Nonlinear Panel Data Models with High-Dimensional Fixed Effects

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:	Amrei Luise Stammann, M. Sc.
	geboren am 18.05.1989 in Engelskirchen
Erstgutachter:	Prof. Dr. Florian Heiß
Zweitgutachter:	Prof. Dr. Joel Stiebale
Abgabedatum:	24. September 2019

Acknowledgment

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

First of all, I would like to thank my supervisor Florian Heiss for the great opportunity to write my thesis at his chair and his support from the beginning on. Thanks to him, I discovered the exciting topic of nonlinear models with high-dimensional fixed effects and my passion for *R*-programming. He always gave me the freedom to pursue my research interests, which contributed greatly to my personal and academic development. Secondly, I thank my second supervisor, Joel Stiebale, for his helpful comments and organizing the Empirical Research Group.

Moreover, I would also like to thank my co-authors – Daniel Czarnowske, Florian Heiss, Julian Hinz, Daniel McFadden, and Joschka Wanner – for the pleasant and inspiring cooperation.

In addition, the quality of some parts of this thesis has been greatly improved by valuable comments from a large number of people. Many thanks for that to: Tanmay Belavadi, Daniel Brunner, Daniel Czarnowske, Miriam Frey, Simen Gaure, Mario Larch, Maximilian Osterhaus, André Romahn, Yuta Watabe, and Thomas Zylkin. I also had the chance to participate in several conferences and seminars, where I received many helpful inspirations and suggestions.

Finally, a lot of thanks to my colleagues, who made the daily routine enjoyable through their friendly and helpful manner. Special thanks to Daniel for his support throughout this long and difficult process. His love and encouragement has always given me the strength to continue my thesis, even when I suffered several setbacks during the writing.

Contents

List of Abbreviations

ABC	Analytical Bias Correction
APE	Average Partial Effect
BCL	Analytically Bias-Corrected UCL
CES	Constant Elasticity of Substitution
CP	Coverage Probability
CML	Multinomial Logit Estimator
CL	Conditional Logit Estimator
CLsub	Conditional Logit Estimator Based on Random Subsets
CU	Currency Union
EMU	European Monetary Union
FTA	Foreign Trade Agreement
FWL	Frisch and Waugh (1933) and Lovell (1963)
GDP	Gross Domestic Product
GLM	Generalized Linear Model
GSOEP	German Socio Economic Panel
HMR	Helpman, Melitz, and Rubinstein (2008)
IPP	Incidental Parameters Problem
LPM	Linear Probability Model
MAP	Method of Alternating Projections
MLE	Maximum Likelihood Estimator
OLS	Ordinary Least Squares Estimator
PPML	Pseudo-Poisson MLE
PUMA	Public Use Microdata Area
RMSE	Root Mean Squared Error
SD	Standard Deviation
SE	Standard Error
SPJ	Split-Panel Jackknife Bias Correction
UCL	Unconditional Fixed Effects Logit Estimator
WTO	World Trade Organization

List of Figures

List of Tables

Chapter 1

Introduction

The careful handling of unobserved heterogeneity – such as individual or time specific effects – is an essential concern in econometrics for causal analyses. If these unobserved effects are related to any explanatory variable, neglecting them causes an omitted variables bias. A great advantage of panel data over pure cross-sections is that they allow to fully control for these unobserved effects and thus offer new possibilities to researchers that go beyond proxy variable and instrumental variable approaches. Another benefit is the possibility to study different sources of persistence (see chapter 1.2 in Baltagi 2013 and Hsiao 2014 for a comprehensive list of further advantages). For example, Roberts and Tybout (1997) and Bernard and Jensen (2004) employed dynamic discrete choice models to disentangle the drivers behind firms' exporting persistence, such as sunk costs and unobserved plant heterogeneity. Within panel data estimators, fixed effects estimators are very popular, because unlike random-effects estimators, they do not impose any distributional assumption on the unobserved heterogeneity. A flourishing part of the theoretical econometric literature is in particular concerned with nonlinear fixed effects models. However, their application confronts practitioners with several problems that will be discussed below.

To illustrate some typical problems, let us consider a popular class of binary choice estimators to control for individual specific unobserved heterogeneity. So-called conditional logit estimators, such as Rasch (1960), Andersen (1970), Chamberlain (1980), and Honoré and Kyriazidou (2000), are consistent under an asymptotic framework, where the number of time periods T is held fixed and the number of individuals N grows. They achieve the desirable fixed T consistency property by conditioning on sufficient statistics such that the likelihood function does not dependent on the individual effects anymore. However, this transformation trick is also one of the major disadvantages of conditional logit estimators. Removing the unobserved heterogeneity precludes the estimation of partial effects, which are important to obtain interpretable values for the ceteris paribus effects on the response probability.

Another limitation of this transformation trick is that sufficient statistics cannot be found for all nonlinear panel estimators, such as probit estimators (see Hahn and Newey 2004; Arellano and Hahn 2007; Fernández-Val and Weidner 2018a).

An alternative approach is a maximum likelihood estimator which jointly estimates the fixed effects and the structural parameters. In the following this estimator is denoted as *nonlinear fixed effects estimator*. The problem associated with this estimator is that, under fixed T asymptotics, it is affected by the so-called *incidental parameters problem* (see Neyman and Scott 1948). Intuitively, the fixed effects are estimated with noise because only a few number of observations contribute to their identification. Due to the nonlinear nature of the estimator, the estimation noise contaminates the estimates of the structural parameters, which is reflected in a bias.

Motivated by the increasing availability of large panel data sets, recent literature gradually switches to asymptotic approximations that require both panel dimensions (usually N and T) to grow with the sample size, i.e. $N, T \to \infty$. Under this asymptotic framework, the nonlinear fixed effects estimator is consistent, but shows a bias in its asymptotic distribution that can be reduced substantially by bias corrections (see Fernández-Val and Weidner 2018a for an overview). The corrected estimators also allow to estimate average partial effects and show desirable finite sample properties in simulations. Since bias corrections can be developed for a variety of nonlinear models (including dynamic models) with different error components, they are broadly applicable.

Despite their wide range of possible applications, nonlinear fixed effects estimators still receive little attention in empirical research and are often substituted by linear ones. As frequently made statements by empirical researchers let suggest, there are two main reasons. The first one is a widespread misbelief that the estimation of nonlinear models with high-dimensional fixed effects is infeasible (see among others Glick and Rose 2016; Markussen and Røed 2017). Another frequently mentioned reason is that researchers want to avoid the incidental parameters problem (see among others Markussen and Røed 2017; Ullman 2017; Sanches, Silva Junior, and Srisuma 2018; Popov and Zaharia 2019).

This thesis is intended to draw empirical researchers' attention to nonlinear fixed effects estimators and to facilitate their applicability. For this purpose, the strand of econometric literature on bias corrections is linked to computational advances. This makes it possible to estimate nonlinear models even with many observations and high-dimensional fixed effects, which is more and more required due to the increasing magnitude of panel data sets. This thesis also contributes to the literature on bias corrections itself by providing further insights on finite sample properties of various corrections and proposing novel corrections for special two- and three-way fixed effects models required in international trade. Further, I offer the

corresponding software routines, *bife* and *alpaca*, to make the methods presented in this thesis ready to use. In the following, I give a detailed overview about the different contributions of this thesis.

In chapter 2, co-authored with Florian Heiss and Daniel McFadden, we derive a computationally efficient maximum likelihood algorithm to estimate logit models with individual fixed effects. For clarification, this algorithm corresponds to a nonlinear fixed effects estimator. The methodological core of this algorithm is the application of the Frisch-Waugh-Lovell (Frisch and Waugh 1933; Lovell 1963) (FWL) theorem in each iteration of the Newton-Raphson optimization routine and establishes the basis for several extensions presented in the subsequent chapters of this thesis. More precisely, since Newton's update is the solution of a weighted regression, the updates of the structural parameters can be separated from the high-dimensional fixed effects. The corresponding projection matrix is sparse and thus allows to derive a straightforward scalar expression, we refer to as *pseudo-demeaning*. The beauty of this approach is its link to the within transformation that is well-known in the context of linear fixed effects models. Another aspect we address, is that this fixed effects estimator suffers from a bias in its asymptotic distribution stemming from the need to estimate incidental parameters. To mitigate this bias, we combine our algorithm with an analytical bias correction proposed by Fernández-Val (2009). Moreover, we propose a novel hybrid approach that allows to estimate average partial effects for conditional logit estimators. In an extensive simulation-based comparison between (bias-corrected) fixed effects estimators and conditional logit estimators, we find that the former are promising candidates both in terms of their statistical properties and in terms of their computation times.

Chapter 3 extends the pseudo-demeaning approach introduced in chapter 2 to the class of nonlinear generalized linear models (GLMs) with multi-way fixed effects. I derive a fast and memory efficient maximum likelihood algorithm by combining the insights on the sparse projection matrix from chapter 2 and the *method of alternating projections* (MAP) tracing back to Neumann (1950) and Halperin (1962). The latter is required, since in nonlinear multi-way fixed effects models the projection matrix loses its sparse structure and thus prevents the formulation of an efficient scalar expression to partial out the fixed effects from Newton's update. The algorithm is highly compatible with the needs of empirical research, because it allows to estimate GLMs with many observations and high-dimensional fixed effects on a standard desktop computer and is directly applicable to unbalanced data. Moreover, it can be easily combined with jackknife bias corrections to mitigate the incidental parameter bias that appears frequently in GLMs with fixed effects. I highlight the relevance of my algorithm by applying it to an example from international trade, where a maximum likelihood estimator that is able to cope with high-dimensional multi-

way fixed effects is urgently needed. The workhorse model in this discipline is the structural gravity model, which is estimated by the pseudo-poisson maximum likelihood estimator (PPML) and, in case of panel data, includes two or three sets of high-dimensional fixed effects.

In chapter 4, co-authored with Daniel Czarnowske, we use the example of binary choice models with individual and time fixed effects to show how the Newton-Raphson pseudo-demeaning algorithm and MAP established in chapter 3, can be adapted to dramatically accelerate analytical bias corrections. Moreover, we study the finite sample properties of several types of bias corrections by conducting extensive simulation experiments. On the one hand, we consider bias corrections that have been proposed by Fernández-Val and Weidner (2016) but not analyzed so far. On the other hand, we introduce different patterns of unbalancedness, giving more realistic insights about the usability of bias corrections for applied work. This aspect has received little attention in the literature so far. Most notably, we find that analytical bias corrections outperform jackknife approaches irrespective of the missing data pattern. Finally, we provide an extended empirical illustration from labor economics to highlight the usefulness of bias corrections in combination with our suggested algorithms. In spirit of Fernández-Val (2009), we estimate a dynamic fixed effects probit model to investigate the inter-temporal labor force participation of women. Contrary to him, we utilize a large scale data set of 10,712 women observed between 1984 and 2013.

Chapter 5, co-authored with Julian Hinz and Joschka Wanner, utilizes and extends the contributions made in chapter 4 in order to estimate the extensive margin of trade. We theoretically motivate a dynamic empirical model by combining a heterogeneous firms model of international trade with bounded productivity and features from the firm dynamics literature to derive expressions for an exporting country's participation in a specific destination market in a given period. Our preferred econometric specification demands a dynamic probit estimator with three sets of high-dimensional fixed effects and thus causes computational and econometric issues. The latter is associated with the incidental parameter bias. Thus, we characterize new analytical and jackknife bias corrections and show how these can be efficiently implemented in the context of high-dimensional fixed effects. Extensive simulation experiments confirm the desirable statistical properties of the bias-corrected estimators. In our empirical application, we demonstrate that controlling for both sources of persistence – market entry dynamics and unobserved heterogeneity – and taking the incidental parameter bias into account makes a substantial difference. To be more specific, we find that among the most frequently studied potential determinants (joint WTO membership, common regional trade agreement, and shared currency), only sharing the same currency has a significant

impact on whether two countries trade with each other or not.

In chapter 6, I briefly summarize the contributions of this thesis and give some outlooks for future research. The user manuals of the *R*-packages *bife* and *alpaca*, developed during this thesis, are listed in the appendices A and B, respectively.

References

Andersen, Erling Bernhard. 1970. “Asymptotic properties of conditional maximum-likelihood estimators.” *Journal of the Royal Statistical Society. Series B:* 283–301.

Arellano, Manuel, and Jinyong Hahn. 2007. “Understanding bias in nonlinear panel models: Some recent developments.” *Econometric Society Monographs* 43:381.

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

Bernard, Andrew B., and J. Bradford Jensen. 2004. “Why some firms export.” *Review of Economics and Statistics* 86 (2): 561–569.

Chamberlain, Gary. 1980. “Analysis of Covariance with Qualitative Data.” *Review of Economic Studies* 47:225–238.

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

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.

Frisch, Ragnar, and Frederick V. Waugh. 1933. “Partial Time Regressions as Compared with Individual Trends.” *Econometrica* 1 (4): 387–401.

Glick, Reuven, and Andrew K. Rose. 2016. “Currency unions and trade: A post-EMU reassessment.” *European Economic Review* 87:78–91.

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 (1-2): 96–99.

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.

Lovell, Michael C. 1963. “Seasonal adjustment of economic time series and multiple regression analysis.” *Journal of the American Statistical Association* 58 (304): 993–1010.

Markussen, Simen, and Knut Røed. 2017. "The gender gap in entrepreneurship–The role of peer effects." *Journal of Economic Behavior & Organization* 134:356–373.

Neumann, John von. 1950. "Functional Operators. Vol. II. The geometry of orthogonal spaces, volume 22 (reprint of 1933 notes) of Annals of Math." *Studies. Princeton University Press.*

Neyman, Jerzy, and Elizabeth L Scott. 1948. "Consistent estimates based on partially consistent observations." *Econometrica* 16 (1): 1–32.

Popov, Alexander, and Sonia Zaharia. 2019. "Credit market competition and the gender gap in labor force participation: Evidence from local markets." *European Economic Review* 115:25–59.

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

Roberts, Mark J., and James R. Tybout. 1997. "The decision to export in Colombia: An empirical model of entry with sunk costs." *The American Economic Review:* 545–564.

Sanches, Fabio, Daniel Silva Junior, and Sorawoot Srisuma. 2018. "Banking privatization and market structure in Brazil: a dynamic structural analysis." *The RAND Journal of Economics* 49 (4): 936–963.

Ullman, Darin F. 2017. "The effect of medical marijuana on sickness absence." *Health economics* 26 (10): 1322–1327.

Chapter 2

Estimation of Fixed Effects Logit Models with Large Panel Data

Co-authored with Florian Heiss and Daniel McFadden

2.1 Introduction

The recent availability of long microeconomic panels like the Panel Study for Income Dynamics constitutes new computational challenges for the estimation of common econometric models. One of these is the logit model with individual fixed effects which is referred to hereinafter as the fixed effects logit model. The fixed effects logit model is a popular specification for analyzing panel data of binary variables, since it allows for unobserved individual heterogeneity like the variation in tastes with an arbitrary distribution.

There are two established approaches for the estimation of fixed effects logit models. On the one hand, it is possible to carry out a standard maximum likelihood estimation in which the regressor set is extended by one dummy variable per cross-sectional unit. We call this estimator the unconditional logit estimator and abbreviate it with UCL. The other estimator, a conditional logit estimator (CL), concentrates the individual heterogeneity out of the likelihood function by conditioning on a sufficient statistic. Both estimators suffer from substantial drawbacks which this article is intended to address.

UCL can become computationally challenging when the number of fixed effects N is large since it requires the computation and inversion of a large Hessian. Apart from the computational challenge, the parameters of most nonlinear fixed effects models suffer from the incidental parameters problem (IPP), which is reflected in a bias, first noted by Neyman and Scott (1948). This incidental parameters bias can be especially severe in models with a small number of observations T per individual. The reason is that only few observations contribute to the estimation of the fixed effects leading to noisy estimates. Due to the nonlinear nature of the logit model, the estimation noise of the fixed effects also contaminates the estimates of the structural parameters. Thus, UCL is inconsistent under fixed T asymptotics (see Arellano and Hahn 2007; Fernández-Val and Weidner 2018a). Even increasing T does not necessary solve the incidental parameters bias because fixed effects estimators are asymptotically biased even if T grows at the same rate as N (see Hahn and Newey 2004).

CL has been derived by Rasch (1960) and Andersen (1970) and later generalized by Chamberlain (1980) as a solution to IPP. They show that CL is a fixed T consistent estimator for structural parameters. However it is not clear how interpretable values, such as average partial effects (APEs), can be estimated since CL does not deliver estimates of the fixed effects (see Hahn and Newey 2004; Arellano and Hahn 2007; Fernández-Val and Weidner 2018a).[1] Another drawback is that CL is computationally very costly if T is large. Even if we use a more efficient recursion

1. Often partial effects are also called marginal or ceteris paribus effects.

method proposed by Gail, Lubin, and Rubinstein (1981), the computational burden increases roughly quadratic in T which makes CL infeasible for panels with large time horizons.

The contributions of our article are manifold. We address the aforementioned problems of the different estimators. With respect to the UCL estimator this means that we first derive an intuitive and efficient algorithm based on the Frisch-Waugh-Lovell (FWL) theorem (Frisch and Waugh 1933; Lovell 1963).[2] We call this approach *pseudo-demeaning* because of its similarity to the within transformation in linear fixed effects models. The remaining challenge of UCL, which is the incidental parameter bias, is reduced by combining the pseudo-demeaning algorithm with an analytical bias correction proposed by Fernández-Val (2009).[3] To tackle the computational burden of CL for large T, we introduce a new estimator which we refer to as CLsub. This estimator is based on an estimator that has been designed by McFadden (1978) to overcome the curse of dimensionality problem in multinomial logit models. CLsub is essentially an adaption of this estimator to a binary dependent variable. The idea is to reduce the computational costs by using only a subset of all permutations of the observed choice sequence in the estimation routine. Furthermore, we propose a novel approach that uses estimates of the fixed effects obtained by an offset algorithm to compute APEs for conditional logit estimators. We also present an appropriate formula, based on a concentrated delta method, which can be used for conditional and unconditional logit estimators to calculate standard errors for APEs without having to use computationally demanding bootstrap methods. In extensive simulation experiments we finally investigate the finite sample properties of the different estimators with respect to structural parameters and APEs. In addition, we empirically verify the theoretical computational complexities that we have derived in advance. Finally, we use an empirical example from labor economics, to demonstrate a relevant field for the application of our pseudo-demeaning algorithm. In this example, T is even so large that conditional logit estimators are not feasible, whereas our pseudo-demeaning approach can easily estimate the model. In order to make our (bias-corrected) pseudo-demeaning algorithm accessible for applied work, we offer it in the *R*-package *bife*.[4]

2. An alternative approach exploits the specific sparse structure of the Hessian (see Hall 1978; Prentice and Gloeckler 1978; Chamberlain 1980; Greene 2004).

3. A comprehensive overview on different bias correction approaches is given by Arellano and Hahn (2007) and Fernández-Val and Weidner (2018a). For our purposes only ex-post bias corrections are of interest, since they can be conveniently combined with our pseudo-demeaning approach. They can be analytical (e.g Hahn and Newey 2004; Fernández-Val 2009) or based on re-sampling methods (e.g Hahn and Newey 2004; Dhaene and Jochmans 2015).

4. The package can estimate structural and incidental parameters, as well as average partial effects of fixed effects logit and probit models and provides the analytical bias correction of Fernández-Val (2009). The package also offers the corresponding standard errors. https://cran.r-project.org/web/packages/bife/.

Our simulation experiments confirm the findings of Greene (2004), who reports large distortions in the UCL estimator of the structural parameters for small T. Furthermore, the bias correction substantially reduces this distortion and, for sufficiently large values of T, it has similar desirable properties like the fixed T consistent CL estimator. Similar results regarding BCL are presented by Fernández-Val (2009), who focuses on probit models. Besides, our results, that UCL shows only little distortions in the APEs even for small values of T and that the bias correction works similarly well, are also in line with Fernández-Val (2009). Furthermore, we find that the CLsub estimator provides consistent estimates for the structural parameters only if the subset is large enough relative to the entire permutation set. However, compared to CL it is less efficient. The simulation results also demonstrate that estimates of the APEs obtained by conditional logit estimators can suffer from severe biases if the contributions of the fixed effects are ignored in their calculation. Our new strategy to estimate APEs for conditional logit estimators based on an offset algorithm is a substantial improvement over the aforementioned approach. However, even CL, which has the best properties among all conditional logit estimators, is slightly outperformed by UCL and BCL in estimating APEs. Moreover, the simulation experiments verify that the computational burden of UCL and BCL, both combined with the pseudo-demeaning approach, increase linearly with T, whereas the burden of recursive CL increases quadratically, which makes a dramatic difference for large T. Besides, we demonstrate that CLsub can further reduce the computation time if the used subset of permutations is small. Considering the trade-off between statistical properties and computation time, we conclude that there is no advantage of using CLsub over CL. Especially if T is large, the speed advantage of a small subset comes at costs of high biases. Overall, UCL and BCL offer a clear computation time advantage over CL, which is particularly evident for samples with large T. Apart from that, (bias-corrected) UCL is also a promising candidate for practical applications in terms of statistical properties, especially when APEs are of main interest.

The paper is organized as follows. Section 2.2 presents a short recap of the fixed effects logit models along with its basic estimators. In section 2.3 we derive the pseudo-demeaning approach and present the entire Newton-Raphson pseudo-demeaning optimization routine. Section 2.4 introduces CLsub. It follows the description of different offset algorithms and the concentrated delta method in section 2.5. In section 2.6, the design and results of a series of Monte Carlo simulations are presented before section 2.7 demonstrates an empirical example. Finally, section 2.8 concludes.

2.2 The Fixed Effects Logit Model and Basic Estimators

2.2.1 The Fixed Effects Logit Model

For the sake of notational simplicity, we assume a balanced panel of $i = 1,\dots,n$ individuals observed for $t = 1,\dots,T$ time periods.[5] Suppose we observe a binary dependent variable y_{it}, such that $y_{it} = 1$ if an event occurs and $y_{it} = 0$ if it does not occur. Let $N = \sum_{i=1}^{n} \mathbf{1}[0 < \sum_{t=1}^{T} y_{it} < T]$ be the number of cross-sectional units for which y_{it} varies over time, where $\mathbf{1}[\cdot]$ is an indicator function. The $n - N$ individuals without varying y_{it} do not contribute to the identification and can be dropped from the analysis without affecting the estimator of the structural parameters. We refer to these observations as perfectly classified.

The fixed effects logit model is defined by the joint probability of observing y_{it}

$$f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) = p_{it}^{y_{it}}(1 - p_{it})^{1-y_{it}} \tag{2.1}$$

with the conditional success probability

$$p_{it} = \Pr(y_{it} = 1|\mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\eta_{it})},$$

where $\eta_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$ is the linear predictor and $\boldsymbol{\beta}$ is a vector of structural parameters corresponding to M regressors $\mathbf{x}_{it}$. The parameter α_i is called a fixed effect which is allowed to be arbitrarily correlated with the regressors. Throughout the paper, we assume that $N \gg M$ and that the regressor matrix $\mathbf{X}$ has full column rank.

The most common approach to estimate the fixed effects logit model is maximum likelihood. In the following subsections we depict the advantages and drawbacks of the two most popular estimators which are the conditional logit estimator (CL) and the unconditional logit estimator (UCL). Further, we address the problem of estimating APEs.

2.2.2 Basic Estimation Approaches for Structural Parameters

Unconditional Logit Estimator via Dummy Variables

The simplest estimator for the fixed effects logit model is a full maximum likelihood estimator which jointly estimates $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = [\alpha_1,\dots,\alpha_N]'$. It can be conveniently estimated with standard statistical software by including a dummy variable for each individual as additional covariates.

5. The same type of model applies to unbalanced data and so-called *pseudo panels* where we include fixed effects for n groups each of size T_i.

This estimator is inconsistent as N increases and T is held constant, which is known as the incidental parameters problem (IPP) noted by Neyman and Scott (1948). However, several bias corrections have been proposed in the literature to reduce this bias (e.g. Hahn and Newey 2004; Carro 2007; Fernández-Val 2009; Dhaene and Jochmans 2015).

Estimates of UCL can be obtained by maximizing the log-likehood function

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \log\big(f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)\big) = \sum_{i=1}^{N} \sum_{t=1}^{T} l_{it} \,. \tag{2.2}$$

The standard routine to optimize (2.2) is the Newton-Raphson algorithm, which has the following parameter update in iteration $(k-1)$

$$(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) = -\mathbf{H}^{-1}\mathbf{g}\,, \tag{2.3}$$

where $\mathbf{H}$ denotes the $(M+N)\times(M+N)$ Hessian, $\mathbf{g}$ denotes the $(M+N)\times 1$ gradient, and $\boldsymbol{\theta} = [\boldsymbol{\beta}', \boldsymbol{\alpha}']'$ is the parameter vector. To be more specific, (2.3) can be reformulated as follows:

$$(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{p})\,, \tag{2.4}$$

where $\mathbf{Z} = [\mathbf{X}, \mathbf{D}]$ denotes the entire regressor matrix, which includes the dummy variable matrix $\mathbf{D}$ and the remaining regressors $\mathbf{X}$, and $\mathbf{W}$ is a positive definite diagonal weighting matrix with $\text{diag}(\mathbf{W}) = \mathbf{p}(1-\mathbf{p})$. After convergence, the standard errors of $\hat{\boldsymbol{\theta}}$ can be obtained as the square-root of the diagonal of the inverse Hessian, $\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = -\mathbf{H}^{-1}$. Details on the implementation are presented in appendix A.

Adding N dummy variables as covariates creates a substantial computational burden if N is large. Especially the computation and inversion of the Hessian needed for a Newton-Raphson optimization is demanding. As shown in appendix B, the computational costs of estimating UCL based on dummy variables is linear in T but cubic in N. This can quickly become prohibitive for large panel data sets. We discuss an algorithm that dramatically reduces the computational burden of this estimator in section 2.3.

The Conditional Logit Estimator

CL uses the individual number of successes $t_{1i} = \sum_t y_{it}$ as sufficient statistics to concentrate the incidental parameters out of the log-likelihood function. Thus, $\boldsymbol{\beta}$ obtained by CL is consistent for $N \to \infty$ and fixed T (see Chamberlain 1980).

The corresponding log-likelihood function is given by

$$L_c(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, t_{1i})\right) , \tag{2.5}$$

where

$$f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, t_{1i}) = \frac{\exp\left(\sum_{t=1}^{T} y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{b_i \in \mathscr{B}(t_{1i})} \exp\left(\sum_{t=1}^{T} b_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)} \tag{2.6}$$

is the joint probability of $\mathbf{y}_i$ conditional on t_{1i}, and $\mathscr{B}(t_{1i})$ is the set of all $c_i = \binom{T}{t_{1i}}$ permutations of $\mathbf{y}_i$. Just like (2.2), (2.5) can be maximized using a standard Newton-Raphson algorithm. In contrast to (2.3) the Hessian corresponding to CL is only of the dimension $M \times M$. Nevertheless, CL can become computationally intensive due to two other problems, stemming from the individual likelihood contributions given by (2.6). First, a large time-series dimension T implies substantial or even prohibitive computational costs, since $\mathscr{B}(t_{1i})$ quickly becomes huge. For example, $c_i = \binom{50}{20}$ is larger than 10^{13}. In total, a brute force implementation of CL requires $\approx O(\sum_{i=1}^{N} t_{1i}\binom{T}{t_{1i}})$ time, which is exponentially increasing in T (see appendix B). Second, the higher the number of permutations, the more likely the denominator in (2.6) becomes numerically hard to deal with.[6]

It is nowadays standard to mitigate the computational burden of CL by using a recursive algorithm proposed by Gail, Lubin, and Rubinstein (1981). As detailed in appendix B, the computational costs of this recursive implementation are $\approx O(\sum_{i=1}^{N} t_{1i}(T - t_{1i}))$. In the worst case, which is $t_{1i} = T/2$, they are quadratic in T.[7] We will discuss another strategy to reduce the computational burden of CL by considering only a random subset of $\mathscr{B}(t_{1i})$ in section 2.4.

2.2.3 Basic Estimation Approaches for Average Partial Effects

Since the structural parameters $\boldsymbol{\beta}$ do not have a direct interpretation, average partial effects (APEs) are often of major interest for applied work. When calculating APEs, a case distinction is made for discrete and continuous regressors. Suppose our k-th regressor is non-binary then we define the partial effect of individual i at time t based on the conditional success probability

$$\begin{aligned} \Delta_{it}^{k} &= \frac{\partial \Pr(y_{it} = 1|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)}{\partial x_{itk}} \\ &= \Pr(y_{it} = 1|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)[1 - \Pr(y_{it} = 1|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)]\beta_k \, . \end{aligned} \tag{2.7}$$

6. For instance, the largest value a computer can handle is $1.797693 \cdot 10^{308}$ in double precision.

7. The recursion can also be accelerated by using a not completely recursive implementation which reuses results and thus decreases the number of arithmetic operations by a factor. This however comes along with a higher memory requirement compared to the fully recursive program (see Gaure 2012).

In the situation where the k-th regressor is binary, we consider the difference between the conditional success probabilities, where once all observations of the regressor are set to one and once all are set to zero

$$\Delta_{it}^{k} = \Pr(y_{it} = 1 | x_{itk} = 1, \mathbf{x}_{it\{-k\}}, \boldsymbol{\beta}, \alpha_i) - \Pr(y_{it} = 1 | x_{itk} = 0, \mathbf{x}_{it\{-k\}}, \boldsymbol{\beta}, \alpha_i) \,. \tag{2.8}$$

An estimator of the APEs can be formed by replacing (2.7) or (2.8) by their sample analogues and taking the average[8]

$$\hat{\delta}_k = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{\Delta}_{it}^{k} \,. \tag{2.9}$$

This is straightforward for UCL because we can simply plug their estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ into the corresponding formulas (2.7) and (2.8).[9] However, CL does not provide any estimates of $\boldsymbol{\alpha}$ to form the plug-in estimator (2.9). A simple but inconsistent approach is to assume that all fixed effects estimates are zero.[10]

Another quantity of interest are the standard errors of APEs. They can be either estimated using bootstrap techniques or the delta method. If at least one of the panel dimensions is large, bootstrapping becomes impractical since we have to re-estimate the model multiple times. Thus the preferred strategy is the delta method. Using this approach, the corresponding covariance matrix for APEs can be estimated as follows:

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\delta}}) = \hat{\mathbf{J}} \widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) \hat{\mathbf{J}}' \,,$$

where $\hat{\mathbf{J}} = \partial \hat{\boldsymbol{\delta}} / \partial \hat{\boldsymbol{\theta}}'$ is the Jacobian and $\hat{\boldsymbol{\delta}} = [\hat{\delta}_1, \ldots, \hat{\delta}_M]'$ is the vector containing estimates of the APEs. In section 2.5 we present solutions to the aforementioned problems that are also feasible in case of large panel data.

2.3 Computationally Efficient Unconditional Logit Estimation

Greene (2004) and Chamberlain (1980), among others, propose an efficient algorithm which results in identical parameter estimates as the dummy variable approach. Their method avoids the inversion of the large Hessian in (2.3) by utilizing the

8. When calculating the average, it is important to include those individuals who do not have a varying response. Since their log-likelihood contributions are zero, these individuals do not contribute to the identification of the structural parameters. However, these individuals are still informative about partial effects. The corresponding partial effects are zero (see appendix C).

9. Note that APEs obtained by UCL are also affected by IPP, but bias corrections are available (e.g. Hahn and Newey 2004; Carro 2007; Fernández-Val 2009; Dhaene and Jochmans 2015).

10. This approach is used for example by the software package *Stata* in post-estimation routines of *clogit* and *xtlogit*.

partitioned inverse formula and exploiting the sparsity of the Hessian. We show how the Frisch-Waugh-Lovell (FWL) theorem (Frisch and Waugh 1933; Lovell 1963) can be applied alternatively.

Our basic idea is to use the fact that the parameter updates of the Newton-Raphson routine is the solution of a weighted least squares problem. This allows to apply the well-known FWL theorem to separate the updates of structural parameters from the ones of the fixed effects. Due to the sparsity of the corresponding projection matrix we can derive a straightforward and computationally efficient update formula based on transformed regressors. This transformation is comparable to the demeaning procedure of a linear fixed effects model. Since in our approach the demeaning involves weights and takes place in each iteration step of the optimization routine, we call the procedure *pseudo-demeaning*.

In order to derive the efficient pseudo-demeaning algorithm we need to reconsider the naive dummy variable approach presented in section 2.2. Since the weighting matrix $\mathbf{W}$ is positive definite and diagonal, (2.4) is equivalent to the solution of a regression of the dependent variable $\tilde{\mathbf{y}} = (\mathbf{y} - \mathbf{p}) \odot \tilde{\mathbf{w}}^{-1}$ on the independent variables $\widetilde{\mathbf{Z}} = \tilde{\mathbf{w}} \odot \mathbf{Z}$, where $\tilde{\mathbf{w}}$ is the square-root of the diagonal of $\mathbf{W}$. The corresponding regression model is

$$\tilde{\mathbf{y}} = \widetilde{\mathbf{X}}(\boldsymbol{\beta}_0^k - \boldsymbol{\beta}_0^{k-1}) + \widetilde{\mathbf{D}}(\boldsymbol{\alpha}_0^k - \boldsymbol{\alpha}_0^{k-1}) + \mathbf{u}\,, \tag{2.10}$$

where the subscript zero denotes the population parameters, $\widetilde{\mathbf{X}} = \tilde{\mathbf{w}} \odot \mathbf{X}$, $\widetilde{\mathbf{D}} = \tilde{\mathbf{w}} \odot \mathbf{D}$, and $\mathbf{u}$ is an error term. Using reformulation (2.10) we can apply the FWL theorem to separate the high-dimensional fixed effects update from the structural parameter update. In terms of our problem, the FWL theorem states that if we regress the residuals obtained from a regression of $\tilde{\mathbf{y}}$ on $\widetilde{\mathbf{D}}$ on the residuals from separate regressions of each column of $\widetilde{\mathbf{X}}$ on $\widetilde{\mathbf{D}}$, we get the same parameter estimates $(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$ as if we estimate the original regression model (2.10). Thus, pre-multiplying (2.10) with the projection matrix $\mathbf{Q} = \mathbf{I}_{NT} - \mathbf{P} = \mathbf{I}_{NT} - \widetilde{\mathbf{D}}(\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}})^{-1}\widetilde{\mathbf{D}}'$ eliminates the fixed effects and residualizes the remaining variables $\tilde{\mathbf{y}}$ and $\widetilde{\mathbf{X}}$. The resulting concentrated regression is

$$\mathbf{Q}\tilde{\mathbf{y}} = \mathbf{Q}\widetilde{\mathbf{X}}(\boldsymbol{\beta}_0^k - \boldsymbol{\beta}_0^{k-1}) + \mathbf{Q}\mathbf{u}$$

and has the solution

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = (\widetilde{\mathbf{X}}'\mathbf{Q}\mathbf{Q}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Q}\mathbf{Q}\tilde{\mathbf{y}}\,. \tag{2.11}$$

Since the matrix $\mathbf{Q}$ is idempotent and symmetric, (2.11) can be further trans-

formed while retaining the same parameter estimates[11]

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\tilde{\mathbf{y}}\,, \tag{2.12}$$

where $\ddot{\mathbf{X}} = \mathbf{Q}\tilde{\mathbf{X}}$. Noticing the special sparse structure of $\mathbf{Q}$, the projection $\mathbf{Q}\tilde{\mathbf{X}}$ can be computed without having to create the $NT \times NT$ projection matrix. In fact, $\mathbf{Q}\tilde{\mathbf{X}}$ translates into an efficiently implementable and intuitive weighted demeaning formula which allows to compute the parameter updates given in (2.12) at minimal computational costs

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\tilde{y}_{it}\right), \tag{2.13}$$

where $\tilde{\mathbf{x}}_{it} = \tilde{w}_{it}\mathbf{x}_{it}$, $\tilde{y}_{it} = (y_{it} - p_{it})/\tilde{w}_{it}$, and $\ddot{\mathbf{x}}_{it} = \tilde{\mathbf{x}}_{it} - (\tilde{w}_{it}\sum_{t=1}^{T}\tilde{w}_{it}\tilde{\mathbf{x}}_{it})/\sum_{t=1}^{T}\tilde{w}_{it}^2$.

