymbol{\lambda}}_i'\mathbf{a}_t - \hat{\mathbf{f}}_t'\mathbf{b}_i)^2\,, \end{aligned} \tag{4.5}$$

$$\begin{aligned} \grave{z}_{it} &:= z_{it} - \hat{\boldsymbol{\lambda}}_i'\hat{\mathbf{a}}_t\,, \text{ where} \\ \widehat{\mathbf{A}} &\in \underset{\mathbf{A}\in\mathbb{R}^{T\times R}}{\arg\min} \sum_{(i,t)\in\mathcal{D}} (z_{it} - \hat{\boldsymbol{\lambda}}_i'\mathbf{a}_t)^2\,, \end{aligned} \tag{4.6}$$

$$\begin{aligned} \acute{z}_{it} &:= z_{it} - \hat{\mathbf{f}}_t'\hat{\mathbf{b}}_i\,, \text{ where} \\ \widehat{\mathbf{B}} &\in \underset{\mathbf{B}\in\mathbb{R}^{N\times R}}{\arg\min} \sum_{(i,t)\in\mathcal{D}} (z_{it} - \hat{\mathbf{f}}_t'\mathbf{b}_i)^2\,. \end{aligned} \tag{4.7}$$

We start with inference under the assumption that e_{it} is homoskedastic. In this case, $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased and the corresponding covariance can be estimated as $\widehat{\mathbf{V}} := \hat{\sigma}^2\widehat{\mathbf{D}}^{-1}$, where $\hat{\sigma}^2 := n^{-1}\sum_{(i,t)\in\mathcal{D}} \hat{e}_{it}^2$, $\hat{e}_{it} := y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\lambda}}_i'\hat{\mathbf{f}}_t$, and $\widehat{\mathbf{D}} := \sum_{(i,t)\in\mathcal{D}} \check{\mathbf{x}}_{it}\check{\mathbf{x}}_{it}'$. In contrast, under the assumption that e_{it} is not homoskedastic, the IFE estimator is asymptotically biased, whereas the inclusion of weakly exogenous regressors introduces an additional Nickell (1981)-type bias. Following Moon and Weidner (2015, 2017), a bias-corrected estimator for $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}} + \widehat{\mathbf{B}} + \widehat{\mathbf{C}}_1 + \widehat{\mathbf{C}}_2\,,$$

where $\widehat{\mathbf{B}} := \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{B}}^{\beta}$, $\widehat{\mathbf{C}}_1 := \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{C}}_1^{\beta}$, and $\widehat{\mathbf{C}}_2 := \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{C}}_2^{\beta}$ are estimators for the asymptotic biases described in

[8] The same augmentation step can also be embedded in the IFE estimator of Bai (2009).

subsection 4.2.1. $\widehat{\mathbf{B}}^{\beta}$, $\widehat{\mathbf{C}}_1^{\beta}$, and $\widehat{\mathbf{C}}_2^{\beta}$ are K-dimensional vectors with their k-th elements defined as

$$\begin{aligned}
\widehat{\mathbf{B}}_k^{\beta} &:= \sum_{(i,t)\in\mathcal{D}} \hat{e}_{it}^2 [\mathfrak{P}_{\mathcal{D}}(\grave{\mathbf{X}}_k)\widehat{\mathbf{\Theta}}']_{ii} \,, \\
\widehat{\mathbf{C}}_{1,k}^{\beta} &:= \sum_{i=1}^{N}\sum_{l=1}^{L}\sum_{t>l\in\mathcal{D}_i} [\mathbb{P}_{\widehat{\mathbf{F}}}]_{t,t-l}\hat{e}_{i,t-l}x_{k,it} \,, \\
\widehat{\mathbf{C}}_{2,k}^{\beta} &:= \sum_{(i,t)\in\mathcal{D}} \hat{e}_{it}^2 [\mathfrak{P}_{\mathcal{D}}(\acute{\mathbf{X}}_k)\widehat{\mathbf{\Theta}}]_{tt} + \\
&\quad\; \sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t>m\in\mathcal{D}_i} \hat{e}_{it}\hat{e}_{i,t-m}\big([\mathfrak{P}_{\mathcal{D}}(\acute{\mathbf{X}}_k)\widehat{\mathbf{\Theta}}]_{t,t-m} + [\mathfrak{P}_{\mathcal{D}}(\acute{\mathbf{X}}_k)\widehat{\mathbf{\Theta}}]_{t-m,t}\big) \,,
\end{aligned}$$

where M and L are bandwidth parameters for the truncation kernel of Newey and West (1987), $\widehat{\mathbf{\Theta}} := \widehat{\mathbf{\Lambda}}(\widehat{\mathbf{\Lambda}}'\widehat{\mathbf{\Lambda}})^{-1}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'$, and $\mathfrak{P}_{\mathcal{D}}(\acute{\mathbf{X}}_k)$ and $\mathfrak{P}_{\mathcal{D}}(\grave{\mathbf{X}}_k)$ are $N \times T$ matrices with elements denoted by $\acute{x}_{k,it}$ and $\grave{x}_{k,it}$ if $(i,t) \in \mathcal{D}$ and zero otherwise. $\widehat{\mathbf{C}}_{1,k}^{\beta}$ estimates the Nickell (1981)-type bias, the first term of $\widehat{\mathbf{C}}_{2,k}^{\beta}$ and $\widehat{\mathbf{B}}_k^{\beta}$ estimate the biases induced by individual and time-serial heteroskedasticity, respectively, and the second term in $\widehat{\mathbf{C}}_{2,k}^{\beta}$ estimates the bias stemming from time-serial correlation (see Bai 2009 remark 6 and Moon and Weidner 2015). Note that, we do not consider cross-sectional correlation in the idiosyncratic error term, which would also introduce an additional bias.[9] Suitable estimators for the covariance of $\tilde{\boldsymbol{\beta}}$ are given by

$$\begin{aligned}
\widetilde{\mathbf{V}}_j &:= \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{\Omega}}_j\widehat{\mathbf{D}}^{-1} \quad \forall\, j \in \{1,2\} \,, \\
\widehat{\mathbf{\Omega}}_1 &:= \sum_{(i,t)\in\mathcal{D}} \hat{e}_{it}^2 \breve{\mathbf{x}}_{it}\breve{\mathbf{x}}_{it}' \,, \\
\widehat{\mathbf{\Omega}}_2 &:= \sum_{i=1}^{N}\Big(\sum_{t\in\mathcal{D}_i}\hat{e}_{it}\breve{\mathbf{x}}_{it}\Big)\Big(\sum_{t\in\mathcal{D}_i}\hat{e}_{it}\breve{\mathbf{x}}_{it}\Big)' \,.
\end{aligned}$$

$\widetilde{\mathbf{V}}_1$ is a White (1980)-type heteroskedasticity robust and $\widetilde{\mathbf{V}}_2$ is a cluster-robust covariance estimator that takes into account arbitrary time-serial correlation. Alternatively, the clustered covariance estimator can be substituted by Newey and West (1987)'s estimator.

In case of balanced panels, the residuals $\breve{\mathbf{z}}$, $\grave{\mathbf{z}}$, and $\acute{\mathbf{z}}$ have straightforward expressions, i. e. $\breve{\mathbf{z}} = \operatorname{vec}(\mathbb{M}_{\widehat{\mathbf{\Lambda}}}\mathbf{Z}\,\mathbb{M}_{\widehat{\mathbf{F}}})$, $\grave{\mathbf{z}} = \operatorname{vec}(\mathbb{M}_{\widehat{\mathbf{\Lambda}}}\mathbf{Z})$, and $\acute{\mathbf{z}} = \operatorname{vec}(\mathbf{Z}\,\mathbb{M}_{\widehat{\mathbf{F}}})$, where $\mathbf{Z}$ is a $N \times T$ matrix with $[\mathbf{Z}]_{it} := z_{it}$. However, in unbalanced panels, we cannot simply augment missing observations with zeros and apply the same expressions as in balanced panels. We can still use a standard ordinary least squares estimator to obtain the residuals, but with increasing sample sizes this problem quickly becomes infeasible. For instance, suppose we have a data set consisting of $N = 200$ individuals observed for $T = 50$ time periods and consider five factors. Even in this moderate example, the rank of the sparse regressor matrix corresponding to the common factors and factor loadings is already $(N + T)R = 1{,}250$. Fortunately, we can use insights from the fixed effects literature and use sparse solvers like Halperin (1962) and Fong and Saunders (2011) to mitigate this problem (see Guimarães and Portugal 2010, Gaure 2013c, and Stammann 2018).

In the presence of missing data, we recommend to compute the residuals with the method of alternating projections (MAP, see Halperin 1962). Let $\mathcal{D}_t = \{i\colon (i,t) \in \mathcal{D}\}$ and $\mathcal{D}_i = \{t\colon (i,t) \in \mathcal{D}\}$, we define the

[9] In the presence of cross-sectional correlation, Bai (2009, in remark 7) presents a partial sample estimator. Alternatively, Bai and Liao (2017) suggest to estimate the cross-sectional correlation using an inverse covariance estimator and incorporate the corresponding weights in the objective function.

following scalar expressions:

$$[\mathbf{M}_{\hat{\lambda}_r}\mathbf{z}]_{it} := z_{it} - \hat{\lambda}_{ir}\frac{\sum_{i\in\mathcal{D}_t}\hat{\lambda}_{ir}z_{it}}{\sum_{i\in\mathcal{D}_t}\hat{\lambda}_{ir}^2} \quad \text{and} \quad [\mathbf{M}_{\hat{\mathbf{f}}_r}\mathbf{z}]_{it} := z_{it} - \hat{f}_{tr}\frac{\sum_{t\in\mathcal{D}_i}\hat{f}_{tr}z_{it}}{\sum_{t\in\mathcal{D}_i}\hat{f}_{tr}^2} .$$

The MAP algorithm can be summarized as follows:

Algorithm 2. MAP algorithm for unbalanced panels

Step 0. Initialize $\mathbf{Mz} = \mathbf{z}$.

Step 1. (If $\widehat{\mathbf{\Lambda}}$ has to be projected out e. g. in (4.5) and (4.6))
For $r = 1, \ldots, R$, set $\mathbf{Mz} = \mathbf{M}_{\hat{\lambda}_r}\mathbf{Mz}$.

Step 2. (If $\widehat{\mathbf{F}}$ has to be projected out e. g. in (4.5) and (4.7))
For $r = 1, \ldots, R$, set $\mathbf{Mz} = \mathbf{M}_{\hat{\mathbf{f}}_r}\mathbf{Mz}$.

Step 3. Repeat step 1 and/or 2 until convergence, e. g. $||\mathbf{Mz}^{\langle i\rangle} - \mathbf{Mz}^{\langle i-1\rangle}||_2 < \epsilon$, where i is the iteration number and ϵ is a tolerance parameter. After convergence $\mathbf{Mz}$ is a close approximation to $\check{\mathbf{z}}$, $\grave{\mathbf{z}}$, or $\acute{\mathbf{z}}$.

4.3 Estimating the Number of Factors

Bai (2009) and Moon and Weidner (2017) show consistency and derive the asymptotic distribution of the IFE estimator under the assumption that the number of factors is known. To avoid ambiguity, we denote the true number of factors as R^0. In practice, this assumption is often very unlikely unless economic theory provides a clear prediction about the number of factors. However, even in this case, it might be necessary to support the theoretical prediction by some empirical evidence. Therefore, we need a reliable method to estimate the number of factors.

For pure factor models, i. e. (4.1) without covariates, there is already an extensive literature on the estimation of the number of factors (see among others Buja and Eyuboglu 1992, Bai and Ng 2002, Hallin and Liška 2007, Alessi, Barigozzi, and Capasso 2010, Onatski 2010, Ahn and Horenstein 2013, and Dobriban and Owen 2019). As pointed out by Bai (2009),

$$y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}} = \boldsymbol{\lambda}_i'\mathbf{f}_t + e_{it} - \mathbf{x}_{it}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

is essentially a pure factor model. Thus given an appropriate estimator for $\boldsymbol{\beta}$, so that the error $\mathbf{x}_{it}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically negligible, we can consistently estimate the number of factors using the estimators developed for pure factor models (see Bai 2009 remark 5 and Appendix).

Bai (2009) argues, without rigorous proof, that the $\hat{\boldsymbol{\beta}}$ is $\sqrt{NT}$ consistent as long as the number of factors is at least R^0. The intuition is very similar to the inclusion of irrelevant regressors in a standard OLS regression. Including redundant common factors does not affect consistency of the IFE estimator, but its precision (see Bai 2009 remark 4). Under stronger assumptions as in Bai (2009) and Moon and Weidner (2017), Moon and Weidner (2015) confirm that the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ with $R > R^0$ is identical to the asymptotic distribution with $R = R_0$.[10] However, imposing assumptions as Bai (2009) and Moon and Weidner (2017), the authors can only show $\sqrt{\min(N, T)}$ consistency of $\hat{\boldsymbol{\beta}}$.

[10] Some of the strong assumptions imposed by Moon and Weidner (2015), like independent and identically standard normally distributed error terms, are mainly due to technical reasons. In simulation experiments the authors violate this assumption and still find support for their theoretical results.

Throughout this paper, we restrict ourselves to the estimators of Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013), and Dobriban and Owen (2019). More specifically, we apply the estimators to $\mathbf{W}(\hat{\boldsymbol{\beta}})$, where $\boldsymbol{\beta}$ is estimated with $R = \overline{R}$ and $\overline{R}$ is a known upper bound on the numbers of factors. Bai and Ng (2002) introduces various model selection criteria based on minimizing the sum of squared residuals plus some penalty function of the number of estimated parameters. Onatski (2010), Ahn and Horenstein (2013), and Dobriban and Owen (2019) segment the eigenvalue spectrum of the covariance of $\mathbf{W}(\hat{\boldsymbol{\beta}})$ to find a cutoff point between the common factors and the remaining noise stemming from the idiosyncratic error term. Onatski (2010) proposed the edge distribution (ED) estimator based on differences of consecutive eigenvalues. Ahn and Horenstein (2013) suggest to use ratios (ER) and growth rates (GR) instead of differences. Buja and Eyuboglu (1992) suggest a specific version of the parallel analysis (PA), which compares the eigenvalues to those obtained of independent data. Intuitively, the eigenvalues of independent data provide a clear cutoff to separate common factors from noise. Independent data is constructed by permuting each column of $\mathbf{W}(\hat{\boldsymbol{\beta}})$, which preserves the variances of the data but breaks the correlation pattern induced by the common factors. Recently, Dobriban (2020) provides the theoretical justification for the accuracy of PA and Dobriban and Owen (2019) propose a deflated version that improves the detection accuracy of smaller but important factors in the presence of large factors.

For unbalanced panels, we follow Gagliardini, Ossola, and Scaillet (2019) and apply the different estimators to $\mathfrak{P}_{\mathcal{D}}(\mathbf{W}(\hat{\boldsymbol{\beta}}))$ instead of $\mathbf{W}(\hat{\boldsymbol{\beta}})$.

4.4 Simulation Experiments

We study the inference drawn from the interactive fixed effects model in unbalanced panels. Remember, in contrast to balanced panels, the IFE estimator requires an additional data augmentation step to estimate the common factors and loadings (see Bai 2009 and Bai, Liao, and Yang 2015). For this purpose, we use EM algorithm proposed by Stock and Watson (1998, 2002). In the first part of our analysis, we analyze whether the inferential theory derived for the IFE estimator in balanced panels is a reasonable approximation of the estimator in unbalanced panels, given we know the true number of factors. We compare relative biases (Bias), average ratios of standard errors and standard deviations, and empirical sizes of z-tests with 5% nominal size (Size) for different patterns of randomly missing data and configurations of the idiosyncratic error term with those from a balanced panel. Because the number of factors is usually unknown, we compare different estimators for the number of factors in our second analysis. In particular, we analyze the estimators suggested by Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013), and Dobriban and Owen (2019). From the various information criteria introduced by Bai and Ng (2002), we focus on IC_2 and BIC_3, which are also used in other studies (see Onatski 2010 and Ahn and Horenstein 2013). To asses their performance, we compare the average estimated number of factors.

Following Moon and Weidner (2015), we consider a static panel data model with one regressor and two factors:

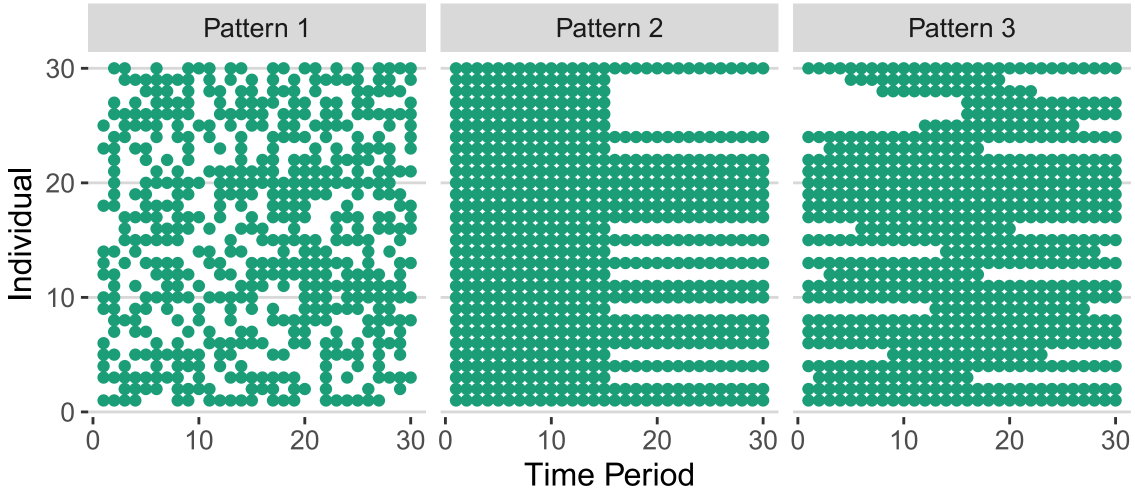
$$\begin{aligned} y_{it} &= \beta x_{it} + \sum_{r=1}^{2} \lambda_{ir} f_{tr} + e_{it} \,, \\ x_{it} &= 1 + \sum_{r=1}^{2} (\lambda_{ir} + \chi_{ir})(f_{tr} + f_{t-1,r}) + w_{it} \,, \end{aligned}$$

$i = 1, \ldots, N$, $t = 1, \ldots, T$, and e_{it} is an idiosyncratic error term. The regressor x_{it} is correlated with the common factors and loadings. Throughout all experiments, we generate $f_{tr}, w_{it} \sim$ iid. $\mathcal{N}(0,1)$ and $\lambda_{ir}, \chi_{ir} \sim$ iid. $\mathcal{N}(1,1)$.

As Bai and Ng (2002) and Ahn and Horenstein (2013), we consider four different configurations for the idiosyncratic error term: i) homoskedastic, ii) homoskedastic with fat tails, iii) cross-sectional heteroskedastic, and iv) cross-sectional heteroskedastic with time-serial correlation. More specifically, i) $e_{it} \sim$ iid. $\mathcal{N}(0,4)$, ii) $e_{it} = \sqrt{6/5}\, v_{it}$, where v_{it} has a t-distribution with five degrees of freedom, iii) $e_{it} \sim$ iid. $\mathcal{N}(0,2)$ if i is odd and $e_{it} \sim$ iid. $\mathcal{N}(0,6)$ else, and iv) $e_{it} = 0.5\, e_{it-1} + v_{it}$, where $v_{it} \sim$ iid. $\mathcal{N}(0,3/2)$ if i is odd and $v_{it} \sim$ iid. $\mathcal{N}(0,9/2)$ else. For configuration iv), we ensure that e_{it} is drawn from its stationary distribution by discarding 1,000 initial time periods. Note that the variance of the idiosyncratic error term is equal across all configurations.

We consider three different patterns where a fraction of $\psi \in \{0, 0.2, 0.4\}$ observations are missing at random. The overall sample size is equal to $NT(1-\psi)$. Figure 4.4.1 illustrates the three missing data patterns. In the first pattern, we irregularly drop $NT\psi$ observations from the entire panel data set. This pattern is also analyzed by Bai, Liao, and Yang (2015) and mimics a situation in surveys where individuals refuse or forget to answer certain questions. The other patterns are borrowed from Czarnowske and Stammann (2020) and reflect situations where individuals are replaced after they drop out from a survey or not. To describe pattern 2 and 3, we divide all individuals into two types. Type 1 consists of $N_1 = 2\psi N$ individuals that are observed for $T_1 = T/2$ time periods. The remaining $N_2 = N - N_1$ individuals are of type 2 and are observed over the entire time horizon ($T_2 = T$). Patterns 2 and 3 differ only in the point in time when the time series of a type 1 individual starts. In pattern 2, all time series start in $t = 1$, whereas in pattern 3, the initial period is chosen randomly with equal probability from $\{0, 1, \ldots, T - T_1\}$. All unbalanced data sets are generated from balanced panels by dropping observations given the corresponding missing data pattern.