Unlike a linear regression model, we also need to recover the estimates of the fixed effects to update the weights of the iterative maximization algorithm. Rearranging (2.10) yields the update formula of the fixed effects estimates

$$(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) = (\tilde{\mathbf{D}}'\tilde{\mathbf{D}})^{-1}\tilde{\mathbf{D}}'\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\right), \tag{2.14}$$

which depends on the previously computed structural parameter updates. Similarly to the updates of the structural parameters, formula (2.14) can be simplified by the block-diagonal structure of $(\tilde{\mathbf{D}}'\tilde{\mathbf{D}})^{-1}\tilde{\mathbf{D}}'$ as follows:

$$(\alpha_i^k - \alpha_i^{k-1}) = \frac{\sum_{t=1}^{T}\tilde{w}_{it}\tilde{y}_{it}}{\sum_{t=1}^{T}\tilde{w}_{it}^2} - \frac{\sum_{t=1}^{T}\tilde{w}_{it}\tilde{\mathbf{x}}_{it}'}{\sum_{t=1}^{T}\tilde{w}_{it}^2}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\,. \tag{2.15}$$

After we have derived all components to update the model parameters $\boldsymbol{\theta}$ efficiently, we can now introduce the entire optimization algorithm, which is linear in N and T.[12] This estimation routine is concisely summarized in algorithm 1.

Finally, we show how to obtain the standard errors of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ after convergence of algorithm 1. Instead of estimating the covariance matrix of $\hat{\boldsymbol{\beta}}$ as the inverse of the entire negative Hessian, it can be easily obtained by its concentrated counterpart

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \left(\ddot{\mathbf{X}}'\ddot{\mathbf{X}}\right)^{-1} = -\ddot{\mathbf{H}}^{-1}\,.$$

11. This transformation would not be useful in a linear regression model, since the residuals of (2.11) and (2.12) differ, and thus the standard errors would be incorrect.

12. A detailed derivation of the computational complexity is presented in appendix B.

Algorithm 1 Newton-Raphson with Pseudo-Demeaning

1: Initialize $\boldsymbol{\beta}^0$, $\boldsymbol{\alpha}^0$, and $k = 0$.
2: **repeat**
3: Set $k = k + 1$.
4: Compute $\mathbf{p}^{k-1}$ (see formula (2.1)).
5: Compute $\tilde{\mathbf{y}}^{k-1}$ and $\ddot{\mathbf{X}}^{k-1}$ to update $\boldsymbol{\beta}^k$ (see formula (2.13)).
6: Update $\boldsymbol{\alpha}^k$ (see formula (2.15)).
7: **until convergence**.

Similarly, the variance of $\hat{\boldsymbol{\alpha}}$ can be computed as

$$\widehat{\text{Var}}(\hat{\alpha}_i) = \frac{1}{\sum_{t=1}^{T} \tilde{w}_{it}^2} + \left(\frac{\sum_{t=1}^{T} \tilde{w}_{it} \tilde{\mathbf{x}}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} \right)' \widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) \left(\frac{\sum_{t=1}^{T} \tilde{w}_{it} \tilde{\mathbf{x}}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} \right).$$

Additionally we would like to draw the reader's attention to the fact that our pseudo-demeaning approach can be combined with different post-estimation bias corrections to reduce the incidental parameters bias; e.g. the analytical ones of Hahn and Newey (2004) and Fernández-Val (2009) or the jack-knife approaches of Hahn and Newey (2004) and Dhaene and Jochmans (2015). Especially if the panel is large the analytical corrections are advantageous because they only require to estimate the model once and the entire estimation procedure remains linear in N and T.

2.4 Conditional Logit with Random Subsets

As discussed above, CL can be attractive since it delivers fixed T consistent estimates for the structural parameters $\boldsymbol{\beta}$. However, it suffers from large computational costs with a long individual time series T. In this section, we introduce a new estimator that reduces this burden at the costs of efficiency.

Similar to the binary case, the multinomial logit estimator (CML) faces a huge computational burden in the presence of many alternatives. McFadden (1978) introduced a consistent but less efficient estimator for the multinomial logit model that overcomes this curse of dimensionality. We denote this estimator as CMLsub. Contrary to CML, it uses only random subsets of all possible permutations. Recently, D'Haultfœuille and Iaria (2016) analyzed the behavior of this estimator for a five-alternative multinomial logit model in a simulation study with respect to bias and computation time. Their key findings are that CMLsub is asymptotically less efficient than CML and that increasing the number of sampled permutations increases the precision. Thus, CMLsub becomes especially attractive when CML is either computationally too costly or not feasible at all.

For the binary fixed effects logit model, this approach is very similar, and we denote the corresponding estimator as CLsub. Instead of using the entire set $\mathscr{B}(t_{1i})$ of all permutations in the denominator of equation (2.6), we only use a random subset $\mathscr{D}(t_{1i})$ which contains m elements of $\mathscr{B}(t_{1i})$ where we make sure that the observed sequence is included. For brevity, we denote $\mathscr{B}(t_{1i})$ and $\mathscr{D}(t_{1i})$ as $\mathscr{B}$ and $\mathscr{D}$, respectively. Suppose that $\mathscr{D}$ is drawn conditionally on the observed choice $\mathbf{y}_i$ according to a probability $\pi(\mathscr{D}|\mathbf{y}_i)$. The key condition that we have to respect when creating the subset is the *uniform conditioning property* of McFadden (1978) which states: if $\mathbf{y}_i, \mathbf{d}_i \in \mathscr{D} \subseteq \mathscr{B}$, then $\pi(\mathscr{D}|\mathbf{y}_i) = \pi(\mathscr{D}|\mathbf{d}_i)$.[13] This condition holds if all remaining possible permutations of the observed choice sequence have the same probability of being selected in the subset, regardless of which choice sequence is observed.

In the following, the log-likelihood function of CLsub is derived. Given the joint success probability $\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta})$ of $\mathbf{y}_i \in \mathscr{B}$ conditioned on covariates $\mathbf{x}_i$ and given the probability $\pi(\mathscr{D}|\mathbf{y}_i)$ of selecting a subset $\mathscr{D} \subseteq \mathscr{B}$, the joint probability of $(\mathbf{y}_i, \mathscr{D})$ is $\pi(\mathscr{D}|\mathbf{y}_i)\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta})$ and hence the conditional probability of $\mathbf{y}_i$ given $\mathscr{D}$ is

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\pi(\mathscr{D}|\mathbf{y}_i)\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta})}{\sum_{\mathbf{d}_i \in \mathscr{D}} \pi(\mathscr{D}|\mathbf{d}_i)\Pr(\mathbf{d}_i|\mathbf{x}_i, \boldsymbol{\beta})} \tag{2.16}$$

with $\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}) = \prod_{t=1}^{T} \exp(y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))/(1 + \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))$. Equation (2.16) can be rewritten to

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\pi(\mathscr{D}|\mathbf{y}_i) \prod_{t=1}^{T} \exp(y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))}{\sum_{\mathbf{d}_i \in \mathscr{D}} \pi(\mathscr{D}|\mathbf{d}_i) \prod_{t=1}^{T} \exp(d_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))}. \tag{2.17}$$

Let $k(\mathbf{y}_i) = \sum_{t=1}^{T} y_{it}$, then equation (2.17) can be further simplified since $k(\mathbf{y}_i) = k(\mathbf{d}_i)$ and thus the fixed effect α_i is conditioned out

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\pi(\mathscr{D}|\mathbf{y}_i) \prod_{t=1}^{T} \exp(y_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}{\sum_{\mathbf{d}_i \in \mathscr{D}} \pi(\mathscr{D}|\mathbf{d}_i) \prod_{t=1}^{T} \exp(d_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}. \tag{2.18}$$

The application of the uniform conditioning property, reduces (2.18) to

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\prod_{t=1}^{T} \exp(y_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}{\sum_{\mathbf{d}_i \in \mathscr{D}} \prod_{t=1}^{T} \exp(d_{it}\mathbf{x}_{it}'\boldsymbol{\beta})},$$

13. The validity of the uniform conditioning property can be shown as follows: $\mathscr{D}$ is selected to contain $\mathbf{y}_i$ plus $m-1$ random permutations $\mathbf{d}_i$ of $\mathbf{y}_i$. There are $c_i = \binom{T}{k(\mathbf{y}_i)}$ ways to place $k(\mathbf{y}_i) = \sum_{t=1}^{T} y_{it}$ ones in T slots, and hence $(c_i - 1)!/((m-1)!(c_i - m)!)$ ways to randomly select $m-1$ permutations of $\mathbf{y}_i$ without replacement. Thus $\pi(\mathscr{D}|\mathbf{y}_i) = ((m-1)!(c_i - m)!)/(c_i - 1)!$ depends only on $k(\mathbf{y}_i)$. Since any permutation $\mathbf{d}_i$ of $\mathbf{y}_i$ has the same c_i, it follows $\pi(\mathscr{D}|\mathbf{d}_i) = \pi(\mathscr{D}|\mathbf{y}_i)$, which is the uniform conditioning property of McFadden (1978).

which can be finally used to form the log-likelihood function of CLsub

$$L_{sub}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(\frac{\exp\left(\sum_{t=1}^{T} \mathbf{x}_{it}' y_{it}\boldsymbol{\beta}\right)}{\sum_{d_i \in \mathscr{D}(t_{1i})} \exp\left(\sum_{t=1}^{T} \mathbf{x}_{it}' d_{it}\boldsymbol{\beta}\right)}\right).$$

Next, we encounter a practical problem with the implementation of CLsub. A naive approach would first generate $\mathscr{B}$ to sample $\mathscr{D}$ from it. However, this approach has two shortcomings: it requires a lot of memory and for data sets with large T the computation of $\mathscr{B}$ is infeasible. Therefore we recommend to randomly shuffle the observed choice sequence $m-1$ times and to store the positions of the successes on each occasion. Multiple permutations are deleted and the process is repeated until the subset contains m unique permutations.

Compared to CL, which uses the entire permutation set $\mathscr{B}$ in the log-likelihood, CLsub reduces the number of arithmetic operations per individual from $c_i t_{1i} - 1$ to $m t_{1i} - 1$. Hence, it can be derived that CLsub requires $O(m \sum_{i=1}^{N} t_{1i})$ time, which means that the shape of the computational complexity depends on the choice of m (see appendix B).[14] Note that the theoretical derivation of the computational complexity is based on the assumption that $\mathscr{D}$ is already generated. From a practical point of view the total computation time, including the sampling of $\mathscr{D}$, is of interest. This will be the subject of our simulation experiments presented in section 2.6.

2.5 Feasible Estimation of Average Partial Effects

2.5.1 Efficient Offset Algorithm

So far, we have dealt with the problems of estimating structural parameters. In this section we tackle the remaining problems associated with the estimation of average partial effects.

Remember that one of the drawbacks of CL and CLsub is that they do not provide estimates of the fixed effects, so that the APE plug-in estimator (2.9) cannot be formed. In the following, we propose a simple ex-post estimation strategy to obtain estimates of the fixed effects. This is usually done by a so-called offset algorithm which in our case maximizes the log-likelihood function (2.2) while keeping the estimates of the structural parameters fixed at their values obtained by any conditional logit estimator.[15] The estimates obtained by this algorithm can in turn be used to calculate the APEs according to (2.9). The same type of algorithm is also

14. For example if m is a linear function of T the computational complexity evolves roughly quadratically in T.

15. In an *offset* algorithm an additional variable is added to the linear predictor whose parameter is constrained to the value one (see Nelder and Wedderburn 1972).

required to alleviate the IPP using analytical bias corrections for average partial effects.[16]

We now turn to the derivation of an efficient offset algorithm, which is linear in N and T. Let $\tilde{\boldsymbol{\beta}}$ denote known estimates of the structural parameters. Maximizing (2.2) with $\mathbf{X}\tilde{\boldsymbol{\beta}}$ being fixed yields the Newton-Raphson update in iteration $(k-1)$

$$(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) = (\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}})^{-1}\widetilde{\mathbf{D}}'\tilde{\mathbf{y}}\,. \tag{2.19}$$

Thus, (2.19) can be efficiently computed according to

$$(\alpha_i^k - \alpha_i^{k-1}) = \frac{\sum_{t=1}^{T} \tilde{w}_{it}\tilde{y}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} \tag{2.20}$$

and the whole procedure is repeated until convergence.[17]

In the context of CL, Bartolucci and Pigini (2019) suggest a refined version of our offset approach presented above. They use a strategy proposed by Firth (1993) to obtain an estimate of $\boldsymbol{\alpha}$ with improved finite sample properties by solving the following modified score equations

$$s^{Firth}(\boldsymbol{\alpha}) = \sum_{t=1}^{T}(y_{it} - p_{it}) + \frac{\sum_{t=1}^{T} p_{it}(1-p_{it})(1-2p_{it})}{2\sum_{t=1}^{T} p_{it}(1-p_{it})} = 0\,, \tag{2.21}$$

where $p_{it} = 1/(\exp(-\alpha_i - \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}))$.

Solving the system (2.21) has the drawback that it becomes computationally demanding if N increases. Therefore, we follow Kosmidis and Firth (2009), who have shown that the solution of (2.21) can be obtained equivalently by using a standard Newton-Raphson algorithm with a modified dependent variable $\mathbf{y}^* = \mathbf{y} + \text{diag}(\mathbf{S})(0.5 - \mathbf{p})$, where $\mathbf{S} = \mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}$. The sparse structure of $\mathbf{S}$ in turn suggests to compute the adjusted dependent variable as follows $y_{it}^* = y_{it} + (\tilde{w}_{it}^2/\sum_{t=1}^{T}\tilde{w}_{it}^2)(0.5 - p_{it})$. Thus, we can use the same kind of efficient offset algorithm described previously by simply replacing the dependent variable in (2.20). Another modification compared to Bartolucci and Pigini (2019) is that we estimate the fixed effects of all n individuals.[18] We draw on a very recent result of Kunz, Staub, and Winkelmann (2018), who have

16. Analytical bias corrections of the APEs require, among other steps, that the fixed effects have to be re-estimated after bias-correcting the structural parameter estimates (see among others Hahn and Newey 2004).

17. Note that $\mathbf{X}\tilde{\boldsymbol{\beta}}$ is still part of the linear predictor and thus has to be incorporated when updating the weights and the adjusted dependent variable.

18. Bartolucci and Pigini (2019) seem to estimate fixed effects only for individuals with varying responses. The specific approach is not clear from the methodological part of their article. However, a replication of their simulation results indicates that they only consider the APEs obtained from non-perfectly classified observations.

proven that Firth's method can be used to obtain finite estimates of the fixed effects in probit models for perfectly classified individuals. It is straightforward to show that the same applies to logit models with fixed effects. Although the article of Kunz, Staub, and Winkelmann (2018) is about predicting fixed effects, we have found that their approach is also useful to obtain non-zero estimates of partial effects in the case of perfect classification.

2.5.2 Concentrated Delta Method

Next, we address the estimation of the standard errors for the APEs. The attentive reader might have noticed that using the brute force delta method as described in section 2.2 is problematic, because it requires the entire covariance matrix of $\hat{\boldsymbol{\theta}}$. However, with our pseudo-demeaning approach for UCL estimation described in section 2.3, we have only a reduced covariance matrix corresponding to the structural parameters and the variance of the fixed effects. The same obstacle occurs when we estimate the APEs for conditional logit estimators using the (modified) offset algorithm.

A solution to this problem consists of a concentrated delta method, which we derive from a combination of the results of Fernández-Val and Weidner (2016) and our pseudo-demeaning approach. To be more precise, Fernández-Val and Weidner (2016) have suggested an estimator for the covariance of APEs for nonlinear models with individual and time fixed effects. Thanks to the fact that our approach is based on the FWL theorem, it is straightforward to translate their estimator to the case of individual fixed effects and to exploit the sparsity of several terms included.[19] Assuming that the individual fixed effects are independent, the variance estimator of the APEs is given by

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\delta}}) = \frac{1}{N^2T^2}\left(\sum_{i=1}^{N}\sum_{t=s=1}^{T}\widehat{\bar{\boldsymbol{\Delta}}}_{it}\widehat{\bar{\boldsymbol{\Delta}}}_{is}' + \sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\boldsymbol{\Gamma}}_{it}\widehat{\boldsymbol{\Gamma}}_{it}'\right),$$

where

$$\widehat{\boldsymbol{\Gamma}}_{it} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial\widehat{\boldsymbol{\Delta}}_{it}}{\partial\beta} - \frac{\bar{\mathbf{x}}_{it}}{\tilde{w}_{it}}\frac{\partial\widehat{\boldsymbol{\Delta}}_{it}}{\partial\alpha_i}\right)'\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\ddot{\mathbf{x}}_{it}\tilde{y}_{it} - \frac{\bar{\boldsymbol{\psi}}_{it}}{\tilde{w}_{it}}\frac{\partial l_{it}}{\partial\alpha_i},$$

$\boldsymbol{\psi}_{it}$ and $\ddot{\boldsymbol{\psi}}_{it}$ are the it-th rows of $\boldsymbol{\Psi}$ and $\mathbf{Q}\boldsymbol{\Psi}$, $\boldsymbol{\psi}_{it} = (\partial\widehat{\boldsymbol{\Delta}}_{it}/\partial\alpha_i)/\tilde{w}_{it}^2$, $\widehat{\boldsymbol{\Delta}}_{it} = [\widehat{\Delta}_{it}^1,\ldots,\widehat{\Delta}_{it}^M]'$, $\bar{\mathbf{x}}_{it} = \mathbf{x}_{it} - \ddot{\mathbf{x}}_{it}$, $\bar{\boldsymbol{\psi}}_{it} = \boldsymbol{\psi}_{it} - \ddot{\boldsymbol{\psi}}_{it}$, and $\widehat{\bar{\boldsymbol{\Delta}}}_{it} = \widehat{\boldsymbol{\Delta}}_{it} - \hat{\boldsymbol{\delta}}$. Note that the first part of the variance estimator takes into account the variation induced by estimating sample instead of population means and the second term is a concentrated version of the delta method. Especially the former is not very well known from the standard textbook

19. This relationship is not so obvious when we use the partitioned inverse formula instead of the FWL theorem.

literature, but it substantially improves the finite sample properties of the estimator (see Fernández-Val and Weidner 2016).

2.6 Simulation Experiments

2.6.1 Simulation Design

In this section we analyze the statistical properties of UCL, BCL, CL, and CLsub in terms of structural parameters and APEs. Further, we investigate the computation times of the different estimation routines. BCL refers to the bias-corrected UCL estimator suggested by Fernández-Val (2009).[20] For CLsub we consider two variants which differ by the size of the random subset $\mathscr{D}$. To be more precise, we choose $m^* \in \{1, T/2\}$, where m^* denotes the size of $\mathscr{D}$ without the observed choice sequence $\mathbf{y}_i$. All estimators analyzed in the simulation study are implemented by ourselves in the same programming language to guarantee comparability.[21] UCL and BCL are estimated using the Newton-Raphson pseudo-demeaning approach introduced in section 2.3. For the recursive CL and the CLsub algorithm we use a standard Newton-Raphson optimization routine with numerical derivatives to make it comparable to the estimation routine used for UCL and BCL without unnecessarily blowing up the memory.[22]

For our simulation experiments we generate the data according to Greene (2004) as follows:

$$y_{it} = \mathbf{1}[\alpha_i + \beta_1 x_{it} + \beta_2 d_{it} + v_{it} > 0]\,, \tag{2.22}$$

where $v_{it} = \log(u_{it}/(1-u_{it}))$, $u_{it} \sim \mathscr{U}(0,1)$, $\beta_1 = \beta_2 = 1$ $x_{it} \sim \mathscr{N}(0,1^2)$, $d_{it} = \mathbf{1}[x_{it} + h_{it} > 0]$, $h_{it} \sim \mathscr{N}(0,1^2)$, $\alpha_i = \sqrt{T}\bar{\mathbf{x}}_i + a_i$, $\bar{\mathbf{x}}_i = T^{-1}\sum_t x_{it}$, $a_i \sim \mathscr{N}(0,1^2)$. This design is well suited to analyze the behavior of the various fixed effects estimators, as it introduces an approximately constant correlation between the unobserved heterogeneity and the regressors for different T.

Throughout our experiments, we analyze several model specifications with different n and T. We also consider panels with unusual large T, which can be justified when we think about so-called *pseudo panels*, where n groups, each consisting of T

20. In an earlier version of this article we use the bias correction of Hahn and Newey (2004). However, we find that the bias correction of Fernández-Val (2009) has better finite sample properties, although both approaches are asymptotically equivalent.

21. All estimators and replication scripts are available on request.

22. We do not investigate the brute-force implementations of UCL and CL because their computational costs are unreasonable high and in most of our analyzed setups they are even infeasible. In addition, we do not use analytical first and second order derivatives of the recursive CL due to its enormous memory requirement especially for large T (see appendix A). Also note that we are not using a full recursive implementation of CL, but an algorithm that exploits the usage of previous results in the recurrence that is substantially faster (see Gaure 2012).

statistical units, are observed. For instance, n could be the number of postal code areas and T the number of households living in each area.

2.6.2 Finite Sample Properties

First of all, we focus on the statistical properties of the different estimators for the structural parameters and APEs. In order to investigate the biases and inference accuracies, all tables report the bias and standard deviations (SD) in percent relative to the true parameter value, the ratio between the average standard errors and the standard deviation, as well as the empirical coverage probabilities at a nominal value of 95%. All results are obtained by 1,000 replications of 9 model specifications with $n = 1{,}000$ and $T \in \{4, 8, 10, 12, 16, 20, 50, 100, 200\}$. For the sake of brevity, we only report the results of the continuous regressor, since we make similar findings for the discrete regressor.[23]

Table 2.1 shows the corresponding results for the structural parameter β_1. The UCL estimator is strongly distorted by the incidental parameter bias, but the distortion decreases as the T increases. At $T = 50$ the estimator still suffers a percentage distortion of 2.51 and even at $T = 200$ the coverage probabilities are too low, although its bias is below one percent. On the other hand, we find that the bias correction considerably reduces the bias of the UCL estimator. However, since the bias correction is based on a large-T expansion, it also requires a sufficiently large T to eliminate most of the distortion (see Fernández-Val 2009). Whereas for $T = 4$ there is still a bias of 13.01 percent, for $T = 8$ it is already only 0.49 percent and finally disappears with increasing T. Furthermore, the bias correction already brings the coverage probabilities close to their nominal level for $T = 8$. As expected, the CL estimator is unaffected by the incidental parameter bias. It delivers almost undistorted estimates across all T and the coverage probabilities are almost at the desired 95 percent. Thus, we can consider CL as a benchmark for the bias correction and find that the properties of CL and BCL for the structural parameters become closer as T increases. CLsub delivers similar results as CL if its subset $\mathscr{D}$ is not too small relative to the entire permutation set $\mathscr{B}$. In the case of $m^* = 1$, CLsub is almost undistorted for $T \leq 12$ but the bias increases rapidly from $T = 16$. While the distortion for $T = 12$ is still 0.25 percent, it rises to 85.12 percent for $T = 200$. We observe a similar behavior for CLsub with $m^* = T/2$, albeit in a delayed form. Since m^* does not depend on the size of the entire permutation set, it is not surprising that the bias of of CLsub increases considerably for large T. The optimization problem also becomes very unstable and produces unreliable results if m^* is small relative to T. Altogether, these findings suggest a careful choice of m^*. In a direct comparison to CL, we also

23. The results of the discrete regressor are available on request.

find that CLsub is less efficient and precise, which is reflected by a larger standard error and higher distortion. Interestingly, CLsub can maintain coverage probabilities of about 95 percent even in cases where it exhibits extreme distortions.

Table 2.1: *Finite sample properties of* $\hat{\beta}_1$

		UCL	BCL	CL	CLsub	
					$m^* = 1$	$m^* = T/2$
$T = 4$	Bias	47.08	-13.01	0.17	0.81	0.35
	SD	11.48	5.13	7.34	11.08	8.78
	SE/SD	0.81	1.42	1.01	0.99	0.99
	CP .95	0.00	0.60	0.96	0.95	0.94
$T = 8$	Bias	19.69	-0.49	0.20	0.74	0.27
	SD	5.51	4.38	4.47	9.62	5.99
	SE/SD	0.91	1.06	1.02	0.98	1.01
	CP .95	0.03	0.96	0.96	0.95	0.95
$T = 10$	Bias	14.94	-0.29	0.03	0.82	0.29
	SD	4.54	3.82	3.85	9.05	5.40
	SE/SD	0.94	1.05	1.03	1.03	1.02
	CP .95	0.06	0.96	0.95	0.95	0.95
$T = 12$	Bias	11.91	-0.33	-0.14	0.25	-0.05
	SD	4.11	3.58	3.60	9.18	5.49
	SE/SD	0.92	1.01	0.99	1.01	0.94
	CP .95	0.14	0.95	0.95	0.95	0.94
$T = 16$	Bias	9.00	0.17	0.26	1.81	0.32
	SD	3.32	3.00	3.01	9.98	4.72
	SE/SD	0.95	1.02	1.00	0.98	1.01
	CP .95	0.20	0.95	0.95	0.95	0.95
$T = 20$	Bias	6.84	-0.03	0.02	1.06	0.07
	SD	2.90	2.68	2.68	10.62	4.71
	SE/SD	0.95	1.00	0.99	0.98	0.98
	CP .95	0.31	0.94	0.94	0.96	0.95
$T = 50$	Bias	2.51	-0.07	-0.06	8.19	0.82
	SD	1.77	1.72	1.72	23.82	5.70
	SE/SD	0.94	0.95	0.95	0.92	1.00
	CP .95	0.66	0.93	0.93	0.97	0.96
$T = 100$	Bias	1.29	0.03	0.03	60.13	2.40
	SD	1.16	1.14	1.14	165.22	10.27
	SE/SD	0.99	1.00	1.00	3.70	0.97
	CP .95	0.80	0.95	0.94	0.97	0.95
$T = 200$	Bias	0.66	0.04	0.04	85.12	5.55
	SD	0.80	0.80	0.80	184.48	22.21
	SE/SD	1.00	1.00	1.00	113.49	0.87
	CP .95	0.87	0.95	0.95	0.95	0.95

Note: Bias and SD denote biases and standard deviations in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; results based on 1,000 repetitions.

Next, we consider the statistical properties of the different estimators regarding the APEs. First, we discuss table 2.2, which summarizes the results for UCL, BCL

and CL. Then we will look at CLsub, whose results are given in table 2.3. For CL we analyze the performance of the three different approaches to estimate APEs. We denote the first approach which neglects the contributions of the fixed effects as *naive*, the second approach which uses the offset algorithm to recover estimates of the fixed effects as *score*, and the third approach which is based on the modified score as *Firth*. Remarkably, the incidental parameter bias present in the UCL estimator of the structural parameters hardly transfers to the APEs. For $T = 4$ we find a distortion of 2.23 percent and for $T \geq 8$ the distortion is close to zero. This notable result is consistent with the finding of Fernández-Val (2009).[24] The bias correction delivers comparatively good results like UCL, with the exception of $T = 4$, where the distortion is substantially higher with 8.16 percent. In addition, both estimators provide coverage probabilities close to the level of 95 percent for $T \geq 8$. We now turn to CL. The *naive* approach has a persistent high bias that ranges between 19.61 and 29.25 percent across all T. The *score* approach leads to a considerable reduction of the distortion with increasing T and also improves the coverage probabilities. At $T = 200$ the bias is only 0.32 percent and the coverage probability is 94 percent. Firth's approach brings a further substantial improvement. Overall, it performs similar as UCL, but always a bit worse for $T \geq 8$. Whereas CL based on Firth's method still has a distortion of 1.01 percent in the case of $T = 8$, UCL is almost undistorted. Next we compare the different conditional logit estimator combined with Firth's method to each other in table 2.3.[25] With regard to the distortion of the APEs, we make similar observations as with the structural parameters. CLsub provides comparable low distortions as CL, as long as $T \leq 20$. However, if T becomes too large, the CLsub collapses and its distortions increase. This is again particularly extreme in the case of $m^* = 1$. A crucial difference to the structural parameters is that the inference of the APEs obtained with CLsub is invalid. We conjecture that this is due to the fact that the higher dispersion of the structural parameters carries over to the estimation of the standard errors of the APEs with the delta method.

24. Fernández-Val (2009) shows that the components that drive the bias of uncorrected APEs are the variation of the individual effects and their impact on the regressors. He finds that the bias is small, even in panels with a short time dimension, for a wide range of different distributions of individual effects and regressors. On the other hand Fernández-Val (2009) motivates the need of bias corrections in models with lagged dependent variables, where the small bias property of static binary-choice models disappears.

25. A complete table with the *naive* and *score* approach can be provided upon request. Overall, they perform substantially worse compared to Firth's approach.

Table 2.2: *Finite sample properties of* $\hat{\delta}_1$

		UCL	BCL	CL		
				naive	score	Firth
$T = 4$	Bias	-2.23	-8.16	29.25	-21.34	1.76
	SD	6.43	5.47	8.28	5.28	6.48
	SE/SD	0.92	0.93	1.24	0.89	0.89
	CP .95	0.92	0.63	0.14	0.01	0.91
$T = 8$	Bias	0.16	-0.22	26.27	-9.71	-1.01
	SD	4.13	4.04	5.08	3.77	3.95
	SE/SD	0.93	0.91	1.14	0.95	0.95
	CP .95	0.93	0.92	0.00	0.24	0.93
$T = 10$	Bias	0.12	-0.15	25.21	-7.68	-1.30
	SD	3.53	3.49	4.39	3.29	3.39
	SE/SD	0.96	0.95	1.14	0.99	0.98
	CP .95	0.94	0.93	0.00	0.34	0.93
$T = 12$	Bias	-0.11	-0.32	24.37	-6.52	-1.53
	SD	3.29	3.27	4.07	3.11	3.18
	SE/SD	0.94	0.93	1.09	0.96	0.96
	CP .95	0.94	0.93	0.00	0.44	0.90
$T = 16$	Bias	0.29	0.13	23.99	-4.47	-1.05
	SD	2.82	2.81	3.43	2.71	2.74
	SE/SD	0.95	0.94	1.09	0.98	0.97
	CP .95	0.93	0.93	0.00	0.60	0.92
$T = 20$	Bias	0.08	-0.04	23.18	-3.68	-1.11
	SD	2.53	2.53	3.01	2.46	2.46
	SE/SD	0.96	0.95	1.09	0.98	0.98
	CP .95	0.95	0.94	0.00	0.65	0.91
$T = 50$	Bias	-0.04	-0.08	21.12	-1.50	-0.65
	SD	1.67	1.67	1.89	1.66	1.65
	SE/SD	0.99	0.98	1.05	0.99	0.99
	CP .95	0.94	0.94	0.00	0.85	0.93
$T = 100$	Bias	0.02	0.00	20.26	-0.70	-0.31
	SD	1.23	1.23	1.26	1.22	1.22
	SE/SD	1.05	1.04	1.09	1.05	1.05
	CP .95	0.96	0.96	0.00	0.93	0.95
$T = 200$	Bias	0.03	0.03	19.61	-0.32	-0.14
	SD	1.05	1.05	0.86	1.05	1.05
	SE/SD	1.01	1.01	1.11	1.01	1.01
	CP .95	0.94	0.94	0.00	0.94	0.94

Note: Bias and SD denote biases and standard deviations in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; results based on 1,000 repetitions.

Table 2.3: *Finite sample properties of* $\hat{\delta}_1$ *(based on Firth's method)*

		CL	CLsub	
			$m^* = 1$	$m^* = T/2$
$T = 4$	Bias	1.76	2.10	1.84
	SD	6.48	9.60	7.70
	SE/SD	0.89	1.24	1.00
	CP .95	0.91	0.98	0.94
$T = 8$	Bias	-1.01	-0.89	-1.02
	SD	3.95	7.73	5.05
	SE/SD	0.95	1.87	1.22
	CP .95	0.93	1.00	0.98
$T = 10$	Bias	-1.30	-1.00	-1.20
	SD	3.39	7.29	4.53
	SE/SD	0.98	2.19	1.30
	CP .95	0.93	1.00	0.98
$T = 12$	Bias	-1.53	-1.45	-1.49
	SD	3.18	7.44	4.52
	SE/SD	0.96	2.39	1.27
	CP .95	0.90	1.00	0.98
$T = 16$	Bias	-1.05	-0.18	-1.01
	SD	2.74	7.54	3.92
	SE/SD	0.97	3.03	1.47
	CP .95	0.92	1.00	0.99
$T = 20$	Bias	-1.11	-0.76	-1.16
	SD	2.46	8.07	3.81
	SE/SD	0.98	3.61	1.57
	CP .95	0.91	1.00	1.00
$T = 50$	Bias	-0.65	2.35	-0.23
	SD	1.65	14.50	4.27
	SE/SD	0.99	13.05	3.32
	CP .95	0.93	1.00	1.00
$T = 100$	Bias	-0.31	8.65	0.66
	SD	1.22	38.56	7.34
	SE/SD	1.05	> 1000	8.22
	CP .95	0.95	1.00	1.00
$T = 200$	Bias	-0.14	17.56	1.54
	SD	1.05	54.44	14.16
	SE/SD	1.01	> 1000	22.44
	CP .95	0.94	1.00	1.00

Note: Bias and SD denote biases and standard deviations in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; results based on 1,000 repetitions.

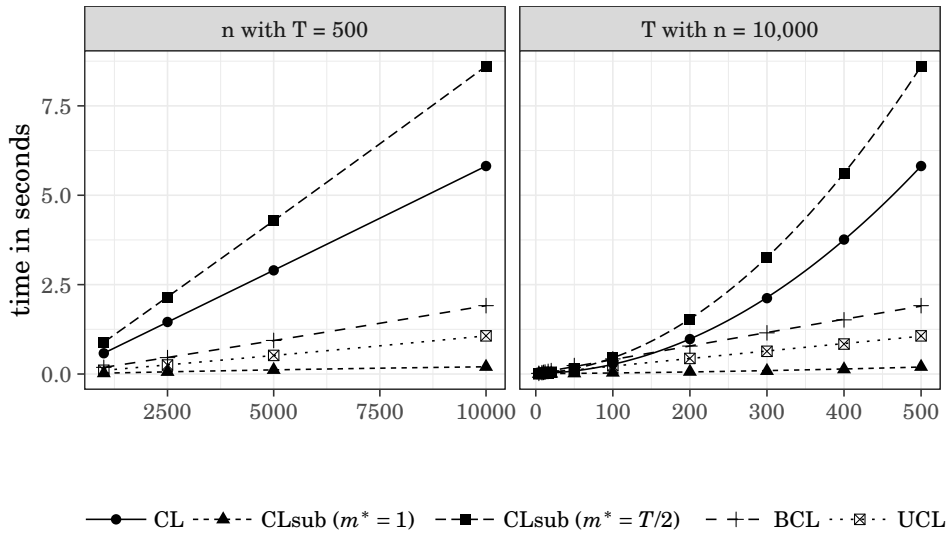
2.6.3 Computational Costs

Aside from the statistical properties of the estimators, their computation times also matter for their application in practice. Whereas the theoretical computational complexities derived earlier in this article give a rough impression of the relationship between the panel dimension and the computation time, they do not reveal anything about the total magnitude of time required by an algorithm.

The computation times reported in table 2.4 are the averages of the respective fitting processes over 30 different data sets per $n - T$ combination generated according to (2.22). Furthermore, we investigate whether the theoretical computational complexities hold up empirically. To this end we measure the average computation times per iteration, since the estimators sometimes require a different number of iterations for each data set and $n - T$ combination. All calculations were done with the software *R* (R Core Team 2019) version 3.6.1 on a Linux Workstation with Intel Xeon E5-2640 v3 and 64 GB RAM.

Altogether our theoretical findings about the shape of the computational complexities are also verified empirically as shown in figure 2.1. The left figure depicts exemplary for $T = 500$ that all estimators evolve linearly in n. Moreover, UCL, BCL, and CLsub with $m^* = 1$ rise linear in T whereas CLsub with $m^* = T/2$ rises quadratically as demonstrated in the right figure for $n = 10{,}000$.

Figure 2.1: *Empirical Computational Complexities*



Note: Dots: average computation times per iteration in seconds; Curves: quadratic polynomial approximations.

Table 2.4 depicts the enormous speed advantage of UCL and BCL compared to CL, especially when T becomes large. BCL takes on average 5.47 seconds when $T = 200$ and $n = 10{,}000$, whereas CL takes 9.77 seconds. The difference becomes even more dramatic when $T = 500$. In this case BCL requires 13.25 seconds and CL

Table 2.4: *Average Computation Times*

		UCL	BCL	CL	CLsub	
					$m^* = 1$	$m^* = T/2$
$n = 10,000$	$T = 4$	0.04	0.11	0.04	0.81 (0.75)	1.30 (1.23)
	$T = 8$	0.10	0.22	0.10	0.91 (0.83)	1.61 (1.49)
	$T = 10$	0.12	0.27	0.13	0.92 (0.83)	1.69 (1.52)
	$T = 12$	0.15	0.32	0.17	0.96 (0.85)	1.79 (1.59)
	$T = 16$	0.22	0.44	0.26	1.02 (0.89)	2.07 (1.76)
	$T = 20$	0.29	0.57	0.35	1.09 (0.91)	2.35 (1.95)
	$T = 50$	0.81	1.42	1.03	1.38 (1.00)	6.73 (4.01)
	$T = 100$	1.56	2.84	2.87	1.91 (1.08)	17.50 (8.55)
	$T = 200$	3.05	5.47	9.77	2.62 (1.20)	50.95 (22.32)
	$T = 300$	4.38	7.94	21.21	3.38 (1.31)	108.40 (42.47)
	$T = 400$	5.78	10.44	37.60	4.37 (1.43)	187.46 (68.35)
	$T = 500$	7.40	13.25	58.17	4.10 (1.50)	287.42 (98.08)
$T = 500$	$n = 1,000$	0.69	1.23	5.81	0.31 (0.15)	31.49 (9.98)
	$n = 2,500$	1.67	3.05	14.55	1.14 (0.37)	73.48 (24.66)
	$n = 5,000$	3.55	6.41	28.99	2.37 (0.75)	144.03 (49.06)
	$n = 10,000$	7.40	13.25	58.17	4.10 (1.50)	287.42 (98.08)

Note: Computation times in seconds; time needed for generating $\mathscr{D}$ in parentheses; results based on 30 repetitions.

roughly 1 minute. As indicated in table 2.4, CLsub with $m^* = 1$ is faster than CL when $T \geq 100$ and CLsub with $m^* = T/2$ is not able to outperform CL. Furthermore, table 2.4 depicts that CLsub with $m^* = 1$ is negligibly slower than CL if T is small, but when T increases it outperforms CL by far. Table 2.4 also reveals two other notable results about CLsub. First, CLsub with $m^* = T/2$ is always much slower than the other two conditional logit estimators. On the other hand, the creation of the subset of the entire permutation set, whose computation time is shown in parentheses, accounts for a large part of the total computation time. Even after subtracting this time from the total computation time, the CL estimator is still much faster, especially at large T.

Summarizing the findings from the simulation experiments, we conclude that CLsub is not an option to CL. If we sample a sufficiently large subset from the entire permutation set, the estimator is still computationally more demanding and additionally less precise than CL. As we have shown in theory and simulation, CL quickly encounters computational challenges when T rises, although we have already employed the efficient recursive implementation. Moreover, the conditional logit estimators are outperformed by UCL and BCL in the estimation of APEs and computation times in general. Thus, UCL and BCL offer attractive alternatives.

2.7 Empirical Illustration

In this section we demonstrate the advantage of or pseudo-demeaning approach by providing an illustration from labor economics, where the brute-force dummy approach as well as the recursive CL approach fail due to computational limitations. We investigate the labor force participation of women using a data set from the *American Community Survey* (2017 ACS 1-YEAR PUMS). The data set can be interpreted as a pseudo panel where the cross-sectional units are *Public Use Microdata Areas* (PUMAs) and the time dimension translates into groups of women in these PUMAs. The data set consists of $1,294,938$ women in $N = 982$ PUMAs, where the smallest PUMA includes $T_i = 230$ and the largest $T_i = 26,772$ women.

We specify our model as follows:

$$\begin{aligned} work_{it} &= \mathbf{1}\left[\eta_{it} \geq v_{it}\right], \\ \eta_{it} &= \alpha_i + \sum_j \gamma_j educ_{jit} + \beta_1 age_{it} + \beta_2 mar_{it} + \beta_3 inc_{it} + \beta_4 kids6_{it}, \end{aligned}$$

where i and t refer to the t-th woman in PUMA i, $work$ denotes the labor force participation status, age refers to the age in years, mar is the marital status, inc is the household income without the labor earnings of the woman in thousand dollars, $educ_j$ are indicators of different educational attainments[26], and $kids6$ is an indicator of the presence of children under the age of 6 years.

Table 2.5 shows the estimates of the structural parameters (left panel), APEs (right panel), and the corresponding standard errors (in parentheses). We observe that the bias-corrected and uncorrected estimates are almost identical, which is as expected due to large T_i. Overall the results are intuitive. For instance, higher education has a significant positive impact on the probability to participate in the labor force. Having a high school degree increases the probability by 26.2 percentage points relative to a woman with no high school degree. Further the presence of young and new born children lowers the probability to participate. Interestingly, the transitory non-labor household income does not affect the participation decision, which is in line with Hyslop (1999).

To demonstrate that the bias correction also works with real data, we extract a subset from the entire data set by randomly drawing $T_i = 8$ observations from each PUMA. Now that T is small enough to be handled by the CL estimator, we can use it as a benchmark for the performance of the bias correction due to its fixed T consistency property in case of structural parameter estimation. Furthermore, the small T makes a bias correction of the UCL estimator necessary. The results

26. More precisely, $educ$ has three levels: no high-school degree, high-school degree, college and/or university degree.

shown in table 2.6 are in line with the findings in the simulation study. Whereas the structural parameter estimates of CL and BCL are close to each other, the corresponding estimates obtained by UCL differ remarkably. However, although the structural parameter estimates obtained by the UCL estimator are clearly distorted, its estimated APEs hardly differ from the bias-corrected ones. With the CL estimator, we get substantially lower estimated APEs in terms of magnitude.

Table 2.5: *Estimation results based on the entire sample*

	$\hat{\boldsymbol{\beta}}$		$\hat{\boldsymbol{\delta}}$	
	UCL	BCL	UCL	BCL
college / university	2.229	2.227	0.375	0.375
	(0.008)	(0.008)	(0.001)	(0.001)
highschool	1.549	1.548	0.262	0.262
	(0.007)	(0.007)	(0.001)	(0.001)
age	-0.057	-0.057	-0.011	-0.011
	(0.000)	(0.000)	(0.000)	(0.000)
married	0.285	0.284	0.053	0.053
	(0.004)	(0.004)	(0.001)	(0.001)
kids6	-0.741	-0.741	-0.139	-0.139
	(0.007)	(0.007)	(0.001)	(0.001)
nlinc	-0.003	-0.003	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)

Note: $\hat{\boldsymbol{\beta}}$ denotes estimates of the structural parameters; $\hat{\boldsymbol{\delta}}$ denotes estimates of APEs; standard errors in parenthesis; standard errors of $\hat{\boldsymbol{\delta}}$ are computed with the delta method.

Table 2.6: *Estimation results based on a subsample*

	$\hat{\boldsymbol{\beta}}$			$\hat{\boldsymbol{\delta}}$		
	UCL	BCL	CL	UCL	BCL	CL Firth
college / university	2.292	1.939	1.952	0.344	0.339	0.323
	(0.108)	(0.102)	(0.099)	(0.013)	(0.013)	(0.012)
highschool	1.619	1.367	1.376	0.247	0.243	0.230
	(0.097)	(0.092)	(0.089)	(0.014)	(0.014)	(0.013)
age	-0.062	-0.053	-0.053	-0.010	-0.010	-0.009
	(0.002)	(0.002)	(0.002)	(0.000)	(0.000)	(0.000)
married	0.363	0.307	0.310	0.060	0.058	0.055
	(0.065)	(0.063)	(0.060)	(0.010)	(0.010)	(0.010)
kids6	-1.003	-0.855	-0.862	-0.164	-0.163	-0.154
	(0.103)	(0.100)	(0.095)	(0.017)	(0.017)	(0.016)
income	-0.003	-0.003	-0.003	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Note: $\hat{\boldsymbol{\beta}}$ denotes estimates of the structural parameters; $\hat{\boldsymbol{\delta}}$ denotes estimates of APEs; standard errors in parenthesis; standard errors of $\hat{\boldsymbol{\delta}}$ are computed with the delta method.

2.8 Conclusion

This paper discussed and addressed the disadvantages of the two most commonly used estimators for logit models with fixed effects, especially in the case of data sets where many cross-sectional units are observed for long time horizons. These are the conditional and the unconditional logit estimators. In a series of simulation experiments we found that the (bias-corrected) unconditional logit estimator has desirable finite sample properties with respect to structural parameters and average partial effects. Furthermore, by combining the estimator with our novel pseudo-demeaning approach, our algorithm is linear in both panel dimensions.

Thus, the (bias-corrected) unconditional logit estimator is a promising candidate for many relevant applications based on large panel data. To allow the readers to use our algorithm in a straightforward and convenient way, we provide an implementation in our *R*-package *bife*.

We would like to draw the attention of our readers to the fact that our pseudo-demeaning paves the way to derive algorithms for more complex nonlinear fixed effects models. Stammann (2018) combines the pseudo-demeaning with the method of alternating projections (MAP) to develop a feasible algorithm for the estimation of generalized linear models with multiple high-dimensional fixed effects. The combination of MAP and pseudo-demeaning can also be extended to bias corrections with multiple fixed effects, as shown by Czarnowske and Stammann (2019) for binary choice models with additive unobservable individual and time effects.

References

Andersen, Erling Bernhard. 1970. “Asymptotic properties of conditional maximum-likelihood estimators.” *Journal of the Royal Statistical Society. Series B:* 283–301.

Arellano, Manuel, and Jinyong Hahn. 2007. “Understanding bias in nonlinear panel models: Some recent developments.” *Econometric Society Monographs* 43:381.

Bartolucci, Francesco, and Claudia Pigini. 2019. “Partial effects estimation for fixed-effects logit panel data models.” *Working Paper*.

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.” *Review of Economic Studies* 47:225–238.

Czarnowske, Daniel, and Amrei Stammann. 2019. “Binary Choice Models with High-Dimensional Individual and Time Fixed Effects.” *arXiv preprint:1904.04217*.

D’Haultfœuille, Xavier, and Alessandro Iaria. 2016. “A convenient method for the estimation of the multinomial logit model with fixed effects.” *Economics Letters* 141:77–79.

Dhaene, Geert, and Koen Jochmans. 2015. “Split-panel jackknife estimation of fixed-effect models.” *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* 150:71–85.

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.

Firth, David. 1993. “Bias reduction of maximum likelihood estimates.” *Biometrika* 80 (1): 27–38.

Frisch, Ragnar, and Frederick V. Waugh. 1933. “Partial Time Regressions as Compared with Individual Trends.” *Econometrica* 1 (4): 387–401.

Gail, Mitchell H., Jay H. Lubin, and Lawrence V. Rubinstein. 1981. “Likelihood calculations for matched case-control studies and survival studies with tied death times.” *Biometrika* 68 (3): 703–707.

Gaure, Simen. 2012. "A Faster Algorithm for Computing the Conditional Logit Likelihood." *Unpublished Note.*

Greene, William. 2004. "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects." *Econometrics Journal* 7:98–119.

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

Hall, Bronwyn H. 1978. "A general framework for the time series-cross section estimation." *Annales de l'INSEE* 30–31:177–202.

Hyslop, Dean R. 1999. "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women." *Econometrica* 67 (6): 1255–1294.

Kosmidis, Ioannis, and David Firth. 2009. "Bias reduction in exponential family nonlinear models." *Biometrika* 96 (4): 793–804.

Kunz, Johannes S., Kevin E. Staub, and Rainer Winkelmann. 2018. "Predicting fixed effects in panel probit models." *Working Paper.*

Lovell, Michael C. 1963. "Seasonal adjustment of economic time series and multiple regression analysis." *Journal of the American Statistical Association* 58 (304): 993–1010.

McFadden, Daniel. 1978. "Modeling the choice of residential location." *Transportation Research Record,* no. 673.

Nelder, John Ashworth, and Robert William Maclagan Wedderburn. 1972. "Generalized linear models." *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–384.

Neyman, Jerzy, and Elizabeth L Scott. 1948. "Consistent estimates based on partially consistent observations." *Econometrica* 16 (1): 1–32.

Prentice, Ross L., and Lynn A. Gloeckler. 1978. "Regression analysis of grouped survival data with application to breast cancer data." *Biometrics:* 57–67.

R Core Team. 2019. *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.*

Reid, Stephen, and Rob Tibshirani. 2014. “Regularization paths for conditional logistic regression: The clogitl1 package.” *Journal of Statistical Software* 58 (12).

Stammann, Amrei. 2018. “Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects.” *arXiv preprint:1707.01815v3*.

Appendix

A Details on the Implementations

A.1 Brute-Force UCL Estimation

Let $\mathbf{Z} = [\mathbf{X}, \mathbf{D}]$ denote the $NT \times (M+N)$ regressor matrix, where $\mathbf{D}$ is the $NT \times N$ dummy variable matrix corresponding to the fixed effects and $\mathbf{X}$ is the $NT \times M$ matrix of the remaining regressors. In a similar way as Greene (2004) we define the gradient and the Hessian of UCL. The $(N+M) \times 1$ gradient is given by

$$\mathbf{g} = [\mathbf{g}'_\beta, \mathbf{g}'_\alpha]'$$

with

$$\mathbf{g}_\beta = \frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - p_{it}),$$

$$\mathbf{g}_{\alpha_i} = \frac{\partial L}{\partial \alpha_i} = \sum_{t=1}^{T} (y_{it} - p_{it}),$$

and the $(N+M) \times (N+M)$ Hessian takes the following form

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{\beta\beta} & \mathbf{h}_{\beta\alpha_1} & \mathbf{h}_{\beta\alpha_2} & \cdots & \mathbf{h}_{\beta\alpha_N} \\ \mathbf{h}_{\alpha_1\beta} & h_{\alpha_1\alpha_1} & 0 & \cdots & 0 \\ \mathbf{h}_{\alpha_2\beta} & 0 & h_{\alpha_2\alpha_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{\alpha_N\beta} & 0 & 0 & \cdots & h_{\alpha_N\alpha_N} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\alpha} \\ \mathbf{H}_{\alpha\beta} & \mathbf{H}_{\alpha\alpha} \end{pmatrix}$$

with

$$\mathbf{H}_{\beta\beta} = \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\partial^2 L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it} \mathbf{x}'_{it} p_{it}(1 - p_{it}),$$

$$\mathbf{h}_{\beta\alpha_i} = \sum_{t=1}^{T} \frac{\partial^2 L}{\partial \boldsymbol{\beta} \partial \alpha_i} = -\sum_{t=1}^{T} \mathbf{x}_{it} p_{it}(1 - p_{it}),$$

$$h_{\alpha_i\alpha_i} = \sum_{t=1}^{T} \frac{\partial^2 L}{\partial \alpha_i^2} = -\sum_{t=1}^{T} p_{it}(1 - p_{it}).$$

Thus, the $(k-1)$-th Newton-Raphson update in (2.3) can be rewritten as follows:

$$(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) = -\mathbf{H}^{-1}\mathbf{g} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{p}), \tag{2.23}$$

where the $NT \times NT$ matrix $\mathbf{W}$ serves as a weighting matrix. $\mathbf{W}$ is a diagonal-matrix with strictly positive weights $w_{it} = p_{it}(1 - p_{it})$ where p_{it} is defined in (2.1). Note that

the weights and all dependent quantities are evaluated at $\boldsymbol{\theta}^{k-1}$.

A.2 Recursive CL Estimation

Gail, Lubin, and Rubinstein (1981) proposed an recursive implementation of CL. This approach accelerates the computation while retaining the exactness of the brute force approach presented in section 2.2.

The individual likelihood contribution in (2.6) can be rewritten as follows:

$$\exp(L_{ci}) = \frac{\prod_{k=1}^{t_{1i}} \exp(\mathbf{x}_k' \boldsymbol{\beta})}{\sum_{h=1}^{c_i} \prod_{k_h=1}^{t_{1i}} \exp(\mathbf{x}_{k_h}' \boldsymbol{\beta})}, \tag{2.24}$$

where the index k denotes the observed data and the index k_h the h-th possible assignment. Lets define the denominator in (2.24) as follows:

$$f_i(t_{1i}, T) = \sum_{h=1}^{c_i} \prod_{k_h=1}^{t_{1i}} \exp(\mathbf{x}_{k_h}' \boldsymbol{\beta}) = \sum_{h=1}^{c_i} \prod_{k_h=1}^{t_{1i}} U_{k_h} .$$

The recursion can be specified by

$$f_i(t_{1i}, T) = f_i(t_{1i}, T-1) + U_T f_i(t_{1i}-1, T-1)$$

with $f_i(0, T) = 1$ for $T \geq 0$, $f_i(t_{1i}, T) = 0$ for $t_{1i} > T$ and $U_T = \exp(\mathbf{x}_{iT}' \boldsymbol{\beta})$. Finally, the conditional log-likelihood in (2.5) can be rewritten to

$$L_c = \sum_{i=1}^{N} L_{ci} = \sum_{i=1}^{N} \left(\sum_{t=1}^{T} y_{it} \mathbf{x}_{it}' \boldsymbol{\beta} - \log(f_i(t_{1i}, T)) \right). \tag{2.25}$$

The maximization of the conditional log-likelihood (2.25) is usually solved iteratively with gradient based maximization techniques. It is possible to apply the recurrence to the computation of the gradient and Hessian as well. However, this has not been proven to be useful since the recurrence is very time and memory consuming. Gail, Lubin, and Rubinstein (1981) proposed to implement the estimator based on numerical first and second order derivatives.

B Computational Complexities

For the following derivations we assume a balanced panel with $N \gg T \gg M$.

B.1 Brute-Force UCL Estimation

Remember the $(k-1)$-th Newton-Raphson step is

$$(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) = -\mathbf{H}^{-1}\mathbf{g} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{p})\,.$$

The most demanding part is the computation of the $(M+N)\times(M+N)$ Hessian.[27] The multiplication of the $NT\times(M+N)$ matrix $\mathbf{Z}$ with the $NT\times NT$ diagonal matrix $\mathbf{W}$ can be done in $\approx O(N^2T)$. Suppose we have already generated the variable $\mathbf{Z}_w = \mathbf{W}\mathbf{Z}$. The matrix multiplication $\mathbf{Z}'\mathbf{Z}_w$ costs $\approx O(N^3T)$, matrix multiplication $\mathbf{Z}'\mathbf{Y}$ costs $\approx O(N^2T)$, matrix inversion $(\mathbf{Z}_w'\mathbf{Z})^{-1}$ costs $\approx O(N^3)$ and finally the product of the Hessian and the gradient costs $\approx O(N^2)$. Since $O(N^3T) > O(N^2T) > O(N^2)$ the computation time increases cubically in N and linear in T.

B.2 Computationally Efficient UCL Estimation

The computational complexity of the pseudo-demeaning can be derived by considering the most extensive part, which is the computation of the structural parameter updates

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\tilde{\mathbf{y}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\tilde{y}_{it}\right).$$

Although $\ddot{\mathbf{X}}$ consists out of MNT elements, its computation requires only $\approx O(MNT)$ time, since $\sum_{t=1}^{T}\tilde{w}_{it}\tilde{x}_{it}$ is different for MN elements and $\sum_{t=1}^{T}\tilde{w}_{it}^2$ is different for N elements. Thus, $\sum_{t=1}^{T}\tilde{w}_{it}\tilde{x}_{it}$ requires $MN(T-1+T)$ arithmetic operations and $\sum_{t=1}^{T}\tilde{w}_{it}^2$ requires $N(T-1)$ arithmetic operations. The matrix multiplication $\ddot{\mathbf{X}}'\ddot{\mathbf{X}}$ costs $\approx O(M^2NT)$, matrix multiplication $\ddot{\mathbf{X}}'\widetilde{\mathbf{Y}}$ costs $\approx O(MNT)$, matrix inversion $(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}$ costs $\approx O(M^3)$ and finally the product of the Hessian and the gradient costs $\approx O(M^2)$. Altogether, $O(M^2NT) > O(MNT) > O(M^3)$. Thus, the computation time of the pseudo-demeaning is linear in T and N.

27. Note that some software routines, such as `glm()` in *R*, include n dummies instead of N and the computation becomes even more costly. Remember, $N = \sum_{i=1}^{n}\mathbf{1}[0 < \sum_{t=1}^{T} y_{it} < T]$ where n denotes the total number of individuals in the data set.

B.3 Brute-Force CL Estimation

Next, we demonstrate the computational complexity of brute force implementation of CL. Taking into account that y_{it} is binary, (2.5) can be rewritten to

$$L_c(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log \left(\frac{\exp\left(\sum_{k=1}^{t_{1i}} \mathbf{x}'_{ik}\boldsymbol{\beta}\right)}{\sum_{h=1}^{c_i} \exp\left(\sum_{k_h=1}^{t_{1i}} \mathbf{x}'_{ik_h}\boldsymbol{\beta}\right)} \right), \tag{2.26}$$

where the index k denotes the observed data and the index k_h the h-th possible assignment. Lets consider the individual likelihood contribution

$$\exp(L_{ci}) = \frac{\prod_{k=1}^{t_{1i}} \exp(\mathbf{x}'_k\boldsymbol{\beta})}{\sum_{h=1}^{c_i} \prod_{k_h=1}^{t_{1i}} \exp(\mathbf{x}'_{k_h}\boldsymbol{\beta})}. \tag{2.27}$$

A direct evaluation of the denominator in (2.27) requires the summation of c_i terms and becomes prohibitive if T increases (see Gail, Lubin, and Rubinstein 1981). The computation of the denominator involves roughly $t_{1i}c_i$ arithmetic operations: there are $(c_i - 1)$ outer additions and $(t_{1i} - 1)$ inner multiplications. Thus, evaluating the log-likelihood, as it is required by a numerical optimization routine, costs $\approx O(\sum_{i=1}^{N} t_{1i}\binom{T}{t_{1i}})$. Since t_{1i} is a proportion of T, which usually grows with T, the complexity is exponential in T.

B.4 Recursive CL Estimation

In order to determine the computational complexity of the recursive implementation of CL, we consider how it tackles the problem of computing the denominator of (2.27). Reid and Tibshirani (2014) have shown that the denominator can be computed in $\approx O(t_{1i}(T-t_{1i})))$ time. Thus evaluating the log-likelihood, takes $\approx O(\sum_{i=1}^{N} t_{1i}(T-t_{1i})))$. Hence, the computational complexity is linear in N and roughly quadratic in T since t_{1i} is a proportion of T, which usually grows with T. Even if one follows Simen Gaure's recommendation not to set up the program completely recursively, but to reuse intermediate results, nothing changes in the form of computational complexity, since it is reduced only by a factor (see Gaure 2012).

B.5 CL Estimation with Random Subsets

CLsub reduces the number of arithmetic operations per individual from roughly $t_{1i}c_i$ with the brute force CL algorithm to $t_{1i}m$, since CLsub requires only $(m-1)$ outer additions and still $(t_{1i}-1)$ inner multiplications. Hence, CLsub requires $\approx O(m\sum_{i=1}^{N} t_{1i})$ to evaluate the log-likelihood function. Therefore, the shape of the computational complexity depends on the choice of m. If m is a function of T the

computational complexity evolves roughly quadratically in T, else linearly.

C Details on Average Partial Effects

Remember that we estimate $N \leq n$ fixed effects since we use the reduced sample in the optimization of the log-likelihood. For those individuals who don't change their status over time (perfectly classified) the estimates of the fixed effects are unbounded. Thus, their estimates of partial effects are zero as shown in the following.

Non-binary regressor:

$$\lim_{\hat{\alpha}_i \to \infty} \hat{\Delta}_{it}^k = \underbrace{\lim_{\hat{\alpha}_i \to \infty} \Pr(y_{it} = 1 | \mathbf{x}_{it}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)}_{=1} \underbrace{\lim_{\hat{\alpha}_i \to \infty} [1 - \Pr(y_{it} = 1 | \mathbf{x}_{it}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)]}_{0} \hat{\beta}_k = 0$$

$$\lim_{\hat{\alpha}_i \to -\infty} \hat{\Delta}_{it}^k = \underbrace{\lim_{\hat{\alpha}_i \to -\infty} \Pr(y_{it} = 1 | \mathbf{x}_{it}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)}_{=0} \underbrace{\lim_{\hat{\alpha}_i \to -\infty} [1 - \Pr(y_{it} = 1 | \mathbf{x}_{it}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)]}_{1} \hat{\beta}_k = 0$$

Binary regressor:

$$\begin{aligned}
\lim_{\hat{\alpha}_i \to \infty} \hat{\Delta}_{it}^k = &\underbrace{\lim_{\hat{\alpha}_i \to \infty} \Pr(y_{it} = 1 | x_{itk} = 1, \mathbf{x}_{it\{-k\}}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)}_{=1} - \\
&\underbrace{\lim_{\hat{\alpha}_i \to \infty} \Pr(y_{it} = 1 | x_{itk} = 0, \mathbf{x}_{it\{-k\}}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)}_{=1} = 0 \\
\lim_{\hat{\alpha}_i \to -\infty} \hat{\Delta}_{it}^k = &\underbrace{\lim_{\hat{\alpha}_i \to -\infty} \Pr(y_{it} = 1 | x_{itk} = 1, \mathbf{x}_{it\{-k\}}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)}_{=0} - \\
&\underbrace{\lim_{\hat{\alpha}_i \to -\infty} \Pr(y_{it} = 1 | x_{itk} = 0, \mathbf{x}_{it\{-k\}}, \hat{\boldsymbol{\beta}}, \hat{\alpha}_i)}_{=0} = 0
\end{aligned}$$

Chapter 3

Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects

3.1 Introduction

Fixed effects models are popular specifications in panel data econometrics to account for unobserved heterogeneity. Some relevant examples can be found in labor economics or in international trade. For instance, models of labor supply incorporate individual and time fixed effects to control for individual specific taste for labor and time specific shifts in preferences (see among others Hyslop 1999). A further example is the estimation of structural gravity models in international trade. Here importer-time and exporter-time fixed effects are required to account for "multilateral resistances" (see Anderson and Van Wincoop 2003) and often additionally dyadic (exporter-importer pairs) fixed effects to control for unobserved bilateral heterogeneity (see Baier and Bergstrand 2007). Due to the rising availability of large micro-level panel data like the U.S. *Panel Study for Income Dynamics* (PSID) or pseudo-panels of trade flows from the *Centre d'Etudes Prospectives et d'Informations Internationales* (CEPII) such model specifications quickly lead to high-dimensional fixed effects that cause a substantial computational burden.

Usually the unobserved heterogeneity is captured by including a dummy variable for each level of each fixed effects category. In linear regression models several approaches exist to handle the computational burden arising from high-dimensional fixed effects. The most common strategy is the within transformation, which for one-way error components only requires to subtract group specific means from all variables. Balazsi, Matyas, and Wansbeek (2018) have derived and revisited various generalizations of the within transformation for the most commonly used two- and three-way error component structures. For some of them it is possible to derive scalar transformations, however, for others a projection matrix is required to within transform the variables. Especially for large data sets, the computation of the projection matrix often becomes infeasible. More flexible approaches have been proposed in the literature that directly calculate the within transformed variables for any k-way error component structure without the need to compute the projection matrix in advance. The two most popular ones are the algorithms by Guimarães and Portugal (2010) and Gaure (2013b). Whereas Guimarães and Portugal (2010) use an efficient version of a Gauss-Seidel algorithm that alternates between the solutions of normal equations, Gaure (2013b) computes the within transformed variables iteratively with the method of alternating projections (MAP).[1] Both approaches are close approximations of the brute-force dummy variable approach and especially

1. Guimarães and Portugal (2010) also sketch an alternative efficient algorithm that is actually the alternating projections approach independently proposed by Gaure (2013b). The latter introduced MAP in the context of linear regression models with high-dimensional fixed effects along with an extensive theoretical foundation.

MAP is widely used by researchers.[2]

In the field of generalized linear models (GLMs) only special cases have been treated so far. One-way fixed effects models can be estimated at low computational costs. Stammann, Heiß, and McFadden (2016) derived computationally efficient parameter update formulas using the FWL theorem (Frisch and Waugh 1933; Lovell 1963). They show that the corresponding projection results essentially in a weighted within transformation. Another approach uses the partitioned inverse formula along with the sparse structure of the Hessian (see among others Hall 1978; Chamberlain 1980; Greene 2004). Guimarães and Portugal (2010) suggest a Gauss-Seidel algorithm to estimate nonlinear models with high-dimensional multi-way fixed effects. Using the example of poisson regression they demonstrate how a closed form of the fixed effects can be exploited to derive a feasible algorithm. However, most GLMs do not have such a closed form. For these cases Guimarães and Portugal (2010) show that the Gauss-Seidel algorithm can be combined with a demanding numerical optimization routine to solve for the fixed effects. Recently, two modifications of the Gauss-Seidel algorithm of Guimarães and Portugal (2010) have been proposed: Larch et al. (2019) provide a *Stata* routine to estimate poisson gravity models with a three-way error component and Bergé (2018) offers an *R*-package to estimate poisson, logit and negative binomial models with a multi-way error structure. Even more recently, Correia, Guimarães, and Zylkin (2019) provide a *Stata* routine *ppmlhdfe* for poisson models with a k-way error component that uses a similar approach as we propose in this article.[3]

For GLMs with a k-way error component a general and memory efficient algorithm is still missing. We close this gap by deriving a straightforward maximum likelihood approach that can be easily incorporated into existing GLM software architectures.[4] Moreover, it is very flexible because it can be directly applied to unbalanced data and linear dependencies between fixed effects do not need to be addressed. Our algorithm combines the work of Gaure (2013b) and Stammann, Heiß, and McFadden (2016) by embedding MAP into a Newton-Raphson optimization

2. Software for linear fixed effects models based on MAP is provided in the *R*-package *lfe* (Gaure 2013a) and in the *Stata* routine *reghdfe* (Correia 2016).

3. The routine *ppmlhdfe* is an extension of Paulo Guimaraes *Stata* routine *poi2hdfe* (Guimarães 2014), which is limited to poisson models with two-way fixed effects. *poi2hdfe* uses the method of alternating projections by incorporating the *Stata* routine *hdfe* of Correia (2016) into an iteratively reweighted least squares algorithm. The underlying approach is similar, albeit different, to the one we present in this article. Both have been independently developed. To the best of our knowledge the routine used by *poi2hdfe* and *ppmlhdfe* has not been presented in an article until February 2019 (Correia, Guimarães, and Zylkin 2019).

4. Our algorithm is made available on CRAN as an *R*-package *alpaca* (co-authored with Daniel Czarnowske): `https://cran.r-project.org/web/packages/alpaca/index.html`. Note that *alpaca* only provides routines for nonlinear GLMs because there is already a comprehensive *R*-package *lfe* by Simen Gaure for linear regression models (Gaure 2013a). Additionally *alpaca* allows to estimate negative binomial models.

algorithm to concentrate out the fixed effects from the parameter update. We refer to this step as pseudo-demeaning to highlight its link to the well-known within transformation for linear regression models.