Figure 4.4.1: *Patterns of Randomly Missing Observations*



We consider panel data sets of different average sizes: $\overline{N} \in \{120, 240\}$ and $\overline{T} \in \{24, 48, 96\}$, where $N = \overline{N}/(1-\psi)$ and $T = \overline{T}/(1-\psi)$. This allows us to compare the results across different fractions of missing data and check whether the conjecture of Fernández-Val and Weidner (2018b) for fixed effects estimators applies to the IFE estimator as well.[11]

[11] All results are based on 1,000 replications and summarized in tables 4.4.1–4.4.6. All computations were done on a Linux Mint

First, we analyze the finite sample properties of the IFE estimator. In configuration iii), we correct for the asymptotic bias induced by cross-sectional heteroskedasticity ($\mathbf{B}$) and use a White (1980)-type covariance estimator. In configuration iv), we additionally correct for the asymptotic bias induced by time-serial correlation ($\mathbf{C_2}$) and use a cluster robust covariance estimator. We choose the bandwidth for the estimation of $\mathbf{C_2}$ according to the rule of thumb proposed by Newey and West (1994), i. e. $M = 4(\overline{T}/100)^{2/9}$. The results are summarized in tables 4.4.1–4.4.3. For configuration i)–iii), we observe biases, ratios, and sizes that are similar to the balanced case irrespective of the fraction and pattern of missing data. Thus, the asymptotic properties of the IFE estimator for balanced data are a fairly well approximation for the unbalanced one in these configurations. This is different for configuration iv). Here we observe biases that are substantially larger compared to the balanced case. Although all ratios are close to one, these larger biases distort the nominal sizes and lead to over-rejection. Contrary to the other configurations, the various missing data patterns affect the finite sample properties of the estimator differently and a larger fraction of missing data leads to worse properties.

Second, we analyze the different estimators for the number of factors suggested by Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013), and Dobriban and Owen (2019). The initial estimator to obtain the pure factor model uses $R = \lceil 12(\min(\overline{N}, \overline{T})/100)^{1/4} \rceil$.[12][13] For ER and GR we use the mock eigenvalue proposed in Ahn and Horenstein (2013) to allow for the possibility to select zero common factors. All results are summarized in tables 4.4.4–4.4.6. First we analyze the case of balanced data. For $\overline{T} \geq 48$, all estimators have little bias. Additionally, for BIC_3, ED, and PA the biases are low irrespective of the sample size whereas ER and GR slightly underestimate the true number of factor. The findings are in line with Ahn and Horenstein (2013) for pure factor models and suggest that the error in estimating $\boldsymbol{\beta}$ is asymptotically negligible. Further, the findings support the conjecture of Moon and Weidner (2015), who expect that their main results also apply to non iid. standard normally distributed error terms. For unbalanced panels, we observe that the missing data patterns as well as the fraction of missing data affect the performance of all estimators differently. In general we find that ER and GR are more likely to underestimate, whereas the others tend to overestimate the number of factors. Further, the performance gets worth as the fraction of missing data increases. While the accuracy of the different estimators in pattern 1 is still very close to that in balanced panels, this is only partially the case in the other two patterns. Intuitively, if the missing data pattern consists of large blocks without any observations, the information used to estimate the common factors and loadings, which are used to augment the missing observations, are substantially lower and lead to noisy estimates. This explains why the performances in patterns 2 and 3, which consist of those large blocks, are relatively worse compared to pattern 1.

Finally, we briefly summarize the key findings. We find that the properties of the IFE estimator in unbalanced panels are fairly well approximated by the asymptotic theory derived by Bai (2009) and Moon and Weidner (2017). Further, the accuracy of the different estimators for the number of factors differs substantially across fractions and patterns of randomly missing data. Overall, these findings are very different from those of conventional fixed effects models where neither the fraction nor the pattern of randomly missing data affect inference, e. g. as shown by Czarnowske and Stammann (2020) for fixed effects binary choice models.

18.1 workstation using R Version 3.6.3 (R Core Team 2021).

[12] The rule of thumb was suggested by Bai and Ng (2002) in footnote 10 and traces back to Schwert (1989).

[13] This choice is different from other studies like Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013), who keep R fixed irrespective of the sample size.

Table 4.4.1: *Properties of* $\hat{\beta}$ *- Missing Data Pattern 1*

$\overline{N}$	$\overline{T}$	$\psi = 0.0$ / $\psi = 0.2$ / $\psi = 0.4$		
		Bias	Ratio	Size
		Homoskedastic		
120	24	0.07 / -0.01 / 0.14	0.91 / 0.88 / 0.92	0.07 / 0.08 / 0.08
120	48	0.05 / 0.00 / 0.03	1.00 / 0.92 / 0.95	0.05 / 0.08 / 0.05
120	96	0.06 / 0.01 / 0.04	0.99 / 1.00 / 0.97	0.05 / 0.06 / 0.06
240	24	0.06 / 0.02 / 0.04	0.91 / 0.96 / 0.90	0.07 / 0.05 / 0.08
240	48	0.03 / 0.01 / 0.02	0.96 / 0.97 / 0.98	0.05 / 0.06 / 0.05
240	96	0.01 / 0.00 / 0.01	0.98 / 0.99 / 0.99	0.05 / 0.04 / 0.05
		Homoskedastic with Fat Tails		
120	24	0.15 / 0.06 / 0.10	0.92 / 0.85 / 0.90	0.07 / 0.09 / 0.08
120	48	-0.01 / 0.03 / 0.05	0.94 / 0.97 / 0.91	0.06 / 0.07 / 0.06
120	96	0.00 / 0.02 / 0.02	0.97 / 0.88 / 0.95	0.06 / 0.06 / 0.06
240	24	0.03 / 0.03 / 0.04	0.92 / 0.89 / 0.93	0.07 / 0.06 / 0.07
240	48	-0.01 / 0.00 / 0.01	0.97 / 0.93 / 0.98	0.06 / 0.07 / 0.06
240	96	-0.01 / -0.03 / 0.00	0.97 / 0.99 / 0.98	0.05 / 0.06 / 0.06
		Cross-Sectional Heteroskedastic		
120	24	0.05 / -0.03 / 0.13	0.90 / 0.92 / 0.92	0.08 / 0.07 / 0.07
120	48	-0.01 / -0.01 / 0.00	0.97 / 0.92 / 0.96	0.06 / 0.07 / 0.06
120	96	0.01 / -0.01 / 0.05	0.95 / 0.94 / 0.96	0.07 / 0.06 / 0.06
240	24	-0.04 / 0.02 / 0.06	0.89 / 0.89 / 0.92	0.07 / 0.08 / 0.08
240	48	0.02 / 0.03 / 0.00	0.97 / 0.95 / 0.95	0.06 / 0.05 / 0.06
240	96	0.02 / -0.01 / 0.01	0.98 / 0.98 / 1.00	0.06 / 0.05 / 0.04
		Cross-Sectional Heteroskedastic with Time-Serial Correlation		
120	24	-1.19 / -1.20 / -1.17	0.87 / 0.93 / 0.97	0.14 / 0.15 / 0.17
120	48	-0.32 / -0.52 / -0.51	1.01 / 0.99 / 0.98	0.05 / 0.09 / 0.09
120	96	-0.16 / -0.24 / -0.25	0.96 / 0.97 / 0.97	0.07 / 0.06 / 0.08
240	24	-1.28 / -1.31 / -1.20	0.86 / 0.92 / 0.98	0.23 / 0.26 / 0.30
240	48	-0.37 / -0.57 / -0.56	0.96 / 0.94 / 0.97	0.08 / 0.15 / 0.17
240	96	-0.12 / -0.24 / -0.29	0.98 / 1.00 / 0.98	0.06 / 0.08 / 0.12

Note: Bias refers to relative biases in percentage, Ratio denotes the average ratios of standard errors and standard deviations, and Size is the empirical sizes of z-tests with 5% nominal size; results are based on 1,000 repetitions.

Table 4.4.2: *Properties of $\hat{\beta}$ - Missing Data Pattern 2*

$\overline{N}$	$\overline{T}$	$\psi = 0.0$ / $\psi = 0.2$ / $\psi = 0.4$		
		Bias	Ratio	Size
		Homoskedastic		
120	24	0.07 / -0.01 / 0.06	0.91 / 0.92 / 0.92	0.07 / 0.08 / 0.07
120	48	0.05 / 0.04 / 0.07	1.00 / 0.96 / 1.00	0.05 / 0.05 / 0.05
120	96	0.06 / 0.00 / 0.04	0.99 / 0.96 / 0.97	0.05 / 0.06 / 0.06
240	24	0.06 / 0.08 / 0.05	0.91 / 0.89 / 0.94	0.07 / 0.08 / 0.06
240	48	0.03 / -0.01 / 0.01	0.96 / 0.94 / 0.98	0.05 / 0.06 / 0.06
240	96	0.01 / 0.00 / 0.02	0.98 / 0.94 / 0.98	0.05 / 0.07 / 0.06
		Homoskedastic with Fat Tails		
120	24	0.15 / 0.05 / 0.12	0.92 / 0.87 / 0.88	0.07 / 0.08 / 0.07
120	48	-0.01 / 0.00 / 0.03	0.94 / 0.93 / 0.94	0.06 / 0.06 / 0.06
120	96	0.00 / -0.01 / 0.00	0.97 / 0.98 / 0.97	0.06 / 0.06 / 0.06
240	24	0.03 / 0.07 / 0.04	0.92 / 0.90 / 0.88	0.07 / 0.07 / 0.07
240	48	-0.01 / -0.01 / 0.05	0.97 / 0.91 / 0.95	0.06 / 0.06 / 0.07
240	96	-0.01 / 0.02 / 0.03	0.97 / 0.96 / 0.96	0.05 / 0.06 / 0.07
		Cross-Sectional Heteroskedastic		
120	24	0.05 / 0.06 / 0.04	0.90 / 0.90 / 0.90	0.08 / 0.07 / 0.07
120	48	-0.01 / 0.01 / 0.05	0.97 / 0.94 / 0.92	0.06 / 0.06 / 0.08
120	96	0.01 / -0.01 / 0.02	0.95 / 0.97 / 0.96	0.07 / 0.05 / 0.06
240	24	-0.04 / 0.06 / 0.01	0.89 / 0.93 / 0.89	0.07 / 0.07 / 0.08
240	48	0.02 / 0.00 / 0.00	0.97 / 0.96 / 0.95	0.06 / 0.06 / 0.06
240	96	0.02 / 0.01 / 0.01	0.98 / 0.93 / 0.97	0.06 / 0.06 / 0.06
		Cross-Sectional Heteroskedastic with Time-Serial Correlation		
120	24	-1.19 / -1.52 / -1.76	0.87 / 0.96 / 0.91	0.14 / 0.19 / 0.30
120	48	-0.32 / -0.69 / -0.82	1.01 / 0.95 / 0.95	0.05 / 0.12 / 0.18
120	96	-0.16 / -0.32 / -0.39	0.96 / 0.98 / 1.00	0.07 / 0.08 / 0.09
240	24	-1.28 / -1.61 / -1.80	0.86 / 0.85 / 0.83	0.23 / 0.37 / 0.53
240	48	-0.37 / -0.65 / -0.83	0.96 / 0.96 / 0.88	0.08 / 0.17 / 0.31
240	96	-0.12 / -0.29 / -0.38	0.98 / 1.02 / 1.00	0.06 / 0.08 / 0.16

Note: Bias refers to relative biases in %, Ratio denotes the average ratios of standard errors and standard deviations, and Size is the empirical sizes of z-tests with 5% nominal size. Results are based on 1,000 repetitions.

Table 4.4.3: *Properties of* $\hat{\beta}$ *- Missing Data Pattern 3*

$\overline{N}$	$\overline{T}$	$\psi = 0.0$ / $\psi = 0.2$ / $\psi = 0.4$		
		Bias	Ratio	Size
		Homoskedastic		
120	24	0.07 / 0.09 / 0.02	0.91 / 0.90 / 0.90	0.07 / 0.08 / 0.08
120	48	0.05 / 0.04 / 0.00	1.00 / 0.92 / 0.96	0.05 / 0.08 / 0.06
120	96	0.06 / 0.03 / 0.04	0.99 / 0.96 / 0.99	0.05 / 0.06 / 0.05
240	24	0.06 / 0.08 / 0.03	0.91 / 0.92 / 0.90	0.07 / 0.08 / 0.08
240	48	0.03 / 0.01 / 0.03	0.96 / 0.92 / 0.95	0.05 / 0.07 / 0.06
240	96	0.01 / 0.03 / 0.00	0.98 / 0.98 / 0.96	0.05 / 0.06 / 0.06
		Homoskedastic with Fat Tails		
120	24	0.15 / 0.14 / 0.07	0.92 / 0.87 / 0.88	0.07 / 0.08 / 0.08
120	48	-0.01 / 0.02 / 0.03	0.94 / 0.86 / 0.89	0.06 / 0.06 / 0.07
120	96	0.00 / 0.03 / 0.00	0.97 / 0.97 / 0.94	0.06 / 0.06 / 0.06
240	24	0.03 / 0.04 / 0.05	0.92 / 0.89 / 0.94	0.07 / 0.07 / 0.06
240	48	-0.01 / 0.01 / 0.01	0.97 / 0.95 / 0.93	0.06 / 0.07 / 0.07
240	96	-0.01 / 0.02 / 0.01	0.97 / 0.94 / 0.94	0.05 / 0.07 / 0.06
		Cross-Sectional Heteroskedastic		
120	24	0.05 / 0.13 / 0.13	0.90 / 0.91 / 0.90	0.08 / 0.07 / 0.09
120	48	-0.01 / 0.07 / 0.05	0.97 / 0.95 / 0.93	0.06 / 0.07 / 0.07
120	96	0.01 / 0.01 / 0.00	0.95 / 0.99 / 0.96	0.07 / 0.05 / 0.06
240	24	-0.04 / 0.00 / 0.01	0.89 / 0.94 / 0.92	0.07 / 0.05 / 0.07
240	48	0.02 / 0.02 / 0.02	0.97 / 0.99 / 1.00	0.06 / 0.06 / 0.05
240	96	0.02 / 0.00 / -0.01	0.98 / 0.95 / 0.98	0.06 / 0.06 / 0.06
		Cross-Sectional Heteroskedastic with Time-Serial Correlation		
120	24	-1.19 / -1.53 / -1.84	0.87 / 0.92 / 0.94	0.14 / 0.21 / 0.33
120	48	-0.32 / -0.66 / -0.87	1.01 / 0.97 / 0.94	0.05 / 0.10 / 0.21
120	96	-0.16 / -0.29 / -0.45	0.96 / 0.98 / 1.01	0.07 / 0.08 / 0.11
240	24	-1.28 / -1.62 / -1.93	0.86 / 0.94 / 0.90	0.23 / 0.37 / 0.59
240	48	-0.37 / -0.70 / -0.92	0.96 / 0.97 / 0.92	0.08 / 0.17 / 0.36
240	96	-0.12 / -0.29 / -0.45	0.98 / 0.97 / 0.97	0.06 / 0.10 / 0.21

Note: Bias refers to relative biases in %, Ratio denotes the average ratios of standard errors and standard deviations, and Size is the empirical sizes of z-tests with 5% nominal size. Results are based on 1,000 repetitions.

Table 4.4.4: *Expected Value of $\widehat{R}$ - Missing Data Pattern 1*

$\overline{N}$	$\overline{T}$	$\psi = 0.0$ / $\psi = 0.2$ / $\psi = 0.4$					
		IC$_2$	BIC$_3$	ER	GR	ED	PA
		Homoskedastic					
120	24	2.00 / 2.00 / 2.04	2.00 / 1.97 / 1.82	1.88 / 1.68 / 1.45	1.97 / 1.83 / 1.57	2.00 / 2.14 / 2.25	1.99 / 1.97 / 1.95
120	48	2.00 / 2.00 / 2.01	2.00 / 2.00 / 1.99	1.98 / 1.92 / 1.72	2.00 / 1.98 / 1.83	2.01 / 2.12 / 2.22	2.00 / 2.00 / 2.01
120	96	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	2.00 / 1.99 / 1.95	2.00 / 2.00 / 1.99	2.02 / 2.12 / 2.13	2.00 / 2.00 / 2.01
240	24	2.00 / 2.00 / 2.05	2.00 / 1.96 / 1.77	1.92 / 1.73 / 1.51	1.98 / 1.87 / 1.64	2.00 / 2.22 / 2.31	1.99 / 1.98 / 1.99
240	48	2.00 / 2.00 / 2.01	2.00 / 2.00 / 2.00	2.00 / 1.96 / 1.84	2.00 / 1.99 / 1.91	2.01 / 2.22 / 2.25	2.00 / 2.00 / 2.02
240	96	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	2.00 / 2.00 / 1.98	2.00 / 2.00 / 1.99	2.01 / 2.20 / 2.23	2.00 / 2.00 / 2.02
		Homoskedastic with Fat Tails					
120	24	2.01 / 2.00 / 2.04	2.01 / 1.98 / 1.84	1.83 / 1.69 / 1.47	1.96 / 1.83 / 1.60	2.07 / 2.13 / 2.21	1.99 / 1.97 / 1.97
120	48	2.01 / 2.00 / 2.00	2.00 / 2.00 / 2.00	1.98 / 1.89 / 1.72	2.00 / 1.97 / 1.84	2.07 / 2.14 / 2.19	2.00 / 2.00 / 2.00
120	96	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	2.00 / 2.00 / 1.96	2.00 / 2.00 / 1.99	2.10 / 2.12 / 2.14	2.00 / 2.01 / 2.01
240	24	2.01 / 2.01 / 2.04	2.00 / 1.97 / 1.76	1.88 / 1.74 / 1.52	1.97 / 1.87 / 1.65	2.06 / 2.21 / 2.31	1.99 / 1.98 / 1.99
240	48	2.01 / 2.00 / 2.01	2.00 / 2.00 / 2.00	2.00 / 1.96 / 1.82	2.00 / 1.99 / 1.90	2.08 / 2.20 / 2.27	2.00 / 2.00 / 2.01
240	96	2.01 / 2.00 / 2.00	2.00 / 2.00 / 2.00	2.00 / 2.00 / 1.99	2.00 / 2.00 / 2.00	2.08 / 2.19 / 2.24	2.00 / 2.00 / 2.02
		Cross-Sectional Heteroskedastic					
120	24	2.00 / 2.01 / 2.03	2.00 / 1.99 / 1.82	1.82 / 1.66 / 1.45	1.95 / 1.81 / 1.57	2.01 / 2.11 / 2.22	1.98 / 1.97 / 1.94
120	48	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	1.98 / 1.88 / 1.72	2.00 / 1.97 / 1.85	2.02 / 2.12 / 2.16	2.00 / 2.00 / 2.01
120	96	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	2.00 / 1.99 / 1.96	2.00 / 2.00 / 1.99	2.01 / 2.10 / 2.15	2.00 / 2.00 / 2.01
240	24	2.00 / 2.00 / 2.04	2.00 / 1.97 / 1.75	1.88 / 1.73 / 1.48	1.97 / 1.87 / 1.62	2.02 / 2.19 / 2.30	1.99 / 1.99 / 1.99
240	48	2.00 / 2.00 / 2.02	2.00 / 2.00 / 2.00	2.00 / 1.96 / 1.81	2.00 / 1.99 / 1.90	2.01 / 2.20 / 2.30	2.00 / 2.00 / 2.02
240	96	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	2.00 / 2.00 / 1.99	2.00 / 2.00 / 2.00	2.01 / 2.16 / 2.24	2.00 / 2.00 / 2.01
		Cross-Sectional Heteroskedastic with Time-Serial Correlation					
120	24	5.50 / 2.01 / 2.04	2.85 / 2.00 / 1.87	1.59 / 1.55 / 1.42	1.78 / 1.73 / 1.56	2.07 / 2.05 / 2.12	1.99 / 1.97 / 1.97
120	48	2.05 / 2.00 / 2.01	2.01 / 2.00 / 2.00	1.87 / 1.85 / 1.68	1.98 / 1.95 / 1.82	2.04 / 2.06 / 2.18	2.00 / 2.00 / 2.01
120	96	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	1.99 / 1.99 / 1.93	2.00 / 2.00 / 1.98	2.05 / 2.04 / 2.12	2.00 / 2.00 / 2.01
240	24	6.75 / 2.02 / 2.07	2.16 / 2.00 / 1.84	1.65 / 1.63 / 1.48	1.82 / 1.79 / 1.60	2.02 / 2.05 / 2.19	2.01 / 1.99 / 2.00
240	48	2.02 / 2.00 / 2.02	2.00 / 2.00 / 2.00	1.94 / 1.93 / 1.80	1.99 / 1.99 / 1.90	2.02 / 2.05 / 2.24	2.00 / 2.00 / 2.02
240	96	2.00 / 2.00 / 2.00	2.00 / 2.00 / 2.00	2.00 / 2.00 / 1.98	2.00 / 2.00 / 2.00	2.01 / 2.05 / 2.25	2.00 / 2.00 / 2.02

Note: IC$_2$ and BIC$_3$ denote the information criteria of Bai and Ng (2002), ER and GR are the estimators of Ahn and Horenstein (2013), ED is the estimator of Onatski (2010), and PA is the deflated parallel analysis suggest by Dobriban and Owen (2019). True number of factors is two. The initial estimator for β uses $R = \lceil 12(\min(\overline{N}, \overline{T})/100)^{1/4} \rceil$ factors. Results are based on 1,000 repetitions.