Another problem besides the computational burden of many nonlinear models with fixed effects is the incidental parameters problem known since Neyman and Scott (1948). As a consequence the estimators of the structural parameters and partial effects are inconsistent. For some selected error components and types of models bias corrections have been proposed (see Fernández-Val and Weidner 2018a for an overview). A very simple type of bias correction is the split-panel jackknife tracing back to Dhaene and Jochmans (2015) that can be directly combined with our suggested algorithm given the researcher knows the order of the bias. A recently proposed heuristic by Fernández-Val and Weidner (2018a) helps to determine the order. With this knowledge one can exploit the relation between sample size and bias to form a suitable correction from multiple estimates based on subsamples. This however makes the split-panel jackknife computationally demanding but combined with our efficient software routine still manageable.

The remainder of the article is organized as follows. First we introduce the k-way fixed effects GLM in section 3.2. In section 3.3 we derive a computationally efficient algorithm based on the FWL theorem and MAP. Afterwards a simulation study in section 3.4 demonstrates the performance of our algorithm and finally an empirical example from international trade highlights its practical relevance in section 3.5. Finally, section 3.6 concludes.

3.2 The Model and Brute-Force Estimation

In this section we introduce a generalized linear model (GLM) with k-way fixed effects and its standard estimation procedure. Every GLM consists of three parts: a stochastic component $\boldsymbol{\mu}$, a systematic component $\boldsymbol{\eta}$, also known as the linear predictor, and a link function $h(\cdot)$ between both components (see McCullagh and Nelder 1989).

In case of a k-way additive separable error component, the linear predictor takes the following specific form:

$$\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\gamma} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} = \sum_{k=1}^{K} \mathbf{D}_k \boldsymbol{\alpha}_k + \mathbf{X}\boldsymbol{\beta},$$

where the regressor matrix $\mathbf{Z}$ can be split into a $n \times p$ matrix $\mathbf{X}$ containing the variables of interest and a sparse $n \times l$ matrix $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_K]$, with n denoting the number of observations. More specifically, the submatrices $\mathbf{D}_k$ arise from dummy encoding K categorical variables that are used to capture different sources of unob-

served heterogeneity. Each dummy matrix is of dimension $n \times l_k$, where l_k is the number of levels of the k-th categorical variable, such that $l = \sum_{k=1}^{K} l_k$. Throughout the article we refer to $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_K']'$ and $\boldsymbol{\beta}$ as fixed effects and structural parameters, respectively. For the sake of clarity we only use notation for balanced data, but all approaches presented in this article are also directly applicable to unbalanced data.

The remaining components of a GLM can be expressed as follows:

$$\mathbb{E}(\mathbf{y}) = \boldsymbol{\mu} = h^{-1}(\boldsymbol{\eta}),$$

where the link function $h(\cdot)$ is a monotonic differentiable function and $\mathbf{y}$ is a realization of an independently distributed random variable from the exponential family $\mathbf{Y}$. The distribution of $\mathbf{Y}$ is given by

$$f_Y(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right),$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are specific functions. In this article ϕ is known and thus θ is a canonical parameter. Table 3.1 summarizes the corresponding functions and parameters of GLMs that are frequently used in economics. Examples of other GLMs are given in McCullagh and Nelder (1989).

Table 3.1: *Common Model Families*

	Logit	Probit	Poisson
Dispersion parameter ϕ	1	1	1
Scale parameter a	1	1	1
Cumulant function $b(\theta)$	$\log(1+\exp(\theta))$	$\log(1+\exp(\theta))$	$\exp(\theta)$
$c(y, \phi)$	0	0	$\log(y!)$
$\mu(\theta)$	$\exp(\theta)/(1+\exp(\theta))$	$\Phi(\theta)$	$\exp(\theta)$
Canonical link $\theta(\mu)$	$\log(\mu/(1-\mu))$	$\log(\mu/(1-\mu))$	$\log(\mu)$
Variance function $V(\mu)$	$\mu(1-\mu)$	$\mu(1-\mu)$	μ

Note: $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.
Source: Modification of table 2.1 in McCullagh and Nelder (1989).

The unknown parameters $\boldsymbol{\gamma} = [\boldsymbol{\alpha}', \boldsymbol{\beta}']'$ are estimated jointly using the method of maximum likelihood. The corresponding log-likelihood function is

$$\mathscr{L}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \frac{y_i \theta_i(\mu_i(\boldsymbol{\gamma})) - b(\theta_i(\mu_i(\boldsymbol{\gamma})))}{a(\phi)} + c(y_i, \phi),$$

which can be maximized iteratively using a standard Newton-Raphson algorithm. The parameter update in iteration $(r-1)$ can be expressed as

$$(\boldsymbol{\gamma}^r - \boldsymbol{\gamma}^{r-1}) = -\left(\mathbf{H}^{r-1}\right)^{-1} \mathbf{g}^{r-1}, \tag{3.1}$$

where the superscript r indicates the iteration number and $\mathbf{g}$ and $\mathbf{H}$ are the gradient and Hessian, respectively. Since $\theta(\boldsymbol{\mu}(\boldsymbol{\gamma}))$ is the canonical link we can apply the chain rule resulting in the following expression of the gradient:

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\gamma}^r} = \mathbf{g}^r = \mathbf{Z}'\mathbf{W}^r \boldsymbol{\nu}^r \,,$$

where $\boldsymbol{\nu}^r = (\mathbf{y} - \boldsymbol{\mu}^r) \odot \partial \boldsymbol{\eta}^r / \partial \boldsymbol{\mu}^r$, $\mathbf{W}^r$ is a positive definite diagonal weighting matrix with its i-th entry equal to $(\partial \mu_i^r / \partial \eta_i^r)^2 / V_i^r$. The Hessian

$$\frac{\partial^2 \mathscr{L}}{\partial \boldsymbol{\gamma}^r \partial \boldsymbol{\gamma}^{r\prime}} = \mathbf{H}^r = -\mathbf{Z}'\mathbf{W}^r\mathbf{Z}$$

can be derived analogously.

At this point we need to assume that $\mathbf{Z}$ has full column rank. If we assume that this holds for $\mathbf{X}$, too, full column rank of $\mathbf{Z}$ can usually be achieved by dropping some reference categories in $\mathbf{D}$.[5] For example in the classical two-way fixed effects model with individual and time fixed effects usually one column associated with a certain time period is removed from $\mathbf{D}$ (given $\mathbf{X}$ does not include an intercept). In models with more complicated error structures this might not be that straightforward, such that in general $\dim(\mathbf{Z}) = n \times (p + l)$, where $l \leq \sum_{k=1}^{K} l_k$. Brute-force estimation of (3.1) requires the computation and inversion of a potentially large Hessian of dimension $(p + l) \times (p + l)$, which quickly becomes computationally demanding or even infeasible.

In the next section we present a new Newton-Raphson pseudo-demeaning algorithm based on the Frisch-Waugh-Lovell (FWL) theorem and MAP, which substantially decreases the computational costs of the optimization problem.

3.3 Estimation with High-Dimensional Fixed Effects

3.3.1 The Newton-Raphson Pseudo-Demeaning Algorithm

In the classical linear fixed effects model the FWL theorem is applied to separate the estimation of the fixed effects from the structural parameters. Recently, Stammann, Heiß, and McFadden (2016) have shown in the context of one-way fixed effects logit models how the FWL theorem can be adapted to separate the Newton-Raphson update of the structural parameters from the fixed effects update.

The same logic can be applied to GLMs with any additive separable k-way error

5. In some cases it might also occur that some columns of $\mathbf{X}$ can be perfectly explained by columns in $\mathbf{D}$.

structure since the parameter update

$$(\boldsymbol{\gamma}^r - \boldsymbol{\gamma}^{r-1}) = (\mathbf{Z}'\mathbf{W}^{r-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}^{r-1}\boldsymbol{\nu}^{r-1} \tag{3.2}$$

is essentially the least-squares solution of

$$\tilde{\boldsymbol{\nu}}^{r-1} = \widetilde{\mathbf{D}}^{r-1}(\boldsymbol{\alpha}_0^r - \boldsymbol{\alpha}_0^{r-1}) + \widetilde{\mathbf{X}}^{r-1}(\boldsymbol{\beta}_0^r - \boldsymbol{\beta}_0^{r-1}) + \tilde{\mathbf{u}}^{r-1}, \tag{3.3}$$

where $\tilde{\boldsymbol{\nu}}^r = \widetilde{\mathbf{W}}^r((\mathbf{y} - \boldsymbol{\mu}^r) \odot \partial\boldsymbol{\eta}^r/\partial\boldsymbol{\mu}^r)$, $\widetilde{\mathbf{D}}^r = \widetilde{\mathbf{W}}^r\mathbf{D}$, $\widetilde{\mathbf{X}}^r = \widetilde{\mathbf{W}}^r\mathbf{X}$, $\widetilde{\mathbf{W}}^r = (\mathbf{W}^r)^{1/2}$, $\tilde{\mathbf{u}}^r$ is an error term, and subscript zero denotes the population parameters.[6]

Since we know that (3.3) is a linear regression model we can apply the FWL theorem to partial out $\widetilde{\mathbf{D}}^{r-1}(\boldsymbol{\alpha}_0^r - \boldsymbol{\alpha}_0^{r-1})$. This yields the following concentrated regression:

$$\begin{aligned}\mathbf{M}_{\widetilde{\mathbf{D}}}^{r-1}\tilde{\boldsymbol{\nu}}^{r-1} &= \mathbf{M}_{\widetilde{\mathbf{D}}}^{r-1}\widetilde{\mathbf{D}}^{r-1}(\boldsymbol{\alpha}_0^r - \boldsymbol{\alpha}_0^{r-1}) + \mathbf{M}_{\widetilde{\mathbf{D}}}^{r-1}\widetilde{\mathbf{X}}^{r-1}(\boldsymbol{\beta}_0^r - \boldsymbol{\beta}_0^{r-1}) + \mathbf{M}_{\widetilde{\mathbf{D}}}^{r-1}\tilde{\mathbf{u}}^{r-1} \\ &= \mathbf{M}_{\widetilde{\mathbf{D}}}^{r-1}\widetilde{\mathbf{X}}^{r-1}(\boldsymbol{\beta}_0^r - \boldsymbol{\beta}_0^{r-1}) + \mathbf{M}_{\widetilde{\mathbf{D}}}^{r-1}\tilde{\mathbf{u}}^{r-1},\end{aligned} \tag{3.4}$$

where the annihilator matrix $\mathbf{M}_{\widetilde{\mathbf{D}}}^r = \mathbf{I}_n - \widetilde{\mathbf{D}}^r(\widetilde{\mathbf{D}}^{r\prime}\widetilde{\mathbf{D}}^r)^{-1}\widetilde{\mathbf{D}}^{r\prime}$ is the projection onto the orthogonal complement of the column space of $\widetilde{\mathbf{D}}^r$.[7] Thus the parameter update of the structural parameters can be obtained as the least-squares solution of (3.4):

$$(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1}) = (\ddot{\mathbf{X}}^{r-1\prime}\ddot{\mathbf{X}}^{r-1})^{-1}\ddot{\mathbf{X}}^{r-1\prime}\ddot{\boldsymbol{\nu}}^{r-1}, \tag{3.5}$$

where $\ddot{\boldsymbol{\nu}}^r = \mathbf{M}_{\widetilde{\mathbf{D}}}^r\tilde{\boldsymbol{\nu}}^r$ and $\ddot{\mathbf{X}}^r = \mathbf{M}_{\widetilde{\mathbf{D}}}^r\widetilde{\mathbf{X}}^r$. Throughout the article we denote $\ddot{\boldsymbol{\nu}}^r$ and $\ddot{\mathbf{X}}^r$ as pseudo-demeaned variables. Equation (3.5) can be interpreted as a Newton-Raphson update based on a concentrated gradient and Hessian. Note that once the pseudo-demeaned variables are computed, the parameter update $(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1})$ only requires to invert a $p \times p$ instead of the $(p+l) \times (p+l)$ matrix in (3.1). However, the computation of the pseudo-demeaned variables is an additional challenge, since the annihilator matrix $\mathbf{M}_{\widetilde{\mathbf{D}}}^r$ has dimension $n \times n$ and is typically non-sparse.[8] This issue will be addressed in the next subsection.

In the following, let us assume that we know a feasible approach to pseudo-

6. The standard approach to estimate GLMs is iteratively reweighted least squares (IRLS) with the following update: $\boldsymbol{\gamma}^r = (\mathbf{Z}'\mathbf{W}^{r-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}^{r-1}\left(\boldsymbol{\nu}^{r-1} + \mathbf{Z}\boldsymbol{\gamma}^{r-1}\right)$. We use the different formulation (3.2) in order to obtain the scores of the log-likelihood directly from the estimation routine. These are required to compute robust and (multi-way) clustered standard errors.

7. Note, $\mathbf{M}$ is idempotent and that (3.4) can be transformed into $\widetilde{\mathbf{W}}^{r-1}\mathbf{P}^{r-1}\boldsymbol{\nu}^{r-1} = \widetilde{\mathbf{W}}^{r-1}\mathbf{P}^{r-1}\mathbf{X}^{r-1}(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1}) + \widetilde{\mathbf{W}}\mathbf{P}^{r-1}\tilde{\mathbf{u}}^{r-1}$, where $\mathbf{P}^r = \mathbf{I}_n - \mathbf{D}(\mathbf{D}'\mathbf{W}^r\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}^r$. Both projection approaches are suitable to concentrate out the high-dimensional fixed effects from (3.3). Throughout the article we restrict ourselves to projection $\mathbf{M}$.

8. One exception is the case $K = 1$ where the block-diagonal structure of the annihilator matrix allows to derive a straightforward scalar expression to compute the pseudo-demeaned variables (see Stammann, Heiß, and McFadden 2016). For $K > 1$ this is not possible since the annihilator matrix loses its sparse structure.

demean the variables. To complete the entire estimation routine, we require to compute a concentrated gradient and Hessian in each iteration of the optimization routine. Since those are functions of the linear predictor $\boldsymbol{\eta}^r = \mathbf{D}\boldsymbol{\alpha}^r + \mathbf{X}\boldsymbol{\beta}^r$ we need to find an efficient way to update $\boldsymbol{\eta}^r$. The naive approach would recover estimates of the fixed effects and use them to update the linear predictor. We present a substantially less costly approach that directly recovers the linear predictor from already computed quantities.[9] Therefore reconsider the reformulation of the Newton-Raphson update into the regression model

$$\tilde{\boldsymbol{\nu}}^{r-1} = \tilde{\mathbf{D}}^{r-1}(\boldsymbol{\alpha}_0^r - \boldsymbol{\alpha}_0^{r-1}) + \tilde{\mathbf{X}}^{r-1}(\boldsymbol{\beta}_0^r - \boldsymbol{\beta}_0^{r-1}) + \tilde{\mathbf{u}}^{r-1}. \tag{3.6}$$

Following Gaure (2013b), it can be shown that the residuals of the projected system (3.4) are identical to the ones of the full system (3.6):

$$\tilde{\boldsymbol{\nu}}^{r-1} - \tilde{\mathbf{X}}^{r-1}(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1}) - \tilde{\mathbf{D}}^{r-1}(\boldsymbol{\alpha}^r - \boldsymbol{\alpha}^{r-1}) = \ddot{\boldsymbol{\nu}}^{r-1} - \ddot{\mathbf{X}}^{r-1}(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1}). \tag{3.7}$$

Solving (3.7) for $\boldsymbol{\eta}^r$ delivers an efficient formula to obtain the linear predictor

$$\boldsymbol{\eta}^r = (\tilde{\mathbf{w}}^{r-1})^{-1} \odot \left(\tilde{\boldsymbol{\nu}}^{r-1} - \ddot{\boldsymbol{\nu}}^{r-1} - \ddot{\mathbf{X}}^{r-1}(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1})\right) + \boldsymbol{\eta}^{r-1}$$

from already computed quantities, where $\tilde{\mathbf{w}}^r = \operatorname{diag}(\tilde{\mathbf{W}}^r)$. Bringing together all previously mentioned components the Newton-Raphson k-way pseudo-demeaning algorithm can be summarized by algorithm 2.

Algorithm 2 Newton-Raphson with Pseudo-Demeaning

1: Initialize $\boldsymbol{\beta}^0$, $\boldsymbol{\eta}^0$, and $r = 0$.
2: **repeat**
3: Set $r = r + 1$.
4: Compute the weights $\tilde{\mathbf{w}}^{r-1}$ and $\boldsymbol{\nu}^{r-1}$.
5: Compute $\tilde{\boldsymbol{\nu}}^{r-1}$ and $\tilde{\mathbf{X}}^{r-1}$.
6: Compute $\ddot{\boldsymbol{\nu}}^{r-1}$ and $\ddot{\mathbf{X}}^{r-1}$.
7: Update $\boldsymbol{\beta}^r$.
8: Update $\boldsymbol{\eta}^r$.
9: **until convergence**.

So far we have only dealt with estimating the structural parameters but usually we are also interested in inference. Since our algorithm is a maximum likelihood approach it facilitates the construction of different covariance estimators and allows for standard testing procedures. Let $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimator of

9. If required, the fixed effects coefficients can be computed ex-post with a numerical solver for linear systems of equations as presented in appendix B. However, most of the times, the fixed effects coefficients are not necessary. For ex-post computations like predictions or partial effects, the linear predictor obtained after convergence of algorithm 2 will suffice.

the structural parameters. Estimates of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be easily computed using the concentrated Hessian $\ddot{\mathbf{H}}$ and/or scores $\ddot{\mathbf{G}}$ after convergence of algorithm 2. Note that $\ddot{\mathbf{H}} = -\ddot{\mathbf{X}}'\ddot{\mathbf{X}}$ and $\ddot{\mathbf{G}} = [\ddot{\mathbf{g}}_1, \dots, \ddot{\mathbf{g}}_p]$, where $\ddot{\mathbf{g}}_j = \ddot{\mathbf{x}}_j \odot \ddot{\boldsymbol{\nu}}$ and $\ddot{\mathbf{x}}_j$ is the j-th column of $\ddot{\mathbf{X}}$. Standard covariance estimators are

$$\begin{aligned}
\widehat{\mathbf{V}}_{emp}(\hat{\boldsymbol{\beta}}) &= -\ddot{\mathbf{H}}^{-1}, \\
\widehat{\mathbf{V}}_{opg}(\hat{\boldsymbol{\beta}}) &= \left(\ddot{\mathbf{G}}'\ddot{\mathbf{G}}\right)^{-1}, \\
\widehat{\mathbf{V}}_{rob}(\hat{\boldsymbol{\beta}}) &= \ddot{\mathbf{H}}^{-1}\ddot{\mathbf{G}}'\ddot{\mathbf{G}}\ddot{\mathbf{H}}^{-1},
\end{aligned}$$

where $\widehat{\mathbf{V}}_{emp}(\hat{\boldsymbol{\beta}})$ is the estimator based on the inverse of the empirical Hessian, $\widehat{\mathbf{V}}_{opg}(\hat{\boldsymbol{\beta}})$ is known as the BHHH estimator, and $\widehat{\mathbf{V}}_{rob}(\hat{\boldsymbol{\beta}})$ is the sandwich estimator for robust standard-errors.

Remember, algorithm 2 is only computationally efficient if we know a feasible approach to compute the pseudo-demeaned variables. In the next subsection, we show how a combination of the one-way pseudo-demeaning proposed by Stammann, Heiß, and McFadden (2016) along with MAP can be used to approximate the pseudo-demeaned variables directly without having to compute the expensive and potentially infeasible annihilator matrix $\mathbf{M}^r_{\widetilde{\mathbf{D}}}$.

3.3.2 The Method of Alternating Projections

An approach to compute the pseudo-demeaned variables efficiently is the method of alternating projections (MAP) tracing back to Neumann (1950) and Halperin (1962). Gaure (2013b) introduced MAP in the context of classical linear models with many fixed effects categories. His solution is widely accepted among researchers as an equivalent to the brute-force dummy variable approach. In the following, we adapt MAP to GLMs.

First of all, we follow Gaure (2013b) and show why MAP is suitable for pseudo-demeaning. Let $\mathbf{A}$ be an arbitrary matrix, $R(\mathbf{A})$ its column space, and $R(\mathbf{A})^\perp$ the orthogonal complement of $R(\mathbf{A})$. Suppose we want to compute $\ddot{\mathbf{v}} = \mathbf{M}_{\widetilde{\mathbf{D}}}\mathbf{v}$ where $\mathbf{v}$ is an arbitrary $n \times 1$ vector. Since $\mathbf{M}_{\widetilde{\mathbf{D}}}$ is the projection onto the orthogonal complement of the column space of $\widetilde{\mathbf{D}}$, it follows that $\ddot{\mathbf{v}} \in R(\widetilde{\mathbf{D}})^\perp$. Further, since $R(\widetilde{\mathbf{D}})^\perp = \cap_{k=1}^K R(\widetilde{\mathbf{D}}_k)^\perp$, the pseudo-demeaned variable lies in the intersection of the subspaces $R(\widetilde{\mathbf{D}}_k)^\perp$, i.e. $\ddot{\mathbf{v}} \in \cap_{k=1}^K R(\widetilde{\mathbf{D}}_k)^\perp$. Since MAP is generally used to approximate a point in the intersection of a finite number of closed subspaces of a Hilbert space (see Escalante and Raydan 2011) it is suitable to find $\ddot{\mathbf{v}}$. In a nutshell, the idea is to approximate $\ddot{\mathbf{v}}$ by repeatedly projecting onto the individual subspaces $R(\widetilde{\mathbf{D}}_k)^\perp$. This is often computationally more efficient than having to compute $\mathbf{M}_{\widetilde{\mathbf{D}}}$ in advance. Another great advantage of MAP is that, unlike the dummy variable approach, the full

column rank assumption of $\mathbf{D}$ and $\widetilde{\mathbf{D}}$ is no longer required. To get an intuition, let $\widetilde{\mathscr{D}}$ denote a rank deficient weighted dummy matrix where no collinear columns have been removed. The structural parameter updates (3.5) are not influenced by the design of the dummy variable matrix, since $R(\widetilde{\mathbf{D}})^{\perp} = R(\widetilde{\mathscr{D}})^{\perp}$, and thus $\tilde{\boldsymbol{\nu}}$ and $\widetilde{\mathbf{X}}$ are projected onto the correct space anyway. For simplicity we do not further distinguish whether $\mathbf{D}$ and $\widetilde{\mathbf{D}}$ are rank deficient or not.[10]

There are basically two methods of alternating projections that differ in how they link the individual projections: Neumann-Halperin and Cimmino. Neumann (1950) developed MAP for the case of two subspaces, and Halperin (1962) extended it to a finite number of subspaces. Originally the method proposed by Cimmino (1938) is intended to solve linear systems of equations. However, as shown by Kammerer and Nashed (1972) it is also suitable for linear operations on subspaces (see Hernández-Ramos, Escalante, and Raydan 2011).

The Neumann-Halperin approach can be summarized as follows:

$$\lim_{N\to\infty} \|(\mathbf{M}^r_{\widetilde{\mathbf{D}}_1}\mathbf{M}^r_{\widetilde{\mathbf{D}}_2}\cdots\mathbf{M}^r_{\widetilde{\mathbf{D}}_K})^N\mathbf{v} - \mathbf{M}^r_{\widetilde{\mathbf{D}}}\mathbf{v}\| = 0\,.$$

This expression means that $\mathbf{v}$ is projected onto $R(\widetilde{\mathbf{D}}_1)^{\perp}$, resulting in vector $\mathbf{v}_1 \in R(\widetilde{\mathbf{D}}_1)^{\perp}$. $\mathbf{v}_1$ is projected onto $R(\widetilde{\mathbf{D}}_2)^{\perp}$, resulting in vector $\mathbf{v}_2 \in R(\widetilde{\mathbf{D}}_2)^{\perp}$ which is projected onto the next subspace, and so on, until we project from $R(\widetilde{\mathbf{D}}_{K-1})^{\perp}$ onto $R(\widetilde{\mathbf{D}}_K)^{\perp}$. The whole procedure is repeated until convergence.

In contrast to Neumann-Halperin's approach, Cimmino's projections are not nested. Instead one projects $\mathbf{v}$ separately onto each of the K subspaces $R(\widetilde{\mathbf{D}}_k)^{\perp}$ and computes the centroid of these projections according to

$$\lim_{N\to\infty} \|(\frac{1}{K}\sum_{k=1}^{K}\mathbf{M}^r_{\widetilde{\mathbf{D}}_k})^N\mathbf{v} - \mathbf{M}^r_{\widetilde{\mathbf{D}}}\mathbf{v}\| = 0\,.$$

With help of MAP the large and non-sparse projection $\mathbf{M}^r_{\widetilde{\mathbf{D}}}\mathbf{v}$ can be decomposed into an iterative procedure based on only sparse projections $\mathbf{M}^r_{\widetilde{\mathbf{D}}_k} = \mathbf{I}_n - \widetilde{\mathbf{D}}^r_k(\widetilde{\mathbf{D}}^{r\prime}_k\widetilde{\mathbf{D}}^r_k)^{-1}\widetilde{\mathbf{D}}^{r\prime}_k$, which translate into one-way pseudo-demeaning over category k. Using the result shown by Stammann, Heiß, and McFadden (2016), the projections

10. What is still required is that $\mathbf{X}$ has column full rank and that none of the regressors is perfectly collinear with the fixed effects. Whereas the former is easy to check the latter implies the need of a well-thought-out model specification by the researcher.

$\mathbf{M}^r_{\widetilde{\mathbf{D}}_k}\mathbf{v}$ can be efficiently computed as follows:[11]

$$(\mathbf{M}^r_{\widetilde{\mathbf{D}}_k}\mathbf{v})_i = v_i - \tilde{w}^r_i \frac{\sum_{j\in g_{k\kappa}} \tilde{w}^r_j v_j}{\sum_{j\in g_{k\kappa}} w^r_j} \quad \forall i \in g_{k\kappa}\,, \tag{3.8}$$

where $g_{k\kappa}$ defines a group consisting of those observations that share the same level κ in category k, and $\tilde{w}^r_i$ and w^r_i are the i-th diagonal entry of $\widetilde{\mathbf{W}}^r$ and $\mathbf{W}^r$, respectively. Equation (3.8) demonstrates that the individual projections essentially subtract "weighted" group means from the dependent variable $\tilde{\boldsymbol{\nu}}^r$ and the regressor matrix $\widetilde{\mathbf{X}}^r$.

In order to approximate $\ddot{\boldsymbol{\nu}}^r$ and $\ddot{\mathbf{X}}^r$, MAP is subsequently applied to $\tilde{\boldsymbol{\nu}}^r$ and each column of $\widetilde{\mathbf{X}}^r$. This could be either the Neumann-Halperin algorithm (algorithm 3) or the Cimmino algorithm (algorithm 4). A suitable selection of the tolerance level allows us to get an arbitrary close approximation of $\ddot{\boldsymbol{\nu}}^r$ and $\ddot{\mathbf{X}}^r$. Finally, the approximations of the pseudo-demeaned variables are used to make algorithm 2 efficient and feasible in the presence of high-dimensional fixed effects.

Algorithm 3 Pseudo-Demeaning: Neumann-Halperin

Let $\mathbf{v} \in \{\tilde{\boldsymbol{\nu}}^r, \tilde{\mathbf{x}}^r_j\}$, $j = 1,\dots,p$.
Set $i = 1$ and $\mathbf{z}_i = \mathbf{v}$.
repeat
    Set $\mathbf{z}_{i0} = \mathbf{z}_i$.
    for $k = 1,\dots,K$ **do**
        Compute $\mathbf{z}_{ik}$ by subtracting "weighted" group means from $\mathbf{z}_{i(k-1)}$ (see formula (3.8)).
    Set $i = i + 1$, $\mathbf{z}_i = \mathbf{z}_{iK}$.
until convergence.
Set $\ddot{\mathbf{v}} = \mathbf{z}_i$.

Algorithm 4 Pseudo-Demeaning: Cimmino

Let $\mathbf{v} \in \{\tilde{\boldsymbol{\nu}}^r, \tilde{\mathbf{x}}^r_j\}$, $j = 1,\dots,p$.
Set $i = 1$, $\mathbf{z}_i = \mathbf{v}$, and $\mathbf{z}_{sum} = \mathbf{0}_p$.
repeat
    Set $\mathbf{z}_{i0} = \mathbf{z}_i$.
    for $k = 1,\dots,K$ **do**
        Compute $\mathbf{z}_{ik}$ by subtracting "weighted" group means from $\mathbf{z}_{i0}$ (see formula (3.8)).
        $\mathbf{z}_{sum} = \mathbf{z}_{ik} + \mathbf{z}_{sum}$
    Set $i = i + 1$, $\mathbf{z}_i = \frac{1}{K}\mathbf{z}_{sum}$.
until convergence.
Set $\ddot{\mathbf{v}} = \mathbf{z}_i$.

11. It would also be possible to use the alternative projection defined in footnote 7. Although this projection seems to be favorable due to fewer operations, we found that it often takes longer to converge such that none of the projections is dominant with respect to total computation time.

3.4 Simulation Experiments

In this section we perform some simulation experiments to demonstrate the capabilities of our algorithm compared to a standard GLM routine based on dummy encoding. To this end we analyse the exactness of the parameter estimates and the corresponding standard errors and measure the computation times given different tolerance levels for MAP.

In order to illustrate the relevance of our algorithm we consider a simulation design that mimics a structural gravity model commonly used in international trade to explain the effect of policy variables on trade flows.[12] A standard structural gravity model for panel data takes the following form:

$$y_{ijt} = \exp(\lambda_{it} + \psi_{jt} + \delta_{ij} + \mathbf{x}_{ijt}'\boldsymbol{\beta})\epsilon_{ijt} \tag{3.9}$$

where exporters and importers are indexed by $i = 1,\dots,I$ and $j = 1,\dots,J$, respectively, and $t = 1,\dots,T$ is a time identifier. Further, y_{ijt} denotes the trade flows from exporter i to importer j at time t, λ_{it} is an exporter-time fixed effect, ψ_{jt} is an importer-time fixed effect, δ_{ij} is an exporter-importer (dyadic) fixed effect, $\mathbf{x}_{ijt}$ is a vector of variables of interest, and $\boldsymbol{\beta}$ the corresponding parameter vector.

The workhorse approach to estimate this kind of model is the so-called pseudo-poisson maximum likelihood estimator (PPML) proposed by Silva and Tenreyro (2006). Because trade flows are positive and continuous, this estimator is basically a poisson maximum likelihood estimator applied to a non-poisson distributed dependent variable.[13] Even in relatively short panels, researchers quickly find themselves confronted with a high computational effort, and thus PPML with three-way fixed effects is a well suited application for our algorithm. For instance, in a balanced panel, where $I = J = N = 30$ and $T = 15$, the number of fixed effects is $N(N-1)+2NT = 1,770$.

For our simulation experiments we generate data according to (3.9), where $\mathbf{x}_{ijt} = [x_{ijt}, d_{ijt}]$, x_{ijt} is generated as iid. standard normal, $d_{ijt} = 1[\psi_{ijt} > 0]$ with ψ_{ijt} beeing generated as iid. standard normal, and ϵ_{ijt} is an iid. log-normal error term with mean zero and variance one (on the log scale). Further, $\boldsymbol{\beta} = \mathbf{1}$, $\lambda_{it} \sim$ iid. $\mathcal{N}(\bar{\mathbf{x}}_{it}, 1)$, $\psi_{jt} \sim$ iid. $\mathcal{N}(\bar{\mathbf{x}}_{jt}, 1)$, and $\delta_{ij} \sim$ iid. $\mathcal{N}(\bar{\mathbf{x}}_{ij}, 1)$, where $\bar{\mathbf{x}}_{\cdot}$ denote the corresponding group means. We consider balanced panels of different sizes ($N = I = J$ and T) and generate 30 different data sets for each size.

All simulations were done on a Linux Mint 18.1 workstation using *R* version 3.6.1

12. We also conduct simulation experiments for a logit model with two-way fixed effects. The corresponding design and results are reported in appendix A. Overall, we make similar findings.

13. See Gourieroux, Monfort, and Trognon (1984) and Silva and Tenreyro (2006) why this estimator is valid.

(R Core Team 2019). We use the pseudo-demeaning algorithm (`feglm()`) provided in our *R*-package *alpaca* (version 0.3.1) and the standard GLM routine `glm()` provided in base *R*. Both routines are very similar, they essentially only differ by the fact that `glm()` uses dummy variables instead of MAP in the optimization routine. The results we report for MAP in this section are based on the Neumann-Halperin algorithm.[14]

At first we investigate the exactness of our Newton-Raphson pseudo-demeaning algorithm. Remember, MAP is an approximate method whose approximation error can be regulated by the choice of the tolerance level in algorithm 3 and 4. Whereas MAP is widely accepted in linear fixed effects models, in nonlinear models there might be suspicion that the approximation error is further exaggerated by the iterative optimization routine. To this end, we use different tolerance levels for MAP and measure how often coefficients and standard errors differ from the exact dummy variable approach which serves as a benchmark. Tables 3.2 and 3.3 summarize the joint relative frequencies of $\hat{\boldsymbol{\beta}}$ and its standard error for different digits. We observe that up to 4 digits the exact dummy approach and our Newton-Raphson pseudo-demeaning algorithm deliver identical coefficients and standard errors for tolerance levels larger than 10^{-4}. As expected, tighter tolerance levels improve precision. Additionally, we observe that the standard errors are more sensitive to the selected tolerance level.

Next we analyze the computation times of the naive dummy variable approach and our Newton-Raphson pseudo-demeaning algorithm for different tolerance levels in MAP. Table 3.4 shows the dramatic increase of the computation time of the dummy variable approach. Whereas the latter approach takes roughly 24 minutes to estimate a three-way fixed effects PPML model with $30,000$ observations and $3,100$ fixed effects, our routine requires roughly 1 second. For higher combinations of N and T we only report the computation times obtained by our Newton-Raphson pseudo-demeaning algorithm. Even in the largest data set consisting of 1.99 million observations and $59,800$ fixed effects our routine is able to estimate the model in roughly 3.3 minutes for the tightest tolerance level of 10^{-8} and only 1.3 minutes for the loosest. Another aspect we observe is that a less strict tolerance level does not necessarily reduce computation time. This is because the larger approximation error associated with MAP causes algorithm 2 to take more iterations until convergence.

Overall, the simulation experiments confirm that our algorithm is able to handle estimation problems with high-dimensional fixed effects. It offers a considerable

14. We also performed simulations using Cimmino's approach and several acceleration schemes, e.g. Hernández-Ramos, Escalante, and Raydan (2011) and Gearhart and Koshy (1989). However, we did not find any algorithm to be superior. It is already well known that acceleration techniques can but do not necessarily accelerate (see among others Hernández-Ramos, Escalante, and Raydan 2011; Escalante and Raydan 2011). Nevertheless, we observed that the classical Neumann-Halperin algorithm never performed worst.

Table 3.2: *PPML: Exactness of* $\hat{\boldsymbol{\beta}}$

	N	T	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
4 digits	10	5	1.00	1.00	1.00	1.00	1.00	0.97
	10	10	1.00	1.00	1.00	1.00	1.00	0.93
	10	25	1.00	1.00	1.00	1.00	1.00	1.00
	10	50	1.00	1.00	1.00	1.00	1.00	0.97
	25	5	1.00	1.00	1.00	1.00	1.00	1.00
	25	10	1.00	1.00	1.00	1.00	1.00	0.97
	25	25	1.00	1.00	1.00	1.00	1.00	1.00
	25	50	1.00	1.00	1.00	1.00	1.00	1.00
6 digits	10	5	1.00	1.00	1.00	1.00	0.73	0.13
	10	10	1.00	1.00	1.00	1.00	0.80	0.10
	10	25	1.00	1.00	1.00	0.97	0.53	0.07
	10	50	1.00	1.00	1.00	1.00	0.67	0.03
	25	5	1.00	1.00	1.00	1.00	0.63	0.20
	25	10	1.00	1.00	1.00	1.00	0.87	0.27
	25	25	1.00	1.00	1.00	1.00	0.87	0.27
	25	50	1.00	1.00	1.00	1.00	0.80	0.40

Note: Three-way fixed effects PPML; measurement of exactness frequencies relative to dummy variable approach up to 4 and 6 digits; used Neumann-Halperin projection with different tolerance levels; results based on 30 repetitions.

Table 3.3: *PPML: Exactness of* $se(\hat{\boldsymbol{\beta}})$

	N	T	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
4 digits	10	5	1.00	1.00	1.00	1.00	0.97	0.93
	10	10	1.00	1.00	1.00	1.00	1.00	0.97
	10	25	1.00	1.00	1.00	1.00	1.00	1.00
	10	50	1.00	1.00	1.00	1.00	1.00	1.00
	25	5	1.00	1.00	1.00	1.00	1.00	1.00
	25	10	1.00	1.00	1.00	1.00	1.00	0.97
	25	25	1.00	1.00	1.00	1.00	1.00	1.00
	25	50	1.00	1.00	1.00	1.00	1.00	1.00
6 digits	10	5	0.83	0.83	0.80	0.40	0.13	0.00
	10	10	1.00	1.00	1.00	0.73	0.57	0.10
	10	25	0.93	0.93	0.97	0.93	0.73	0.30
	10	50	1.00	1.00	0.97	0.90	0.70	0.50
	25	5	1.00	1.00	1.00	0.97	0.83	0.30
	25	10	1.00	1.00	0.97	0.97	0.83	0.53
	25	25	1.00	1.00	1.00	1.00	0.97	0.87
	25	50	1.00	1.00	1.00	1.00	1.00	0.97

Note: Three-way fixed effects PPML; measurement of exactness frequencies relative to dummy variable approach up to 4 and 6 digits; used Neumann-Halperin projection with different tolerance levels; results based on 30 repetitions.

Table 3.4: *PPML: Average Computation Times*

			feglm()					
N	T	glm()	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
10	5	0.09	0.03	0.03	0.03	0.02	0.04	0.04
10	10	0.35	0.05	0.04	0.04	0.03	0.05	0.06
10	25	3.45	0.10	0.09	0.07	0.06	0.09	0.11
10	50	24.25	0.18	0.15	0.13	0.10	0.15	0.16
25	5	9.42	0.13	0.11	0.09	0.08	0.10	0.12
25	10	34.17	0.20	0.17	0.14	0.11	0.14	0.18
25	25	239.99	0.47	0.40	0.33	0.26	0.31	0.39
25	50	1421.84	1.00	0.84	0.67	0.53	0.65	0.78
50	5	-	0.46	0.39	0.32	0.25	0.25	0.31
50	10	-	0.71	0.60	0.50	0.40	0.36	0.47
50	25	-	1.75	1.47	1.21	0.94	0.98	1.26
50	50	-	3.65	3.05	2.51	1.95	1.96	2.38
100	5	-	1.99	1.68	1.37	1.08	1.03	1.29
100	10	-	3.37	2.81	2.28	1.76	1.61	2.03
100	25	-	8.20	6.82	5.54	4.23	4.26	5.04
100	50	-	22.20	18.58	14.94	11.39	10.07	11.83
200	5	-	7.62	6.41	5.24	4.15	3.66	4.30
200	10	-	18.30	15.29	12.26	9.47	7.88	9.22
200	25	-	86.46	71.73	56.15	42.24	35.47	37.71
200	50	-	197.29	161.61	126.28	91.75	80.38	78.59

Note: Three-way fixed effects PPML; average computation times in seconds; `glm()` refers to the standard GLM routine provided in *R*; `feglm()` refers to the Newton-Raphson pseudo-demeaning algorithm (Neumann-Halperin projection with different tolerance levels); results based on 30 repetitions.

computation time advantage over `glm()` while maintaining the same accuracy up to relevant digits. Taking all aspects into account we recommend to use a tolerance level of at least 10^{-5}.

3.5 Empirical Illustration

In this section we emphasize the practical relevance of our algorithm by applying it to an example from international trade. As in section, 3.4 we estimate a structural gravity model with three high-dimensional fixed effects using PPML. We replicate parts of Larch et al. (2019), who reassessed some results of Glick and Rose (2016) and showed that PPML and OLS can lead to different conclusions. We extend the reassessment by two aspects. First, we test the assumptions of symmetry between the effects of entries and exists from currency unions. Testing these assumptions implies computationally demanding hypothesis tests because it requires to estimate models with high-dimensional fixed effects and many regressors. Second, we address an econometric issue that arises from using only specific time intervals of the entire data set. Interval data are used due to concerns that trade flows need some time to adjust to changes in trade costs (see among others Cheng and Wall 2005; Weidner

and Zylkin 2018; Larch et al. 2019). However, this strategy leads to a special incidental parameters problem which occurs only with small T and can be mitigated using a bias correction recently proposed by Weidner and Zylkin (2018).

Glick and Rose (2016) analyzed the effect of being in a currency union (CU) on export flows, with particular interest to quantifying the effect of membership in the European Monetary Union (EMU). For this purpose they considered three specifications with different levels of aggregation with respect to the currency unions. They used a data set with 879,794 observations (after dropping all zero trade flows) where roughly 200 countries trade for 65 years (1948 – 2013). Their model specification required roughly 11,000 exporter-time and importer-time fixed effects, respectively, as well as roughly 34,000 dyadic fixed effects. Due to the lack of a feasible software routine at that time, Glick and Rose (2016) estimated a log-linear specification instead of the desired PPML counterpart. Recently, Larch et al. (2019) proposed a feasible PPML algorithm based on the Gauss-Seidel algorithm by Guimarães and Portugal (2010) that can handle high-dimensional three-way fixed effects. With this tool at hand, they were able to estimate the model of Glick and Rose (2016) by PPML with the entire set of fixed effects.

We start by replicating some estimation results of Larch et al. (2019) to show that our Newton-Raphson pseudo-demeaning algorithm produces identical results as their routine. They estimated the following theory-consistent gravity model:

$$y_{ijt} = \exp(\gamma CU_{ijt} + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \lambda_{it} + \psi_{jt} + \delta_{ij})\epsilon_{ijt}\,,$$

where y_{ijt} denotes the nominal value of bilateral exports from exporter i to importer j at year t, CU_{ijt} is dummy variable, specifying whether i and j use the same currency at time t, $\mathbf{x}_{ijt}$ are further control variables, λ_{it} denotes a time-varying exporter fixed effect, ψ_{jt} a time-varying importer fixed effect, and δ_{ij} is a dyadic fixed effect. Unlike the log-linear specification, PPML is able to deal with zero trade flows. Thus we follow Larch et al. (2019) and replace all missing trade flows with zeros resulting in a data set of roughly 3 million observations. Table 3.5 reproduces table 2 from Larch et al. (2019) using our Newton-Raphson pseudo-demeaning algorithm.[15] Depending on the aggregation level of the currency unions, we are able to estimate the model in 63 up to 160 seconds. A detailed discussion of the estimation results is given in Larch et al. (2019).

15. Larch et al. (2019) also reported multi-way clustered standard errors as motivated by Egger and Tarlea (2015). For this purpose, they adjusted the approach suggested by Figueiredo, Guimarães, and Woodward (2015) to get the residuals of an auxiliary regression. Afterwards they used them to compute multi-way clustered standard errors following Cameron, Gelbach, and Miller (2011). Contrary to the post-estimation procedure of Larch et al. (2019) our approach directly builds on the suggestion of Cameron, Gelbach, and Miller (2011) to compute multi-way clustered standard errors based on the Hessian and scores. In our case we can simply use their concentrated counterparts.

Table 3.5: *Empirical Results*

	All CUs	Disagg. EMU	Disagg. CUs
All Currency Unions	0.1531		
	(0.0102, 0.0828)		
All Non-EMU Currency Unions		0.7276	
		(0.0255, 0.1789)	
EMU		0.0521	0.0489
		(0.0103, 0.0946)	(0.0103, 0.0946)
CFA Franc Zone			-0.1256
			(0.0997, 0.3522)
East Caribbean Currency Union			-0.8773
			(0.0835, 0.2949)
Aussie			0.3845
			(0.1188, 0.2235)
British £			1.0600
			(0.0347, 0.2377)
French Franc			2.0957
			(0.0630, 0.3063)
Indian Ruppee			0.1697
			(0.1470, 0.3009)
US $			0.0183
			(0.0215, 0.0509)
Other CUs			0.7660
			(0.0533, 0.2493)
Importer-time fixed effects		11,277	
Exporter-time fixed effects		11,227	
Dyadic fixed effects		34,104	
time (in sec.)	63	73	160
iterations	13	13	13

Note: After dropping observations that do not contribute to the log-likelihood, we end up with roughly 1.6 million observations; two further control variables: regional FTA membership and current colony/colonizer; standard errors in parentheses: robust (sandwich estimator) and multi-way clustered standard errors by importer, exporter, and time in parentheses; Neumann-Halperin projection with tolerance level 10^{-5}.

In the following we extend the reassessment of Larch et al. (2019). First we show that our routine is also able to deal with many regressors. For this reason, we reassess table 6 of Glick and Rose (2016) where they tested, based on their log-linear specification, the assumption of symmetric effects of entries and exits from a currency union using joint hypotheses. The authors used the following specification:

$$y_{ijt} = \exp\Bigg(\sum_{k=-14}^{14} \theta_k \text{CUENTRY}_{ij(t-k)} + \sum_{k=-14}^{14} \phi_k \text{CUEXIT}_{ij(t-k)} + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \lambda_{it} + \psi_{jt} + \delta_{ij}\Bigg)\epsilon_{ijt},$$

where $\text{CUEXIT}_{ij(t-k)}$ is one if country i and j entered a currency union at time $t-k$ and $\text{CUEXIT}_{ij(t-k)}$ is one if country i and j exited a currency union at time $t-k$. The tests require to estimate unrestricted models with 60 and 89 regressors. Table 3.6 summarizes the seven hypotheses and the corresponding results of the

Table 3.6: *Entry - Exit Symmetry*

	Wald test	
Hypothesis	robust	clustered
After any CU Entry = - After any CU Exit?	27.5 (0.0165)	7.5 (0.9119)
Before any CU Entry = - Before any CU Exit?	16.8 (0.2665)	6.0 (0.9669)
Both	149.8 (0.0000)	6.8 (1.0000)
Number of regressors: 60, Time: 16 minutes		
After non-EMU CU Entry = After EMU Entry?	29.0 (0.0106)	7.8 (0.9010)
Before non-EMU CU Entry = Before EMU Entry?	52.5 (0.0000)	24.2 (0.0430)
Both	80.3 (0.0000)	21.0 (0.8252)
After non-EMU CU Exit = - After EMU Entry?	28.3 (0.0132)	7.7 (0.9049)
Number of regressors: 89, Time: 22 minutes		

Note: Wald tests based on robust standard errors and multi-way clustered standard errors by importer, exporter, and time; reported test-statistics and p-values in parentheses; two further control variables: regional FTA membership and current colony/colonizer; Neumann-Halperin projection with tolerance level 10^{-5}.

Wald tests based on PPML. The test statistics are constructed from robust and multi-way clustered estimates of covariance matrices. The results are ambiguous. If the test statistics are based on the robust covariances, we reject the null hypotheses of symmetry in all except one case, assuming a significance level of 5%. Using the multi-way clustered covariances we get completely opposite results. Since clustered inference is common practice in structural gravity estimation, our results seem to validate the symmetry assumption of Glick and Rose (2016).

Our second extension of the reassessment addresses concerns among trade economists that 1 year is not enough for trade flows to adjust to changes in trade costs. In this case, data should not be taken annually but for example at five year intervals (see among others Cheng and Wall 2005; Weidner and Zylkin 2018; Larch et al. 2019). Restricting our sample to 5-year intervals, $t \in \{1948, 1953, \ldots, 2013\}$,

results in a data set with 14 time periods. This however, raises econometric concerns. Very recently, Weidner and Zylkin (2018) study the statistical properties of the PPML estimator in a three-way fixed effects gravity model. They show that although the estimator is fixed T consistent, it exhibits a bias in its asymptotic distribution. This incidental parameter bias is especially severe for panels with small T and may lead to invalid inference. Thus, the authors propose a jackknife bias correction in the spirit of Dhaene and Jochmans (2015).

As shown by Weidner and Zylkin (2018), under fixed T, the leading order bias has the following structure $B/I+D/J$, which implies two suitable splitting strategies. The first strategy, proposed by Weidner and Zylkin (2018), splits the panel simultaneously along exporters and importers, leading to the following bias-corrected estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{spj1} &= 2\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{I/2,J/2,T}, \quad \text{with} \\ \widehat{\boldsymbol{\beta}}_{I/2,J/2,T} &= \frac{1}{4}\Big[\widehat{\boldsymbol{\beta}}_{\{i:i\leq\lceil I/2\rceil\},\{j:j\leq\lceil J/2\rceil\},T} + \widehat{\boldsymbol{\beta}}_{\{i:i\geq\lfloor I/2+1\rfloor\},\{j:j\leq\lceil J/2\rceil\},T} \\ &\qquad + \widehat{\boldsymbol{\beta}}_{\{i:i\leq\lceil I/2\rceil\},\{j:j\geq\lfloor J/2+1\rfloor\},T} + \widehat{\boldsymbol{\beta}}_{\{i:i\geq\lfloor I/2+1\rfloor\},\{j:j\geq\lfloor J/2+1\rfloor\},T}\Big],\end{aligned}$$

where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the floor and ceiling functions, respectively.[16] The subscript $\{i:i\leq\lceil I/2\rceil\},\{j:j\geq\lfloor J/2+1\rfloor\},T$ indicates that the estimator is based on a subsample containing the first half of all exporters, and the second half of all importers, but all time periods. Another valid splitting strategy splits the panel sequentially along exporters and importers. In this case the corresponding bias-corrected estimator is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{spj2} &= 3\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{I/2,J,T} - \widehat{\boldsymbol{\beta}}_{I,J/2,T}, \quad \text{with} \\ \widehat{\boldsymbol{\beta}}_{I/2,J,T} &= \frac{1}{2}\Big[\widehat{\boldsymbol{\beta}}_{\{i:i\leq\lceil I/2\rceil\},J,T} + \widehat{\boldsymbol{\beta}}_{\{i:i\geq\lfloor I/2+1\rfloor\},J,T}\Big], \\ \widehat{\boldsymbol{\beta}}_{I,J/2,T} &= \frac{1}{2}\Big[\widehat{\boldsymbol{\beta}}_{I,\{j:j\leq\lceil J/2\rceil,T} + \widehat{\boldsymbol{\beta}}_{I,\{j:j\geq\lfloor J/2+1\rfloor,T}\Big].\end{aligned}$$

We reestimate column 1 of table 3.5 using the sample on 5-year intervals and the suggested bias corrections. Table 3.7 reports estimates of the currency union effect based on the uncorrected estimator (column 1), and the two split-panel jackknife estimators (column 2 and 3).[17] The two bias corrections correct the PPML estimate of the currency union effect downward, but they are still substantially higher than those based on annual data (see column 1 of table 3.5). Interestingly, the estimates of both splitting strategies differ, making a difference of roughly 0.5% points in the

16. The application of floor and ceiling functions generates overlapping subpanels in case of a odd number of exporters and/or importers as suggested by Fernández-Val and Weidner (2016) and Cruz-Gonzalez, Fernández-Val, and Weidner (2017).

17. Since there is no natural ordering of the exporters and importers, we follow the suggestion of Fernández-Val and Weidner (2016) to compute the average over different split-panel jackknife estimates, where the indices of the exporters and importers are randomly shuffled. We average over 100 different estimates.

effect of membership in a currency union.[18] Another aspect that is of importance for the practical application of bias corrections is the overall computation time. As expected, it increases considerably due to the number of repetitions and the necessity to estimate multiple subpanels. However, the overall computation time of 14 and 24 minutes is still manageable with our algorithm.

Table 3.7: *Empirical Results (5-Year Intervals)*

	PPML	SPJ 1	SPJ 2
All Currency Unions	0.1732	0.1691	0.1652
		(0.0226 , 0.0850)	
Importer-time fixed effects		2378	
Exporter-time fixed effects		2382	
Dyadic fixed effects		30789	
time (in sec.)	7.56	844.52	1461.67

Note: PPML, SPJ1, and SPJ2 denote the (bias-corrected) estimators; after dropping observations that do not contribute to the log-likelihood, we end up with roughly 311,563 observations; two further control variables: regional FTA membership and current colony/colonizer; standard errors in parentheses: robust (sandwich estimator) and multi-way clustered standard errors by importer, exporter, and time in parentheses; Neumann-Halperin projection with tolerance level 10^{-5}.

To sum up, the application demonstrates the practical relevance of our Newton-Raphson pseudo-demeaning routine, since it allows to estimate models with many high-dimensional fixed effects in a reasonable amount of time even in the presence of many regressors.

3.6 Conclusion

We presented a new algorithm for the maximum likelihood estimation of generalized linear models (GLMs) with a high-dimensional k-way error component. Our approach is straightforward since it resembles the classical within transformation used in linear regression models. To be more precise the algorithm incorporates a special pseudo-demeaning procedure into a standard Newton-Raphson estimation routine such that the updates of the structural parameters are separated from the high-dimensional fixed effects. Given an appropriate tolerance level in the pseudo-demeaning procedure we are able to obtain estimates and standard errors that are arbitrarily close to those of a classical maximum likelihood estimation with dummy variables. Whereas the latter approach quickly becomes either time demanding or even infeasible, our algorithm is fast and memory efficient. Although this article

18. Further research is required on the statistical properties of both split panel jackknife approaches. Intuitively, one would expect that the second splitting strategy should be more precise, since it is based on larger subpanels.

focuses on GLMs the proposed procedures might be adjustable to other nonlinear models.

Many nonlinear models with fixed effects are affected by the incidental parameters problem. Whereas an extensive literature on one-way bias corrections exists, further research is required for models with multi-way fixed effects. Fortunately, our algorithm can be easily combined with split-panel bias corrections in spirit of Dhaene and Jochmans (2015), which require only knowledge of the order of bias. Fernández-Val and Weidner (2018a) propose a simple heuristic to determine this order. As we demonstrate in the empirical application, jackknife bias corrections can become computationally demanding emphasizing the relevance of our algorithm. At the same time it encourages the development of analytical bias corrections for commonly used multi-way error components. Recent developments include analytical bias corrections of Fernández-Val and Weidner (2016), Weidner and Zylkin (2018), and Hinz, Stammann, and Wanner (2019) for nonlinear models with two-way fixed effects, and pseudo-poisson and binary choice models with three-way fixed effects, respectively. Czarnowske and Stammann (2019) show how the computational burden of analytical bias corrections can be reduced considerably using the algorithm described in this article.

To sum up, our Newton-Raphson pseudo-demeaning routine offers new possibilities and reliefs to researchers since it allows to estimate models with many observations and high-dimensional fixed effects even on a standard computer. We provide the routine in an *R*-package *alpaca* to make it available for empirical research.

References

Anderson, James E., and Eric Van Wincoop. 2003. "Gravity with gravitas: a solution to the border puzzle." *The American Economic Review* 93 (1): 170–192.

Baier, Scott L., and Jeffrey H. Bergstrand. 2007. "Do free trade agreements actually increase members' international trade?" *Journal of International Economics* 71 (1): 72–95.

Balazsi, Laszlo, Laszlo Matyas, and Tom Wansbeek. 2018. "The estimation of multidimensional fixed effects panel data models." *Econometric Reviews* 37 (3): 212–227.

Bergé, Laurent. 2018. "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm." *Working Paper*.

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.

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

Cheng, I-Hui, and Howard J. Wall. 2005. "Controlling for heterogeneity in gravity models of trade and integration." *Working Paper*.

Cimmino, Gianfranco. 1938. "Cacolo approssimato per le soluzioni dei systemi di equazioni lineari." *La Ricerca Scientifica (Roma)* 1:326–333.

Correia, Sergio. 2016. "A feasible estimator for linear models with multi-way fixed effects." *Working Paper*.

Correia, Sergio, Paulo Guimarães, and Thomas Zylkin. 2019. "PPMLHDFE: Fast poisson estimation with high-dimensional fixed effects." *arXiv preprint:1903.01690*.

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.

Czarnowske, Daniel, and Amrei Stammann. 2019. "Binary Choice Models with High-Dimensional Individual and Time Fixed Effects." *arXiv preprint:1904.04217*.

Dhaene, Geert, and Koen Jochmans. 2015. "Split-panel jackknife estimation of fixed-effect models." *Review of Economic Studies* 82 (3): 991–1030.

Egger, Peter H., and Filip Tarlea. 2015. "Multi-way clustering estimation of standard errors in gravity models." *Economics Letters* 134:144–147.

Escalante, Renâ, and Marcos Raydan. 2011. *Alternating projection methods.* Vol. 8. SIAM.

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.

Figueiredo, Octávio, Paulo Guimarães, and Douglas Woodward. 2015. "Industry localization, distance decay, and knowledge spillovers: Following the patent paper trail." *Journal of Urban Economics* 89:21–31.

Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1 (4): 387–401.

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

———. 2013b. "OLS with multiple high dimensional category variables." *Computational Statistics & Data Analysis* 66:8–18.

Gearhart, William B., and Mathew Koshy. 1989. "Acceleration schemes for the method of alternating projections." *Journal of Computational and Applied Mathematics* 26 (3): 235–249.

Glick, Reuven, and Andrew K. Rose. 2016. "Currency unions and trade: A post-EMU reassessment." *European Economic Review* 87:78–91.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52 (3): 681–700.

Greene, William. 2004. "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects." *Econometrics Journal* 7:98–119.

Guimarães, Paulo. 2014. "POI2HDFE: Stata module to estimate a Poisson regression with two high-dimensional fixed effects." *Statistical Software Components: Boston College Department of Economics.*

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

Hall, Bronwyn H. 1978. "A general framework for the time series-cross section estimation." *Annales de l'INSEE* 30–31:177–202.

Halperin, Israel. 1962. "The product of projection operators." *Acta Sci. Math.(Szeged)* 23 (1-2): 96–99.

Hernández-Ramos, Luis M., René Escalante, and Marcos Raydan. 2011. "Unconstrained optimization techniques for the acceleration of alternating projection methods." *Numerical functional analysis and optimization* 32 (10): 1041–1066.

Hinz, Julian, Amrei Stammann, and Joschka Wanner. 2019. "Persistent Zeros: The Extensive Margin of Trade." *Working Paper*.

Hyslop, Dean R. 1999. "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women." *Econometrica* 67 (6): 1255–1294.

Kaczmarz, Stefan. 1937. "Angenaherte auflosung von systemen linearer gleichungen." *Bull. Int. Acad. Sci. Pologne, A* 35:355–357.

Kammerer, William J., and M. Zuhair Nashed. 1972. "Iterative methods for best approximate solutions of linear integral equations of the first and second kinds." *Journal of Mathematical Analysis and Applications* 40 (3): 547–573.

Larch, Mario, Joschka Wanner, Yoto V. Yotov, and Thomas Zylkin. 2019. "Currency Unions and Trade: A PPML Re-assessment with High-dimensional Fixed Effects." *Oxford Bulletin of Economics and Statistics* 81 (3): 487–510.

Lovell, Michael C. 1963. "Seasonal adjustment of economic time series and multiple regression analysis." *Journal of the American Statistical Association* 58 (304): 993–1010.

McCullagh, Peter, and James A. Nelder. 1989. *Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability*.

Neumann, John von. 1950. "Functional Operators. Vol. II. The geometry of orthogonal spaces, volume 22 (reprint of 1933 notes) of Annals of Math." *Studies. Princeton University Press*.

Neyman, Jerzy, and Elizabeth L Scott. 1948. "Consistent estimates based on partially consistent observations." *Econometrica* 16 (1): 1–32.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Silva, JMC Santos, and Silvana Tenreyro. 2006. "The log of gravity." *The Review of Economics and Statistics* 88 (4): 641–658.

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

Weidner, Martin, and Thomas Zylkin. 2018. “Bias and Consistency in Three-way Gravity Models.” *Working Paper*.

Appendix

A Further Monte Carlo Experiments

For the logit model with two-way fixed effects we generate data according to

$$y_{it} = \mathbf{1}[\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma_t + \epsilon_{it} > 0],$$

where $i = 1,\dots,N$, $t = 1,\dots,T$, $\mathbf{x}_{it} = [x_{it1}, x_{it2}, x_{it3}]$ with x_{itp} beeing generated as iid. standard normal for $p = 1,\dots,3$, ϵ_{it} is an iid. logistic error term with location zero and scale one, and $\boldsymbol{\beta} = [1,-1,1]'$. To introduce a correlation between the fixed effects and the regressors, they are generated according to $\alpha_i \sim$ iid. $\mathcal{N}(\sum_{p=1}^{3} \bar{\mathbf{x}}_{ip}, 1)$ and $\gamma_t \sim$ iid. $\mathcal{N}(\sum_{p=1}^{3} \bar{\mathbf{x}}_{tp}, 1)$, where $\bar{\mathbf{x}}_{\cdot}$ denote the corresponding group means.

Table 3.8: *Logit: Exactness of* $\hat{\boldsymbol{\beta}}$

	N	T	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
4 digits	250	50	1.00	1.00	1.00	1.00	1.00	1.00
	250	100	1.00	1.00	1.00	1.00	1.00	1.00
	500	50	1.00	1.00	1.00	1.00	1.00	1.00
	500	100	1.00	1.00	1.00	1.00	1.00	1.00
	500	250	1.00	1.00	1.00	1.00	1.00	1.00
6 digits	250	50	1.00	1.00	1.00	1.00	1.00	1.00
	250	100	1.00	1.00	1.00	1.00	1.00	1.00
	500	50	1.00	1.00	1.00	1.00	1.00	1.00
	500	100	1.00	1.00	1.00	1.00	1.00	1.00
	500	250	1.00	1.00	1.00	1.00	1.00	1.00

Note: Two-way fixed effects logit; measurement of exactness frequencies relative to dummy variable approach up to 4 and 6 digits; used Neumann-Halperin projection with different tolerance levels; results based on 30 repetitions.

Table 3.9: *Logit: Exactness of* $se(\hat{\boldsymbol{\beta}})$

	N	T	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
4 digits	250	50	0.90	0.90	0.90	0.90	0.90	0.90
	250	100	0.97	0.97	0.97	0.97	0.97	0.97
	500	50	1.00	1.00	1.00	1.00	1.00	1.00
	500	100	1.00	1.00	1.00	1.00	1.00	1.00
	500	250	1.00	1.00	1.00	1.00	1.00	1.00
6 digits	250	50	0.10	0.10	0.10	0.10	0.10	0.03
	250	100	0.17	0.17	0.17	0.17	0.20	0.13
	500	50	0.13	0.13	0.13	0.13	0.13	0.13
	500	100	0.27	0.27	0.27	0.27	0.27	0.30
	500	250	0.63	0.63	0.63	0.63	0.63	0.67

Note: Two-way fixed effects logit; measurement of exactness frequencies relative to dummy variable approach up to 4 and 6 digits; used Neumann-Halperin projection with different tolerance levels; results based on 30 repetitions.

Table 3.10: *Logit: Average Computation Times*

			feglm()					
N	T	glm()	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
250	50	4.60	0.06	0.05	0.05	0.05	0.05	0.05
250	100	13.26	0.11	0.10	0.10	0.09	0.09	0.08
500	50	32.42	0.11	0.10	0.10	0.09	0.09	0.08
500	100	77.44	0.21	0.20	0.19	0.18	0.17	0.16
500	250	299.06	0.52	0.50	0.46	0.45	0.42	0.39
1,000	50	-	0.22	0.22	0.20	0.18	0.17	0.16
1,000	100	-	0.44	0.40	0.38	0.36	0.35	0.31
1,000	250	-	1.06	1.01	0.94	0.94	0.84	0.78
1,000	500	-	2.61	2.50	2.25	2.16	1.95	1.78
5,000	50	-	1.19	1.08	1.03	0.97	0.89	0.84
5,000	100	-	2.75	2.59	2.37	2.30	2.07	1.87
5,000	250	-	8.78	8.36	7.52	7.15	6.27	5.58
5,000	500	-	17.51	17.09	15.28	14.22	12.67	11.27
5,000	1,000	-	32.75	32.18	28.35	27.19	24.25	21.95
1,0000	50	-	2.57	2.39	2.27	2.12	1.93	1.80
1,0000	100	-	7.06	6.64	6.04	5.71	4.99	4.46
10,000	250	-	17.77	17.22	15.05	14.23	12.84	11.27
10,000	500	-	33.35	31.92	28.26	26.95	24.24	22.00
10,000	1,000	-	65.78	65.00	57.77	55.16	49.93	44.87

Note: Two-way fixed effects logit; average computation times in seconds; `glm()` refers to the standard GLM routine provided in *R*; `feglm()` refers to the Newton-Raphson pseudo-demeaning algorithm (Neumann-Halperin projection with different tolerance levels); results based on 30 repetitions.

B Recovering the Fixed Effects Ex-Post

Sometimes researchers might not only require estimates of the structural parameters but also of the fixed effects. In the following we present two algorithms to efficiently recover estimates of the fixed effects in a post-estimation routine.

Rearranging the linear predictor $\boldsymbol{\eta} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}$ after convergence of algorithm 2 yields a large and sparse system of linear equations

$$\mathbf{D}\boldsymbol{\alpha} = \underbrace{\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}}_{\mathbf{b}}, \tag{3.10}$$

where $\mathbf{b}$ can be computed at low computational cost from already available quantities. Since the analytical solution of (3.10) is inefficient and often infeasible, we propose two numerical routines to efficiently solve the linear system of equations.[19]

The first solver we present is in spirit of the Gauss-Seidel algorithm used by Guimarães and Portugal (2010). We apply the same idea in order to compute the fixed effects by alternating between the solution of the normal equations corresponding to (3.10). Consider the case with three high-dimensional fixed effects $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$.

19. Note that the numerical solvers do not require $\mathbf{D}$ to have full column rank. In order to get meaningful estimates for the fixed effects it is necessary to apply an estimable function to the solution (see Gaure 2013b).

The normal equations of system (3.10) are

$$\begin{aligned}
\mathbf{D}_1'\mathbf{D}_1\boldsymbol{\alpha}_1 + \mathbf{D}_1'\mathbf{D}_2\boldsymbol{\alpha}_2 + \mathbf{D}_1'\mathbf{D}_3\boldsymbol{\alpha}_3 &= \mathbf{D}_1'\mathbf{b} \\
\mathbf{D}_2'\mathbf{D}_1\boldsymbol{\alpha}_1 + \mathbf{D}_2'\mathbf{D}_2\boldsymbol{\alpha}_2 + \mathbf{D}_2'\mathbf{D}_3\boldsymbol{\alpha}_3 &= \mathbf{D}_2'\mathbf{b} \\
\mathbf{D}_3'\mathbf{D}_1\boldsymbol{\alpha}_1 + \mathbf{D}_3'\mathbf{D}_2\boldsymbol{\alpha}_2 + \mathbf{D}_3'\mathbf{D}_3\boldsymbol{\alpha}_3 &= \mathbf{D}_3'\mathbf{b}
\end{aligned}$$

and can be rearranged to

$$\begin{aligned}
\boldsymbol{\alpha}_1 &= (\mathbf{D}_1'\mathbf{D}_1)^{-1}\mathbf{D}_1'(\mathbf{b} - \mathbf{D}_2\boldsymbol{\alpha}_2 - \mathbf{D}_3\boldsymbol{\alpha}_3) \\
\boldsymbol{\alpha}_2 &= (\mathbf{D}_2'\mathbf{D}_2)^{-1}\mathbf{D}_2'(\mathbf{b} - \mathbf{D}_1\boldsymbol{\alpha}_1 - \mathbf{D}_3\boldsymbol{\alpha}_3) \\
\boldsymbol{\alpha}_3 &= (\mathbf{D}_3'\mathbf{D}_3)^{-1}\mathbf{D}_3'(\mathbf{b} - \mathbf{D}_1\boldsymbol{\alpha}_1 - \mathbf{D}_2\boldsymbol{\alpha}_2)
\end{aligned}$$

The solver works as follows: given some starting values for the fixed effects, we alternate between the solutions of the three normal equations. Fortunately, the single equations can be computed easily. The vector $\boldsymbol{\alpha}_k = (\mathbf{D}_k'\mathbf{D}_k)^{-1}\mathbf{D}_k'(\mathbf{b} - \mathbf{D}_{-k}\boldsymbol{\alpha}_{-k})$ contains the group means of $(\mathbf{b} - \mathbf{D}_{-k}\boldsymbol{\alpha}_{-k})$ with respect to group k, where $\mathbf{D}_{-k}\boldsymbol{\alpha}_{-k}$ denotes all fixed effects contributions without the k-th. Algorithm 5 summarizes the procedure for an arbitrary number of fixed effects.

Algorithm 5 Alternating Between Solutions of Normal Equations

1: Set $j = 1$, $\boldsymbol{\rho}_j = (\boldsymbol{\alpha}_{1j}, \ldots, \boldsymbol{\alpha}_{Kj}) = \mathbf{0}_l$, where $\boldsymbol{\alpha}_{kj}$ is the vector of coefficients corresponding to group k at iteration j, $\boldsymbol{\rho}_{j-1} = \boldsymbol{\rho}_j - \mathbf{1}_l$, and tolerance level ϵ.
2: **while** $||\boldsymbol{\rho}_j - \boldsymbol{\rho}_{j-1}||_2 \geq \epsilon$ **do**
3: **for** $k = 1, \ldots, K$ **do**
4: Compute $\boldsymbol{\alpha}_{kj}$ as the group means of $(\mathbf{b} - \mathbf{D}_{-k}\boldsymbol{\alpha}_{-kj})$ w.r.t. group k.
5: Update $\boldsymbol{\rho}_j$ with new $\boldsymbol{\alpha}_{kj}$.
6: Set $j = j + 1$.
7: Set $\boldsymbol{\alpha} = \boldsymbol{\rho}_j$.