Table 4.4.5: *Expected Value of $\widehat{R}$ - Missing Data Pattern 2*

$\overline{N}$	$\overline{T}$	$\psi = 0.0$ / $\psi = 0.2$ / $\psi = 0.4$					
		IC$_2$	BIC$_3$	ER	GR	ED	PA
		Homoskedastic					
120	24	2.00 / 2.89 / 3.07	2.00 / 2.33 / 2.76	1.88 / 1.45 / 1.73	1.97 / 1.86 / 2.21	2.00 / 3.26 / 3.27	1.99 / 2.19 / 2.59
120	48	2.00 / 3.04 / 3.41	2.00 / 2.79 / 2.99	1.98 / 1.61 / 1.88	2.00 / 2.26 / 2.52	2.01 / 3.87 / 3.94	2.00 / 2.88 / 2.99
120	96	2.00 / 3.24 / 3.88	2.00 / 3.00 / 3.10	2.00 / 1.65 / 2.17	2.00 / 2.64 / 3.24	2.02 / 4.00 / 4.00	2.00 / 3.05 / 3.46
240	24	2.00 / 2.93 / 3.11	2.00 / 2.20 / 2.64	1.92 / 1.46 / 1.76	1.98 / 1.96 / 2.27	2.00 / 3.58 / 3.49	1.99 / 2.41 / 2.66
240	48	2.00 / 3.12 / 3.60	2.00 / 2.81 / 3.00	2.00 / 1.60 / 2.01	2.00 / 2.38 / 2.76	2.01 / 3.99 / 3.99	2.00 / 2.97 / 3.05
240	96	2.00 / 3.72 / 3.99	2.00 / 3.00 / 3.20	2.00 / 1.81 / 2.83	2.00 / 3.16 / 3.61	2.01 / 4.00 / 4.00	2.00 / 3.27 / 3.69
		Homoskedastic with Fat Tails					
120	24	2.01 / 2.90 / 3.10	2.01 / 2.38 / 2.74	1.83 / 1.43 / 1.66	1.96 / 1.84 / 2.10	2.07 / 3.16 / 3.21	1.99 / 2.23 / 2.55
120	48	2.01 / 3.05 / 3.40	2.00 / 2.83 / 2.99	1.98 / 1.50 / 1.85	2.00 / 2.19 / 2.50	2.07 / 3.80 / 3.88	2.00 / 2.89 / 2.99
120	96	2.00 / 3.25 / 3.86	2.00 / 2.99 / 3.11	2.00 / 1.66 / 2.15	2.00 / 2.57 / 3.13	2.10 / 4.07 / 4.06	2.00 / 3.05 / 3.43
240	24	2.01 / 2.95 / 3.12	2.00 / 2.23 / 2.66	1.88 / 1.44 / 1.79	1.97 / 1.97 / 2.23	2.06 / 3.47 / 3.40	1.99 / 2.43 / 2.67
240	48	2.01 / 3.15 / 3.61	2.00 / 2.80 / 3.00	2.00 / 1.56 / 1.97	2.00 / 2.33 / 2.69	2.08 / 4.03 / 4.00	2.00 / 2.99 / 3.05
240	96	2.01 / 3.74 / 3.99	2.00 / 3.00 / 3.20	2.00 / 1.79 / 2.68	2.00 / 3.01 / 3.54	2.08 / 4.08 / 4.05	2.00 / 3.27 / 3.70
		Cross-Sectional Heteroskedastic					
120	24	2.00 / 2.88 / 3.07	2.00 / 2.36 / 2.76	1.82 / 1.45 / 1.68	1.95 / 1.84 / 2.12	2.01 / 3.12 / 3.19	1.98 / 2.21 / 2.57
120	48	2.00 / 3.04 / 3.44	2.00 / 2.84 / 2.99	1.98 / 1.53 / 1.81	2.00 / 2.22 / 2.46	2.02 / 3.69 / 3.84	2.00 / 2.88 / 3.01
120	96	2.00 / 3.24 / 3.88	2.00 / 2.99 / 3.16	2.00 / 1.67 / 2.06	2.00 / 2.49 / 3.05	2.01 / 3.99 / 4.00	2.00 / 3.06 / 3.44
240	24	2.00 / 2.92 / 3.10	2.00 / 2.23 / 2.63	1.88 / 1.47 / 1.81	1.97 / 1.93 / 2.24	2.02 / 3.44 / 3.38	1.99 / 2.41 / 2.65
240	48	2.00 / 3.15 / 3.65	2.00 / 2.81 / 2.99	2.00 / 1.57 / 1.90	2.00 / 2.37 / 2.61	2.01 / 3.97 / 3.98	2.00 / 2.98 / 3.06
240	96	2.00 / 3.75 / 4.00	2.00 / 3.00 / 3.25	2.00 / 1.72 / 2.51	2.00 / 2.89 / 3.48	2.01 / 4.00 / 4.00	2.00 / 3.28 / 3.72
		Cross-Sectional Heteroskedastic with Time-Serial Correlation					
120	24	5.50 / 5.57 / 7.95	2.85 / 2.95 / 3.54	1.59 / 1.36 / 1.37	1.78 / 1.52 / 1.61	2.07 / 2.43 / 2.56	1.99 / 2.35 / 2.71
120	48	2.05 / 3.34 / 7.75	2.01 / 2.99 / 3.16	1.87 / 1.45 / 1.64	1.98 / 1.84 / 2.14	2.04 / 2.98 / 3.02	2.00 / 2.92 / 3.02
120	96	2.00 / 3.33 / 3.92	2.00 / 3.02 / 3.40	1.99 / 1.54 / 1.81	2.00 / 2.39 / 2.60	2.05 / 3.46 / 3.72	2.00 / 3.07 / 3.51
240	24	6.75 / 6.19 / 7.99	2.16 / 2.76 / 3.11	1.65 / 1.40 / 1.37	1.82 / 1.57 / 1.60	2.02 / 2.52 / 2.58	2.01 / 2.53 / 2.84
240	48	2.02 / 3.54 / 9.82	2.00 / 2.99 / 3.09	1.94 / 1.42 / 1.67	1.99 / 2.01 / 2.26	2.02 / 3.04 / 3.06	2.00 / 3.00 / 3.11
240	96	2.00 / 3.83 / 4.28	2.00 / 3.02 / 3.55	2.00 / 1.59 / 1.90	2.00 / 2.50 / 2.66	2.01 / 3.79 / 3.92	2.00 / 3.33 / 3.75

Note: IC$_2$ and BIC$_3$ denote the information criteria of Bai and Ng (2002), ER and GR are the estimators of Ahn and Horenstein (2013), ED is the estimator of Onatski (2010), and PA is the deflated parallel analysis suggest by Dobriban and Owen (2019). True number of factors is two. The initial estimator for β uses $R = \lceil 12(\min(\overline{N}, \overline{T})/100)^{1/4} \rceil$ factors. Results are based on 1,000 repetitions.

Table 4.4.6: *Expected Value of $\widehat{R}$ - Missing Data Pattern 3*

$\overline{N}$	$\overline{T}$	$\psi = 0.0$ / $\psi = 0.2$ / $\psi = 0.4$					
		IC$_2$	BIC$_3$	ER	GR	ED	PA
				Homoskedastic			
120	24	2.00 / 2.28 / 3.72	2.00 / 2.00 / 2.68	1.88 / 1.58 / 1.26	1.97 / 1.76 / 1.56	2.00 / 2.94 / 3.93	1.99 / 1.98 / 2.99
120	48	2.00 / 2.60 / 4.40	2.00 / 2.03 / 3.22	1.98 / 1.68 / 1.23	2.00 / 1.87 / 1.57	2.01 / 3.71 / 4.77	2.00 / 2.12 / 3.95
120	96	2.00 / 3.22 / 4.90	2.00 / 2.21 / 3.91	2.00 / 1.79 / 1.16	2.00 / 1.92 / 1.63	2.02 / 4.46 / 5.68	2.00 / 3.00 / 4.80
240	24	2.00 / 2.29 / 3.90	2.00 / 2.00 / 2.63	1.92 / 1.62 / 1.26	1.98 / 1.76 / 1.53	2.00 / 3.28 / 4.30	1.99 / 2.00 / 3.21
240	48	2.00 / 3.01 / 4.71	2.00 / 2.01 / 3.26	2.00 / 1.75 / 1.17	2.00 / 1.90 / 1.55	2.01 / 4.33 / 4.97	2.00 / 2.29 / 4.25
240	96	2.00 / 3.85 / 5.27	2.00 / 2.30 / 4.10	2.00 / 1.87 / 1.13	2.00 / 1.97 / 1.64	2.01 / 5.47 / 6.84	2.00 / 3.59 / 5.06
				Homoskedastic with Fat Tails			
120	24	2.01 / 2.30 / 3.74	2.01 / 2.01 / 2.69	1.83 / 1.58 / 1.26	1.96 / 1.74 / 1.53	2.07 / 2.82 / 3.81	1.99 / 1.99 / 2.98
120	48	2.01 / 2.64 / 4.41	2.00 / 2.05 / 3.23	1.98 / 1.71 / 1.21	2.00 / 1.84 / 1.58	2.07 / 3.53 / 4.63	2.00 / 2.12 / 3.94
120	96	2.00 / 3.20 / 4.92	2.00 / 2.23 / 3.91	2.00 / 1.81 / 1.17	2.00 / 1.94 / 1.60	2.10 / 4.35 / 5.40	2.00 / 2.95 / 4.85
240	24	2.01 / 2.35 / 3.92	2.00 / 2.00 / 2.59	1.88 / 1.62 / 1.24	1.97 / 1.78 / 1.53	2.06 / 3.20 / 4.27	1.99 / 2.00 / 3.21
240	48	2.01 / 2.98 / 4.75	2.00 / 2.02 / 3.27	2.00 / 1.77 / 1.21	2.00 / 1.91 / 1.60	2.08 / 4.02 / 4.98	2.00 / 2.26 / 4.24
240	96	2.01 / 3.84 / 5.29	2.00 / 2.30 / 4.12	2.00 / 1.86 / 1.13	2.00 / 1.98 / 1.54	2.08 / 5.31 / 6.58	2.00 / 3.57 / 5.08
				Cross-Sectional Heteroskedastic			
120	24	2.00 / 2.30 / 3.77	2.00 / 2.01 / 2.73	1.82 / 1.60 / 1.23	1.95 / 1.75 / 1.48	2.01 / 2.79 / 3.78	1.98 / 1.98 / 3.02
120	48	2.00 / 2.65 / 4.45	2.00 / 2.06 / 3.23	1.98 / 1.70 / 1.22	2.00 / 1.86 / 1.58	2.02 / 3.47 / 4.67	2.00 / 2.11 / 3.95
120	96	2.00 / 3.27 / 4.90	2.00 / 2.28 / 3.98	2.00 / 1.79 / 1.18	2.00 / 1.93 / 1.57	2.01 / 4.06 / 5.15	2.00 / 3.03 / 4.83
240	24	2.00 / 2.30 / 3.92	2.00 / 1.99 / 2.60	1.88 / 1.64 / 1.23	1.97 / 1.77 / 1.49	2.02 / 3.13 / 4.29	1.99 / 1.99 / 3.23
240	48	2.00 / 3.00 / 4.71	2.00 / 2.02 / 3.28	2.00 / 1.74 / 1.19	2.00 / 1.90 / 1.57	2.01 / 4.01 / 4.96	2.00 / 2.28 / 4.25
240	96	2.00 / 3.83 / 5.27	2.00 / 2.31 / 4.12	2.00 / 1.87 / 1.15	2.00 / 1.96 / 1.62	2.01 / 5.11 / 6.61	2.00 / 3.59 / 5.09
				Cross-Sectional Heteroskedastic with Time-Serial Correlation			
120	24	5.50 / 3.46 / 5.14	2.85 / 2.23 / 3.03	1.59 / 1.49 / 1.23	1.78 / 1.65 / 1.43	2.07 / 2.12 / 2.68	1.99 / 2.00 / 3.13
120	48	2.05 / 2.94 / 4.72	2.01 / 2.29 / 3.57	1.87 / 1.66 / 1.19	1.98 / 1.82 / 1.45	2.04 / 2.45 / 3.55	2.00 / 2.20 / 4.05
120	96	2.00 / 3.35 / 4.98	2.00 / 2.57 / 4.11	1.99 / 1.79 / 1.14	2.00 / 1.92 / 1.44	2.05 / 3.38 / 4.60	2.00 / 3.11 / 4.87
240	24	6.75 / 3.70 / 5.68	2.16 / 2.08 / 2.91	1.65 / 1.53 / 1.19	1.82 / 1.70 / 1.36	2.02 / 2.13 / 2.77	2.01 / 2.03 / 3.40
240	48	2.02 / 3.34 / 5.05	2.00 / 2.16 / 3.56	1.94 / 1.73 / 1.14	1.99 / 1.88 / 1.40	2.02 / 2.68 / 3.83	2.00 / 2.38 / 4.34
240	96	2.00 / 3.92 / 5.40	2.00 / 2.67 / 4.26	2.00 / 1.85 / 1.11	2.00 / 1.96 / 1.48	2.01 / 3.79 / 4.94	2.00 / 3.69 / 5.14

Note: IC$_2$ and BIC$_3$ denote the information criteria of Bai and Ng (2002), ER and GR are the estimators of Ahn and Horenstein (2013), ED is the estimator of Onatski (2010), and PA is the deflated parallel analysis suggest by Dobriban and Owen (2019). True number of factors is two. The initial estimator for β uses $R = \lceil 12(\min(\overline{N}, \overline{T})/100)^{1/4} \rceil$ factors. Results are based on 1,000 repetitions.

4.5 Empirical Illustration

Whether democracy causes economic growth is a long standing and very controversial question among economists. Recently Acemoglu et al. (2019) provide evidence that democratization has a very substantial positive impact on GDP per capita.[14] Using annual data of 175 countries observed between 1960–2010, their main findings suggest a long-run effect of about 20%. The data set constructed by the authors is very well suited for our purposes as it is naturally unbalanced and covers a very long time horizon with several unobserved common shocks triggered by technological progress and financial crises. Overall the sample consists of 6,934 observations, where 3,558 are classified as democratic. From 88 different countries, 122 transit to democracy and 71 to non-democracy. The average GDPs, measured in year 2000 dollars, are 8,150 for democratic and 2,074 for non-democratic countries. 71 countries were observed over the entire time horizon such that, on average, the data set covers 136 countries and 40 years. The fraction and pattern of missing data are comparable to the setting with $\psi = 0.2$ and pattern 3 in our simulation study.

We reassess the baseline analysis of Acemoglu et al. (2019) using the IFE estimator and the following specification:

$$y_{it} = \beta\, D_{it} + \sum_{j=1}^{p} \gamma_j\, y_{it-j} + \alpha_i + \delta_t + \boldsymbol{\lambda}_i' \mathbf{f}_t + e_{it}\,,$$

where i and t are country and time specific indexes, D_{it} is an indicator for being a democracy, and y_{it} is the corresponding natural logarithm of GDP per capita. α_i and δ_t are additive fixed effects that capture time-invariant country characteristics and control for the global business cycle, respectively. Contrary to Acemoglu et al. (2019), we further decompose the error term into a factor structure $\boldsymbol{\lambda}_i'\mathbf{f}_t$ and a remaining idiosyncratic component e_{it}. This allows us to capture unobserved common shocks ($\mathbf{f}_t$), which simultaneously affect the growth and democratization of a country in different ways ($\boldsymbol{\lambda}_i$). The dynamic specification permits a distinction between short- and long-run effects of democratization, where the former is $\hat{\beta}$ and the latter can be computed as

$$\hat{\phi} := \frac{\hat{\beta}}{1 - \sum_{j=1}^{p} \hat{\gamma}_j}\,.$$

We use $p \in \{1, 2, 4\}$, where $p = 4$ is the preferred specification of Acemoglu et al. (2019). To reduce the number of parameters during the optimization, we project the country and time fixed effects out before estimating β and $\boldsymbol{\gamma}$ (see Bai 2009 section 8). The model after the projection becomes

$$\ddot{y}_{it} = \beta\, \ddot{D}_{it} + \sum_{j=1}^{p} \gamma_j\, \ddot{y}_{it-j} + \ddot{\boldsymbol{\lambda}}_i' \ddot{\mathbf{f}}_t + e_{it}\,,$$

where the two dots above denote variables after projecting out both fixed effects. In the absence of any common factors, i. e. $R = 0$, this is simply a conventional fixed effects model.

For valid inference it is important to know the true number of factors, or at least an estimate that is larger but close to the true number. Because the true number of factors is unknown, we proceed as follows: First, we estimate each specification with $R = 10$ to obtain the pure factor models $\mathfrak{P}_{\mathcal{D}}(\mathbf{W}(\hat{\beta}, \hat{\boldsymbol{\gamma}}))$, where $[\mathbf{W}(\hat{\beta}, \hat{\boldsymbol{\gamma}})]_{it} := \hat{\beta}\, \ddot{D}_{it} + \sum_{j=1}^{p} \hat{\gamma}_j\, \ddot{y}_{it-j}$.[15] Afterwards, we apply the estimators suggested by Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013), and Dobriban and Owen (2019) to estimate the number of factors.

[14] The data are part of the replication package provided by the authors.

[15] The number of factors chosen is equal to the rule-of-thumb used during the simulation experiments.

Table 4.5.1 summarizes the results. The estimates are almost identical across different specifications. Both model selection criteria of Bai and Ng (2002) predict a substantially larger number of factors compared to the other estimators, where IC_2 always predicts the upper bound. This is in line with Ahn and Horenstein (2013) who find that the information criteria are quite sensitive to the chosen upper bound on the number of factors and tend to overestimate. We partially observe the same behavior during our simulation experiments. Contrary to the model selection criteria, the estimators of Onatski (2010), Ahn and Horenstein (2013), and Dobriban and Owen (2019) all suggest one or three common factors. Additionally, figure 4.5.1 shows the singular values of the pure factor models and those for permuted versions. More precisely, we randomly shuffle each column of $\mathfrak{P}_{\mathcal{D}}(\mathbf{W}(\hat{\beta}, \hat{\boldsymbol{\gamma}}))$ and compute the maximum value of each singular value across 199 randomized samples.[16] The large gap between the first and the second common factor, explains why most of the estimators that try to decompose the eigenvalue spectrum predict one factor. However, if we compare the spectra with those of permuted data, we find that factor two and three have some additional explanatory power even if it is quite low in terms of variance explained. If we additionally consider the results of Moon and Weidner (2015), who showed that overestimating the number of factors is better than underestimating, then $R = 3$ is our preferred choice.

Table 4.5.1: *Estimated Number of Factors*

Specification	IC_2	BIC_3	ER	GR	ED	PA
$p = 1$	10	7	1	1	1	3
$p = 2$	10	8	1	1	1	3
$p = 4$	10	8	1	1	3	3

Note: IC_2 and BIC_3 denote the information criteria of Bai and Ng (2002), ER and GR are the estimators of Ahn and Horenstein (2013), ED is the estimator of Onatski (2010), and PA is the deflated parallel analysis suggest by Dobriban and Owen (2019). Estimators applied to $\mathfrak{P}_{\mathcal{D}}(\mathbf{W}(\hat{\beta}, \hat{\boldsymbol{\gamma}}))$. The initial estimator for β and $\boldsymbol{\gamma}$ uses $R = 10$.
Source: Acemoglu et al. (2019).

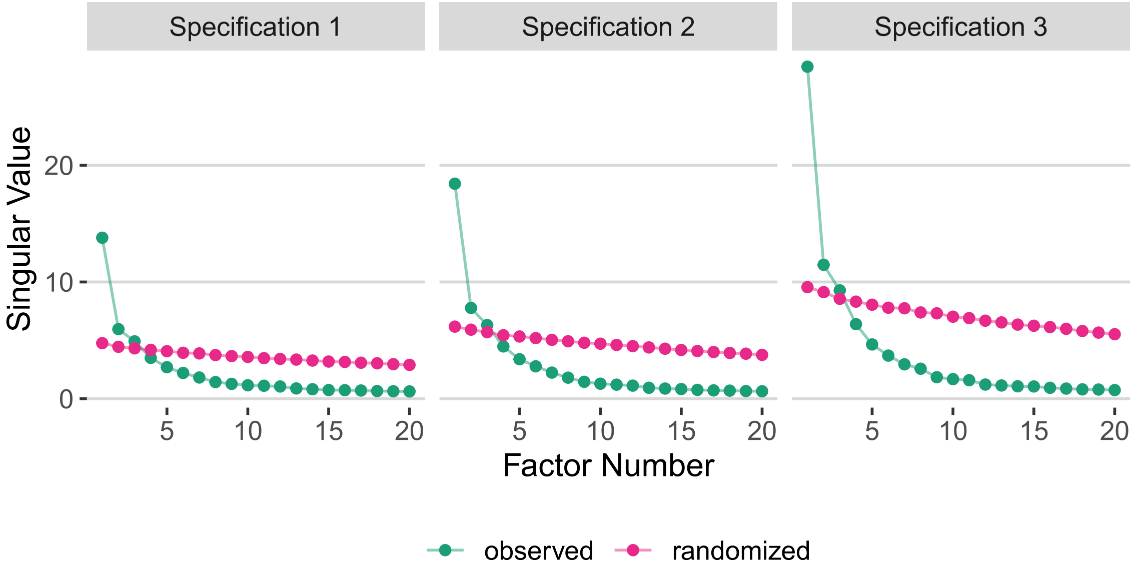
Table 4.5.2 summarizes the results of different conventional and interactive fixed effects estimators (Interactive). As Acemoglu et al. (2019), we report results for the fixed effects estimator (Within), the Arellano-Bond estimator (AB, see Arellano and Bond 1991), and the Hahn-Hausman-Kuersteiner estimator (HHK, see Hahn, Hausman, and Kuersteiner 2004). However, instead of the fixed effects estimator used by Acemoglu et al. (2019), we report results of a bias-corrected within estimator with bandwidth $L = 5$ that addresses the Nickell (1981) bias.[17] For Interactive, we report results for $R \in \{1, 2, 3\}$. To correct for the Nickell (1981)-type bias and those biases induced by cross-sectional heteroskedasticity and time-serial correlation, we use the asymptotic bias corrections proposed by Bai (2009) and Moon and Weidner (2017) with $L = 5$ and $M = 3$. Similar to Acemoglu et al. (2019), we report estimates and standard errors of the short- and long-run effects of democratization and the persistence of GDP processes. Further, all standard errors are heteroskedasticity robust and clustered at the country level to allow for arbitrary patterns of time-serial correlation.[18] We find that all estimators reveal a strong and significant persistence of GDP processes across

[16] This is essentially a graphical illustration of the parallel analysis without the deflation proposed by Dobriban and Owen (2019).

[17] This estimator was also used by Chen, Chernozhukov, and Fernández-Val (2019) for the same illustration, but on a balanced subset of the data. The authors also proposed a split-panel jackknife bias correction to reduce the many moment bias of the Arellano-Bond estimator (see Newey and Smith 2004).

[18] All covariance estimators use a degrees-of-freedom adjustment to improve their finite sample properties.

Figure 4.5.1: *Largest Singular Values in Descending Order*



Note: Singular values for permuted data are based on 199 replications; initial estimator for β and γ uses $R = 10$.
Source: Acemoglu et al. (2019).

all specifications. The coefficients of democratization obtained by Within and Interactive with $R \geq 2$ are always significant at the 5% level, whereas those of AB and HHK are only significant for $p = 4$. If we focus on $p = 4$, which is Acemoglu et al. (2019)'s preferred specification, the fixed effects models used by the authors reveal short-run effects of a transition to democracy between 0.828% and 1.178% and long-run effects between 16.448% and 29.262%. However, after controlling for additional time-varying unobserved heterogeneity, we find short- and long-run effects that are substantially lower compared to those reported by the authors. Our preferred Interactive with $R = 3$, yields short- and long-run estimates of 0.622% and 18.264%.

Next, we consider two different sensitivity checks. First, the estimation of the asymptotic biases requires different bandwidth choices. We check the sensitivity of the results by analyzing all combinations of the following bandwidth choices: $L \in \{1, \ldots, 7\}$ and $M \in \{1, \ldots, 7\}$. Second, we report estimates of Interactive for $R \in \{4, \ldots, 10\}$. As shown by Moon and Weidner (2015), the inclusion of additional redundant common factors should only affect the precision of the IFE estimator after controlling for all relevant common factors. Table 4.5.3 and 4.5.4 summarize the results. With respect to the different bandwidth choices, we find that the results of Interactive are very robust to all combinations of bandwidth choices as indicated by the narrow intervals reported in Table 4.5.3. Contrary, the estimates of Within are more sensitive. For $p = 4$, we find long-run effects between 21.571% and 32.289%. With respect to the number of factors, we find that after controlling for more than three factors, the estimated persistence of the GDP process reported in Table 4.5.4 starts declining. The same pattern was also recognized in the empirical illustration of Moon and Weidner (2015). The authors argue that the dynamic specification might be incorrectly specified in the sense that the lagged outcome variables simply capture time-serial correlation in the idiosyncratic error term instead of true state dependence. Because the factor structure also captures time-serial correlation, this might indicate that there is no true state dependence. In contrast, the coefficients of democratization become larger and remain significant in most specifications.

Table 4.5.2: *Effect of Democracy on Logarithmic GDP per Capita* (×100)

	Within	AB	HHK	Interactive		
				$R = 1$	$R = 2$	$R = 3$
			Specification 1 - $p = 1$			
Democracy	1.051	0.959	0.781	0.745	0.755	0.809
	(0.293)	(0.477)	(0.455)	(0.329)	(0.287)	(0.305)
Persistence of	0.983	0.946	0.938	0.960	0.973	0.967
GDP process	(0.006)	(0.009)	(0.011)	(0.007)	(0.006)	(0.007)
Long-run effect	60.489	17.608	12.644	18.644	27.690	24.546
of democracy	(28.073)	(10.609)	(8.282)	(9.460)	(12.250)	(11.131)
			Specification 2 - $p = 2$			
Democracy	0.671	0.797	0.582	0.488	0.554	0.559
	(0.247)	(0.417)	(0.387)	(0.292)	(0.263)	(0.279)
Persistence of	0.975	0.946	0.941	0.956	0.968	0.966
GDP process	(0.005)	(0.009)	(0.010)	(0.007)	(0.005)	(0.006)
Long-run effect	26.513	14.882	9.929	11.030	17.258	16.557
of democracy	(12.026)	(9.152)	(7.258)	(7.174)	(8.881)	(9.005)
			Specification 3 - $p = 4$			
Democracy	0.828	0.875	1.178	0.513	0.600	0.622
	(0.225)	(0.374)	(0.370)	(0.267)	(0.259)	(0.249)
Persistence of	0.972	0.947	0.953	0.958	0.964	0.966
GDP process	(0.005)	(0.009)	(0.009)	(0.006)	(0.006)	(0.005)
Long-run effect	29.262	16.448	25.032	12.226	16.749	18.264
of democracy	(10.281)	(8.436)	(10.581)	(6.780)	(7.956)	(8.030)

Note: Within, AB, HHK, and Interactive denote the bias-corrected fixed effects estimator, the Arellano-Bond estimator, the Hahn-Hausman-Kuersteiner estimator, and the IFE estimator. Standard errors in parentheses are heteroskedasticity robust and clustered at the country level. Within and Interactive use bandwidths $L = 5$ and $M = 3$ for the estimation of the asymptotic biases. The results of AB and HHK are taken from table 2 in Acemoglu et al. (2019).
Source: Acemoglu et al. (2019).

Table 4.5.3: *Sensitivity to Different Bandwidth Choices*

	Within	Interactive		
		$R = 1$	$R = 2$	$R = 3$
	Specification 1 - $p = 1$			
Democracy	[0.986; 1.057]	[0.744; 0.775]	[0.755; 0.806]	[0.806; 0.867]
Persistence of GDP process	[0.974; 0.984]	[0.958; 0.960]	[0.972; 0.973]	[0.966; 0.968]
Long-run effect of democracy	[38.127; 64.280]	[18.551; 18.872]	[27.336; 29.065]	[24.368; 25.713]
	Specification 2 - $p = 2$			
Democracy	[0.633; 0.674]	[0.488; 0.522]	[0.554; 0.609]	[0.548; 0.583]
Persistence of GDP process	[0.968; 0.976]	[0.954; 0.956]	[0.967; 0.968]	[0.966; 0.967]
Long-run effect of democracy	[19.509; 27.723]	[10.987; 11.369]	[17.231; 18.721]	[16.158; 17.150]
	Specification 3 - $p = 4$			
Democracy	[0.773; 0.849]	[0.504; 0.547]	[0.596; 0.635]	[0.620; 0.661]
Persistence of GDP process	[0.964; 0.974]	[0.957; 0.958]	[0.964; 0.964]	[0.965; 0.966]
Long-run effect of democracy	[21.571; 32.289]	[11.945; 12.672]	[16.548; 17.715]	[18.113; 19.352]

Note: Effect of democracy on logarithmic GDP per capita (×100). Within and Interactive denote the bias-corrected fixed effects estimator and the IFE estimator. The intervals denote the ranges of all estimates across different combinations of $L \in \{1, \ldots, 7\}$ and $M \in \{1, \ldots, 7\}$.
Source: Acemoglu et al. (2019).

Table 4.5.4: *Sensitivity to the Number of Factors*

	$R = 4$	$R = 5$	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$
	Specification 1 - $p = 1$						
Democracy	0.417	0.160	1.101	1.553	1.615	1.619	2.197
	(0.511)	(0.486)	(0.516)	(0.564)	(0.602)	(0.634)	(0.718)
Persistence of	0.848	0.854	0.754	0.702	0.600	0.554	0.453
GDP process	(0.018)	(0.020)	(0.030)	(0.035)	(0.034)	(0.039)	(0.040)
Long-run effect	2.749	1.095	4.476	5.220	4.037	3.635	4.019
of democracy	(3.363)	(3.339)	(2.140)	(1.939)	(1.551)	(1.479)	(1.394)
	Specification 2 - $p = 2$						
Democracy	0.388	0.868	1.178	1.424	1.905	1.150	1.418
	(0.469)	(0.446)	(0.511)	(0.565)	(0.614)	(0.651)	(0.705)
Persistence of	0.810	0.689	0.586	0.485	0.369	0.342	0.247
GDP process	(0.019)	(0.031)	(0.036)	(0.031)	(0.038)	(0.042)	(0.056)
Long-run effect	2.041	2.795	2.845	2.767	3.019	1.747	1.882
of democracy	(2.478)	(1.458)	(1.266)	(1.129)	(1.013)	(1.018)	(0.968)
	Specification 3 - $p = 4$						
Democracy	0.763	1.091	1.452	1.474	1.513	1.474	0.476
	(0.455)	(0.446)	(0.534)	(0.567)	(0.557)	(0.596)	(0.589)
Persistence of	0.628	0.593	0.380	0.213	0.174	0.187	-0.202
GDP process	(0.025)	(0.038)	(0.049)	(0.050)	(0.056)	(0.092)	(0.075)
Long-run effect	2.053	2.683	2.341	1.874	1.831	1.814	0.396
of democracy	(1.211)	(1.124)	(0.883)	(0.729)	(0.692)	(0.766)	(0.493)

Note: Effect of democracy on logarithmic GDP per capita (×100). Results obtained by the interactive fixed effects estimator for $R \in \{4, \dots, 10\}$. Standard errors in parentheses are heteroskedasticity robust and clustered at the country level. Bandwidths $L = 5$ and $M = 3$ for the estimation of the asymptotic biases.

Source: Acemoglu et al. (2019).

Finally, we consider an additional specification without predetermined regressors, i. e. $p = 0$. Again we estimate the number of factors from a pure factor model, where the initial estimate is based on $R = 10$. The estimates are identical to those of $p = 4$ and provide further support for our preferred choice of $R = 3$. The corresponding estimate of democratization is - 1.251% (standard error = 1.286%) and is in line with Barro (1996) who reports a negative and/or insignificant effect of democracy on growth.

To sum up, we find some additional support for the hypothesis of Acemoglu et al. (2019), i. e. democracy does cause growth. Using the IFE estimator to control for time-varying unobserved heterogeneity, we obtain results that are qualitatively similar to those of the authors. If we compare HHK to Interactive with $R = 3$ in the authors preferred specification $p = 4$, we find that the short-run effect of democratization is halved. However the corresponding long-run effect of 18.264% is still pretty close to the 20% reported by Acemoglu et al. (2019).

4.6 Extensions

Although we analyzed the IFE estimator of Bai (2009) and Moon and Weidner (2015, 2017), we want to briefly discuss two extensions where our findings may also be helpful. First, in the presence of endogenous regressors, Moon and Weidner (2017) and Moon, Shum, and Weidner (2018) suggest a minimum distance estimator in spirit of Chernozhukov and Hansen (2006, 2008). Second, because the objective function of the IFE estimator is generally non-convex, Moon and Weidner (2019) suggest an alternative estimator that avoids the potentially difficult optimization problem with multiple local minima and results in optimizing a convex objective function.

Example 1. Minimum Distance Estimator

Suppose that $\mathbf{x}_{it}$ can be further decomposed into K_1 endogenous and K_2 exogenous regressors, i. e. $K = K_1 + K_2$. To avoid ambiguity, we label endogenous and exogenous regressors with an appropriate superscript. Further, let $\mathbf{z}_{it} = (z_{1,it}, \dots, z_{M,it})$ be a vector of excluded exogenous instruments, where $M \geq K_1$. Moon and Weidner (2017) suggest the following minimum distance estimator. In a first step, an estimator for $\boldsymbol{\beta}^{\text{end}}$ is obtained by

$$\hat{\boldsymbol{\beta}}^{\text{end}} \in \underset{\boldsymbol{\beta}^{\text{end}} \in \mathbb{R}^{K_1}}{\arg\min}\; \hat{\boldsymbol{\pi}}(\boldsymbol{\beta}^{\text{end}})' \, \boldsymbol{\Sigma} \, \hat{\boldsymbol{\pi}}(\boldsymbol{\beta}^{\text{end}}) \, ,$$

where $\hat{\boldsymbol{\pi}}(\boldsymbol{\beta}^{\text{end}})$ is the IFE estimator of

$$y_{it} - \mathbf{x}_{it}^{\text{end}\,\prime} \boldsymbol{\beta}^{\text{end}} = \mathbf{x}_{it}^{\text{exo}\,\prime} \boldsymbol{\beta}^{\text{exo}} + \mathbf{z}_{it}' \boldsymbol{\pi} + \boldsymbol{\lambda}_i' \mathbf{f}_t + e_{it}$$

and $\boldsymbol{\Sigma}$ is a positive definite $M \times M$ weighting matrix. At the true value of $\boldsymbol{\beta}^{\text{end}}$, $\boldsymbol{\pi}$ is zero given the instrumental variable moment conditions. In a second step, $\hat{\boldsymbol{\beta}}^{\text{exo}}$ is the IFE estimator of

$$y_{it} - \mathbf{x}_{it}^{\text{end}\,\prime} \hat{\boldsymbol{\beta}}^{\text{end}} = \mathbf{x}_{it}^{\text{exo}\,\prime} \boldsymbol{\beta}^{\text{exo}} + \boldsymbol{\lambda}_i' \mathbf{f}_t + e_{it} \, .$$

The properties of the minimum distance estimator are studied in Moon, Shum, and Weidner (2018), where the authors extend the random coefficient demand model of Berry, Levinsohn, and Pakes (1995) with interactive fixed effects to account for unobserved product-market specific heterogeneity, e. g. perceived utility by advertisement at the product-market level. Under similar assumptions as in Moon and Weidner

(2017), the authors show consistency and derive the asymptotic distribution of the minimum distance estimator. Lee, Moon, and Weidner (2012) use the same estimator to account for measurement errors in the dependent variable in dynamic interactive fixed effects models.

Example 2. Nuclear Norm Minimizing Estimator

Moon and Weidner (2019) show that the imposed rank constraint on the factor structure leads to a non-convex optimization problem. The authors suggest an alternative estimator based on a convex relaxation of this constraint. More precisely, they show that an estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{\star} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^K} \frac{1}{2n} \sum_{r=1}^{\min(N,T)} \sigma_r\left(\mathfrak{P}_{\mathcal{D}}(\mathbf{W}(\boldsymbol{\beta}))\right),$$

where $\sigma_r(\cdot)$ denotes the r-th largest singular value. Moon and Weidner (2019) show consistency of this estimator, but only at a rate of $\sqrt{\min(N,T)}$. As a consequence, the convex relaxation leads to a certain loss of efficiency compared to the IFE estimator.

To recover the properties of the IFE estimator, Moon and Weidner (2019) suggest to estimate the number of factors from $\mathfrak{P}_{\mathcal{D}}(\mathbf{W}(\hat{\boldsymbol{\beta}}^{\star}))$ and afterwards apply an iterative post estimation routine. After a finite number of iterations the estimator has the same limiting distribution as the IFE estimator. The post estimation routine can be summarized as follows:

Algorithm 3. Post nuclear norm estimation

Given $\hat{\boldsymbol{\beta}}^{\star}$ and R, initialize $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\star}$ and repeat the following steps a finite number of times

Step 1. Compute $\widehat{\mathbf{F}}(\hat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\beta}})$ from $\widetilde{\mathbf{W}}(\hat{\boldsymbol{\beta}})$

Step 2. Compute $\check{\mathbf{y}}$ and $\check{\mathbf{x}}_k$ for all $k \in \{1, \ldots, K\}$

Step 3. Update $\hat{\boldsymbol{\beta}} = (\check{\mathbf{X}}'\check{\mathbf{X}})^{-1}\check{\mathbf{X}}'\check{\mathbf{y}}$, where $\check{\mathbf{X}} = (\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_K)$

4.7 Concluding Remarks

The assumption that unobserved heterogeneity is constant over time is often too restrictive. Especially in panels that cover a long time horizon, like macroeconomic panels of countries, it is unlikely that an unobserved global shock affects all countries equally. Thus, interactive fixed effects estimators offer researchers new possibilities to consider this form of heterogeneity in their analysis (see among others Holtz-Eakin, Newey, and Rosen 1988, Pesaran 2006, and Bai 2009). However, these panels are often naturally unbalanced, demanding an additional data augmentation step for the estimator of Bai (2009) (see Appendix of Bai 2009 and Bai, Liao, and Yang 2015).

In this paper, we analyzed the finite sample behavior of Bai (2009)'s interactive fixed effects estimator in unbalanced panels. Simulation experiments confirmed that the inferential theory derived by Bai (2009) and Moon and Weidner (2017) for balanced panels also provides a reasonable approximation for unbalanced panels. However, we also found that the finite sample performance can be affected by the fraction and pattern of missing data.

Future research could address this issue and provide an inferential theory, which takes the additional uncertainty induced by the data augmentation into account. This might help to improve the finite sample behavior of Bai (2009)'s estimator in unbalanced panels.

References

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A. Robinson. 2019. “Democracy Does Cause Growth.” *Journal of Political Economy* 127 (1): 47–100.

Ahn, Seung C., and Alex R. Horenstein. 2013. “Eigenvalue Ratio Test for the Number of Factors.” *Econometrica* 81 (3): 1203–1227.

Alessi, Lucia, Matteo Barigozzi, and Marco Capasso. 2010. “Improved penalization for determining the number of factors in approximate factor models.” *Statistics & Probability Letters* 80 (23): 1806–1813.

Anderson, T.W., and Cheng Hsiao. 1982. “Formulation and estimation of dynamic models using panel data.” *Journal of Econometrics* 18 (1): 47–82.

Arellano, Manuel, and Stephen Bond. 1991. “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations.” *The Review of Economic Studies* 58 (2): 277–297.

Bai, Jushan. 2003. “Inferential Theory for Factor Models of Large Dimensions.” *Econometrica* 71 (1): 135–171.

———. 2009. “Panel Data Models with Interactive Fixed Effects.” *Econometrica* 77 (4): 1229–1279.

Bai, Jushan, and Yuan Liao. 2017. “Inferences in panel data with interactive effects using large covariance matrices.” *Journal of Econometrics* 200 (1): 59–78.