A second approach to solve the system (3.10) is the Kaczmarz method (see Kaczmarz 1937). The Kaczmarz method belongs to the so-called row-action methods and is suitable to solve large and sparse systems (see Escalante and Raydan 2011). The idea is similar to the alternating projection methods described in section 3.3. Each equation of (3.10) defines a hyperplane and by alternating orthogonal projections onto hyperplanes we can find the intersection. In our application the intersection translates into the fixed effects coefficients. Each projection of the i-th hyperplane onto the $(i+1)$-th hyperplane can be summarized as follows:

$$\boldsymbol{\rho}_{i+1} = \boldsymbol{\rho}_i + \frac{(\mathbf{b}_i - \langle \mathbf{d}_i, \boldsymbol{\rho}_i \rangle)}{||\mathbf{d}_i||_2^2}\mathbf{d}_i\,, \tag{3.11}$$

where $\boldsymbol{\rho}$ is a l-dimensional vector of coefficients, $\mathbf{d}_i$ and $\mathbf{b}_i$ denote the i-th row of $\mathbf{D}$

and $\mathbf{b}$ respectively, and $||\cdot||_2^2$ is the squared euclidean norm. Each row of $\mathbf{D}$ contains K times the value one, such that the denominator in (3.11) can be simplified as follows:

$$\boldsymbol{\rho}_{i+1} = \boldsymbol{\rho}_i + \frac{(\mathbf{b}_i - \langle \mathbf{d}_i, \boldsymbol{\rho}_i \rangle)}{K} \mathbf{d}_i \,. \tag{3.12}$$

Since $\mathbf{D}$ is sparse, the Kaczmarz updates can be computed at minimum memory. Algorithm 6 summarizes the procedure. During the development of our software package *alpaca* we found that algorithm 5 performs much faster than algorithm 6.

Algorithm 6 Kaczmarz

1: Set $j = 1$, $\boldsymbol{\rho}_j = \mathbf{0}_l$, $\boldsymbol{\rho}_{j-1} = \boldsymbol{\rho}_j - \mathbf{1}_l$, and tolerance level ϵ.
2: **while** $||\boldsymbol{\rho}_j - \boldsymbol{\rho}_{j-1}||_2 \geq \epsilon$ **do**
3: Set $\boldsymbol{\rho}_{0j} = \boldsymbol{\rho}_j$.
4: **for** $i = 1, \ldots, n$ **do**
5: Compute $\boldsymbol{\rho}_{ij}$ (see formula (3.12)).
6: Set $j = j + 1$, $\boldsymbol{\rho}_j = \boldsymbol{\rho}_{nj}$.
7: Set $\boldsymbol{\alpha} = \boldsymbol{\rho}_j$.

Chapter 4

Binary Choice Models with High-Dimensional Individual and Time Fixed Effects

Co-authored with Daniel Czarnowske

4.1 Introduction

The increasing number and availability of large panel data sets offers 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 heterogeneity such as unobserved individual and/or time specific effects. So-called fixed effects models treat these 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 parameters problem (IPP) first mentioned by Neyman and Scott (1948). Intuitively, only T observations contribute to the identification of one individual effect, resulting in potentially noisy estimates that bias the estimation of the other model parameters (see Arellano and Hahn 2007; Fernández-Val and Weidner 2016, 2018a). This strand of literature is therefore particularly interested in deriving fixed T consistent estimators. For instance, so-called conditional logit estimators have been proposed for static and dynamic binary choice models with individual fixed effects (see Rasch 1960; Andersen 1970; Chamberlain 1980; 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 often of interest in economics (see Arellano and Hahn 2007; 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 longitudinal data, a growing literature now focuses on large N and T asymptotics. The beauty of this asymptotic framework is that IPP turns into an asymptotic bias problem which is easier to deal with than an inconsistency problem. In the meantime, this strand of literature proposed several bias-corrected estimators for nonlinear models with different error structures, which substantially reduce this asymptotic bias. We refer the reader to Arellano and Hahn (2007) and Fernández-Val and Weidner (2018a) for comprehensive overviews. The rest of this article focuses on binary choice models with individual and time fixed effects and the appropriate bias corrections proposed

by Fernández-Val and Weidner (2016).

Another apparent challenge that discourages researchers from using nonlinear fixed effects models is the computational burden associated with the estimation. This issue is especially severe if the model specification leads to high-dimensional fixed effects, which happens quite often if we think about longitudinal micro data of individuals or firms. Models with only one source of unobserved heterogeneity are easy to handle, thanks to the partitioned inverse formula (see Chamberlain 1980; Greene 2004), or an approach introduced as pseudo-demeaning by Stammann, Heiß, and McFadden (2016). Even the estimation of nonlinear panel models with multiple fixed effects is feasible, using algorithms such as Guimarães and Portugal (2010) and Stammann (2018).

In this article, we offer new insights that facilitate and validate the usage of binary choice models with individual and time fixed effects in empirical research. First of all, we show how the computational obstacles which often preclude the application of bias corrections, can be tackled by combining them with the method of alternating projections (MAP). This approach is very well suited to our problem, because MAP is the work-horse method in linear models to deal with high-dimensional fixed effects and is easily adjustable to generalized linear models, for instance logit and probit models, as shown by Stammann (2018).[1] Apart from the computational improvements, we extend the simulation experiments of Fernández-Val and Weidner (2016) by several aspects to gain deeper insights into the statistical properties of various bias-corrected estimators. More precisely, we analyze further analytical and split-panel jackknife bias-corrected estimators which have been suggested but not studied by the authors. We additionally consider alternative estimators of average partial effects based on bias-corrected linear fixed effects models which are frequently used in empirical research to avoid the aforementioned pitfalls of nonlinear models. Furthermore, because many real world data sets are initially unbalanced, we add different patterns of unbalancedness to our analysis. This aspect has received little attention in the literature so far. Finally, we provide an illustrative example using an unbalanced panel data set drawn from the German Socio Economic Panel (see Wagner, Frick, and Schupp 2007) to investigate the inter-temporal labor force participation of 10,712 women between 1984 and 2013. Our suggested algorithm reduces the computational burden of this application dramatically. For instance, obtaining analytically bias-corrected estimates takes roughly two seconds on a standard desktop computer. To encourage the application of analytical bias corrections, we provide our routines in the *R*-package *alpaca*.[2]

1. The corresponding *R* and *Stata* routines for linear models (*lfe* and *reghdfe*) provided by Gaure (2013a) and Correia (2016), respectively, are widely used in empirical research. Together they have about 170 citations on *Google Scholar* (checked at 2019-04-18).

2. Until now, the analytical bias correction proposed by Fernández-Val and Weidner (2016) was

Overall, we find that analytical bias corrections are preferable to split-panel jackknife approaches. In general, the latter show higher distortion and dispersion and they are less robust to different missing data patterns. In addition, our findings suggest a prudent use of (bias-corrected) linear probability models. Although their application is very simple, their inference might be severely misleading.

The remainder of this article is organized as follows. Section 4.2 introduces the model and different bias corrections. Section 4.3 demonstrates how to handle high-dimensional fixed effects. Section 4.4 provides results of extensive simulation experiments. Section 4.5 applies the different bias-corrected estimators to an empirical example from labor economics. Finally section 4.6 concludes.

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

4.2 Bias Corrections for Fixed Effects Binary Choice Models

4.2.1 Fixed Effects Binary Choice Models and the Incidental Parameters Problem

At first, we derive the fixed effects binary choice model, studied in this article, from a latent variable model with an additive separable two-way error component for the disturbance. Let

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

be the latent variable, where $i = 1,\dots,N$ and $t = 1,\dots,T$ are individual and time specific identifiers, $\mathbf{x}_{it}$ is a J-dimensional vector of explanatory variables equal to the it-th row of the regressor matrix $\mathbf{X}$, $\boldsymbol{\beta}$ are the corresponding parameters, and e_{it} is an idiosyncratic error term. Note that $\mathbf{x}_{it}$ might also include predetermined variables. Further, let α_i and γ_t denote unobserved individual and time specific heterogeneity, respectively. Throughout this article, we refer to $\boldsymbol{\beta}$ as structural and $\boldsymbol{\phi} = (\boldsymbol{\alpha}', \boldsymbol{\gamma}')'$ as incidental parameters. However, instead of the latent variable, we observe $y_{it} = 1$ if $y_{it}^{*} \geq 0$ and $y_{it} = 0$ otherwise which leads to the nonlinear nature of the binary choice model.

The most popular way to derive an parametric estimator for fixed effects binary choice models is the principle of maximum likelihood. Suppose the idiosyncratic

only provided in a *Stata* routine, which is not adapted to large panel data (see Cruz-Gonzalez, Fernández-Val, and Weidner 2017).

error term is drawn independently from a specific distribution. Then

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

is the log-likelihood contribution of individual i at time t, where F_{it} is the cumulative distribution function of the idiosyncratic error term evaluated at $\eta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma_t$. Note that in the literature of generalized linear models (GLMs), η_{it} is known as the linear predictor (see McCullagh and Nelder 1989). Common choices for F_{it}, in economics, are the standard normal, the logistic, and the complementary log-log distribution. The corresponding maximum likelihood estimator is

$$\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\gamma}}'\right)' = \underset{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}}{\arg\max}\, \mathscr{L}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) , \tag{4.1}$$

where

$$\mathscr{L}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} l_{it}\left(\boldsymbol{\beta}, \alpha_i, \gamma_t\right) .$$

Because (4.1) does not have a closed form solution, it has to be solved numerically. The standard approach to estimate these models is to use any available standard software routine and add indicators for each individual and time period to the list of explanatory variables, also known as dummy encoding. However, if N and T increases this estimation approach quickly becomes very time consuming or even infeasible (see Stammann 2018 for a recent treatment of this issue).

Beside some computational obstacles, fixed effects estimators also suffer from the so-called incidental parameters problem (IPP) known since the article of Neyman and Scott (1948). To get an intuition of IPP suppose that T is small. In this case only a few observations per individual provide information that contribute to the identification of $\boldsymbol{\alpha}$. Thus the estimation error with respect to these incidental parameters can be very severe. Due to the nonlinear nature of binary choice models, this estimation error carries over to $\hat{\boldsymbol{\beta}}$, which is known as IPP (see among others Arellano and Hahn 2007; Fernández-Val and Weidner 2018a). To deal with this problem, several bias-corrected estimators have been proposed (see among others Hahn and Newey 2004; Fernández-Val 2009; Dhaene and Jochmans 2015; Fernández-Val and Weidner 2016; Kim and Sun 2016).

Next, we briefly summarize the key findings of Fernández-Val and Weidner (2016), who developed bias-corrected estimators for nonlinear models with a two-way error component. Using the same asymptotic framework as Hahn and Kuersteiner (2011), the authors show that under certain conditions, most notably additive separability and concavity, and under asymptotic sequences where $N/T \to \kappa^2$ and $0 < \kappa < \infty$, the

fixed effects estimator has the following asymptotic distribution:

$$\sqrt{NT}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{d} \overline{\mathbf{W}}_{\infty}^{-1} \mathcal{N}\left(\kappa \overline{\mathbf{B}}_{\infty}^{\beta}+\kappa^{-1} \overline{\mathbf{C}}_{\infty}^{\beta}, \overline{\mathbf{W}}_{\infty}\right),$$

where $\overline{\mathbf{B}}_{\infty}^{\beta}$ and $\overline{\mathbf{C}}_{\infty}^{\beta}$ are leading bias terms, stemming from the inclusion of individual and time specific fixed effects, and $\overline{\mathbf{W}}_{\infty}$ is the Hessian of the concentrated log-likelihood:

$$\mathscr{L}^{*}\left(\boldsymbol{\beta}\right)=\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \sum_{i=1}^{N} \sum_{t=1}^{T} l_{it}\left(\boldsymbol{\beta}, \alpha_{i}, \gamma_{t}\right). \tag{4.2}$$

Despite that $\hat{\boldsymbol{\beta}}$ is consistent ($\operatorname{plim}_{N,T\to\infty} \hat{\boldsymbol{\beta}}=\boldsymbol{\beta}$), its distribution reveals an asymptotic bias which can lead to severe consequences for inference even in moderately large panels (see Fernández-Val and Weidner 2016, 2018a).

Often researchers are not directly interested in estimates of $\boldsymbol{\beta}$, but rather in so-called partial effects. Let $\boldsymbol{\Delta}_{itj}$ denote the partial effect of a change in x_{itj} corresponding to individual i at time t, where x_{itj} is the j-th element in $\mathbf{x}_{it}$. This yields

$$\boldsymbol{\Delta}_{itj}=\beta_{j} \partial_{\eta} F_{it} \tag{4.3}$$

for continuous and

$$\boldsymbol{\Delta}_{itj}=F_{it}|_{x_{itj}=1}-F_{it}|_{x_{itj}=0} \tag{4.4}$$

for binary regressors, where $\partial_{\eta} F_{it}$ is the first-order partial derivative of F_{it} with respect to η_{it}. Because $\boldsymbol{\Delta}_{itj}$ is most likely different across individuals and time periods, a common strategy is to compute the average such that $\delta_{j}=(NT)^{-1} \sum_{i} \sum_{t} \boldsymbol{\Delta}_{itj}$. This quantity is known as the average partial effect of a change in x_{itj}.

Imposing further sampling conditions, Fernández-Val and Weidner (2016) derive the asymptotic distribution of the average partial effects estimator $\hat{\boldsymbol{\delta}}$:

$$r\left(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}-T^{-1} \overline{\mathbf{B}}_{\infty}^{\delta}-N^{-1} \overline{\mathbf{C}}_{\infty}^{\delta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \overline{\mathbf{V}}_{\infty}^{\delta}\right),$$

where r is a convergence rate which depends on the sampling assumptions of the unobserved heterogeneity, and $\overline{\mathbf{V}}_{\infty}^{\delta}$ is the asymptotic covariance matrix. Again, $\overline{\mathbf{B}}_{\infty}^{\delta}$ and $\overline{\mathbf{C}}_{\infty}^{\delta}$ are asymptotic bias terms stemming from the inclusion of individual and time specific fixed effects. Thus similar to $\hat{\boldsymbol{\beta}}$ there is an asymptotic bias in the distribution of $\hat{\boldsymbol{\delta}}$.

In the next subsection, we review the various bias-corrected estimators proposed by Fernández-Val and Weidner (2016). We use modified notation to ensure that it is consistent with that of the acceleration techniques presented in section 4.3.

4.2.2 Asymptotic Bias Corrections

Before we present the different bias-corrected estimators proposed by Fernández-Val and Weidner (2016), we introduce some additional notation. Let $\partial_\eta G_{it}$, $\partial_{\eta^2} G_{it}$, and $\partial_{\eta^3} G_{it}$ denote the first-, second-, and third-order partial derivative of an arbitrary function G_{it} with respect to η_{it}. $\partial_\eta \widehat{G}_{it}$, $\partial_{\eta^2} \widehat{G}_{it}$, and $\partial_{\eta^3} \widehat{G}_{it}$ are the corresponding sample analogues. For clarification, 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}. Further, let $\partial_\eta \hat{l}_{it} = \widehat{H}_{it}(y_{it} - \widehat{F}_{it})$, $\hat{\omega}_{it} = \widehat{H}_{it}\partial_\eta \widehat{F}_{it}$, $\widehat{H}_{it} = \partial_\eta \widehat{F}_{it}/(\widehat{F}_{it}(1-\widehat{F}_{it}))$, and $\hat{v}_{it} = (y_{it} - \widehat{F}_{it})/\partial_\eta \widehat{F}_{it}$. Finally, we define the residual projection $\widehat{\mathbb{M}} = \mathbb{1}_{NT} - \widehat{\mathbb{P}} = \mathbb{1}_{NT} - \mathbf{D}(\mathbf{D}'\widehat{\boldsymbol{\Omega}}\mathbf{D})^{-1}\mathbf{D}'\widehat{\boldsymbol{\Omega}}$, where $\mathbb{1}_{NT}$ is an identity matrix of dimension $NT \times NT$, $\mathbf{D}$ is a sparse indicator matrix arising from dummy encoding of individual and time identifiers, and $\widehat{\boldsymbol{\Omega}}$ is a diagonal weighting matrix with $\text{diag}(\widehat{\boldsymbol{\Omega}}) = \hat{\boldsymbol{\omega}}$. Table 4.1 provides explicit expressions of some frequently used distributions for binary choice models.

Table 4.1: *Common Distributions and 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.

Throughout this article we distinguish between two types of bias corrections: analytical and re-sampling. The latter exploits the relation between sample size and bias to construct estimators of the bias terms, whereas the former relies on explicit expressions. A general expression for a bias-corrected estimator of the structural parameter is

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{\mathbf{b}}^\beta, \tag{4.5}$$

where $\hat{\mathbf{b}}^\beta$ is an estimator of the composite bias term such that

$$\sqrt{NT}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \overline{\mathbf{W}}_\infty^{-1}\right).$$

Next, we describe one of the analytical bias corrections proposed by Fernández-Val and Weidner (2016). The corresponding estimator of the composite bias term is

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

where

$$\widehat{\mathbf{B}}^{\beta} = -\frac{1}{2}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T}\widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{it} + 2\sum_{l=1}^{L}(T/(T-l))\sum_{t=l+1}^{T}\partial_{\eta}\hat{l}_{it-l}\hat{\omega}_{it}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{it}}{\sum_{t=1}^{T}\hat{\omega}_{it}},$$

$$\widehat{\mathbf{C}}^{\beta} = -\frac{1}{2}\sum_{t=1}^{T}\frac{\sum_{i=1}^{N}\widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{it}}{\sum_{i=1}^{N}\hat{\omega}_{it}},$$

$$\widehat{\mathbf{W}} = \sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\omega}_{it}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{it}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{it}'.$$

Note that $\widehat{\mathbf{W}}$ is the Hessian of (4.2) evaluated at $\hat{\boldsymbol{\beta}}$, L is a bandwidth parameter proposed by Hahn and Kuersteiner (2007) required for the estimation of spectral densities, and $T/(T-l)$ is a finite sample adjustment suggested by Fernández-Val and Weidner (2016). If all explanatory variables are assumed to be strictly exogenous, we can set $L = 0$ and the second term in $\widehat{\mathbf{B}}^{\beta}$ drops out, leading to symmetric bias terms. If not, Fernández-Val and Weidner (2016, 2018a) suggest to do a sensitivity analysis reporting estimates for $L \in \{1,\dots,4\}$. The authors also note that the analytical bias-corrected estimator can be further iterated. More precisely, for a given $\tilde{\boldsymbol{\beta}}$, we can compute $\hat{\mathbf{b}}^{\text{abc}}$ and update $\tilde{\boldsymbol{\beta}}$ again and again. Although the asymptotic distribution of (4.5) is not affected by the iteration, its finite-sample performance might improve (see among others Hahn and Newey 2004; Arellano and Hahn 2007).

Fernández-Val and Weidner (2016) also extend the split-panel jackknife bias correction of Dhaene and Jochmans (2015) to nonlinear models with a two-way error component. The idea is to split the panel into smaller subpanels and use these to form an estimator of the composite bias term. Those subpanels are extracted as blocks, to maintain the dependency structure of the panel. Next, we describe two estimators of the bias term that are based on different splitting strategies to generate subpanels. The first one is described in Fernández-Val and Weidner (2016). Let

$$\hat{\mathbf{b}}^{\beta}_{\text{spj1}} = \hat{\boldsymbol{\beta}}^{N} + \hat{\boldsymbol{\beta}}^{T} - 2\hat{\boldsymbol{\beta}} \tag{4.6}$$

be an estimator of the composite bias term, where

$$\hat{\boldsymbol{\beta}}^{N} = \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_{\{i\leq\lceil N/2\rceil\}} + \hat{\boldsymbol{\beta}}_{\{i\geq\lfloor N/2+1\rfloor\}}\right), \quad \hat{\boldsymbol{\beta}}^{T} = \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_{\{t\leq\lceil T/2\rceil\}} + \hat{\boldsymbol{\beta}}_{\{t\geq\lfloor T/2+1\rfloor\}}\right),$$

$\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ are floor and ceiling functions, and the subscript in curly brackets indicates the corresponding subpanel. For instance, $\{i \leq \lceil N/2\rceil\}$ means that we only use the first half of all individuals in the sample to compute $\hat{\boldsymbol{\beta}}$. Note that if N and/or T are odd this leads to overlapping subpanels that introduce an additional variance inflation (see Dhaene and Jochmans 2015).[3] Cruz-Gonzalez, Fernández-Val, and

3. The authors also describe how to construct non-overlapping subpanels.

Weidner (2017) propose another splitting strategy. The corresponding estimator of the composite bias term is

$$\hat{\mathbf{b}}^{\beta}_{\text{spj2}} = \hat{\boldsymbol{\beta}}^{NT} - \hat{\boldsymbol{\beta}}, \tag{4.7}$$

where

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

Contrary to the first strategy, the panel is split simultaneously along both dimensions. Thus $\{i \leq \lceil N/2\rceil; t \leq \lceil T/2\rceil\}$ indicates that $\hat{\boldsymbol{\beta}}$ is computed based on the first half of all individuals in the first half of all time periods. The second splitting strategy is computationally less intense because the four subpanels are significantly smaller. However this strategy might lead to larger dispersion compared to the first one. Further 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 details). For instance, this condition rules out time-trends or structural breaks in the explanatory variables. Intuitively, if the subpanels stem from very different data generating processes, for instance due to non-stationarity, this will result in a poor estimate of the bias term because the subpanel estimates are very different from each other (see Dhaene and Jochmans 2015; Fernández-Val and Weidner 2016, 2018a).

So far the bias corrections are applied at the level of the estimator. Fernández-Val and Weidner (2016) also show how to apply the analytical correction at the level of score. The corresponding bias-corrected estimates can be obtained by solving

$$\left(\widehat{\mathbb{M}}\mathbf{X}\left(\tilde{\boldsymbol{\beta}}\right)\right)' \widehat{\boldsymbol{\Omega}}\left(\tilde{\boldsymbol{\beta}}\right) \hat{\boldsymbol{v}}\left(\tilde{\boldsymbol{\beta}}\right) = \widehat{\mathbf{B}}^{\beta} + \widehat{\mathbf{C}}^{\beta}$$

for $\tilde{\boldsymbol{\beta}}$ using any nonlinear solver. Note that the left hand side is the gradient of (4.2) evaluated at $\tilde{\boldsymbol{\beta}}$. The authors also suggest a continuously updated score correction by replacing $\widehat{\mathbf{B}}^{\beta}$ and $\widehat{\mathbf{C}}^{\beta}$ with $\widehat{\mathbf{B}}^{\beta}(\tilde{\boldsymbol{\beta}})$ and $\widehat{\mathbf{C}}^{\beta}(\tilde{\boldsymbol{\beta}})$, respectively.

Additionally Fernández-Val and Weidner (2016) derive bias corrections for average partial effects. Let $\hat{\boldsymbol{\delta}} = (NT)^{-1}\sum_i\sum_t \widehat{\boldsymbol{\Delta}}_{it}$, where $\widehat{\boldsymbol{\Delta}}_{it}$ is the sample analogue of (4.3) or (4.4). Similar to the structural parameters, a bias-corrected estimator for the average partial effects is

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

where $\hat{\mathbf{b}}^{\delta}$ is an estimator of the composite bias term such that

$$r\left(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \overline{\mathbf{V}}^{\delta}_{\infty}\right).$$

Again, we can either use analytical expressions to construct an estimator of the composite bias term or we can use re-sampling techniques. Because the adjustment of the different split-panel jackknife strategies to average partial effects is generic and straightforward, we omit it for brevity.

Next, we describe the analytical bias-corrected estimator of the averaged partial effects proposed by Fernández-Val and Weidner (2016, 2018b) and assume that $\hat{\boldsymbol{\Delta}}_{it}$ and $\hat{\boldsymbol{\delta}}$ are constructed from bias-corrected estimates of $\boldsymbol{\beta}$. The corresponding estimator of the composite bias term is

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

where

$$\widehat{\mathbf{B}}^{\delta} = \frac{1}{2}\sum_{i=1}^{N} \frac{\sum_{t=1}^{T} -\widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}\left(\widehat{\mathbb{P}}\widehat{\boldsymbol{\Psi}}\right)_{it} + \partial_{\eta^2}\hat{\boldsymbol{\Delta}}_{it} + 2\sum_{l=1}^{L}(T/(T-l))\sum_{t=l+1}^{T}\partial_{\eta}\hat{l}_{it-l}\hat{\omega}_{it}\left(\widehat{\mathbb{M}}\widehat{\boldsymbol{\Psi}}\right)_{it}}{\sum_{t=1}^{T}\hat{\omega}_{it}}$$

$$\widehat{\mathbf{C}}^{\delta} = \frac{1}{2}\sum_{t=1}^{T} \frac{\sum_{i=1}^{N} -\widehat{H}_{it}\partial_{\eta^2}\widehat{F}_{it}\left(\widehat{\mathbb{P}}\widehat{\boldsymbol{\Psi}}\right)_{it} + \partial_{\eta^2}\hat{\boldsymbol{\Delta}}_{it}}{\sum_{i=1}^{N}\hat{\omega}_{it}},$$

and $\widehat{\boldsymbol{\Psi}}_{it} = \partial_{\eta}\hat{\boldsymbol{\Delta}}_{it}/\hat{\omega}_{it}$, where $\partial_{\eta}\hat{\boldsymbol{\Delta}}_{it}$ is the first-order partial derivative of $\hat{\boldsymbol{\Delta}}_{it}$ with respect to $\hat{\eta}_{it}$. An estimator of the asymptotic covariance is

$$\widehat{\mathbf{V}}^{\delta} = \frac{r^2}{N^2T^2}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\bar{\hat{\boldsymbol{\Delta}}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\bar{\hat{\boldsymbol{\Delta}}}_{it}\right)' + \sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\boldsymbol{\Gamma}}_{it}\widehat{\boldsymbol{\Gamma}}_{it}' + 2\sum_{i=1}^{N}\sum_{s>t}^{T}\bar{\hat{\boldsymbol{\Delta}}}_{it}\widehat{\boldsymbol{\Gamma}}_{is}'\right],$$

where

$$\widehat{\boldsymbol{\Gamma}}_{it} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\partial_{\beta}\hat{\boldsymbol{\Delta}}_{it} - \left(\widehat{\mathbb{P}}\mathbf{X}\right)_{it}\partial_{\eta}\hat{\boldsymbol{\Delta}}_{it}\right)'\widehat{\mathbf{W}}^{-1}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{it}\hat{\omega}_{it}\hat{v}_{it} - \left(\widehat{\mathbb{P}}\widehat{\boldsymbol{\Psi}}\right)_{it}\partial_{\eta}\hat{l}_{it},$$

$\bar{\hat{\boldsymbol{\Delta}}}_{it} = \hat{\boldsymbol{\Delta}}_{it} - \hat{\boldsymbol{\delta}}$, and $\partial_{\beta}\hat{\boldsymbol{\Delta}}_{it}$ is the first-order partial derivative of $\hat{\boldsymbol{\Delta}}_{it}$ with respect to $\hat{\boldsymbol{\beta}}$. Note that the first term takes into account the variation induced by estimating sample instead of population means, the second term captures variation due to parameter estimation also known as delta method, and the last term is a covariance between both sources of variation that can be dropped if all explanatory variables are assumed to be strictly exogenous. Fernández-Val and Weidner (2016, 2018b) also derive an alternative estimator of $\overline{\mathbf{V}}^{\delta}$ by imposing additional sampling conditions with respect to the unobserved heterogeneity. Given that $\{\alpha_i\}_N$ and $\{\gamma_t\}_T$ are sequences of independent random variables and that $\alpha_i \perp \gamma_t \,\forall\, i,t$, the estimator of the asymptotic covariance simplifies to

$$\widehat{\mathbf{V}}^{\delta} = \frac{r^2}{N^2T^2}\sum_{i=1}^{N}\left(\sum_{t,s=1}^{T}\bar{\hat{\boldsymbol{\Delta}}}_{it}\bar{\hat{\boldsymbol{\Delta}}}_{is}' + \sum_{j\neq i}^{N}\sum_{t=1}^{T}\bar{\hat{\boldsymbol{\Delta}}}_{it}\bar{\hat{\boldsymbol{\Delta}}}_{jt}' + \sum_{t=1}^{T}\widehat{\boldsymbol{\Gamma}}_{it}\widehat{\boldsymbol{\Gamma}}_{it}' + 2\sum_{s>t}^{T}\bar{\hat{\boldsymbol{\Delta}}}_{it}\widehat{\boldsymbol{\Gamma}}_{is}'\right).$$

So far we know how to construct bias-corrected estimators for binary choice models with a two-way error component. However, the estimation of these models themselves is computationally challenging even in moderately large panels. The same applies to the computation of all quantities based on $\widehat{\mathbb{M}}$ and $\widehat{\mathbb{P}}$ that are required for the different bias corrections. In the next section, we present three algorithms that help to overcome those computational obstacles.

4.3 Computation in Large Panel Data

Recently Stammann (2018) proposed a fast and feasible algorithm to estimate all GLMs with a multi-way error component that is also directly applicable to unbalanced data. We briefly review the algorithm for binary choice models with individual and time fixed effects and show how parts of the estimation algorithm can be used to accelerate analytical bias corrections as well.

Remember, (4.1) has no closed form solution and has to be solved numerically. Using Newton's method, the parameter update in iteration r is

$$\left(\hat{\boldsymbol{\theta}}_{r+1} - \hat{\boldsymbol{\theta}}_r\right) = \left(\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{\nu}}, \tag{4.8}$$

where $\mathbf{Z} = (\mathbf{X}, \mathbf{D})$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\phi}}')'$.[4] Because increasing the number of observations also increases the rank of $\mathbf{D}$, the computation of the parameter update quickly becomes infeasible. Fortunately, a closer look reveals that (4.8) is essentially the solution of the following weighted least-squares problem:

$$\hat{\boldsymbol{\nu}} = \mathbf{X}\left(\boldsymbol{\beta}_{r+1} - \boldsymbol{\beta}_r\right) + \mathbf{D}\left(\boldsymbol{\phi}_{r+1} - \boldsymbol{\phi}_r\right) + \mathbf{u}, \tag{4.9}$$

where $\widehat{\boldsymbol{\Omega}}$ is the corresponding weighting matrix. This implies that the normal equations of (4.9) are

$$\mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{X}\left(\hat{\boldsymbol{\beta}}_{r+1} - \hat{\boldsymbol{\beta}}_r\right) + \mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{D}\left(\hat{\boldsymbol{\phi}}_{r+1} - \hat{\boldsymbol{\phi}}_r\right) = \mathbf{X}'\widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{\nu}}, \tag{4.10}$$

$$\mathbf{D}'\widehat{\boldsymbol{\Omega}}\mathbf{X}\left(\hat{\boldsymbol{\beta}}_{r+1} - \hat{\boldsymbol{\beta}}_r\right) + \mathbf{D}'\widehat{\boldsymbol{\Omega}}\mathbf{D}\left(\hat{\boldsymbol{\phi}}_{r+1} - \hat{\boldsymbol{\phi}}_r\right) = \mathbf{D}'\widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{\nu}}. \tag{4.11}$$

Re-arranging (4.11) yields

$$\mathbf{D}\left(\hat{\boldsymbol{\phi}}_{r+1} - \hat{\boldsymbol{\phi}}_r\right) = \widehat{\mathbb{P}}\left(\hat{\boldsymbol{\nu}} - \mathbf{X}\left(\hat{\boldsymbol{\beta}}_{r+1} - \hat{\boldsymbol{\beta}}_r\right)\right). \tag{4.12}$$

4. It is actually a particular variant of Newton's method known as Fisher scoring where the observed Hessian is replaced by its expectation (see chapter 2.5 in McCullagh and Nelder 1989).