Bai, Jushan, Yuan Liao, and Jisheng Yang. 2015. “Unbalanced Panel Data Models with Interactive Effects.” In *The Oxford Handbook of Panel Data,* 149–170.

Bai, Jushan, and Serena Ng. 2002. “Determining the Number of Factors in Approximate Factor Models.” *Econometrica* 70 (1): 191–221.

———. 2013. “Principal components estimation and identification of static factors.” *Journal of Econometrics* 176 (1): 18–29.

Barro, Robert J. 1996. “Democracy and Growth.” *Journal of Economic Growth* 1 (1): 1–27.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. “Automobile Prices in Market Equilibrium.” *Econometrica* 63 (4): 841–890.

Bonhomme, Stéphane, and Elena Manresa. 2015. “Grouped Patterns of Heterogeneity in Panel Data.” *Econometrica* 83 (3): 1147–1184.

Buja, Andreas, and Nermin Eyuboglu. 1992. “Remarks on Parallel Analysis.” *Multivariate Behavioral Research* 27 (4): 509–540.

Chamberlain, Gary. 1982. “Multivariate regression models for panel data.” *Journal of Econometrics* 18 (1): 5–46.

———. 1984. “Chapter 22 Panel data,” 2:1247–1318. Handbook of Econometrics.

Chen, Mingli, Iván Fernández-Val, and Martin Weidner. 2019. “Nonlinear Factor Models for Network and Panel Data.” *arXiv preprint arXiv: 1412.5647.*

Chen, Shuowen, Victor Chernozhukov, and Iván Fernández-Val. 2019. "Mastering Panel Metrics: Causal Impact of Democracy on Growth." *AEA Papers and Proceedings* 109:77–82.

Chernozhukov, Victor, and Christian Hansen. 2006. "Instrumental quantile regression inference for structural and treatment effect models." *Journal of Econometrics* 132 (2): 491–525.

———. 2008. "Instrumental variable quantile regression: A robust inference approach." *Journal of Econometrics* 142 (1): 379–398.

Choi, In, and Hanbat Jeong. 2019. "Model selection for factor analysis: Some new criteria and performance comparisons." *Econometric Reviews* 38 (6): 577–596.

Czarnowske, Daniel, and Amrei Stammann. 2020. "Fixed Effects Binary Choice Models: Estimation and Inference with Long Panels." *arXiv preprint arXiv:1904.04217.*

Dobriban, Edgar. 2020. "Permutation methods for factor analysis and PCA." *The Annals of Statistics* 48 (5): 2824–2847.

Dobriban, Edgar, and Art B. Owen. 2019. "Deterministic parallel analysis: an improved method for selecting factors and principal components." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81 (1): 163–183.

Fernández-Val, Iván, and Martin Weidner. 2018b. "Fixed Effects Estimation of Large-T Panel Data Models." *Annual Review of Economics* 10 (1): 109–138.

Fong, David Chin-Lung, and Michael Saunders. 2011. "LSMR: An Iterative Algorithm for Sparse Least-Squares Problems." *SIAM Journal on Scientific Computing* 33 (5): 2950–2971.

Gagliardini, Patrick, Elisa Ossola, and Olivier Scaillet. 2019. "A diagnostic criterion for approximate factor structure." *Journal of Econometrics* 212 (2): 503–521.

Gaure, Simen. 2013c. "OLS with multiple high dimensional category variables." *Computational Statistics & Data Analysis* 66:8–18.

Guimarães, Paulo, and Pedro Portugal. 2010. "A simple feasible procedure to fit models with high-dimensional fixed effects." *Stata Journal* 10 (4): 628–649.

Hahn, Jinyong, Jerry Hausman, and Guido Kuersteiner. 2004. "Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations." *The Econometrics Journal* 7 (1): 272–306.

Hallin, Marc, and Roman Liška. 2007. "Determining the Number of Factors in the General Dynamic Factor Model." *Journal of the American Statistical Association* 102 (478): 603–617.

Halperin, Israel. 1962. "The product of projection operators." *Acta Sci. Math. (Szeged)* 23:96–99.

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen. 1988. "Estimating Vector Autoregressions with Panel Data." *Econometrica* 56 (6): 1371–1395.

Lee, Nayoung, Hyungsik Roger Moon, and Martin Weidner. 2012. "Analysis of interactive fixed effects dynamic linear panel regression with measurement error." *Economics Letters* 117 (1): 239–242.

Moon, Hyungsik Roger, Matthew Shum, and Martin Weidner. 2018. "Estimation of random coefficients logit demand models with interactive fixed effects." *Journal of Econometrics* 206 (2): 613–644.

Moon, Hyungsik Roger, and Martin Weidner. 2015. "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects." *Econometrica* 83 (4): 1543–1579.

———. 2017. "DYNAMIC LINEAR PANEL REGRESSION MODELS WITH INTERACTIVE FIXED EFFECTS." *Econometric Theory* 33 (1): 158–195.

———. 2019. "Nuclear Norm Regularized Estimation of Panel Regression Models." *arXiv preprint arXiv: 1810.10987.*

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46 (1): 69–85.

Newey, Whitney K., and Richard J. Smith. 2004. "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators." *Econometrica* 72 (1): 219–255.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (3): 703–708.

———. 1994. "Automatic Lag Selection in Covariance Matrix Estimation." *The Review of Economic Studies* 61 (4): 631–653.

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–1426.

Onatski, Alexei. 2010. "Determining the Number of Factors from Empirical Distribution of Eigenvalues." *The Review of Economics and Statistics* 92 (4): 1004–1016.

Pesaran, M. Hashem. 2006. "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica* 74 (4): 967–1012.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schwert, G. William. 1989. "Tests for Unit Roots: A Monte Carlo Investigation." *Journal of Business & Economic Statistics* 7 (2): 147–159.

Stammann, Amrei. 2018. "Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects." *arXiv preprint arXiv:1707.01815.*

Stock, James H., and Mark W. Watson. 1998. "Diffusion indexes." *NBER Working Paper No. 6702.*

———. 2002. "Macroeconomic forecasting using diffusion indexes." *Journal of Business & Economic Statistics* 20 (2): 147–162.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817–838.

Chapter 5

A Classifier-Lasso Approach for Estimating Production Functions with Latent Group Structures

5.1 Introduction

Production functions are one of the central concepts in economic theory, with origins dating back at least to the mid-18th century (see Humphrey 1997 for the history of production functions). It is all the more remarkable that the identification of production functions is still an econometric challenge.[1] The main challenge is an endogeneity problem first mentioned by Marschak and Andrews (1944). It arises when firms make input decisions based on their productivity, while productivity is usually unobservable to the researcher. Since the seminal work of Olley and Pakes (1996) and Levinsohn and Petrin (2003), all recent identification strategies, such as those of Blundell and Bond (2000), Ackerberg, Caves, and Frazer (2015), and Gandhi, Navarro, and Rivers (2020), use panel data along with assumptions about firm behavior to solve the endogeneity problem.

Although these identification strategies are frequently applied in practice, a remaining challenge is latent firm heterogeneity in the data. For instance, firms may differ in their production technologies, which can be simply thought of parameter heterogeneity. Therefore, it is common practice to classify firms into groups based on prior information, like an industry classification, and then estimate group-specific production functions. This allows the parameters to be different between groups but identical within groups. Throughout this paper, I refer to this type of classification procedure as *ex-ante* classification. However, as pointed out by Griliches and Mairesse (1999), even in a narrowly defined industry where firms produce nearly identical products, these firms may use their inputs very differently, which in turn may indicate different production technologies. On the other hand, firms in different industries may nevertheless use very similar production technologies. Thus neglecting this similarity reduces the precision of the estimation strategies.

To overcome the drawbacks of ex-ante classifications, I propose an estimation procedure for identifying production functions with latent group structures. More specifically, using examples of three recent identification strategies, I show how to embed them into the classifier-Lasso (*C-Lasso*) proposed by Su, Shi, and Phillips (2016). Importantly, my estimation procedure is fully *data-driven* and has no additional data requirements compared to the embedded identification strategies, except that the panel must be sufficiently long. Furthermore, my estimator can be used in situations where prior information is not available at all or only in insufficient detail. I demonstrate the good finite-sample performance of my estimator in simulation experiments. Even in moderately long panels, my estimation procedure is able to assign all firms to the correct latent group with probability close to one. The properties of my estimator are asymptotically identical to an infeasible estimator that knows the latent group structure. In addition, because the iterative estimation algorithm proposed by Su, Shi, and Phillips (2016) is computationally intensive or even infeasible for panels with many firms, I present an extension to the algorithm that overcomes this problem. Finally, I apply my estimation procedure to the Chilean panel data set used in Gandhi, Navarro, and Rivers (2020) and find that an ex-ante classification by industry does not match the data-driven classification of my estimator. In particular, I classify firms from five industries into three latent groups, where each of the five industries is composed of firms from all three latent groups. A simple comparison of the mean squared residuals, reveals that the estimation based on the data-driven classification explains the data better than the ex-post classification based on an industry classification. Thus, estimating separate production functions for each industry is not only less efficient, it also leads to different conclusions.

Besides the C-Lasso of Su, Shi, and Phillips (2016), there are at least two other approaches to deal with latent group structures. The first approach is a finite mixture specification, which models the probability

[1] Gandhi, Navarro, and Rivers (2020) and Shenoy (2020) are two very recent examples that solve this identification problem for gross output production functions.

of belonging to a latent group as a function of the explanatory variables. Some applications of parametric and semi-parametric finite mixtures are Sun (2005), Kasahara and Shimotsu (2009), and Browning and Carro (2014). The second approach is a modification of the k-means clustering algorithm. Examples are Lin and Ng (2012), Bonhomme and Manresa (2015), and Sarafidis and Weber (2015). Both approaches have been recently adapted to the estimation of production functions with latent group structures. Kasahara, Schrimpf, and Suzuki (2017) use a finite mixture specification to extend the identification strategy of Gandhi, Navarro, and Rivers (2020). They also provide empirical evidence of latent group structures in a panel of Japanese firms. Cheng, Schorfheide, and Shao (2019) extend the k-means clustering algorithm to account for multidimensional heterogeneity and combined it with the identification strategy of De Loecker and Warzynski (2012) to reassess the rise of markups reported by De Loecker, Eeckhout, and Unger (2020). They find that the level and growth of markups are lower when additional heterogeneity within industries is taken into account. I contribute to this literature by presenting an alternative estimation procedure based on the C-Lasso and providing additional empirical evidence that classification by industry is not sufficient to account for firm heterogeneity.

The paper is organized as follows. I introduce production functions with latent group structures in Section 5.2. I show how these production functions can be estimated in Section 5.3. I demonstrate the performance of my estimation procedure in Section 5.4. I provide an empirical illustration using Chilean panel data in Section 5.5. Finally, I give some concluding remarks in Section 5.6.

Throughout this paper, I follow conventional notation: scalars are represented in standard type, vectors and matrices in boldface, and all vectors are column vectors. Further, $\|\cdot\|$ denotes the Euclidean norm.

5.2 Production Functions with Latent Group Structures

5.2.1 Structural Panel Model for Production Functions

I consider panel data of N firms observed for $T+1$ periods $\{\xi_{it} : i \in \{1,\dots,N\}, t \in \{0,\dots,T\}\}$, where $\xi_{it} := (Y_{it}, \mathbf{X}_{it}, \mathbf{Z}_{it})$ is a collection of variables for firm i at time t. Y_{it} is the output, $\mathbf{X}_{it}$ is a vector of inputs, like capital K_{it}, labor L_{it}, or intermediate inputs M_{it}, and $\mathbf{Z}_{it}$ is a vector of additional firm characteristics, like input and output prices or a firm's export status. I follow standard convention and denote the corresponding natural log values in lowercase letters, i. e. y_{it} is the natural logarithm of Y_{it}.

To allow for heterogeneous production technologies, I introduce a latent group structure where each firm belongs to exactly one of J^0 latent groups. As in Lin and Ng (2012), Bonhomme and Manresa (2015), and Su, Shi, and Phillips (2016), I assume a time-homogeneous latent group structure, i. e. the group membership of a firm is not allowed to change over time. I collect all firms that are members of a latent group $j \in \{1,\dots,J^0\}$ in a set G_j^0, where $\bigcup_{j=1}^{J^0} G_j^0 = \{1,\dots,N\}$ and $G_j^0 \cap G_{j'}^0 = \varnothing$ for all $j \neq j'$. For the moment, I assume that J^0 is known to the researcher, but the group membership of each firm is not. All firms in G_j^0 have the same smooth and time-homogeneous production function in logs

$$y_{i(j)t} := f_j^0(\mathbf{x}_{i(j)t}) + \omega_{i(j)t} + \epsilon_{i(j)t}\,, \tag{5.1}$$

where $\exp(\omega_{i(j)t} + \epsilon_{i(j)t})$ is an additively separable Hicksian neutral efficiency level. In particular, $\omega_{i(j)t}$ is a predictable productivity shock, unobserved by the researcher, and $\epsilon_{i(j)t}$ is an unanticipated productivity shock or a measurement error. Throughout the paper, I will use the index $i(j)$ to highlight the membership of firm i in latent group j. Furthermore, I assume $f_j^0(\mathbf{x}_{i(j)t}) \neq f_{j'}^0(\mathbf{x}_{i(j)t})$ for all $j \neq j'$, i. e. firms in different latent groups have to use different production technologies.

Besides the unknown latent group membership of the firms, the estimation of (5.1) poses further econometric challenges. Some important issues are related to the measurements of output and inputs, the functional form of the production function, and the endogeneity problem (see Ackerberg et al. 2007 and Aguirregabiria 2019 for more details).[2] In particular, these problems led to very different estimation strategies that can be readily applied if the researcher knows the latent group membership of the firms.[3] More traditional identification strategies rely on instrumental variables and fixed effects. Both strategies are not very popular in recent work, because the former requires additional excluded instruments, like output and input prices, and the latter requires the restrictive assumption $\omega_{i(j)t} = \omega_{i(j)}$. More recent identification strategies rely on modeling firm behavior in a dynamic environment. The key assumptions are related to the timing of firms' input decisions and the information available to firms at the time of these decisions. There are three popular estimation strategies: control function / proxy variable approach (Olley and Pakes 1996, Levinsohn and Petrin 2003, and Ackerberg, Caves, and Frazer 2015), dynamic panel estimator (Arellano and Bond 1991, Blundell and Bond 1998, and Blundell and Bond 2000), and first-order condition approach (McElroy 1987 and Gandhi, Navarro, and Rivers 2020). All these approaches derive moment conditions from the underlying structural model and use these moments to estimate the model parameters by the generalized method of moments (GMM). In the remainder of this paper, I will show how these three recent estimation strategies can be adapted to the latent group structure in (5.1).[4]

5.2.2 Identifying Moment Conditions for Three Baseline Strategies

Because there are numerous modifications and extensions of the different estimation strategies, I focus on the three corresponding baseline strategies. In particular, I briefly summarize the control function approach of Ackerberg, Caves, and Frazer (2015), a dynamic panel estimator in spirit of Blundell and Bond (2000), and the first-order condition approach of Gandhi, Navarro, and Rivers (2020). These three approaches provide a good starting point and a sufficient ground to understand the estimator.

To keep the estimation strategies comprehensible and the notation simple, I restrict myself to the case of three inputs (k_{it}, l_{it}, and m_{it}), and make the following assumptions. i) Firms are homogeneous in their Cobb and Douglas (1928)-type production technology. ii) Each firm chooses its period t inputs based on available information $\mathcal{I}_{it}$, where $\omega_{it} \in \mathcal{I}_{it}$ is known before and $\epsilon_{it} \notin \mathcal{I}_{it}$ is realized after each firm's input decisions. iii) Firms predict their future productivity by $\omega_{it} = h^0_\omega(\omega_{it-1}) + \eta_{it}$, where η_{it} is unknown before t. iv) Input choices for capital and labor have dynamic implications, e. g. capital is accumulated by past investment decisions and labor is partially predetermined by hiring and firing costs.

In the following, I denote the P-dimensional vector with model parameters as $\boldsymbol{\theta}^0$. Furthermore, I refer to $\mathbf{g}(\xi_i^t, \boldsymbol{\theta}^0)$ as P'-dimensional vector of moments, where $\xi_i^t := \{\xi_{it'} : t' \in \{0, \ldots, t\}\}$, so that $P' \geq P$ moment conditions of the form $\mathbb{E}[\mathbf{g}(\xi_i^t, \boldsymbol{\theta}^0)] = 0$ can be used to estimate $\boldsymbol{\theta}^0$ by GMM.

[2] It is known since Marschak and Andrews (1944) that when firms choose their inputs optimally, at least to some extent, $\omega_{i(j)t}$ and $\mathbf{x}_{i(j)t}$ are correlated, leading to inconsistent estimators of output elasticities (see Griliches and Mairesse 1999).

[3] For instance, it is a common strategy to classify firms by an industry classification.

[4] The traditional strategies can already be applied using the estimators presented in Section 2 and 3 of Su, Shi, and Phillips (2016). The instrumental variable strategy requires a slight modification of their GMM estimation for linear panel structure models.

Example 1. Ackerberg, Caves, and Frazer (2015) propose a control function approach for a specific type of value-added production function, e. g.

$$y_{it} = \beta_0^0 + \beta_1^0 k_{it} + \beta_2^0 l_{it} + \omega_{it} + \epsilon_{it} = \beta_3^0 m_{it} + \omega_{it} + \epsilon_{it} \,.$$

This is a reasonable specification if there is perfect complementarity between intermediate (m_{it}) and the other two inputs (k_{it} and l_{it}), e. g.

$$y_{it} = \min(\beta_0^0 + \beta_1^0 k_{it} + \beta_2^0 l_{it}, \beta_3^0 m_{it}) + \omega_{it} + \epsilon_{it} \,.$$

For identification, it is necessary to make two additional assumptions. First, firms choose m_{it} according to $m_{it} = h_m^0(k_{it}, l_{it}, \omega_{it})$. Second, $h_m^0(k_{it}, l_{it}, \omega_{it})$ must be a strictly monotonically increasing function in ω_{it}, so that the only unobservable in this function (ω_{it}) can be expressed as $\omega_{it} = h_m^{0\langle -1\rangle}(k_{it}, l_{it}, m_{it})$, where $h_m^{0\langle -1\rangle}(\cdot)$ denotes the inverse function of $h_m^0(\cdot)$. Putting all together yields the following system of two equations:

$$\begin{aligned} y_{it} &= \beta_0^0 + \beta_1^0 k_{it} + \beta_2^0 l_{it} + h_m^{0\langle -1\rangle}(k_{it}, l_{it}, m_{it}) + \epsilon_{it} = h^0(k_{it}, l_{it}, m_{it}) + \epsilon_{it} \,, \\ y_{it} &= \beta_0^0 + \beta_1^0 k_{it} + \beta_2^0 l_{it} + h_\omega^0(h^0(k_{it-1}, l_{it-1}, m_{it-1}) - \beta_0^0 - \beta_1^0 k_{it-1} - \beta_2^0 l_{it-1}) + v_{it} \,, \end{aligned}$$

where $h^0(\cdot)$ and $h_\omega^0(\cdot)$ are unknown functions that can be approximated by a sequence of sieves, e. g. polynomials, and $v_{it} := \eta_{it} + \epsilon_{it}$. The model parameters can then be estimated by the following identifying moment conditions: $\mathbb{E}[\epsilon_{it} \mid \mathcal{I}_{it}] = 0$ and $\mathbb{E}[v_{it} \mid \mathcal{I}_{it-1}] = 0$. For instance, if $h^0(k_{it}, l_{it}, m_{it}) = \alpha_0^0 + \alpha_1^0 k_{it} + \alpha_2^0 l_{it} + \alpha_3^0 m_{it}$ and $h_\omega(\omega_{it-1}) = \delta^0 \omega_{it-1}$ then $\boldsymbol{\theta}^0 = (\alpha_0^0, \alpha_1^0, \alpha_2^0, \alpha_3^0, \beta_0^0, \beta_1^0, \beta_2^0, \delta^0)$ and $\mathbf{g}(\xi_i^t, \boldsymbol{\theta}^0) = (\epsilon_{it}, \epsilon_{it} k_{it}, \epsilon_{it} l_{it}, \epsilon_{it} m_{it}, v_{it}, v_{it} k_{it}, v_{it} k_{it-1}, v_{it} l_{it-1}, v_{it} m_{it-1})$.

Example 2. A dynamic panel estimator in spirit of Blundell and Bond (2000) can be used to estimate the model parameters by imposing $h_\omega(\omega_{it-1}) = \delta^0 \omega_{it-1}$. To see this, consider the production function

$$y_{it} = \beta_0^0 + \beta_1^0 k_{it} + \beta_2^0 l_{it} + \beta_3^0 m_{it} + \omega_{it} + \epsilon_{it}$$

and its reformulation

$$\Delta_\delta^0(y_{it}) = (1 - \delta^0)\beta_0^0 + \beta_1^0 \Delta_\delta^0(k_{it}) + \beta_2^0 \Delta_\delta^0(l_{it}) + \beta_3^0 \Delta_\delta^0(m_{it}) + w_{it} \,,$$

where $\Delta_\delta^0(x_{it}) := x_{it} - \delta^0 x_{it-1}$ and $w_{it} := \eta_{it} + \Delta_\delta^0 \epsilon_{it}$. Under the assumption that m_{it} has dynamic implications, e. g. due to adjustment costs as in Gandhi, Navarro, and Rivers (2020), or that firms intermediate input choices are distorted, e. g. due to input market frictions as in Shenoy (2020), the moment conditions $\mathbb{E}[w_{it} \mid \mathcal{I}_{it-1}]$ can be used to estimate the model parameters. For instance, the identification strategy of Shenoy (2020) implies $\boldsymbol{\theta}^0 = (\beta_0^0, \beta_1^0, \beta_2^0, \beta_3^0, \delta^0)$ and $\mathbf{g}(\xi_i^t, \boldsymbol{\theta}^0) = (w_{it}, w_{it} k_{it}, w_{it} l_{it}, w_{it} k_{it-1}, w_{it} l_{it-1}, w_{it} m_{it-1})$.

Example 3. Gandhi, Navarro, and Rivers (2020) propose an estimation strategy for gross output productions functions, e. g.

$$y_{it} = \beta_0^0 + \beta_1^0 k_{it} + \beta_2^0 l_{it} + \beta_3^0 m_{it} + \omega_{it} + \epsilon_{it} \,.$$

For their baseline strategy, they assume that firms maximize their expected profits in a perfectly competitive environment. The basic idea is to identify β_3^0 from firms' optimal intermediate input decisions. Afterwards they exploit the dynamic structure of the model to identify the remaining model parameters. Under the additional assumption that M_{it} is a flexible input with a linear cost function $P^M_t M_{it}$, i. e. the choice of M_{it} is undistorted and has no dynamic implications, firms choose M_{it} as the solution to

$$\max_{M_{it} \in \mathbb{R}^{>0}} \mathbb{E}[P^Y_t \exp(\beta^0_0 + \beta^0_1 k_{it} + \beta^0_2 l_{it} + \beta^0_3 m_{it} + \omega_{it} + \epsilon_{it}) - P^M_t M_{it} \mid \mathcal{I}_{it}] \, ,$$

where P^Y_t and P^M_t are common output and intermediate input prices, respectively. Reformulating the corresponding first-order condition

$$P^Y_{it} \exp(\beta^0_0 + \beta^0_1 k_{it} + \beta^0_2 l_{it} + \beta^0_3 m_{it} + \omega_{it}) \beta^0_3 \mathcal{E}^0 = P^M_{it} M_{it}$$

and exploiting the dynamic structure yields the following system of two equations:

$$\begin{aligned} s_{it} &= \log(\beta^0_3 \mathcal{E}^0) - \epsilon_{it} \, , \\ y^r_{it} &= \beta^0_0 + \beta^0_1 k_{it} + \beta^0_2 l_{it} + h^0_\omega (y^r_{it-1} - \beta^0_1 k_{it-1} - \beta^0_2 l_{it-1}) + \eta_{it} \, , \end{aligned}$$

where $s_{it} := \log((P^M_{it} M_{it})/(P^Y_{it} Y_{it}))$ is the logarithm of intermediate inputs expenditures divided by revenues, $\mathcal{E}^0 := \mathbb{E}[\exp(\epsilon_{it}) \mid \mathcal{I}_{it}] = \mathbb{E}[\exp(\epsilon_{it})]$ is a positive constant, and $y^r_{it} := y_{it} - \beta^0_3 m_{it} - \epsilon_{it}$. One difference to the other two strategies is that the researcher additionally needs data on prices (P^Y_t and P^M_t), or at least on the ratio of prices (P^M_t / P^Y_t). The model parameters can then be estimated by the following identifying moment conditions: $\mathbb{E}[\epsilon_{it} \mid \mathcal{I}_{it}] = 0$, $\mathbb{E}[\exp(\epsilon_{it})] = \mathcal{E}^0$, and $\mathbb{E}[\eta_{it} \mid \mathcal{I}_{it-1}] = 0$. For instance, if $h_\omega(\omega_{it-1}) = \delta^0 \omega_{it-1}$ then $\boldsymbol{\theta}^0 = (\beta^0_3, \mathcal{E}^0, \beta^0_0, \beta^0_1, \beta^0_2, \delta^0)$ and $\mathbf{g}(\xi^t_i, \boldsymbol{\theta}^0) = (\epsilon_{it}, \exp(\epsilon_{it}) - \mathcal{E}^0, \eta_{it}, \eta_{it} k_{it}, \eta_{it} l_{it}, \eta_{it} y^r_{it-1})$.

5.3 Identifying Latent Group Structures

5.3.1 Penalized Generalized Method of Moments Estimator

To identify the unknown latent group structure, I suggest to use the penalized GMM (PGMM) estimator of Su, Shi, and Phillips (2016). The authors propose a novel Lasso technique where an additive-multiplicative penalty term is added to a GMM objective function. The basic idea of their penalization approach is to achieve classification by shrinking firm-specific model parameters $\boldsymbol{\pi}^0_i$ to group-specific parameters $\boldsymbol{\theta}^0_j$. In the following, I show how the PGMM approach can be used to estimate production functions with latent group structures. For the moment, I assume that the true number of groups J^0 is known. However, since J^0 is rarely known in practice, I will show how to estimate it in the next subsection. The full estimation procedure consists of three subsequent steps. First, the firm- and group-specific model parameters are estimated by PGMM. Second, based on the PGMM estimates, each firm is assigned to one latent group. Third, the group-specific model parameters are re-estimated by a Post-Lasso estimator that estimates a separate production function for each of the estimated latent groups.

First, I suggest to estimate $\boldsymbol{\pi}^0_i$ and $\boldsymbol{\theta}^0_j$ by minimizing the following PGMM objective function:

$$Q^{J^0}_\lambda(\boldsymbol{\Pi}, \boldsymbol{\Theta}) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i(\xi^T_i, \boldsymbol{\pi}_i)' \mathbf{g}_i(\xi^T_i, \boldsymbol{\pi}_i) + \lambda \prod_{j=1}^{J^0} \| \boldsymbol{\pi}_i - \boldsymbol{\theta}_j \| \, ,$$

where $\mathbf{\Pi} := (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N)$, $\mathbf{\Theta} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{J^0})$, $\lambda > 0$ is a tuning parameter, and

$$\mathbf{g}_i(\xi_i^T, \boldsymbol{\theta}) := \frac{1}{T} \sum_{t=1}^{T} \mathbf{g}(\xi_i^t, \boldsymbol{\theta}) \,. \tag{5.2}$$

Some examples for $\boldsymbol{\pi}_i$ and $\mathbf{g}(\xi_i^t, \boldsymbol{\pi}_i)$ are given in Section 5.2.2, e. g. $\boldsymbol{\pi}_i = (\beta_{3i}, \mathcal{E}_i, \beta_{0i}, \beta_{1i}, \beta_{2i}, \delta_i)$ and $\mathbf{g}(\xi_i^t, \boldsymbol{\pi}_i) = (\epsilon_{it}, \exp(\epsilon_{it}) - \mathcal{E}_i, \eta_{it}, \eta_{it} k_{it}, \eta_{it} l_{it}, \eta_{it} y_{it-1}^r)$. The first part of the sum in $Q_\lambda^{J^0}(\mathbf{\Pi}, \mathbf{\Theta})$ is the GMM objective function for estimating the firm-specific model parameters, and the second part is the penalty term scaled by λ. The former is minimal when the moments are close to zero, whereas the latter is minimal when $\boldsymbol{\pi}_i$ is close to any $\boldsymbol{\theta}_j$. To put it simple, while the GMM part ensures that the model fits the data, the penalty term constraints the firm-specific model parameters to be close to any group-specific parameter.

Second, I assign all firms to one of J^0 latent groups. For this purpose, I use a classification rule proposed by Su, Shi, and Phillips (2016):

$$\widehat{G}_j := \{i \in \{1, \ldots, N\} \colon \min(\{\|\hat{\boldsymbol{\pi}}_i - \hat{\boldsymbol{\theta}}_{j'}\| \colon j' \in \{1, \ldots, J^0\}) = \|\hat{\boldsymbol{\pi}}_i - \hat{\boldsymbol{\theta}}_j\|\} \text{ for } j \in \{1, \ldots, J^0\} \,, \tag{5.3}$$

where $\hat{\boldsymbol{\pi}}_i$ and $\hat{\boldsymbol{\theta}}_j$ are the firm- and group-specific PGMM estimators, respectively. Intuitively, firms are assigned to the latent group to which they are most similar, where similarity is defined here as the euclidean distance between $\hat{\boldsymbol{\pi}}_i$ and $\hat{\boldsymbol{\theta}}_j$. Note that even if all firms were correctly classified, it does not follow that $\widehat{G}_j = G_j^0$, because the classification rule only ensures that $\widehat{G}_j \in \{G_1^0, \ldots, G_{J^0}^0\}$. To keep the notation simple, I pretend that $\widehat{G}_j = G_j^0$ follows.

Third, I suggest to estimate $\boldsymbol{\theta}_j^0$ by the following Post-Lasso estimator:

$$\tilde{\boldsymbol{\theta}}_j \in \underset{\boldsymbol{\theta}_j \in \mathbb{R}^P}{\arg\min} \Big(\frac{1}{|\widehat{G}_j|} \sum_{i \in \widehat{G}_j} \mathbf{g}_i(\xi_i^T, \boldsymbol{\theta}_j) \Big)' \mathbf{W}_j \Big(\frac{1}{|\widehat{G}_j|} \sum_{i \in \widehat{G}_j} \mathbf{g}_i(\xi_i^T, \boldsymbol{\theta}_j) \Big) \,, \tag{5.4}$$

where $\mathbf{W}_j$ is a positive definite $P' \times P'$ weighting matrix that may depend on the data. Note that $\tilde{\boldsymbol{\theta}}_j$ is simply a standard GMM estimator applied to the subsample of firms that are member of the estimated group j. Thus, the estimation strategy is essentially based on the idea that if I can correctly assign all firms, my Post-Lasso estimator has the same properties as an infeasible oracle estimator that exploits the unknown latent group structure.

Su, Shi, and Phillips (2016) develop a limiting theory for a PGMM estimator that can be applied to dynamic linear panel models with endogeneity, such as those that are typically estimated by the Arellano and Bond (1991) estimator. Under asymptotics where N and T pass jointly to infinity, but not necessary at the same rate, the authors show that the PGMM estimator is able to assign all firms belonging to a given latent group to the same group with probability approaching one. This classification consistency property allows them to prove that the resulting Post-Lasso estimator is asymptotically equivalent to the infeasible oracle estimator. Because of the close relationship between the dynamic panel and the other approaches for production function estimation, as noted by Ackerberg, Caves, and Frazer (2015), I conjecture, that the properties derived by Su, Shi, and Phillips (2016) also hold for my proposed nonlinear estimator, such that the asymptotic distribution of $\tilde{\boldsymbol{\theta}}_j$ can be approximated by $\mathcal{N}(\boldsymbol{\theta}_j^0, \mathbf{V}_j)$ for all $j \in \{1, \ldots, J^0\}$.[5] For instance, a

[5] Although the production function itself is linear, the assumption about the evolution of productivity causes the entire model to become nonlinear. A rigorous proof for my suggested estimation strategy will be added later. My conjecture is further supported by simulation experiments in Section 5.4.

White (1980)-type heteroskedasticity consistent estimator for the covariance is

$$\widetilde{\mathbf{V}}_j := \frac{1}{|\widehat{G}_j|T}\Big((\widetilde{\mathbf{\Gamma}}_j'\mathbf{W}_j\widetilde{\mathbf{\Gamma}}_j)^{-1}\widetilde{\mathbf{\Gamma}}_j'\mathbf{W}_j\widetilde{\mathbf{\Omega}}_j\mathbf{W}_j\widetilde{\mathbf{\Gamma}}_j(\widetilde{\mathbf{\Gamma}}_j'\mathbf{W}_j\widetilde{\mathbf{\Gamma}}_j)^{-1}\Big)\,, \tag{5.5}$$

where

$$\widetilde{\mathbf{\Gamma}}_j := \frac{1}{|\widehat{G}_j|T}\sum_{i\in\widehat{G}_j}\sum_{t=1}^{T}\frac{\partial\mathbf{g}(\xi_i^t,\tilde{\boldsymbol{\theta}}_j)}{\partial\tilde{\boldsymbol{\theta}}_j} \quad \text{and}$$

$$\widetilde{\mathbf{\Omega}}_j := \frac{1}{|\widehat{G}_j|T}\sum_{i\in\widehat{G}_j}\sum_{t=1}^{T}\mathbf{g}(\xi_i^t,\tilde{\boldsymbol{\theta}}_j)\mathbf{g}(\xi_i^t,\tilde{\boldsymbol{\theta}}_j)'\,.$$

Other estimators, such as the heteroskedasticity and autocorrelation consistent estimator of Newey and West (1987), are obtained by modifying $\widetilde{\mathbf{\Omega}}_j$. Note that the choice of the weighting matrix $\mathbf{W}_j$ affects the efficiency of the GMM estimator only when the number of moments P' is larger than the number of model parameters P. Hansen (1982) and Hansen and Singleton (1982) show that choosing $\mathbf{W}_j = \mathbf{\Omega}_j^{-1}$ yields the most efficient GMM estimator in this case. However since $\mathbf{\Omega}_j^{-1}$ is not known, the authors propose a two-step procedure in which $\mathbf{\Omega}_j^{-1}$ is estimated in the first step. For $P = P'$, $\mathbf{W}_j = \mathbf{1}_P$ is a standard choice.

Finally, from a practical point of view, two things are particularly important. First, N and T have to be sufficiently large. In particular, a large number of time periods reduces the influence of a single firm-specific moment in the objective function, which contributes significantly to the classification accuracy. Second, the tuning parameter λ must satisfy certain conditions that hold, for instance, if $\lambda \in \{T^{-a}: a \in (0, 0.5)\}$.

Remark 1. The estimation procedure can be readily used for unbalanced panels. I assume that the observations for each firm are consecutive, i. e. $\{\xi_{it}: i \in \{1, \dots, N\}, t \in \{t_i, \dots, T_i\}, 0 \le t_i < T_i \le T\}$. Adjusting the estimators requires only replacing (5.2) with

$$\mathbf{g}_i(\xi_i^{T_i}, \boldsymbol{\theta}) = \frac{1}{T_i - t_i}\sum_{t=t_i+1}^{T_i}\mathbf{g}(\xi_i^t, \boldsymbol{\theta})\,. \tag{5.6}$$

Unfortunately, the asymptotic theory developed by Su, Shi, and Phillips (2016) is only for balanced panels. Their asymptotic theory has recently been extended by Su, Wang, and Jin (2019) to unbalanced panels. From a practical point of view, the crucial difference is that $\min(\{T_i - t_i : i \in 1, \dots, N\})$ has to be sufficiently large in unbalanced panels.

Remark 2. The classification rule presented in (5.3) ensures that all firms are assigned to a latent group. It is also possible to use a stricter rule that may leave some of the firms unclassified, e. g.

$$\widehat{G}_j := \{i \in \{1, \dots, N\}: \ \min(\{\|\hat{\boldsymbol{\pi}}_i - \hat{\boldsymbol{\theta}}_{j'}\|: j' \in \{1, \dots, J^0\}) = \|\hat{\boldsymbol{\pi}}_i - \hat{\boldsymbol{\theta}}_j\| \le \varepsilon\} \text{ for } j \in \{1, \dots, J^0\}\,, \tag{5.7}$$

where ε is small positive constant. Intuitively, firms are only assigned to the most similar group if the distance is sufficiently small. This stricter rule could help to deal with outlier firms and thus improve the finite sample properties of the Post-Lasso estimator. For clarification, the unclassified firms are simply excluded from the Post-Lasso estimation.

5.3.2 Determining the Number of Latent Groups

So far, I have assumed that the true number of latent groups is known to the researcher. Since this is very unlikely in practice, I follow Su, Shi, and Phillips (2016) and suggest to estimate J^0 by minimizing the following BIC-type information criterion:

$$\mathrm{IC}_p(J,\lambda) := \log\Big(\frac{1}{NT}\sum_{j=1}^{J}\sum_{i\in\widehat{G}_j}\sum_{t=1}^{T}(\tilde{r}_{i(j)t}(J,\lambda))^2\Big) + JPp(N,T)\,, \tag{5.8}$$

where J is a guess for the number of latent groups, $\tilde{r}_{i(j)t}(J,\lambda)$ are the corresponding Post-Lasso residuals of an estimation given λ and J, e. g. $\tilde{r}_{i(j)t}(J,\lambda) = \tilde{\eta}_{i(j)t}(J,\lambda) + \tilde{\epsilon}_{i(j)t}(J,\lambda)$, and $p(N,T)$ is a penalty term that satisfies $p(N,T) \to 0$ and $NTp(N,T) \to \infty$ as $N,T \to \infty$. Su, Shi, and Phillips (2016) suggest two different penalty terms: $p(N,T) = 2/3\,(NT)^{-0.5}$ and $p(N,T) = 0.25\log(\log(T))/T$. Further examples of penalty terms can be found in Bai and Ng (2002), Lin and Ng (2012), and Bonhomme and Manresa (2015). Thus, given an appropriate choice of λ, the number of latent groups can be estimated as follows:

$$\widehat{J}_p(\lambda) \in \underset{J\in\{1,\ldots,\overline{J}\}}{\arg\min}\ \mathrm{IC}_p(J,\lambda)\,, \tag{5.9}$$

where $\overline{J}$ is a known upper bound on the number of latent groups.

Remark 3. As noted by Su, Shi, and Phillips (2016), the information criterion can also be used to jointly determine λ and J, e. g.

$$\widehat{J}_p \in \underset{\lambda\in\mathcal{L}}{\arg\min}\ \mathrm{IC}_p(\widehat{J}_p(\lambda),\lambda)\,, \tag{5.10}$$

where $\mathcal{L} := \{T^{-a}\colon a \in (0,0.5)\}$. In practice, a grid of candidate values between 0 and 0.5 can be used to keep the number of estimates tractable, e. g. $a \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}$.

5.3.3 Estimation Algorithm

In general, the objective function is not jointly convex in $\mathbf{\Pi}$ and $\mathbf{\Theta}$ and therefore very costly to minimize. To reduce the computational costs, Su, Shi, and Phillips (2016) propose an iterative algorithm that exploits the specific structure of the optimization problem. Their basic idea is to split the problem into a sequence of J convex subproblems that are solved sequentially until convergence. The corresponding objective function of the j-th subproblem is

$$Q_\lambda^{\langle j,J\rangle}(\mathbf{\Pi}^j,\boldsymbol{\theta}_j) := \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\xi_i^T,\boldsymbol{\pi}_i^j)'\mathbf{g}_i(\xi_i^T,\boldsymbol{\pi}_i^j) + \lambda\|\boldsymbol{\pi}_i^j - \boldsymbol{\theta}_j\|\zeta_i^j\,, \tag{5.11}$$

where $\mathbf{\Pi}^j := (\boldsymbol{\pi}_1^j,\ldots,\boldsymbol{\pi}_N^j)$, $\mathbf{\Theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J)$, and $\zeta_i^j := \prod_{j'\neq j}^{J}\|\boldsymbol{\pi}_i^{j'} - \boldsymbol{\theta}_{j'}\|$. Note that $\boldsymbol{\zeta}_i := (\zeta_1^j,\ldots,\zeta_N^j)$ is the part of the additive-multiplicative penalty term that is fixed when solving the j-th subproblem. Their entire algorithm can be sketched as follows.

Algorithm 1. Iterative Algorithm of Su, Shi, and Phillips (2016)

Step 0. Given λ and J, e. g. $\lambda = T^{-0.25}$ and $J = J^0$, initialize $\{\mathbf{\Pi}^j\colon j \in \{1,\ldots,J\}\}$ and $\mathbf{\Theta}$. For instance, natural candidates for initial values are $\mathbf{\Pi}^1 = \ldots = \mathbf{\Pi}^J = \mathbf{\Pi}^{\langle 0\rangle}$, where $\mathbf{\Pi}^{\langle 0\rangle}$ is a $P \times N$ matrix whose i-th column is a time series estimate of firm i's production function, and $\mathbf{\Theta} = \mathbf{0}_{P\times J}$.