Substituting (4.12) in (4.10) and exploiting that $\widehat{\mathbb{M}}$ is idempotent reveals that

$$(\hat{\boldsymbol{\beta}}_{r+1} - \hat{\boldsymbol{\beta}}_r) = \left((\widehat{\mathbb{M}}\mathbf{X})' \widehat{\boldsymbol{\Omega}} (\widehat{\mathbb{M}}\mathbf{X})\right)^{-1} (\widehat{\mathbb{M}}\mathbf{X})' \widehat{\boldsymbol{\Omega}} (\widehat{\mathbb{M}}\hat{\boldsymbol{\nu}})$$

is the weighted least-squares solution of

$$\widehat{\mathbb{M}}\hat{\boldsymbol{\nu}} = \widehat{\mathbb{M}}\mathbf{X}(\boldsymbol{\beta}_{r+1} - \boldsymbol{\beta}_r) + \mathbf{u}. \tag{4.13}$$

Consequently, as for the linear model, we can separate the estimation of the structural from the incidental parameters.[5]

However, we also need to update $\hat{\boldsymbol{\nu}}$ and $\widehat{\boldsymbol{\Omega}}$ in each iteration. Both are functions of the linear predictor $\hat{\boldsymbol{\eta}}$ which is a function of the incidental parameters as well. Either we need to use a numerical solver to find estimates of the incidental parameters for a given $\hat{\boldsymbol{\beta}}$, which can be very computationally demanding, or we need to find a way to update the linear predictor itself. Fortunately, $\hat{\boldsymbol{\eta}}$ can be updated quite easily using already computed quantities. From the linear fixed effects model it is well known that the residuals of (4.9) and (4.13) are equal see (see Gaure 2013b). Some rearrangements yield

$$(\hat{\boldsymbol{\eta}}_{r+1} - \hat{\boldsymbol{\eta}}_r) = \hat{\boldsymbol{\nu}} - \widehat{\mathbb{M}}\hat{\boldsymbol{\nu}} + \widehat{\mathbb{M}}\mathbf{X}(\hat{\boldsymbol{\beta}}_{r+1} - \hat{\boldsymbol{\beta}}_r).$$

Summing up, the entire algorithm can be sketched as follows:

Definition. *Newton's Method*

Initialize $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$; repeat the following steps until convergence

Step 1: Given $\hat{\boldsymbol{\eta}}$ compute $\hat{\boldsymbol{\nu}}$ and $\widehat{\boldsymbol{\Omega}}$
Step 2: Given $\hat{\boldsymbol{\nu}}$ and $\widehat{\boldsymbol{\Omega}}$ update $\hat{\boldsymbol{\beta}}$
Step 3: Given $\hat{\boldsymbol{\beta}}$ update $\hat{\boldsymbol{\eta}}$

So far we have re-arranged the optimization problem such that it abstains from the estimation of potentially many incidental parameters. Unfortunately, a remaining challenge is the computation of $\widehat{\mathbb{M}}$ itself. Because the residual projection is of dimension $NT \times NT$, the computation and storage quickly becomes infeasible. Let $\mathbf{v}$ be an arbitrary vector and $\widehat{\mathbb{M}}\mathbf{v}$ the corresponding weighted within transformation. In case of a one-way error component, $\widehat{\mathbb{M}}\mathbf{v}$ can be efficiently computed by subtracting weighted group means from $\mathbf{v}$. Throughout this article we refer to any $\widehat{\mathbb{M}}\mathbf{v}$ as centered

5. Note that Stammann (2018) proposes an additional valid residual projection. Let $\widetilde{\mathbb{M}} = \mathbb{1}_{NT} - \widetilde{\mathbb{P}} = \mathbb{1}_{NT} - \widetilde{\mathbf{D}}(\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}})^{-1}\widetilde{\mathbf{D}}'$, where $\widetilde{\mathbf{D}} = \widehat{\boldsymbol{\Omega}}^{1/2}\mathbf{D}$. An estimate of $(\boldsymbol{\beta}_{r+1} - \boldsymbol{\beta}_r)$ can be obtained by regressing $\widetilde{\mathbb{M}}\tilde{\boldsymbol{\nu}}$ on $\widetilde{\mathbb{M}}\widetilde{\mathbf{X}}$, where $\tilde{\boldsymbol{\nu}} = \widehat{\boldsymbol{\Omega}}^{1/2}\hat{\boldsymbol{\nu}}$ and $\widetilde{\mathbf{X}} = \widehat{\boldsymbol{\Omega}}^{1/2}\mathbf{X}$. Thus $\widehat{\boldsymbol{\Omega}}^{1/2}\widehat{\mathbb{M}}\hat{\boldsymbol{\nu}} = \widetilde{\mathbb{M}}\tilde{\boldsymbol{\nu}}$ and $\widehat{\boldsymbol{\Omega}}^{1/2}\widehat{\mathbb{M}}\mathbf{X} = \widetilde{\mathbb{M}}\widetilde{\mathbf{X}}$. During extensive studies in the development of our *R*-package *alpaca*, we did not find any projection to be superior in terms of computation time. In this article, we use $\widehat{\mathbb{M}}$ because it is in line with notation used in Fernández-Val and Weidner (2016, 2018a).

vector. However, because $\widehat{\mathbb{M}}$ loses its sparse structure for models with a multi-way error component, we cannot derive a simple scalar expression for these cases.

In the context of linear models, Guimarães and Portugal (2010) and Gaure (2013b) propose a computationally efficient approach to obtain centered vectors for any multi-way error component. Combining the results of Neumann (1950) and Halperin (1962), they suggest an iterative procedure known as the method of alternating projections (MAP) which results in an arbitrary close approximation of the within transformation. Gaure (2013b) gives a detailed theoretical foundation of this approach in the context of linear models. Stammann (2018) shows how to extend MAP to GLMs.

To get an intuition how MAP works, we briefly describe an algorithm for GLMs with a two-way error component. Let $\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2)$, where $\mathbf{D}_1$ and $\mathbf{D}_2$ are submatrices indicating individuals and time periods, respectively. Further, we introduce the following centered vectors $\widehat{\mathbb{M}}_k \mathbf{v} = \mathbb{1}_{NT} - \mathbf{D}_k (\mathbf{D}_k' \widehat{\mathbf{\Omega}} \mathbf{D}_k)^{-1} \mathbf{D}_k' \widehat{\mathbf{\Omega}} \mathbf{v}$, where $k \in \{1, 2\}$. The appropriate scalar expressions for the weighted within transformations are

$$\left(\widehat{\mathbb{M}}_1 \mathbf{v}\right)_{it} = v_{it} - \frac{\sum_{t=1}^{T} \hat{\omega}_{it} v_{it}}{\sum_{t=1}^{T} \hat{\omega}_{it}} \quad \text{and} \quad \left(\widehat{\mathbb{M}}_2 \mathbf{v}\right)_{it} = v_{it} - \frac{\sum_{i=1}^{N} \hat{\omega}_{it} v_{it}}{\sum_{i=1}^{N} \hat{\omega}_{it}}.$$

The centering algorithm using MAP can be described as follows:

Definition. *Centering Algorithm using MAP (von Neumann / Halperin)*
Initialize $\widehat{\mathbb{M}} \mathbf{v} = \mathbf{v}$*; repeat the following steps until convergence*

Step 1: Compute $\widehat{\mathbb{M}}_1 \widehat{\mathbb{M}} \mathbf{v}$ *and update* $\widehat{\mathbb{M}} \mathbf{v}$ *such that* $\widehat{\mathbb{M}} \mathbf{v} = \widehat{\mathbb{M}}_1 \widehat{\mathbb{M}} \mathbf{v}$
Step 2: Compute $\widehat{\mathbb{M}}_2 \widehat{\mathbb{M}} \mathbf{v}$ *and update* $\widehat{\mathbb{M}} \mathbf{v}$ *such that* $\widehat{\mathbb{M}} \mathbf{v} = \widehat{\mathbb{M}}_2 \widehat{\mathbb{M}} \mathbf{v}$

Because this algorithm only needs to evaluate scalar expressions, it is memory efficient and quite fast. Further, given an appropriate tolerance level, it returns an arbitrary close approximation to $\widehat{\mathbb{M}} \mathbf{v}$, that can be used to accelerate Newton's method as well as analytical bias corrections (see Stammann 2018 for further details on MAP). More precisely, we can use MAP to approximate $\widehat{\mathbb{M}} \hat{\boldsymbol{\nu}}$ and $\widehat{\mathbb{M}} \mathbf{X}$, where the latter is obtained by sequentially applying the algorithm to each column of $\mathbf{X}$. These approximations can be used afterwards to compute updates of the structural parameters and estimates of the leading bias terms.

Next, we present a further algorithm that is required in the context of bias corrections. Suppose we have bias-corrected the structural parameter estimates and want to re-estimate our model given we already know that $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$. For instance, this is required if we want to apply an analytical bias correction at the level of the score or bias-correct the average partial effects. In the literature of GLMs, this type of algorithm is known as *offset* tracing back to Nelder and Wedderburn (1972). In the

following, we derive a computationally efficient *offset* algorithm based on MAP. To do this, we have to re-formulate the maximization problem in (4.1) as

$$\hat{\boldsymbol{\phi}} = \left(\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\gamma}}'\right)' = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \sum_{i}^{N} \sum_{t}^{T} l_{it}(\tilde{\boldsymbol{\beta}}, \alpha_i, \gamma_t),$$

where $\tilde{\boldsymbol{\beta}}$ is assumed to be known. This yields the following update step in iteration r:

$$\left(\hat{\boldsymbol{\phi}}_{r+1} - \hat{\boldsymbol{\phi}}_r\right) = (\mathbf{D}'\widehat{\boldsymbol{\Omega}}\mathbf{D})^{-1}\mathbf{D}'\widehat{\boldsymbol{\Omega}}\hat{\mathbf{v}}$$

with

$$\hat{\boldsymbol{\eta}}_{r+1} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{D}\hat{\boldsymbol{\phi}}_{r+1} \quad \text{and} \quad \mathbf{D}\hat{\boldsymbol{\phi}}_{r+1} = \widehat{\mathbb{P}}\hat{\mathbf{v}} + \mathbf{D}\hat{\boldsymbol{\phi}}_r .$$

Note that the linear predictor is a sufficient quantity to compute, for example, standard errors, partial effects, or predictions. The entire *offset* algorithm can be summarized as follows:

Definition. *Newton's Method (Offset)*

Given $\tilde{\boldsymbol{\beta}}$ initialize $\hat{\boldsymbol{\eta}}$; repeat the following steps until convergence

Step 1: Given $\hat{\boldsymbol{\eta}}$ compute $\hat{\mathbf{v}}$ and $\widehat{\boldsymbol{\Omega}}$
Step 2: Given $\hat{\mathbf{v}}$ and $\widehat{\boldsymbol{\Omega}}$ update $\hat{\boldsymbol{\eta}}$

Finally, we give a short impression about the capabilities of the algorithms presented in this section. Therefore, we estimate a fixed effects probit model with three explanatory variables and a two-way error component and compare the overall computation time of different *R* commands. More precisely, we use `feglm()` provided in our *R*-package *alpaca*, which is based on the algorithms described in this section, and compare it to `speedglm()` (Enea 2017) and `glm()` (R Core Team 2019). Used on a data set consisting of 2,000 individuals observed for 52 time periods, our routine takes about half a second, while `speedglm()` and `glm()` require 22 and 1,120 seconds, respectively.[6]

In summary, we have presented three algorithms that help to speed up the computation of binary choice models with two-way error components and the corresponding bias corrections. In the next subsequent sections, we use these algorithms in an extensive simulation study and an empirical example from labor economics.

4.4 Simulation Experiments

We analyze the finite sample behavior of different uncorrected and bias-corrected fixed effects estimators for binary choice models. The quantities of interest are the

6. All computations were done on a Linux Mint 18.1 workstation using *R* version 3.6.1 and an Intel Xeon E5-2640 v3s.

structural parameters and average partial effects. Beside the different nonlinear estimators introduced in this article, we additionally consider the linear fixed effects estimator as an alternative to obtain estimates of the average partial effects. We restrict ourselves to the analysis of dynamic models, because the statistical properties with respect to the exogenous regressor are similar in static and dynamic designs (see Fernández-Val and Weidner 2016).

Next, we describe all estimators analyzed in this simulation study. Beside the uncorrected probit estimator (MLE), we consider four different analytical bias corrections for the structural parameters. Two of them correct the estimator itself, whereas the others are obtained by minimizing modified score equations. ABC1 is the analytical bias correction analyzed by Fernández-Val and Weidner (2016, 2018a). ABC2 is essentially ABC1, but additionally iterated until convergence. Arellano and Hahn (2007) refer to this approach as infinitely repeated analytical bias correction. ABC3 and ABC4 are the score-corrected estimators. They only differ in that ABC4 updates the bias terms in each iteration of the nonlinear solver, whereas ABC3 treats them as fixed. The analytical bias-corrected estimators of the average partial effects are labeled analogously. Further, we consider two split-panel jackknife bias-corrected estimators that differ in their splitting strategy. SPJ1 and SPJ2 refer to the strategies used in (4.6) and (4.7), respectively. Finally, we use an analytical bias-corrected estimator for dynamic linear fixed effects models that was initially proposed by Nickell (1981) (see among others Hahn and Kuersteiner 2002; Hahn and Moon 2006; Fernández-Val and Weidner 2018a; Chen, Chernozhukov, and Fernández-Val 2019). Throughout this article, we denote the (bias-corrected) linear fixed effects estimator as LPM.

We use the dynamic model design of Fernández-Val and Weidner (2016) and generate

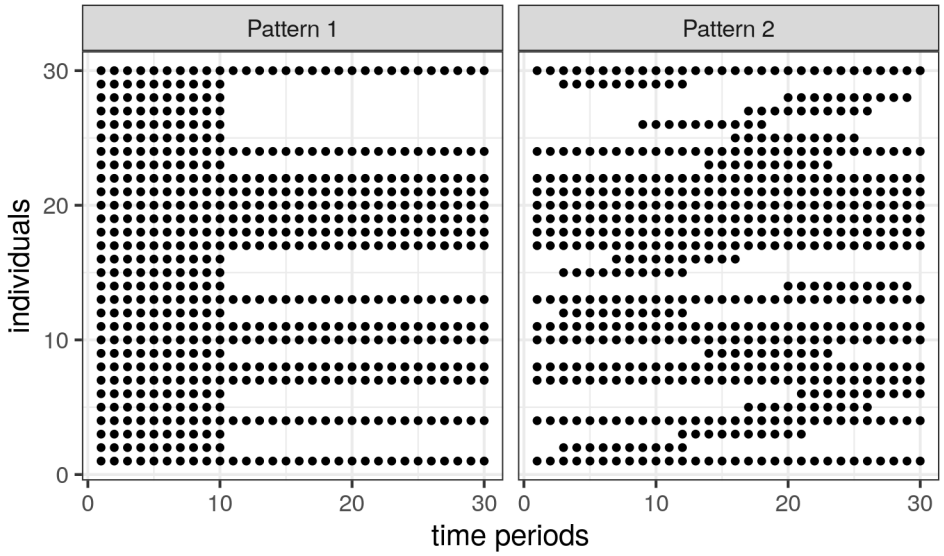
$$y_{it} = \mathbf{1}\left[\rho y_{it-1} + \beta x_{it} + \alpha_i + \gamma_t \geq \epsilon_{it}\right],$$
$$y_{i0} = \mathbf{1}\left[\beta x_{i0} + \alpha_i + \gamma_0 \geq \epsilon_{i0}\right],$$

where $i = 1,\dots,N$, $t = s_i,\dots,T_i$, $\alpha_i \sim$ iid. $\mathcal{N}(0,1/16)$, $\gamma_t \sim$ iid. $\mathcal{N}(0,1/16)$, $\epsilon_{it} \sim$ iid. $\mathcal{N}(0,1)$, and $\mathbf{1}[\cdot]$ is an indicator function. Furthermore, we assume that the exogenous regressor follows an AR-1 process: $x_{it} = 0.5x_{it-1} + \alpha_i + \gamma_t + \nu_{it}$, where $\nu_{it} \sim$ iid. $\mathcal{N}(0,0.5)$ and $x_{i0} \sim$ iid. $\mathcal{N}(0,1)$. The corresponding structural parameters are $\rho = 0.5$ and $\beta = 1$.

Contrary to Fernández-Val and Weidner (2016), we analyze three different panel structures and use sample sizes that better reflect commonly used data sets (much more individuals than time periods). More precisely, the first structure is a balanced panel, whereas the others mimic different patterns of randomly missing observations.

To describe the different patterns, we introduce two types of individuals: type 1 and type 2. Let N_1 and N_2 denote the number of type 1 and type 2, such that $N = N_1 + N_2$. Further, we assume that type 1 and type 2 are observed for T_1 and T_2 consecutive time periods, respectively. In the first pattern, the time series of both types starts at $t = 1$ but type 1 leaves the panel at an earlier point of time such that $T_1 < T_2$. The second pattern is identical in the sense that type 2 is observed longer than type 1. However, the time series of any type 1 does not necessary start at $t = 1$. Instead, an initial period is chosen randomly for each type 1 such that $t = s_i, \ldots, s_i + T_1$, where s_i is sampled with equal probability from $\{0, 1, \ldots, T_2 - T_1\}$. Figure 4.1 provides a graphical illustration for both of the missing

Figure 4.1: *Patterns of Randomly Missing Observations*



data patterns. Further, we generate panel data sets of different sizes. In case of balanced data $N = 200$ and $T_i = T \in \{10, 15, 20, 25, 30\}$, whereas in case of unbalanced data $\{N_1, N_2\} \in \{\{300, 100\}, \{150, 150\}, \{60, 180\}\}$, $T_1 = 10$, and $T_2 = 30$. The different pairs of $\{N_1, N_2\}$ are chosen such that the average number of individuals ($\overline{N}$) and time periods ($\overline{T}$) allow comparisons between the different panel structures. More precisely, $\overline{N} = N = 200$ and $\overline{T} \subset T_i \in \{15, 20, 25\}$.

To analyze the finite-sample properties and ensure comparability, we follow Fernández-Val and Weidner (2016) and compute the following statistics: biases, standard deviations (*SD*), root mean squared errors (*RMSE*), average ratios of standard errors and standard deviations (*SE*/*SD*), and empirical coverage probabilities of 95 % confidence intervals (*CP .95*). Throughout this article we report biases, *SD*, and *RMSE* in percentage relative to the truth. The average partial effects and the corresponding standard errors are computed based on (4.3) and (4.4) and the simplified expression of the asymptotic covariance. Additionally, we consider different choices

of the bandwidth parameter for the analytical bias corrections, $L \in \{1,2,3,4\}$. To get insights how joint hypothesis testing is affected by IPP, we analyze sizes of different Wald tests with $H_0 : \rho = 0.5 \wedge \beta = 1$ at a nominal level of 5 % using test statistics constructed from different estimators. All results are based on 1,000 replications using *R* version 3.6.1 (R Core Team 2019) on a Linux Mint 18.1 workstation with an Intel Xeon E5-2640 v3s.[7] A complete summary of all statistics can be found in the supplementary material.[8]

We start with the comparison between the different analytical corrections. Table 4.2 reports the relative biases of the estimators of the structural parameters and average partial effects along with different choices of the bandwidth parameter. For brevity, we only present results for balanced panels where $T \in \{10,20,30\}$ and note that the findings are similar in unbalanced panels. The relative biases of estimators corresponding to the predetermined variable are more severe than their exogenous counterpart. As expected, all corrections reduce a larger fraction of the bias as T increases. Furthermore, the differences between the estimators are most apparent in the case of T =10, where ABC2–ABC4 are clearly dominated by ABC1. This also holds for $T = 20$ and $T = 30$, but the differences in relative biases become negligible small. This is in line with findings of Juodis (2015) who analyzed an iterated analytical bias correction for static probit models with one-way error component. If we additionally take into account that the other analytical bias corrections are much more computationally demanding, ABC1 is clearly preferable. Further, we find that values of $L \in \{1,2\}$ are the most appropriate bandwidth choices for our chosen panel dimensions.

Next, we compare the two different split-panel jackknife estimators described in this article. Again, we restrict ourselves to the case of balanced panels and note that we have similar findings for unbalanced. The results are reported in table 4.3. Similar to the analytical correction, the bias reduction improves as T increases. We find almost identical properties of both estimators which is remarkably, because we would expect that the splitting strategy of SPJ2 leads to higher dispersion due to the use of significantly smaller subpanels to construct the composite bias term. Only for estimators of the structural parameters and $T = 10$, we observe that the relative bias and dispersion of SPJ2 is slightly higher. For the average partial effects we observe that the properties of both estimators are indistinguishable irrespective of the sample size. Further, note that SPJ2 is computationally less demanding because

7. Additionally, we use the *R*-package *lfe* of Gaure (2013a) for the estimation of linear fixed effects models and the nonlinear equations solver (*nleqslv*) provided by Hasselman (2018) for the score-corrected analytical bias corrections.

8. We also report results of a static data generating process and different designs of the exogenous regressor following Fernández-Val and Weidner (2016). Additionally, we provide a replication of the authors simulation study. The complete summary of all statistics is available from the authors upon request.

Table 4.2: *Analytical Bias Corrections and Bandwidth Parameters*

	Coefficients				APE			
	$L=1$	$L=2$	$L=3$	$L=4$	$L=1$	$L=2$	$L=3$	$L=4$
	$N=200; T=10$							
	Lagged Dependent Variable							
ABC1	-7.31	-8.75	-17.52	-26.82	-9.26	-10.67	-19.70	-29.21
ABC2	-13.94	-13.91	-20.99	-29.00	-16.94	-16.88	-24.05	-32.09
ABC3	-9.34	-10.73	-19.18	-28.13	-11.82	-13.18	-21.77	-30.84
ABC4	-11.13	-10.96	-18.39	-26.86	-13.60	-13.40	-21.00	-29.58
	Exogenous Regressor							
ABC1	1.38	1.20	1.37	1.54	-0.01	-0.14	-0.15	-0.20
ABC2	4.90	4.74	4.76	4.70	2.37	2.22	2.08	1.87
ABC3	3.08	2.90	2.95	3.02	1.12	0.99	0.89	0.78
ABC4	3.04	2.85	2.95	2.92	1.14	0.97	0.88	0.69
	$N=200; T=20$							
	Lagged Dependent Variable							
ABC1	-4.25	-2.20	-3.85	-5.91	-4.99	-2.81	-4.54	-6.70
ABC2	-6.07	-3.85	-5.21	-7.03	-7.16	-4.84	-6.26	-8.16
ABC3	-4.63	-2.61	-4.24	-6.27	-5.51	-3.37	-5.06	-7.18
ABC4	-5.36	-3.05	-4.45	-6.33	-6.27	-3.84	-5.31	-7.27
	Exogenous Regressor							
ABC1	0.86	0.72	0.78	0.88	0.28	0.18	0.24	0.32
ABC2	1.80	1.69	1.75	1.82	0.95	0.87	0.91	0.96
ABC3	1.22	1.10	1.15	1.23	0.52	0.44	0.49	0.56
ABC4	1.29	1.16	1.22	1.31	0.60	0.50	0.55	0.61
	$N=200; T=30$							
	Lagged Dependent Variable							
ABC1	-3.07	-1.15	-1.72	-2.58	-3.32	-1.29	-1.88	-2.79
ABC2	-3.92	-1.97	-2.43	-3.20	-4.35	-2.29	-2.78	-3.59
ABC3	-3.22	-1.32	-1.88	-2.73	-3.53	-1.52	-2.11	-3.00
ABC4	-3.60	-1.60	-2.07	-2.86	-3.93	-1.82	-2.32	-3.15
	Exogenous Regressor							
ABC1	0.33	0.20	0.23	0.27	0.16	0.05	0.09	0.12
ABC2	0.78	0.67	0.70	0.73	0.48	0.39	0.42	0.45
ABC3	0.48	0.36	0.39	0.42	0.26	0.17	0.20	0.23
ABC4	0.53	0.41	0.44	0.48	0.31	0.21	0.24	0.27

Note: All entries are biases in percentage relative to the truth; results based on 1,000 repetitions.

Table 4.3: *Split-Panel Jackknife Bias Corrections*

	Coefficients				APE			
	SPJ1		SPJ2		SPJ1		SPJ2	
	Bias	SD	Bias	SD	Bias	SD	Bias	SD
	Lagged Dependent Variable							
$N = 200; T = 10$	19.82	21.20	22.15	21.93	-12.05	19.64	-11.88	19.68
$N = 200; T = 15$	-0.07	15.06	0.52	15.11	-10.67	15.14	-10.62	15.11
$N = 200; T = 20$	3.38	12.38	3.75	12.42	-2.82	13.37	-2.80	13.38
$N = 200; T = 25$	0.34	10.64	0.57	10.68	-3.21	11.41	-3.18	11.42
$N = 200; T = 30$	1.44	9.79	1.63	9.82	-0.89	10.70	-0.86	10.71
	Exogenous Regressor							
$N = 200; T = 10$	-7.12	9.05	-9.86	9.55	5.67	7.60	5.26	7.65
$N = 200; T = 15$	-1.08	6.00	-1.98	6.07	2.55	5.79	2.39	5.79
$N = 200; T = 20$	-1.70	4.71	-2.31	4.74	1.43	4.68	1.31	4.69
$N = 200; T = 25$	-0.61	4.13	-0.99	4.13	0.93	4.17	0.85	4.17
$N = 200; T = 30$	-1.00	3.51	-1.32	3.51	0.54	3.77	0.48	3.77

Note: Bias and SD denote biases and standard deviations in percentage relative to the truth; results based on 1,000 repetitions.

the model is re-estimated the same amount of times but on smaller subpanels.

So far, we found that the statistical properties of the different analytical and split-panel bias corrections barely differ from each other. To allow for some comparisons with the study of Fernández-Val and Weidner (2016), we focus on the small sample properties of MLE, ABC1, SPJ1, and LPM, where values in parentheses indicate the corresponding choice of the bandwidth parameter. Table 4.4 and 4.5 report the results based on balanced panel data sets. First, we find that the properties of the estimators that refer to effect of the predetermined variable are worse than those that are related to the exogenous regressor. For instance, we observe larger relative biases and dispersion as well as coverage probabilities further away from their nominal level. The relative distortion we find in the coefficients is also reflected in the estimates of the average partial effects. That is contrary to the results we observe with regard to the average partial effects of the exogenous regressor, where we can only find negligibly small relative biases.[9] Generally, the bias corrections work as expected. They reduce the relative biases and improve the coverage probabilities. As in Fernández-Val and Weidner (2016), the properties of SPJ1 are worse than those of ABC1. Another interesting insight can be learned from LPM. In case of the exogenous regressor, the estimators do not show any distortion, but valid inference is questionable, because standard errors are underestimated and coverage probabilities are lower than their nominal level. For the predetermined variable, there is also the curiosity that the relative bias increases in T.[10] Our explanation for

9. Hahn and Newey (2004), Fernández-Val (2009), and Fernández-Val and Weidner (2016) also find only small biases in average partial effects of the exogenous regressor.

10. To ensure that this is not due to a weird programming error, we add some simulation experiments in appendix A where we apply the bias-corrected linear fixed effects estimator to a standard data

Table 4.4: *Properties: Balanced - Lagged Dependent Variable*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	$N = 200; T = 10$									
MLE	-64	18	66	0.96	0.05	-70	15	72	1.05	0.01
ABC1 (1)	-7	16	17	1.09	0.95	-9	16	19	1.10	0.94
ABC1 (2)	-9	17	19	1.01	0.92	-11	18	21	1.02	0.91
SPJ1	20	21	29	0.79	0.76	-12	20	23	0.99	0.88
LPM (1)						6	18	18	0.95	0.92
LPM (2)						7	19	20	0.88	0.89
	$N = 200; T = 15$									
MLE	-42	14	44	1.01	0.12	-50	12	51	1.05	0.03
ABC1 (1)	-6	12	14	1.09	0.95	-7	13	15	1.08	0.94
ABC1 (2)	-4	13	14	1.03	0.95	-5	14	15	1.03	0.94
SPJ1	-0	15	15	0.89	0.91	-11	15	19	0.94	0.87
LPM (1)						10	14	17	0.95	0.87
LPM (2)						13	15	20	0.90	0.82
	$N = 200; T = 20$									
MLE	-31	12	33	0.98	0.23	-38	11	39	0.98	0.09
ABC1 (1)	-4	11	12	1.04	0.94	-5	12	13	1.01	0.94
ABC1 (2)	-2	11	12	1.00	0.95	-3	12	13	0.97	0.94
SPJ1	3	12	13	0.92	0.91	-3	13	14	0.91	0.93
LPM (1)						12	13	17	0.92	0.80
LPM (2)						15	13	20	0.88	0.71
	$N = 200; T = 25$									
MLE	-24	10	26	1.03	0.35	-30	10	32	1.02	0.15
ABC1 (1)	-3	9	10	1.08	0.95	-4	10	11	1.05	0.95
ABC1 (2)	-1	10	10	1.04	0.96	-2	10	11	1.02	0.95
SPJ1	0	11	11	0.95	0.94	-3	11	12	0.94	0.93
LPM (1)						14	11	18	0.96	0.71
LPM (2)						17	11	20	0.93	0.61
	$N = 200; T = 30$									
MLE	-20	10	23	0.97	0.41	-25	10	27	0.95	0.22
ABC1 (1)	-3	9	10	1.01	0.94	-3	10	10	0.97	0.93
ABC1 (2)	-1	9	9	0.98	0.94	-1	10	10	0.95	0.93
SPJ1	1	10	10	0.94	0.93	-1	11	11	0.91	0.92
LPM (1)						15	11	18	0.89	0.65
LPM (2)						17	11	20	0.87	0.56

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; values in parentheses after ABC1 and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 4.5: *Properties: Balanced - Exogenous Regressor*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	$N = 200; T = 10$									
MLE	22	8	24	0.87	0.14	3	6	7	1.10	0.93
ABC1 (1)	1	6	7	1.00	0.95	-0	6	6	1.11	0.97
ABC1 (2)	1	6	7	0.99	0.94	-0	6	6	1.10	0.97
SPJ1	-7	9	12	0.69	0.68	6	8	9	0.91	0.83
LPM (1)						-0	6	6	0.81	0.88
LPM (2)						-0	6	6	0.81	0.88
	$N = 200; T = 15$									
MLE	14	6	15	0.95	0.23	3	5	6	0.97	0.90
ABC1 (1)	1	5	5	1.04	0.95	0	5	5	0.98	0.94
ABC1 (2)	1	5	5	1.04	0.95	0	5	5	0.97	0.94
SPJ1	-1	6	6	0.82	0.88	3	6	6	0.87	0.87
LPM (1)						-0	5	5	0.78	0.88
LPM (2)						-0	5	5	0.77	0.87
	$N = 200; T = 20$									
MLE	11	5	12	0.95	0.32	2	4	5	0.98	0.90
ABC1 (1)	1	4	4	1.02	0.95	0	4	4	0.99	0.94
ABC1 (2)	1	4	4	1.02	0.95	0	4	4	0.98	0.95
SPJ1	-2	5	5	0.88	0.89	1	5	5	0.90	0.90
LPM (1)						-0	4	4	0.78	0.88
LPM (2)						-0	4	4	0.78	0.88
	$N = 200; T = 25$									
MLE	8	4	9	0.95	0.41	2	4	4	0.96	0.91
ABC1 (1)	0	4	4	1.01	0.96	0	4	4	0.96	0.94
ABC1 (2)	0	4	4	1.01	0.96	0	4	4	0.96	0.94
SPJ1	-1	4	4	0.88	0.92	1	4	4	0.91	0.91
LPM (1)						-0	4	4	0.77	0.87
LPM (2)						-0	4	4	0.77	0.86
	$N = 200; T = 30$									
MLE	7	4	8	0.95	0.48	2	4	4	0.95	0.90
ABC1 (1)	0	3	3	1.00	0.95	0	4	4	0.96	0.93
ABC1 (2)	0	3	3	1.00	0.95	0	4	4	0.95	0.93
SPJ1	-1	4	4	0.94	0.93	1	4	4	0.92	0.93
LPM (1)						-0	4	4	0.74	0.84
LPM (2)						-1	4	4	0.74	0.84

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; values in parentheses after ABC1 and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

this phenomenon is that for larger values of T, the predicted probabilities of LPM are more frequently outside of the unit interval. Because the average partial effects of binary regressors are simply differences in the predicted probabilities, this might explain the increase in the relative bias for the effect of the predetermined variable. Note that Hinz, Stammann, and Wanner (2019) use a slight modification of the data generating process used in this article and have very similar findings.

Table 4.6: *Properties: Unbalanced 1 - Lagged Dependent Variable*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	$\overline{N} = 200; \overline{T} = 15$									
MLE	-40	10	42	0.94	0.01	-48	9	49	0.99	0.00
ABC1 (1)	-5	9	11	1.02	0.91	-7	10	12	1.03	0.90
ABC1 (2)	-5	10	11	0.96	0.90	-6	10	12	0.97	0.89
SPJ1	-31	12	33	0.81	0.14	-37	11	39	0.86	0.05
LPM (1)						10	11	15	0.90	0.78
LPM (2)						12	11	17	0.85	0.69
	$\overline{N} = 200; \overline{T} = 20$									
MLE	-30	9	31	1.02	0.10	-37	9	38	1.02	0.02
ABC1 (1)	-4	9	10	1.09	0.95	-5	9	11	1.04	0.92
ABC1 (2)	-3	9	9	1.05	0.96	-4	10	10	1.00	0.92
SPJ1	-14	10	17	0.97	0.68	-19	10	21	0.95	0.52
LPM (1)						13	10	16	0.97	0.73
LPM (2)						15	10	18	0.93	0.63
	$\overline{N} = 200; \overline{T} = 25$									
MLE	-24	10	26	0.98	0.28	-30	9	31	0.98	0.13
ABC1 (1)	-3	9	10	1.03	0.94	-4	10	10	1.01	0.93
ABC1 (2)	-1	9	9	0.99	0.94	-2	10	10	0.97	0.95
SPJ1	-4	10	11	0.91	0.91	-7	11	13	0.91	0.85
LPM (1)						14	10	18	0.91	0.66
LPM (2)						17	11	20	0.88	0.58

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; values in parentheses after ABC1 and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Next, we analyze how the two patterns of unbalancedness affect the properties of the different estimators. First of all our results, summarized in table 4.6–4.9, support the conjecture of Fernández-Val and Weidner (2018a) that the order of the bias in the asymptotic distribution of MLE, in case of randomly missing observations, depends on $\overline{N}$ and $\overline{T}$. This can be confirmed by comparing the statistical properties of MLE in balanced and unbalanced settings where $N = \overline{N}$ and $T = \overline{T}$. We observe that in these cases the properties of MLE are almost identical. Whereas the different missing data patterns do not affect MLE, ABC1, and LPM, they worsen the statistical properties

generating process for dynamic linear fixed effects models. This small simulation study confirms that the bias correction works as intended.