Step 1. For each $j \in \{1, \ldots, J\}$, compute $\boldsymbol{\zeta}_i$ given the current values of $\{\boldsymbol{\Pi}^j : j \in \{1, \ldots, J\}\}$ and $\boldsymbol{\Theta}$, and afterwards update $\boldsymbol{\Pi}^j$ and $\boldsymbol{\theta}_j$, i. e. the j-th column in $\boldsymbol{\Theta}$, by solving the j-th subproblem.

Step 2. Repeat Step 1 until convergence. For instance, stop if $|Q_1^{\langle r \rangle} - Q_1^{\langle r-1 \rangle}|/(Q_1^{\langle r-1 \rangle} + 1) < \varepsilon$, where $Q_1^{\langle r \rangle} := \sum_{j=1}^{J} Q_\lambda^{\langle j,J \rangle}(\boldsymbol{\Pi}^{j\langle r \rangle}, \boldsymbol{\theta}_j^{\langle r \rangle})$ is the sum of the function values of all subproblems in the r-th iteration and ε is a small positive constant.

Although convexity significantly reduces the computational cost, the overall optimization is still challenging because the solution of each subproblem in Step 1 involves $(N + 1) \times P$ parameters. For instance, in my empirical illustration, I consider a panel of $N = 571$ firms and a production function with $P = 6$ model parameters, which turns solving each subproblem into a high-dimensional optimization problem (i. e. 3,432 parameters per subproblem). To further reduce the computational costs, especially for large N, I suggest to exploit the separable structure of (5.11). The idea is that instead of minimizing jointly over $\boldsymbol{\Pi}$ and $\boldsymbol{\theta}_j$, I can alternate between solving two optimization problems, i. e. a minimization over $\boldsymbol{\Pi}^j$ holding $\boldsymbol{\theta}_j$ fixed and a minimization over $\boldsymbol{\theta}_j$ holding $\boldsymbol{\Pi}^j$ fixed. The second optimization problem involves only P parameters and can be interpreted as Fermat-Weber location problem for which there are numerous efficient algorithms, e. g. the fixed-point algorithm of Weiszfeld (1937). Unfortunately, the first minimization is still over $N \times P$ parameters. However, holding $\boldsymbol{\theta}_j$ fixed, the first minimization problem can be perfectly separated into N independent optimization problems involving only P parameters. My suggested algorithm is essentially an *Alternate Convex Search* (see Hastie, Tibshirani, and Wainwright 2015, chapter 5.9) and can be sketched as follows.

Algorithm 2. Alternate Convex Search for Step 1 in Algorithm 1

Step 0. Given λ and $j \in \{1, \ldots, J\}$, initialize $\boldsymbol{\Pi}^j$ and $\boldsymbol{\theta}_j$.

Step 1. Given $\boldsymbol{\Pi}^j$ update $\boldsymbol{\theta}_j$ as

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^P}{\arg\min} \; \frac{\lambda}{N} \sum_{i=1}^{N} \|\boldsymbol{\pi}_i^j - \boldsymbol{\theta}\| \zeta_i^j \, .$$

Step 2. For each $i \in \{1, \ldots, N\}$, given $\boldsymbol{\theta}_j$ update $\boldsymbol{\pi}_i^j$ as

$$\underset{\boldsymbol{\pi} \in \mathbb{R}^P}{\arg\min} \; \mathbf{g}_i(\xi_i^T, \boldsymbol{\pi})' \mathbf{g}_i(\xi_i^T, \boldsymbol{\pi}) + \lambda \|\boldsymbol{\pi} - \boldsymbol{\theta}_j\| \zeta_i^j \, .$$

Step 3. Repeat Step 1 and Step 2 until convergence. For instance, stop if $|Q_2^{\langle r \rangle} - Q_2^{\langle r-1 \rangle}|/(Q_2^{\langle r-1 \rangle} + 1) < \varepsilon$, where $Q_2^{\langle r \rangle} := Q_\lambda^{\langle j,J \rangle}(\boldsymbol{\Pi}^{j\langle r \rangle}, \boldsymbol{\theta}_j^{\langle r \rangle})$ is the function value of the j-th subproblem in the r-th iteration and ε is a small positive constant.

Remark 4. First, although each minimization problem in Step 2 of Algorithm 2 is convex, the associated objective function is not differentiable due to the Euclidean norm in the penalty function. For these types of minimization problems, there are special optimization algorithms, such as those presented in Hastie, Tibshirani, and Wainwright (2015, chapter 5). Second, because the N optimization problems are independent of each other, they can also be easily parallelized.

5.4 Simulation Experiments

To analyze the classification accuracy and statistical properties of my proposed estimator, I adjust the representative firm model of Syverson (2001) to generate theory-consistent data of firms with latent group structures. His model has the advantage that each firm's decisions problem can be solved analytically and thus avoids the computational obstacles of using numerical tools. This specific representative firm model has been widely used and extended in the related literature, for instance by van Biesebroeck (2007), Ackerberg, Caves, and Frazer (2015), and Collard-Wexler and De Loecker (2016).

Next, I describe the representative firm model used to simulate balanced panels of $N = 200$ firms. Each firm is observed for T time periods and belongs to one of $J^0 = 3$ latent groups. Each group consists of N_j firms. All firms are homogeneous within a group, but differ both in their parameter configuration and in their relative occurrence between groups. At the beginning of period t, a firm $i(j)$ with rational expectations faces the following input decisions problem:

$$\begin{aligned} &\max_{M_{i(j)t},\, I_{i(j)t}} \; \mathbb{E}\left[\sum_{t=0}^{\infty} b^t (Y_{i(j)t} - M_{i(j)t} - \phi_i I_{i(j)t}^2/2) \mid \mathcal{I}_{i(j)t}\right] \qquad (5.12) \\ &\text{subject to} \\ &Y_{i(j)t} := K_{i(j)t}^{\beta_j} M_{i(j)t}^{\gamma_j} \exp(\omega_{i(j)t} + \epsilon_{i(j)t}) \,, \\ &K_{i(j)t} := (1-d)K_{i(j)t-1} + I_{i(j)t-1} \,, \\ &\omega_{i(j)t} := \alpha_j + \delta_j \omega_{i(j)t-1} + \eta_{i(j)t} \,, \end{aligned}$$

where $Y_{i(j)t}$, $K_{i(j)t}$, $M_{i(j)t}$, $I_{i(j)t}$ are output, capital, intermediate input, and investment, $\mathcal{I}_{i(j)t}$ is a set of information available at the beginning of period t, $\omega_{i(j)t}$ is a persistent and anticipated productivity shock, and $\epsilon_{i(j)t}$ is a productivity shock realized after each firm's input decision. Furthermore, $\log(\phi_i^{-1}) \sim$ iid. $\mathcal{N}(0, 1)$, $\eta_{i(j)t} \sim$ iid. $\mathcal{N}(0, \sigma^2_{\eta(j)})$, $\epsilon_{i(j)t} \sim$ iid. $\mathcal{N}(0, \sigma^2_{\epsilon(j)})$, $\omega_{i(j)0} \sim$ iid. $\mathcal{N}(\alpha_j/(1-\delta_j), \sigma^2_{\eta(j)}/(1-\delta_j^2))$, and $K_{i(j)0} := \mathbf{0}_{N_j}$. All model parameters are described and defined in Table 5.4.1.

Table 5.4.1: *Model Parameters: Description and Definition*

Parameter	Description	Latent Group		
		1	2	3
Heterogeneous Parameters:				
N_j/N	Relative Occurrence	0.30	0.40	0.30
γ_j	Output Elasticity of Intermediate Input	0.35	0.50	0.65
$\beta_j := 1 - \gamma_j$	Output Elasticity of Capital	0.65	0.50	0.35
$\sigma_{\epsilon(j)}$	Standard Deviation of Ex-Post Productivity Shock	0.02	0.04	0.02
α_j	Constant of AR(1) Process	0.00	0.20	0.40
δ_j	Slope Parameter of AR(1) Process	0.90	0.80	0.70
$\sigma_{\eta(j)}$	Standard Deviation of Innovation in AR(1) Process	0.01	0.01	0.01
Homogeneous Parameters:				
b	Discount Factor		0.985	
d	Depreciation Rate		0.100	

I briefly summarize the core features of my model. Firms maximize their expected future profits with respect to their intermediate input and investment choices. Future profits are discounted and all firms face firm-specific but time constant quadratic capital adjustment costs, as in Ackerberg, Caves, and Frazer (2015). Within a latent group, all firms share the same time-homogeneous Cobb-Douglas production function with constant returns to scale. The current capital stock is accumulated through a dynamic process determined by depreciation and past investment decisions. Productivity is additively separable and can be decomposed into a persistent component and an ex-post shock. The persistent component is modeled as exogenous AR(1) process.

Next, I present the optimal input decisions of firm $i(j)$ at time t. Because the intermediate input is fully flexible to adjust, i. e. there are no adjustment costs or other dynamic implications, its optimal choice follows directly from the first-order condition:

$$M^*_{i(j)t} := (\gamma_j \exp(\omega_{i(j)t})\mathcal{E}_j)^{\frac{1}{\beta_j}} K_{i(j)t} ,$$

where $\mathcal{E}_j = \exp(\sigma^2_{\epsilon(j)}/2)$. Using the Euler equation for investment along with forward substitution yields the following optimal investment decision:

$$I^*_{i(j)t} := b\beta_j(\gamma_j\mathcal{E}_j)^{\frac{\gamma_j}{\beta_j}}\phi_i^{-1}\sum_{\tau=0}^{\infty}\left((b(1-d))^\tau \exp\left(\frac{\alpha_j\sum_{s=0}^{\tau}\delta_j^s+\delta_j^{\tau+1}\omega_{i(j)t}}{\beta_j}+\frac{\sigma^2_{\eta(j)}\sum_{s=0}^{\tau}\delta_j^{2s}}{2\beta_j^2}\right)\right).$$

As first noted by Syverson (2001), the constant returns to scale assumption together with convex adjustment costs ensure that the capital stock drops out of the Euler equation, leading to a fully deterministic optimal investment decision. To ensure that all firms are sampled from their steady state distribution, I extend each time span by 1,000 initial periods that are dropped from the final sample. For clarification, after dropping the initial periods, the final sample consists of $N(T+1)$ observations. The additional time period per firm is used to generate lagged values of the output and input variables so that I can use NT observations for the estimation. Because the optimal investment decision is a convergent series, I approximate it considering only the first 1,001 terms of the sum.[6]

Because the length of the panel is the key determinant of the classification accuracy of my estimator, I focus on the analysis of panels with different time spans $T \in \{15, 25, 50\}$ and a fixed number of firms $N = 200$. The estimation is based on the identifying moments proposed by Gandhi, Navarro, and Rivers (2020). I use moments implied by the functional forms in the data generating process, i. e. a Cobb-Douglas production function, an AR(1) process for the persistent component of productivity, and $s_{i(j)t} = m_{i(j)t} - y_{i(j)t}$. However, I do not exploit the constant returns to scale assumption that would allow me to obtain both output elasticities from the share regression. My analysis consists of two parts. First, I compare the accuracy of my proposed estimator for the number of latent groups for two specifications of the penalty term taken from Su, Shi, and Phillips (2016). Second, I assume that $J = J^0 = 3$ is known and compare the finite sample performance of my post-lasso estimator (Post-Lasso) to an infeasible oracle estimator (Oracle) that exploits the true group membership of each firm.[7] For brevity, I restrict my analysis to results for $\lambda = T^{-0.25}$.[8] All reported results

[6] Formally, I split the series into two parts:

$$I^*_{i(j)t} \approx c_{i(j)}\Big(\sum_{\tau=0}^{1000} a_{i(j)\tau} + \sum_{\tau=1001}^{\infty} a_{i(j)\tau}\Big),$$

where $c_{i(j)}$ is a factor in front of the series and $a_{i(j)\tau}$ are the terms of the series. I assume that the first 1,001 terms are sufficient to approximate the optimal level of investment and thus that all remaining terms are negligible small.

[7] For clarification, Oracle is simply the estimator of Gandhi, Navarro, and Rivers (2020) applied separately to each latent group.

[8] In a preliminary analysis, I tried different candidate values for $\lambda \in \{T^{-a} : a \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}\}$.

are bases on 100 simulated samples for each T.

In the first part of my analysis, I consider $J \in \{1, \ldots, 5\}$ and the following specifications of the penalty terms: $p_{1,r}(N,T) := r\,(NT)^{-0.5}$ and $p_{2,r}(N,T) := r \log(\log(T))/T$, where $r \in \{0.25, 0.5, 0.75, 1\}$ is a finite sample adjustment factor.[9] To compare the selection accuracy, I compute the following quantities: expected value and probabilities to select exactly or at least $J^0 = 3$ latent groups. Table 5.4.2 reports the results. For $T \in \{25, 50\}$, all estimators perfectly predict the true number of latent groups. Only in the shortest panel $T = 15$, I find noticeable differences in the performance. While the estimator with $p_{1,r}(N,T)$ slightly overestimates the number of latent groups for small r, the estimator with $p_{2,r}(N,T)$ underestimates it for large r. However, given a suitable adjustment factor, estimators based on both specifications of the penalty term are able to perfectly predict the number of latent groups. In this particular example, my simulation experiments suggest a larger adjustment factor for $p_{1,r}(N,T)$ and a smaller factor for $p_{2,r}(N,T)$, e. g. $r = 1$ for the former and $r = 0.25$ for the latter.

Table 5.4.2: *Point Estimation of $\widehat{J}_p(\lambda)$*

T	Quantity	$p_{1,r}(N,T)$ with $r =$				$p_{2,r}(N,T)$ with $r =$			
		0.25	0.5	0.75	1	0.25	0.5	0.75	1
15	$\mathbb{E}[\widehat{J}_p(\lambda)]$	3.05	3.00	3.00	3.00	3.00	3.00	2.94	2.04
	$\Pr(\widehat{J}_p(\lambda) = 3)$	0.95	1.00	1.00	1.00	1.00	1.00	0.94	0.04
	$\Pr(\widehat{J}_p(\lambda) \geq 3)$	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.04
25	$\mathbb{E}[\widehat{J}_p(\lambda)]$	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
	$\Pr(\widehat{J}_p(\lambda) = 3)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\Pr(\widehat{J}_p(\lambda) \geq 3)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	$\mathbb{E}[\widehat{J}_p(\lambda)]$	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
	$\Pr(\widehat{J}_p(\lambda) = 3)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\Pr(\widehat{J}_p(\lambda) \geq 3)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: $N = 200$, $J^0 = 3$, $\lambda = T^{-0.25}$, and $J \in \{1, \ldots, 5\}$; $\widehat{J}_p(\lambda)$ is defined in (5.9) with $p_{1,r}(N,T) = r\,(NT)^{-0.5}$ and $p_{2,r}(N,T) = r \log(\log(T))/T$; Results are based on 100 simulated samples.

In the second part of my analysis, I take the true number of latent groups as given and compare the finite sample performance of Post-Lasso with a benchmark estimator Oracle. To evaluate the finite sample performance, I compute the following quantities: bias, standard deviation, root mean squared error, ratio of standard error and standard deviation, and coverage rates of confidence intervals with 95% nominal level. The statistics are computed separately for each latent group. The bias, standard deviation, and root mean squared error are all relative to the group specific true parameter value in percent. Because the performance of Post-Lasso is directly related to the classification accuracy in the first stage, I additionally report the proportion of correctly classified firms. To keep the analysis concise, I focus on the output elasticities $\boldsymbol{\gamma} := (\gamma_j)_{j=1}^3$ and $\boldsymbol{\beta} := (\beta_j)_{j=1}^3$, which are parameters of interest in most studies. Further, I follow Su, Shi, and Phillips (2016) and report aggregate rather than latent group specific statistics. Each aggregate statistic is the sum of the corresponding group specific statistics weighted by their relative occurrence. For instance, I compute the relative bias in percent as $100/N \sum_{j=1}^{J} N_j(\bar{\hat{\gamma}}_j - \gamma_j)/\gamma_j$, where $\bar{\hat{\gamma}}_j$ is the average estimate of γ_j over all

I found that the classification accuracy is very robust to different choices of λ.

[9] I also tried some of the penalty terms suggested by Bai and Ng (2002), but I did not find any better alternatives.

simulated samples. Table 5.4.3 reports the results. I start with the classification accuracy of Post-Lasso. The fractions of correctly classified firms in the first stage are always close to one, even in the shortest panel $T = 15$. As expected, the classification accuracy improves with increasing T. For $T = 50$, the fraction of correctly classified firms reaches one, which means that Post-Lasso and Oracle are identical for sufficiently large T. The very precise classification is also reflected in the finite sample properties. Post-Lasso's performance is close to that of the benchmark estimator Oracle and improves as T increases. This is expected because high classification accuracy means that there are very few misclassified firms, which in turn have little impact on the finite sample performance of Post-Lasso. In general, I find biases smaller than 1%, ratios of standard errors and standard deviation near one, and coverage rates close to their nominal values. The misclassification in the first step mainly results in larger dispersions. Therefore, the largest differences in performance are also apparent in the standard deviation and the root mean squared error.

Table 5.4.3: *Classification Accuracy and Point Estimation of β and γ*

T	Quantity	Post-Lasso		Oracle (Infeasible)	
		γ	β	γ	β
15	% of Correct Classification	0.9959		-	
	Bias (relative in %)	-0.0476	0.8605	-0.0439	-0.0491
	Standard Deviation (relative in %)	0.5304	4.3252	0.5117	3.6788
	Root Mean Squared Error (relative in %)	0.5301	4.4658	0.5136	3.6775
	Standard Error / Standard Deviation	0.9891	0.9513	1.0184	1.0112
	Coverage Rate (nominal level = 0.95)	0.9450	0.9340	0.9560	0.9410
25	% of Correct Classification	0.9995		-	
	Bias (relative in %)	0.0194	0.1695	0.0222	0.0216
	Standard Deviation (relative in %)	0.4087	2.8645	0.4053	2.7452
	Root Mean Squared Error (relative in %)	0.4104	2.8554	0.4068	2.7367
	Standard Error / Standard Deviation	0.9950	1.0349	1.0033	1.0662
	Coverage Rate (nominal level = 0.95)	0.9690	0.9560	0.9720	0.9590
50	% of Correct Classification	1.0000		-	
	Bias (relative in %)	-0.0083	-0.0260	-0.0083	-0.0260
	Standard Deviation (relative in %)	0.2696	2.0006	0.2696	2.0006
	Root Mean Squared Error (relative in %)	0.2689	2.0048	0.2689	2.0048
	Standard Error / Standard Deviation	1.0509	1.0339	1.0509	1.0339
	Coverage Rate (nominal level = 0.95)	0.9570	0.9500	0.9570	0.9500

Note: $N = 200$, $J = 3$, and $\lambda = T^{-0.25}$; Post-Lasso is defined in (5.4), Oracle exploits the true latent group membership of each firm; results are based on 100 simulated samples.

5.5 Empirical Illustration

It has become standard practice to estimate separate production functions for each industry. In most studies, industries are delineated by an industry classification, like ISIC, SIC, NACE, or NAICS.[10] The underlying assumption here is that firms operating in the same economic environment, like an industry, use the same production technology. Although the ex-ante classification by industry is very convenient and intuitive, the question arises whether this classification is sufficient to account for latent firm heterogeneity. I will

[10] Some recent and influential examples of these studies include Autor et al. (2020) and De Loecker, Eeckhout, and Unger (2020). The former use a 2-digit SIC and the latter use a 2- and 4-digit NAICS classification to define an industry.

demonstrate the usefulness of my estimator by investigating this question. More specifically, I will use my estimator to identify latent firm heterogeneity in the data and then analyze to which extent the identified group structure matches the industry classification.

I use a balanced sub sample of the Chilean panel data set used by Gandhi, Navarro, and Rivers (2020).[11] Their data set stems from the *Instituto Nacional de Estadística de Chile* and includes all manufacturing plants with more than ten employees in the five largest industrial sectors, i. e. food products (311), textiles (321), apparel (322), wood products (331), and fabricated metals (381), that were in operation between 1979 and 1996.[12] In addition to the output and input variables $(Y_{it}, K_{it}, L_{it}, M_{it}, S_{it})$, the data set contains further firm characteristics. I know whether a firm is an exporter, is an importer of intermediate goods, pays above median industry wages, or is an advertiser. Further, Y_{it} is measured as deflated revenue, K_{it} is the capital stock constructed by the perpetual inventory method, L_{it} is a weighted sum of unskilled and skilled number of workers, M_{it} is measured as the sum of several intermediate input expenditures, e. g. raw materials, energy, and services, and S_{it} is intermediate input expenditure relative to revenue.[13] My final sample consists of $N = 571$ Chilean firms in five industries that are consecutively observed for $T = 17$ years.

To identify latent firm heterogeneity, I consider the following time-homogeneous but possibly group-heterogeneous Cobb-Douglas production function in natural logs:

$$\begin{aligned} y_{i(j)t} &= \beta_{0j} + \beta_{1j} k_{i(j)t} + \beta_{2j} l_{i(j)t} + \beta_{3j} m_{it} + \omega_{i(j)t} + \epsilon_{i(j)t} \, , \\ \omega_{i(j)t} &= \delta_{0j} + \delta_{1j} \omega_{i(j)t-1} + \eta_{i(j)t} \, , \end{aligned} \tag{5.13}$$

where the subscript $i(j)$ indicates that a firm i belongs to a latent group j, $\omega_{i(j)t}$ is a persistent productivity shock that follows a stationary AR(1) process, $\eta_{i(j)t}$ is an unexpected innovation, and $\epsilon_{i(j)t}$ is an ex-post productivity shock. To estimate the latent group membership and the output elasticities, I use the identifying moments of Gandhi, Navarro, and Rivers (2020) adjusted to my functional form assumptions.[14]

Because the true number of latent groups J^0 is unknown, I have to estimate it. Further, I need to choose a suitable tuning parameter λ. To do both simultaneously, I use two information criteria: IC1 ($p(N,T) = (NT)^{-0.5} \approx 0.0101$) and IC2 ($p(N,T) = 0.25 \log(\log(T))/T \approx 0.0153$), with the corresponding specification of the penalty terms in parentheses. I compute both criteria for all combinations of $J \in \{1, \ldots, 8\}$ and $\lambda \in \{T^{-a} \colon a \in \{0.2, 0.25, 0.3, 0.35, 0.4\}\}$, where $J = 1$ refers to absence of latent firm heterogeneity. Figure 5.5.1 visualizes the results. Both information criteria suggest to estimate (5.13) with three latent groups and $a = 0.4$. Thus, there is strong evidence for the presence of a latent group structure, i. e. latent firm heterogeneity, as indicated by the significantly improved fit of the structural model (5.13) for $J > 1$. I follow the suggestion and choose $J = 3$ and $\lambda = 17^{-0.4} \approx 0.3220$ for the remaining part of the analysis.

Although I find that the estimated number of latent groups is smaller than the number of industry sectors, it is still unclear to which extent the estimated latent group structure matches the three-digit industry classification. The chord diagram in Figure 5.5.2 visualizes the matching. First, I find firms from all five industries in each of the three latent groups. Or, to say it conversely, I find firms from three latent groups in each of the five industries. Second, I find that firms from different industries split very unequally among the latent groups. Firms in industries 311 and 381 split mainly between two latent groups, whereas firms in the

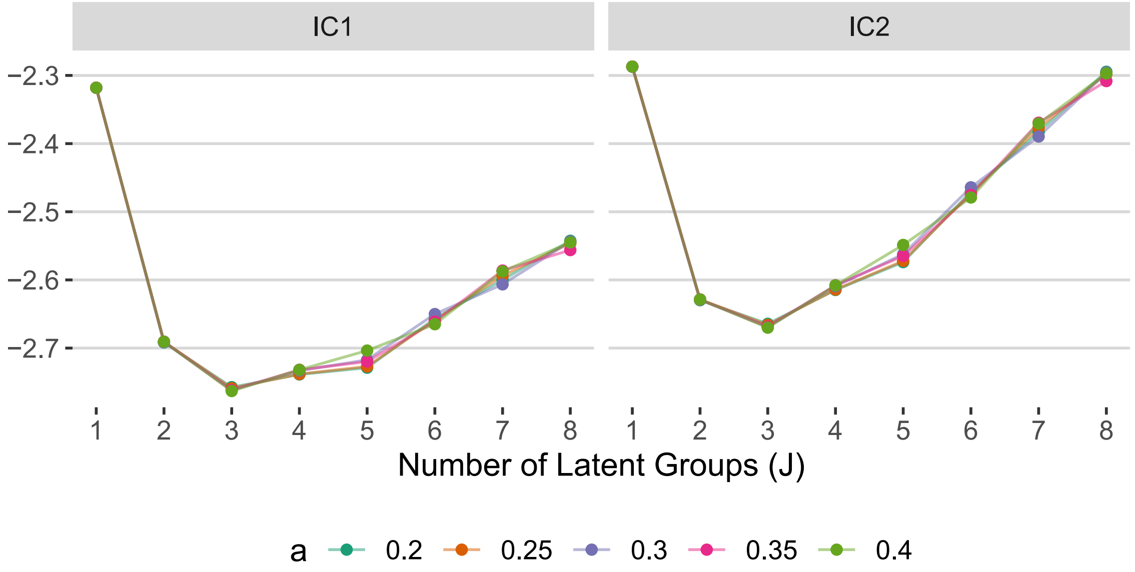
[11] Previous studies that also use the Chilean data set include Pavcnik (2002) and Levinsohn and Petrin (2003).

[12] The data are part of the replication package provided by the authors. *ISIC Rev. 2* industry classification.

[13] Gandhi, Navarro, and Rivers (2020) provide further details about the construction of their data set in footnote 42.

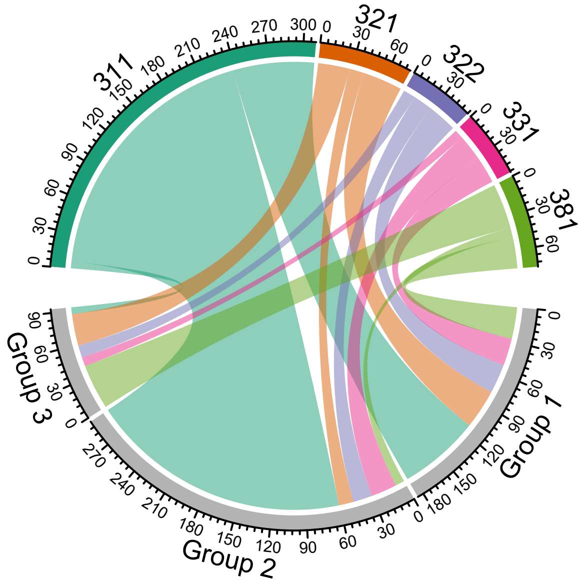
[14] These are exactly the moments defined at the end of Example 3 in Section 5.2.2.

Figure 5.5.1: *Information Criterion 1 and 2 for Different Values of λ and J*



Note: $N = 571$, $T = 17$, $\lambda = T^{-a}$, and $J \in \{1, \ldots, 8\}$; IC1 and IC2 are computed based on (5.8) with $p(N,T) = (NT)^{-0.5} \approx 0.0101$ and $p(N,T) = 0.25\log(\log(T))/T \approx 0.0153$, respectively.
Source: Gandhi, Navarro, and Rivers (2020).

Figure 5.5.2: *Relationship between Ex-Ante and Data-Driven Classification*



Note: $N = 571$, $T = 17$, $\lambda = 17^{-0.4} \approx 0.3220$, and $J = 3$; 311, 321, 322, 331, and 381 are industry sectors; Group 1–3 are the estimated latent groups; the smaller numbers between 0 and 300 refer to firms in each group; the links in different colors represent the connection between an industry sector and a latent group; colors are based on industry sectors.
Source: Gandhi, Navarro, and Rivers (2020).

other industries split more evenly between all three latent groups. For instance, about three quarters of the firms in industry 311 are assigned to Group 2. Interestingly, three quarters of the firms in Group 2 are also from industry 311. Thus, there is evidence that an ex-ante classification by industry sectors is not sufficient to fully address the latent firm heterogeneity in the data.[15]

Finally, I analyze to which extent the estimates from the ex-ante (Industry Classification) and data-driven classification (Post-Lasso) differ. I start with the evaluation of the model fit. To do this, I compare the mean squared residuals (MSR) for both classification approaches. The residuals are defined as $\hat{\eta}_{i(j)t} + \hat{\epsilon}_{i(j)t}$. The MSR for Industry Classification (≈ 0.0703) is larger than the MSR for Post-Lasso (≈ 0.0528). Thus, the data-driven classification improves the model fit and does so even with fewer estimated parameters. Next, I analyze the differences in the estimated output elasticities and how these differences translate into heterogeneity in total factor productivity (TFP).[16] I follow Olley and Pakes (1996) and estimate TFP in levels as $\exp(y_{i(j)t} - \hat{\beta}_{1j}k_{i(j)t} - \hat{\beta}_{2j}l_{i(j)t} - \hat{\beta}_{3j}m_{i(j)t})$. I analyze the heterogeneity in TFP through excluded firm characteristics. I use a pseudo-poisson model with the following conditional mean specification:

$$\exp(\pi_{1j}\,\text{trade}_{i(j)t} + \pi_{2j}\,\text{highwage}_{i(j)t} + \pi_{3j}\,\text{advertiser}_{i(j)t} + \alpha_{i(j)} + \gamma_t)\,, \tag{5.14}$$

where $\text{trade}_{i(j)t} := \max(\text{exporter}_{i(j)t}, \text{importer}_{i(j)t})$, $\text{highwage}_{i(j)t}$, and $\text{advertiser}_{i(j)t}$ are indicator variables equal to one, if firm $i(j)$ at time t engages in international trade, either through exporting and/or importing, pays above median wages, and is an advertiser, respectively, and $\alpha_{i(j)}$ and γ_t are firm and year fixed effects. Table 5.5.1 reports the estimates. I find sizable differences, relative to the magnitude of the standard errors, between the estimated output elasticities for Industry Classification and Post-Lasso. Nevertheless, there are some similarities between the estimates for Industry 311 and Group 2 and the estimates for Industry 381 and Group 3. This similarities might be explained by the large overlap in the ex-ante and data-driven classification of these firms. The estimated capital elasticities for Post-Lasso are substantially larger than those for Industry Classification whereas the ranges of the estimated labor and intermediate input elasticities are wider. The differences in the estimated output elasticities also lead, in some cases, to different conclusions about the heterogeneity in TFP. For Industry 331, I find negative but insignificant TFP differences between firms that engage in international trade and those who do not. The differences in all other industries are positive, but only for industry 321 significantly different from zero.[17] The remaining conclusions about heterogeneity in TFP remain qualitatively the same for both classification approaches. For instance, I find that firms with above median wages are significantly more productive, while firms that advertise are not significantly different from those that do not advertise.

[15] This is in line with Kasahara, Schrimpf, and Suzuki (2017) who report substantial firm heterogeneity even in narrowly defined industries using Japanese data.

[16] In addition to output elasticities, TFP is another quantity of interest in some empirical studies. More specifically, these studies aim to better understand the persistent differences in TFP that are frequently reported in related studies (see Bartelsman and Doms 2000 and Syverson 2011 for comprehensive overviews).

[17] This is surprising, as there is wide agreement that firms that engage in international trade, such as exporters, are more productive. Two reasons mentioned by Bernard and Wagner (1997) and Bernard and Jensen (1999) are self selection into exporting, e. g. because entering foreign markets is costly, and learning by exporting, e. g. knowledge spillovers.

Table 5.5.1: *Estimation Results: Output Elasticities and Heterogeneity in TFP*

	Industry Classification					Post-Lasso		
	311	321	322	331	381	Group 1	Group 2	Group 3
Output Elasticities:								
Capital	0.163	0.096	0.149	0.095	0.186	0.158	0.137	0.199
	(0.012)	(0.027)	(0.033)	(0.025)	(0.034)	(0.014)	(0.010)	(0.023)
Labor	0.182	0.346	0.249	0.284	0.472	0.299	0.155	0.503
	(0.018)	(0.037)	(0.046)	(0.037)	(0.056)	(0.022)	(0.015)	(0.041)
Intermediates	0.674	0.494	0.550	0.583	0.439	0.557	0.720	0.386
	(0.003)	(0.007)	(0.007)	(0.007)	(0.006)	(0.003)	(0.002)	(0.005)
Heterogeneity in Total Factor Productivity:								
Trade	0.027	0.047	0.043	-0.036	0.028	0.042	0.023	0.070
	(0.016)	(0.026)	(0.032)	(0.039)	(0.025)	(0.024)	(0.011)	(0.036)
Wages > median	0.042	0.053	0.085	0.070	0.070	0.043	0.048	0.076
	(0.008)	(0.024)	(0.033)	(0.033)	(0.027)	(0.016)	(0.009)	(0.039)
Advertiser	-0.008	0.008	-0.060	-0.003	-0.027	-0.031	-0.003	-0.022
	(0.009)	(0.023)	(0.035)	(0.019)	(0.022)	(0.015)	(0.008)	(0.028)

Note: Output Elasticities are estimated based on the identifying moments of Gandhi, Navarro, and Rivers (2020) adjusted to the functional form assumptions in (5.13); asymptotic standard errors obtained by (5.5) in parentheses; Industry Classification and Post-Lasso refer to estimates for ex-ante classification by industry sector and data-driven classification, respectively; Heterogeneity in Total Factor Productivity is analyzed through excluded firm characteristics using a poisson model with conditional mean (5.14); TFP is estimated in levels as $\exp(y_{i(j)t} - \hat{\beta}_{1j} k_{i(j)t} - \hat{\beta}_{2j} l_{i(j)t} - \hat{\beta}_{3j} m_{i(j)t})$; pseudo-poisson estimates can be interpreted as semi-elasticities and are relative to firms that do not export or import, pay below median wages and do not advertise; for instance, a firm in industry 311 engaged in international trade is $100(\exp(0.027) - 1)\% \approx 2.7368\%$ more productive than a firm with the same characteristics that does not engage; robust standard errors in parentheses.

Source: Gandhi, Navarro, and Rivers (2020).

5.6 Concluding Remarks

I proposed a fully data-driven estimation procedure for production functions with latent group structures. My approach combines recent identification strategies with the classifier-Lasso of Su, Shi, and Phillips (2016). Simulation experiments show good finite sample properties and the practical relevance is illustrated by an empirical example.

My estimation procedure assumes that the model parameters are time-homogeneous within a latent group. This assumption could be relaxed by introducing smooth time-varying model parameters as in Su, Wang, and Jin (2019). Another possible extension could be to think of the model parameters not as group-specific fixed constants, but as firm-specific random coefficients with unknown distribution function. Group-specific estimates of the model parameters could then be considered as representatives point, i. e. support points, of this unknown distribution. The challenge here would be to derive a limiting theory that incorporates the resulting approximation error.

References

Ackerberg, Daniel, C. Lanier Benkard, Steven Berry, and Ariel Pakes. 2007. "Chapter 63: Econometric Tools for Analyzing Market Outcomes," edited by James J. Heckman and Edward E. Leamer, 6:4171–4276. Handbook of Econometrics. Elsevier.

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer. 2015. "Identification Properties of Recent Production Function Estimators." *Econometrica* 83 (6): 2411–2451.

Aguirregabiria, Victor. 2019. "Empirical Industrial Organization: Models, Methods, and Applications." *University of Toronto (Version: December 2019).*

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58 (2): 277–297.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." *The Quarterly Journal of Economics* 135 (2): 645–709.

Bai, Jushan, and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70 (1): 191–221.

Bartelsman, Eric J., and Mark Doms. 2000. "Understanding Productivity: Lessons from Longitudinal Microdata." *Journal of Economic Literature* 38 (3): 569–594.

Bernard, Andrew B., and J. Bradford Jensen. 1999. "Exceptional exporter performance: cause, effect, or both?" *Journal of International Economics* 47 (1): 1–25.

Bernard, Andrew B., and Joachim Wagner. 1997. "Exports and Success in German Manufacturing." *Weltwirtschaftliches Archiv* 133 (1): 134–157.

Blundell, Richard, and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics* 87 (1): 115–143.

———. 2000. "GMM Estimation with persistent panel data: an application to production functions." *Econometric Reviews* 19 (3): 321–340.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica* 83 (3): 1147–1184.

Browning, Martin, and Jesus M. Carro. 2014. "Dynamic binary outcome models with maximal heterogeneity." *Journal of Econometrics* 178 (2): 805–823.

Cheng, Xu, Frank Schorfheide, and Peng Shao. 2019. "Clustering for Multi-Dimensional Heterogeneity." *Working Paper.*

Cobb, Charles W., and Paul H. Douglas. 1928. "A Theory of Production." *The American Economic Review* 18 (1): 139–165.

Collard-Wexler, Allan, and Jan De Loecker. 2016. "Production Function Estimation with Measurement Error in Inputs." *Working Paper.*

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. “The Rise of Market Power and the Macroeconomic Implications.” *The Quarterly Journal of Economics* 135 (2): 561–644.

De Loecker, Jan, and Frederic Warzynski. 2012. “Markups and Firm-Level Export Status.” *American Economic Review* 102 (6): 2437–2471.

Gandhi, Amit, Salvador Navarro, and David A. Rivers. 2020. “On the Identification of Gross Output Production Functions.” *Journal of Political Economy* 128 (8): 2973–3016.

Griliches, Zvi, and Jacques Mairesse. 1999. “Production Functions: The Search for Identification.” In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium,* edited by Steinar Strøm, 169–203. Econometric Society Monographs. Cambridge University Press.

Hansen, Lars Peter. 1982. “Large Sample Properties of Generalized Method of Moments Estimators.” *Econometrica* 50 (4): 1029–1054.

Hansen, Lars Peter, and Kenneth J. Singleton. 1982. “Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models.” *Econometrica* 50 (5): 1269–1286.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical learning with sparsity: the lasso and generalizations.* CRC press.

Humphrey, Thomas M. 1997. “Algebraic production functions and their uses before Cobb-Douglas.” *Federal Reserve Bank of RichmondEconomic Quarterly* 83 (1): 51–83.

Kasahara, Hiroyuki, Paul Schrimpf, and Michio Suzuki. 2017. “Identification and Estimation of Production Function with Unobserved Heterogeneity.” *Working Paper.*

Kasahara, Hiroyuki, and Katsumi Shimotsu. 2009. “Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices.” *Econometrica* 77 (1): 135–175.

Levinsohn, James, and Amil Petrin. 2003. “Estimating Production Functions Using Inputs to Control for Unobservables.” *The Review of Economic Studies* 70 (2): 317–341.

Lin, Chang-Ching, and Serena Ng. 2012. “Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown.” *Journal of Econometric Methods* 1 (1): 42–55.

Marschak, Jacob, and William H. Andrews. 1944. “Random Simultaneous Equations and the Theory of Production.” *Econometrica* 12 (3 & 4): 143–205.

McElroy, Marjorie B. 1987. “Additive General Error Models for Production, Cost, and Derived Demand or Share Systems.” *Journal of Political Economy* 95 (4): 737–757.

Newey, Whitney K., and Kenneth D. West. 1987. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica* 55 (3): 703–708.

Olley, G. Steven, and Ariel Pakes. 1996. “The Dynamics of Productivity in the Telecommunications Equipment Industry.” *Econometrica* 64 (6): 1263–1297.

Pavcnik, Nina. 2002. “Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants.” *The Review of Economic Studies* 69 (1): 245–276.

Sarafidis, Vasilis, and Neville Weber. 2015. “A Partially Heterogeneous Framework for Analyzing Panel Data.” *Oxford Bulletin of Economics and Statistics* 77 (2): 274–296.

Shenoy, Ajay. 2020. “Estimating the Production Function Under Input Market Frictions.” *The Review of Economics and Statistics:* 1–45.

Su, Liangjun, Zhentao Shi, and Peter C. B. Phillips. 2016. “Identifying Latent Structures in Panel Data.” *Econometrica* 84 (6): 2215–2264.

Su, Liangjun, Xia Wang, and Sainan Jin. 2019. “Sieve Estimation of Time-Varying Panel Data Models With Latent Structures.” *Journal of Business & Economic Statistics* 37 (2): 334–349.

Sun, Yixiao. 2005. “Estimation and inference in panel structure models.” *Working Paper.*

Syverson, Chad. 2001. “Market Structure and Productivity.” Ph.D. Dissertation, University of Maryland.

———. 2011. “What Determines Productivity?” *Journal of Economic Literature* 49 (2): 326–365.

van Biesebroeck, Johannes. 2007. “ROBUSTNESS OF PRODUCTIVITY ESTIMATES.” *The Journal of Industrial Economics* 55 (3): 529–569.

Weiszfeld, Endre. 1937. “Sur le point pour lequel la somme des distances de n points donnés est minimum.” *Tohoku Mathematical Journal, First Series* 43:355–386.

White, Halbert. 1980. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica* 48 (4): 817–838.

Eidesstattliche Versicherung

Ich, Daniel Czarnowkse, versichere an Eides statt, dass die vorliegende Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der “Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf” erstellt worden ist.

Düsseldorf, 07.06.2021

Daniel Czarnowske