Table 4.7: *Properties: Unbalanced 1 - Exogenous Regressor*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	$\overline{N} = 200; \overline{T} = 15$									
MLE	14	4	14	0.90	0.05	2	4	5	1.00	0.90
ABC1 (1)	1	4	4	0.98	0.94	0	4	4	1.00	0.94
ABC1 (2)	1	4	4	0.98	0.94	-0	4	4	0.99	0.95
SPJ1	9	5	10	0.76	0.32	3	5	6	0.85	0.81
LPM (1)						-0	4	4	0.72	0.84
LPM (2)						-0	4	4	0.72	0.84
	$\overline{N} = 200; \overline{T} = 20$									
MLE	10	4	11	0.91	0.22	2	4	4	0.99	0.91
ABC1 (1)	1	3	4	0.98	0.95	-0	4	4	0.99	0.95
ABC1 (2)	0	3	4	0.97	0.95	-0	4	4	0.98	0.95
SPJ1	3	4	5	0.86	0.80	1	4	4	0.93	0.91
LPM (1)						-0	4	4	0.76	0.86
LPM (2)						-1	4	4	0.76	0.85
	$\overline{N} = 200; \overline{T} = 25$									
MLE	8	4	9	0.97	0.36	2	4	4	0.95	0.91
ABC1 (1)	0	3	3	1.02	0.95	0	4	4	0.96	0.94
ABC1 (2)	0	3	3	1.02	0.95	-0	4	4	0.96	0.94
SPJ1	1	4	4	0.93	0.93	1	4	4	0.91	0.92
LPM (1)						-0	4	4	0.75	0.86
LPM (2)						-0	4	4	0.75	0.86

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; values in parentheses after ABC1 and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 4.8: *Properties: Unbalanced 2 - Lagged Dependent Variable*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	$\overline{N} = 200; \overline{T} = 15$									
MLE	-40	10	41	0.95	0.01	-47	9	48	0.99	0.00
ABC1 (1)	-5	9	10	1.04	0.93	-6	10	12	1.03	0.92
ABC1 (2)	-5	10	11	0.98	0.92	-6	10	12	0.98	0.91
SPJ1	-20	10	22	0.91	0.46	-27	10	29	0.93	0.24
LPM (1)						11	11	15	0.91	0.76
LPM (2)						13	11	17	0.86	0.69
	$\overline{N} = 200; \overline{T} = 20$									
MLE	-30	9	32	1.00	0.10	-37	9	38	0.99	0.03
ABC1 (1)	-4	9	10	1.06	0.93	-5	10	11	1.02	0.92
ABC1 (2)	-3	9	10	1.02	0.94	-4	10	11	0.98	0.93
SPJ1	-9	10	13	0.96	0.83	-14	10	18	0.94	0.67
LPM (1)						12	10	16	0.95	0.73
LPM (2)						15	11	18	0.91	0.64
	$\overline{N} = 200; \overline{T} = 25$									
MLE	-24	10	26	0.94	0.28	-30	10	32	0.94	0.12
ABC1 (1)	-4	9	10	0.99	0.93	-4	10	11	0.97	0.91
ABC1 (2)	-2	10	10	0.96	0.94	-2	10	11	0.94	0.92
SPJ1	-3	10	11	0.90	0.91	-6	11	13	0.89	0.87
LPM (1)						14	11	17	0.88	0.67
LPM (2)						16	11	20	0.86	0.58

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; values in parentheses after ABC1 and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 4.9: *Properties: Unbalanced 2 - Exogenous Regressor*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	$\overline{N} = 200; \overline{T} = 15$									
MLE	13	4	14	0.92	0.05	2	4	4	1.01	0.91
ABC1 (1)	1	3	4	1.01	0.95	-0	4	4	1.01	0.95
ABC1 (2)	1	3	4	1.01	0.95	-0	4	4	1.01	0.95
SPJ1	6	4	7	0.89	0.62	3	4	5	0.96	0.86
LPM (1)						-0	4	4	0.79	0.88
LPM (2)						-0	4	4	0.79	0.88
	$\overline{N} = 200; \overline{T} = 20$									
MLE	10	4	11	0.95	0.19	2	4	4	0.98	0.90
ABC1 (1)	1	3	3	1.01	0.95	0	4	4	0.98	0.95
ABC1 (2)	1	3	3	1.01	0.95	0	4	4	0.98	0.95
SPJ1	2	4	4	0.93	0.89	2	4	4	0.94	0.90
LPM (1)						-0	4	4	0.75	0.86
LPM (2)						-0	4	4	0.75	0.85
	$\overline{N} = 200; \overline{T} = 25$									
MLE	8	4	9	0.96	0.32	2	4	4	0.95	0.91
ABC1 (1)	1	3	3	1.02	0.95	0	4	4	0.95	0.94
ABC1 (2)	0	3	3	1.01	0.95	0	4	4	0.95	0.94
SPJ1	0	4	4	0.95	0.94	1	4	4	0.91	0.92
LPM (1)						-0	4	4	0.74	0.85
LPM (2)						-0	4	4	0.74	0.85

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; values in parentheses after ABC1 and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

of SPJ1 to some extend substantially, especially for smaller values of $\overline{T}$. *Pattern 1* stands out in particular, because it clearly shows that the reduction of distortion decreases and the dispersion increases. An intuitive explanation is that the splitting strategy leads to subpanels of widely differing sizes. This issue is not that severe in *pattern 2*, but the performance is still worse than in the balanced case.

Table 4.10: *Sizes of Different Wald Tests*

	MLE	ABC1		SPJ1
		$L=1$	$L=2$	
		Balanced		
$N=200; T=10$	0.99	0.05	0.07	0.39
$N=200; T=15$	0.97	0.05	0.05	0.13
$N=200; T=20$	0.90	0.06	0.06	0.11
$N=200; T=25$	0.80	0.04	0.04	0.08
$N=200; T=30$	0.74	0.05	0.05	0.08
		Unbalanced 1		
$\overline{N}=200; \overline{T}=15$	1.00	0.08	0.08	0.92
$\overline{N}=200; \overline{T}=20$	0.97	0.06	0.05	0.34
$\overline{N}=200; \overline{T}=25$	0.85	0.05	0.05	0.10
		Unbalanced 2		
$\overline{N}=200; \overline{T}=15$	1.00	0.06	0.07	0.63
$\overline{N}=200; \overline{T}=20$	0.97	0.06	0.06	0.17
$\overline{N}=200; \overline{T}=25$	0.86	0.06	0.06	0.08

Note: All entries refer to sizes of different Wald tests: $H_0: \rho = 0.5 \wedge \beta = 1$; nominal size 5%; results based on 1,000 repetitions.

Table 4.10 reports different sizes of Wald tests. Overall the results are in line with the insights we have gained so far. Whereas sizes of tests based on MLE are heavily distorted, using bias-corrected estimators to construct test-statistics brings them closer to their nominal level, irrespective of the missing data pattern. But as in our previous analysis of the different missing data patterns, we find that the performance of SPJ1 worsens by randomly missing observations while MLE and ABC1 remain unaffected. This is especially apparent when we look at $T = \overline{T} = 15$ where the sizes based on SPJ1 range between 0.13 and 0.92 whereas those of MLE and ABC1 are almost identical. Again the distortion is less severe in *pattern 2*. Overall ABC1 strictly dominates MLE and SPJ1 as its sizes are always very close to their nominal level.

Finally, we conclude that the various analytically bias-corrected and the different split-panel jackknife estimators work similarly well with each other. Further, we find that analytical bias corrections are clearly preferable to split-panel jackknife approaches. Although the latter have the advantage that they are relatively easy to implement, this convenience is associated with considerable performance losses. More precisely, split-panel jackknife estimators have higher distortion and react sensitive to different missing data patterns. Lastly, we suggest a cautious use

of linear probability models, because its inference can be misleading as in our considered designs.

In the next section, we apply MLE, ABC1, SPJ1, and LPM to an empirical example of labor economics where we investigate the inter-temporal labor force participation of 10,712 German women between 1984 and 2013.

4.5 Empirical Illustration

In the following, we illustrate one possible area of application by analyzing the inter-temporal labor-force participation of women using longitudinal micro data from the German Socio Economic Panel (*GSOEP*). More precisely, we want to examine how fertility decisions and the availability of non-labor income jointly affect women's participation decisions in the labor market.

For a long time labor economists are concerned with fertility decisions being endogenous due to correlation with multiple unobserved variables. Most studies use cross-sectional data along with an instrumental variable strategy to deal with this problem (see among others Angrist and Evans 1998). However, the availability of comprehensive panel data sets offers new reliefs to researchers. For instance, Heckman and MaCurdy (1980, 1982), Hyslop (1999), and Carro (2007) use panel data from the *Panel Study of Income Dynamics* (*PSID*) which allows them to tackle this omitted variables problem by controlling for individual specific unobserved effects.

For our illustration, we use an empirical strategy adopted from Hyslop (1999) and estimate the following dynamic binary choice model:

$$y_{it} = \mathbf{1}\left[\rho y_{it-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\boldsymbol{\pi} + \alpha_i + \gamma_t + e_{it} \geq 0\right]$$

where $i = 1,\ldots,N$ and $t = s_i,\ldots,T_i$ are individual and time specific identifiers, y_{it} is an indicator equal to one if woman i is in labor-force at time period t, $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are vectors of explanatory and further control variables, γ, $\boldsymbol{\beta}$, and $\boldsymbol{\pi}$ are the corresponding parameters, and e_{it} is an idiosyncratic error term assumed to be independently and identically distributed standard normal. More precisely, we consider the following explanatory variables: number of children in different age groups, non-labor income, and an indicator that is equal to one if a birth occurs in the next time period. Further controls are squared age, martial status, regional identifier, number of children between zero and one in the previous period, and number of other household members. Additionally, we include individual and time specific intercepts to control for unobserved heterogeneity. For instance, α_i captures individual specific taste for labor and permanent income, whereas γ_t controls for the business cycle and other time specific shifts in preferences.

For our analysis, we extract an unbalanced panel data set of 10,712 women between 1984 and 2013 from the *GSOEP*.[11] Because we want to estimate a dynamic model of labor supply, we restrict the sample to women between 16 and 65 that are observed consecutively for at least five 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. Further, a proxy for transitory non-labor income is constructed from post-government household income minus woman's individual labor earnings. Note that all income variables are converted to constant 2010 EURO using a consumer price index and that labor earnings are reported before taxes. Thus we additionally correct 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). More precisely, we divide the transitory non-labor income by the square root of household members. To analyze whether the effect of transitory non-labor income on participation decisions differs across groups, we define the following 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.[12] The class distinction is taken from the *Armuts- und Reichtumsbericht* of the federal government.[13] Further, we follow Grabka (2014) and construct regional identifiers. Therefor the federal states are grouped in four geographic regions (north, south, west, and east) which allows us to control for regional differences in preferences for labor.[14]

The descriptive statistics of our data set are reported in table 4.11. The average participation rate is 72 % in the full sample and 65 % for women who change their labor-force participation decision at least once. We refer to the latter group as movers. Further, 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 prefers to live in the west instead of the east. Contrary, women who always participate have less children and live in smaller households. Note that identification in binary choice models with two-way error component is solely based on the group of movers, which consist of 5,346 women observed for roughly 13 time periods on average. Because our model specification requires to estimate roughly 5,400 fixed

11. More precisely, we use the $PEQUV-File from 1984–2013 (version 30 of the *GSOEP*).
12. https://www.destatis.de
13. https://www.armuts-und-reichtumsbericht.de/
14. North: Schleswig-Holstein, Hamburg, Lower-Saxony, Bremen; South: Hessen, Baden-Wuerttemberg, Bavaria; West: North-Rhine-Westfalia, Rheinland-Pfalz, Saarland; East: Berlin, Brandenburg, Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt, Thueringia.

Table 4.11: *Descriptive Statistics*

	Full		Always		Never		Movers	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Participation	0.72	0.45	1.00	0.00	0.00	0.00	0.65	0.48
Age	40.10	11.76	42.47	10.38	46.83	12.73	37.40	11.66
Married	0.66	0.47	0.64	0.48	0.85	0.36	0.65	0.48
Middle Class	0.44	0.50	0.42	0.49	0.45	0.50	0.45	0.50
Upper Class	0.01	0.09	0.01	0.09	0.01	0.11	0.01	0.08
North	0.13	0.34	0.13	0.33	0.16	0.37	0.13	0.34
East	0.22	0.41	0.27	0.44	0.08	0.27	0.21	0.41
South	0.36	0.48	0.35	0.48	0.35	0.48	0.37	0.48
#Children 0-1	0.04	0.21	0.02	0.13	0.05	0.22	0.06	0.25
#Children 2-4	0.12	0.35	0.05	0.23	0.13	0.39	0.16	0.41
#Children 5-18	0.68	0.94	0.53	0.81	0.74	1.10	0.76	0.98
#HH older	2.27	0.86	2.18	0.81	2.55	0.97	2.28	0.87
Birth_{t+1}	0.03	0.18	0.01	0.12	0.03	0.17	0.04	0.21
#Observations	127,736		46,398		11,644		69,694	
#Individuals (N)	10,712		4,220		1,146		5,346	
Avg. #Individuals ($\overline{N}$)	4,562		1,657		416		2,489	
Avg. #Years ($\overline{T}$)	12		11		10		13	

effects, it is a suitable candidate for the application of our algorithms.

Table 4.12 reports estimates of the structural parameters and average partial effects obtained by different linear and probit fixed effects estimators. The labels are identical to the ones used in section 4.4. All results are intuitive and in line with the theoretical model of Hyslop (1999). We find strong positive state-dependence and negative effects with respect to transitory non-labor income, number of children, and expectations about future fertility. Remarkably, the estimated average partial effects obtained from dynamic probit models are all very close to each other. An exception is the state dependence which ranges from roughly 0.20 up to 0.29. All effects are significant at the 5 % level, except being in the upper income class. Estimates obtained by the bias-corrected linear probability models are also very close to their nonlinear counterparts. Two exceptions are the average partial effects with respect to the lagged dependent variable and number of children between zero and one. However the standard errors obtained by the linear probability models are unreasonable low.

Our final conclusions are based on the results obtained by the different fixed effects probit estimators. First, we detect strong persistence in womens' participation decisions. A woman who has currently a job increases her probability to participate in the future by 20-29 percentage points. Second, we find that women only respond weakly to changes in transitory non-labor income. More precisely, being in the middle class reduces the participation probability by roughly two percentage points compared to a woman in the lower income class. The reduction associated with

Table 4.12: *Empirical Results: Labor-Force Participation Decision*

	MLE	ABC1		SPJ1	LPM	
		L = 1	L = 2		L = 1	L = 2
	Coefficients					
Participation$_{t-1}$	1.315	1.476	1.546	1.577	-	-
	(0.015)	(0.015)	(0.015)	(0.015)	-	-
Middle Class	-0.122	-0.103	-0.111	-0.093	-	-
	(0.020)	(0.020)	(0.021)	(0.021)	-	-
Upper Class	-0.368	-0.318	-0.311	-0.363	-	-
	(0.111)	(0.112)	(0.113)	(0.112)	-	-
#Children 0-1	-1.839	-1.606	-1.591	-1.684	-	-
	(0.033)	(0.032)	(0.032)	(0.032)	-	-
#Children 2-4	-0.480	-0.351	-0.339	-0.425	-	-
	(0.023)	(0.023)	(0.023)	(0.024)	-	-
#Children 5-18	-0.186	-0.133	-0.127	-0.176	-	-
	(0.011)	(0.011)	(0.011)	(0.012)	-	-
Birth$_{t+1}$	-0.564	-0.518	-0.509	-0.573	-	-
	(0.034)	(0.034)	(0.034)	(0.034)	-	-
	Average Partial Effects					
Participation$_{t-1}$	0.198	0.277	0.292	0.267	0.492	0.521
	(0.038)	(0.043)	(0.045)	(0.046)	(0.003)	(0.003)
Middle Class	-0.014	-0.014	-0.015	-0.011	-0.014	-0.016
	(0.005)	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)
Upper Class	-0.042	-0.044	-0.044	-0.048	-0.037	-0.040
	(0.027)	(0.026)	(0.026)	(0.026)	(0.013)	(0.013)
#Children 0-1	-0.203	-0.214	-0.216	-0.204	-0.305	-0.303
	(0.037)	(0.033)	(0.032)	(0.033)	(0.004)	(0.004)
#Children 2-4	-0.053	-0.047	-0.046	-0.052	-0.047	-0.040
	(0.011)	(0.009)	(0.008)	(0.010)	(0.003)	(0.003)
#Children 5-18	-0.021	-0.018	-0.017	-0.022	-0.016	-0.010
	(0.005)	(0.004)	(0.004)	(0.004)	(0.001)	(0.001)
Birth$_{t+1}$	-0.066	-0.074	-0.073	-0.074	-0.085	-0.085
	(0.014)	(0.013)	(0.013)	(0.014)	(0.005)	(0.005)

Note: Standard errors in parentheses; additional covariates: squared age, married, regional identifiers, number of children between zero and one in the previous period, and number of household members above 18; estimates relative to lower income class.

belonging to the upper income class is stronger (five percentage points), but not significantly different from zero at any usual level. Finally, the number of children reduces the likelihood of current participation decision significantly. As expected, the effect is negative and declining in age of children. Each additional child between zero and one reduces current participation probability by roughly 20 percentage points. For children older than four, the reduction is only one percentage point. The results presented in this illustration are largely consistent with the empirical findings of Hyslop (1999). However, contrary to him, we find that future birth always negatively affects current participation decision across different models. This might confirm the author's perfect foresight assumption with respect to life-cycle fertility decisions.

4.6 Conclusion

In this article, we offered new reliefs and guidance for empirical researchers who would be otherwise deterred from using binary choice models with fixed effects. First, we showed how to overcome computational obstacles that arise both in estimating these models themselves and in applying appropriate bias corrections to mitigate the incidental parameters problem. Beyond that, we have carried out extensive simulation experiments to gain further insights into the statistical properties of various bias corrections. Analytical bias corrections performed particularly well, even in unbalanced panel data. An empirical illustration from labor economics gave a first impression about the applicability of bias corrections in longitudinal data sets. To encourage the usage of bias-corrected binary choice models, we embedded the analytical bias correction of Fernández-Val and Weidner (2016) in our *R*-package *alpaca*.

Although we focused on binary choice models, remember that Fernández-Val and Weidner (2016) derived bias corrections for any nonlinear model with a two-way error component. It is straightforward to apply the same acceleration techniques described in this article to other generalized linear models, such as poisson models. Further, note that the bias corrections proposed by Fernández-Val and Weidner (2016) are not limited to classical panel structures. For instance, Cruz-Gonzalez, Fernández-Val, and Weidner (2017) applied some of bias corrections to cross-sectional data of bilateral trade flows such as those used by Helpman, Melitz, and Rubinstein (2008).

Other related research projects, which dealt with bias corrections in the presence of multiple high-dimensional fixed effects, are Weidner and Zylkin (2018) and Hinz, Stammann, and Wanner (2019). Both adapted and extended the bias corrections of Fernández-Val and Weidner (2016) to special two- and three-way error components which are particularly relevant in the context of international trade. Whereas the

former dealt with pseudo poisson models, the latter have treated binary choice.

References

Andersen, Erling Bernhard. 1970. “Asymptotic properties of conditional maximum-likelihood estimators.” *Journal of the Royal Statistical Society. Series B:* 283–301.

Angrist, Joshua D., and William N. Evans. 1998. “Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size.” *The American Economic Review* 88 (3): 450–477.

Arellano, Manuel, and Jinyong Hahn. 2007. “Understanding bias in nonlinear panel models: Some recent developments.” *Econometric Society Monographs* 43:381.

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

Buhmann, Brigitte, Lee Rainwater, Günther Schmaus, and Timothy M. Smeeding. 1988. “EQUIVALENCE SCALES, WELL-BEING, INEQUALITY, AND POVERTY: SENSITIVITY ESTIMATES ACROSS TEN COUNTRIES USING THE LUXEMBOURG INCOME STUDY (LIS) DATABASE.” *Review of Income and Wealth* 34 (2): 115–142.

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.” *Review of Economic Studies* 47:225–238.

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.

Correia, Sergio. 2016. “A feasible estimator for linear models with multi-way fixed effects.” *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.” *Review of Economic Studies* 82 (3): 991–1030.

Enea, Marco. 2017. “speedglm: Fitting Linear and Generalized Linear Models to Large Data Sets.” *R Software Package (version 0.3-2).*

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

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.

———. 2018b. “Individual and time effects in nonlinear panel models with large N, T.” *arXiv preprint:1311.7065*.

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

———. 2013b. “OLS with multiple high dimensional category variables.” *Computational Statistics & Data Analysis* 66:8–18.

Grabka, Markus. 2014. “SOEP 2013 – Codebook for the $PEQUIV File 1984-2013: CNEF Variables with Extended Income Information for the SOEP.” *SOEP Survey Papers 204: Series D. Berlin: DIW/SOEP*.

Greene, William. 2004. “The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects.” *Econometrics Journal* 7:98–119.

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*.

———. 2011. “Bias reduction for dynamic nonlinear panel models with fixed effects.” *Econometric Theory* 27 (06): 1152–1191.

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 (1-2): 96–99.

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

Heckman, James J., and Thomas E. MaCurdy. 1980. “A Life Cycle Model of Female Labour Supply.” *The Review of Economic Studies* 47 (1): 47–74.

———. 1982. “Corrigendum on A Life Cycle Model of Female Labour Supply.” *The Review of Economic Studies* 49 (4): 659–660.

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. 2019. “Persistent Zeros: The Extensive Margin of Trade.” *Working Paper*.

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.

Hyslop, Dean R. 1999. “State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women.” *Econometrica* 67 (6): 1255–1294.

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

McCullagh, Peter, and James A. Nelder. 1989. *Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability*.

Nelder, John Ashworth, and Robert William Maclagan Wedderburn. 1972. “Generalized linear models.” *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–384.

Neumann, John von. 1950. “Functional Operators. Vol. II. The geometry of orthogonal spaces, volume 22 (reprint of 1933 notes) of Annals of Math.” *Studies. Princeton University Press.*

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 Roger Moon. 1999. "Linear regression limit theory for nonstationary panel data." *Econometrica* 67 (5): 1057–1111.

R Core Team. 2019. *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:1707.01815v3*.

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." *SOEPpaper*.

Weidner, Martin, and Thomas Zylkin. 2018. "Bias and Consistency in Three-way Gravity Models." *Working Paper*.

Appendix

A Further Simulation Experiments

To demonstrate that the analytical bias corrections for dynamic linear models work as intended, we adjust the data generating process used in section 4.4 to linear models. More precisely, we change the data generating process to

$$y_{it} = \rho y_{it-1} + \beta x_{it} + \alpha_i + \gamma_t + \epsilon_{it},$$
$$y_{i0} = \beta x_{i0} + \alpha_i + \gamma_0 + \epsilon_{i0},$$

and keep everything else unchanged.

Table 4.13: *Properties: Balanced - Dynamic Linear Model*

	Coefficients ($\hat{\rho}$)					Coefficients ($\hat{\beta}$)				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 200; T = 10									
LM	-17	3	18	0.99	0.00	3	3	4	1.00	0.84
BC (1)	-8	3	8	1.03	0.34	1	3	3	1.00	0.93
BC (2)	-4	3	6	0.99	0.72	0	3	3	0.99	0.95
	N = 200; T = 15									
LM	-11	3	11	0.98	0.01	3	3	4	0.95	0.80
BC (1)	-5	2	6	1.01	0.44	1	3	3	0.96	0.92
BC (2)	-3	3	4	0.99	0.81	0	3	3	0.95	0.92
	N = 200; T = 20									
LM	-8	2	9	0.95	0.03	2	2	3	0.98	0.84
BC (1)	-4	2	5	0.98	0.55	1	2	2	0.98	0.93
BC (2)	-2	2	3	0.97	0.82	0	2	2	0.98	0.94
	N = 200; T = 25									
LM	-7	2	7	0.97	0.07	2	2	3	0.99	0.83
BC (1)	-3	2	4	0.99	0.60	1	2	2	0.99	0.93
BC (2)	-2	2	3	0.98	0.87	0	2	2	0.99	0.94
	N = 200; T = 30									
LM	-5	2	6	0.99	0.13	2	2	2	1.02	0.87
BC (1)	-3	2	3	1.00	0.68	1	2	2	1.02	0.94
BC (2)	-1	2	2	1.00	0.88	0	2	2	1.02	0.95

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; LM and BC denote (bias-corrected) fixed effects estimators; values in parentheses after BC indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 4.13 reports the results of the simulation experiments. As expected, the bias correction reduces the distortion considerably and brings the coverage probabilities closer to their nominal level.

Chapter 5

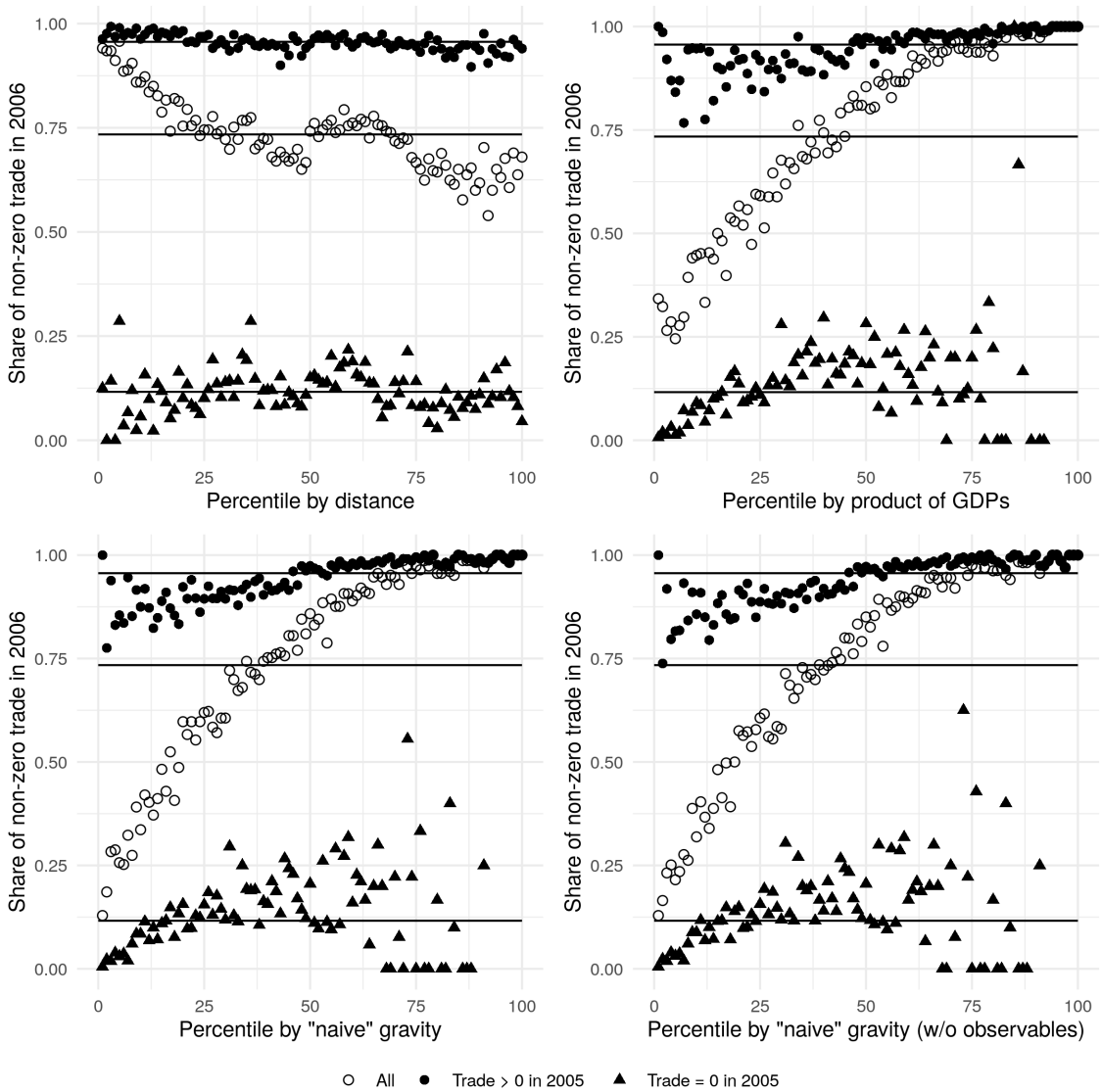
Persistent Zeros: The Extensive Margin of Trade

Co-authored with Julian Hinz and Joschka Wanner

5.1 Introduction

What induces country pairs to trade? In 2006, still more than one quarter of potential bilateral trade relations reported zero trade flows. Figure 5.1 breaks down the share of nonzero trade flows in 2006 along the percentiles of four different ad-hoc indicators of "trade potential": Bilateral distance; product of GDPs; "naive" gravity, i.e. the product of GDPs divided by their bilateral distance; and the latter when excluding country pairs in FTAs, with common currencies or common colonial history. The x-axis indicates the potential trade volume, i.e. the joint economic size and/or proximity of any two countries. All four plots paint a common picture: The circles, covering all country pairs, show a strong general relationship between trade potential and actual nonzero trade. The filled dots and triangles split the country pairs according to whether the two did or did not engage in trade in the previous year. The clearly

Figure 5.1: *Determinants of the Extensive Margin - Gravity and Persistence*



separated pattern for the two groups highlights a remarkable persistence of trade relations, even after controlling for differences in trade potential in terms of distance, size, and bilateral trade policy. More than 75 percent of those country pairs in the lowest percentile of trade potential trade again in 2006, provided they already did so in 2005. On the other hand, even comparably large and close pairs are likely not to trade in 2006 if they did not trade in 2005 either.[1,2]

In this paper we examine the determinants of the extensive margin of international trade, taking explicitly into account its persistence. We combine a heterogeneous firms model of international trade with bounded productivity with features from the firm dynamics literature to derive expressions for an exporting country's participation on a specific destination market in a given period. These expressions depend on partly unobserved (i) exporter-time, (ii) destination-time, and (iii) exporter-destination specific components, as well as on (iv) whether the exporter has already served the market in the previous period, and on (v) exporter-destination-time specific gravity-type trade cost determinants. We estimate the model making use of recent advances in the estimation of binary choice estimators with high-dimensional fixed effects to address (i)-(iii). The inclusion of fixed effects in a binary choice setting induces an incidental parameters problem that is potentially aggravated by the dynamics introduced by (iv). To mitigate this bias, we characterize and implement new analytical and jackknife bias corrections for coefficients and estimates of average partial effects in our specifications with two- and three-way fixed effects. Extensive simulation experiments demonstrate the desirable statistical properties of our proposed bias-corrected two- and three-way fixed effect logit and probit estimators. The empirical results provide evidence that both unobserved bilateral factors and true state dependence due to entry dynamics contribute strongly to the high persistence. Taking this persistence into account changes the coefficients considerably: out of the most commonly studied potential determinants (joint WTO membership, common regional trade agreement, and shared currency), only sharing a common currency has a significant effect on whether two countries trade with each other at all.

Our paper builds on recent insights from three flourishing strands of literature. First, our paper is related to the literature on the extensive margin of international trade. A number of theoretical frameworks have sought to propose mechanisms behind the decisions of firms to export, and their aggregate implications of zero or nonzero trade flows at the country pair level. Analogous to the intensive margin counterpart, these theories have established *gravity*-like determinants, such as two

1. Note that throughout the paper, "country pair" refers to a *directed* pair of countries, i.e. Germany-France and France-Germany are two distinct country pairs.

2. The years 2005–2006 are the last available in our data set. A very similar pattern emerges for other points in time. If longer time intervals are considered (e.g. 10 years), a similar picture remains, but the relationship gets considerably weaker.

countries' bilateral distance, a free trade agreement, a common currency and joint membership in the WTO. Egger and Larch (2011) and Egger et al. (2011) append an extensive margin to a Anderson and Van Wincoop (2003)-type model by assuming export participation to be determined by (homogeneous) firms weighing operating profits and bilateral fixed costs of exporting. This results in a two-part model in which given a country's participation in exporting to any given destination, trade flows follow structural gravity. Helpman, Melitz, and Rubinstein (2008) build a model of international trade with heterogeneous firms. Here the volume of trade between two countries can change either because incumbent firms expand their operations, or due to new competitors entering into a market. Eaton, Kortum, and Sotelo (2013) move away from the arguably simplifying notion of a continuum of firms and develop a model of a finite set of heterogeneous firms. Here no firm may export to a given market due to their individual efficiency draws. Our model proposed in this paper directly builds on Helpman, Melitz, and Rubinstein (2008) and extends it by features from the literature on firm dynamics. In this firm-level literature, Das, Roberts, and Tybout (2007) develop a dynamic discrete-choice model in which current export participation depends on previous exporting, and hence sunk costs, and observable characteristics of profits from exporting. Alessandria and Choi (2007) extend this line of research and develop a general equilibrium framework that takes sunk costs and "period-by-period" fixed costs into account, showing that contrary to previous partial equilibrium evidence, aggregate effects are negligible for the US. More recent works have looked at *new* exporter dynamics (see Ruhl and Willis 2017), emphasizing that sunk costs may be relatively smaller and continuation costs relatively larger than previously assumed. Bernard et al. (2017) stand somewhat in contrast to this finding, showing that first and second year growth rates may suffer from a bias due to different entry dates throughout the year. Berman, Rebeyrol, and Vicard (2019) note the important role of "demand learning" and firms' updating of their future demand and market participation. In a similar vein, Piveteau (2019) develops a model in which new firms accumulate consumers — or fail to do so — determining entry and exit. While these newer models feature rich firm-level predictions, they require tailor-made econometric models for their estimation. Our model abstracts from the specific role of *new* firms and has the advantage of yielding an econometric specification and demanding an estimator that remains general and flexible to be applied in other contexts.

Second, our paper builds on advances in the literature on the gravity equation and the *intensive* margin of international trade. With the advent of what has now been coined *structural* gravity (see Head and Mayer 2014) the gravity framework has gained rich microfoundations. Anderson and Van Wincoop (2003) and Eaton and Kortum (2002) each formulate an underlying structure for exporting and importing

countries that in estimations can easily be captured by appropriate two-way country(-time) fixed effects, as first noted by Feenstra (2015) and Redding and Venables (2004). Although not theoretically motivated, since Baier and Bergstrand (2007) it has furthermore become standard to include country pair fixed effects to tackle unobservable bilateral characteristics. Estimating the model introduced in this paper similarly calls for *at least* two sets of fixed effects, specific to exporters and importers in a given year. Additionally, and following Baier and Bergstrand (2007), there is no reason to believe that bilateral unobservables should not be a problem in the context of the extensive margin. Our preferred estimation of the model thus includes the "full set" of fixed effects that has become standard in the estimation of gravity models of the intensive margin of trade: exporter-year, importer-year and bilateral fixed effects that leave only bilateral-time-specific variation for the estimation of parameters of interest.

Third, the paper builds on and contributes to the literature on the econometrics of generalized linear models (GLMs) with fixed effects. Recent advances in this literature have made it possible to go beyond ordinary linear models in the context of high-dimensional fixed effects by providing fast and feasible algorithms (see Guimarães and Portugal 2010; Stammann 2018; Hinz, Hudlet, and Wanner 2019).[3] As known since Neyman and Scott (1948) the inclusion of fixed effects potentially introduces an incidental parameters problem, leading to biased estimates. In the last few years, there have been a number of advances to correct this bias, and a variety of approaches have been proposed (see Fernández-Val and Weidner 2018a for a recent overview). Fernández-Val and Weidner (2016) develop analytical and jackknife bias corrections for nonlinear maximum likelihood estimators in static and dynamic models with individual and time effects for structural parameters and average partial effects. In Fernández-Val and Weidner (2018a) they generalize their previous findings and show that the order of the bias induced by fixed effects in a wide family of models translates into a simple heuristic p/n, with n being the sample size and p the number of estimated parameters. Recently, Czarnowske and Stammann (2019) show how analytical bias corrections can be efficiently implemented in a high-dimensional fixed effects setting using the methods described by Stammann (2018). Our paper is complementary to computational and econometric contributions on the estimation of the intensive margin of trade. Larch et al. (2019) present a feasible procedure to estimate pseudo-poisson (PPML) models with three high-dimensional

3. Stammann, Heiß, and McFadden (2016) have shown in the context of binary choice models with individual fixed effects that a weighted version of the Frisch-Waugh-Lovell theorem (Frisch and Waugh 1933; Lovell 1963) can be incorporated in a standard Newton-Raphson optimization procedure. This result paved the way to derive a computationally efficient algorithm for all GLMs with high-dimensional multi-way fixed effects (see Stammann 2018). More recently, Hinz, Hudlet, and Wanner (2019) offer a different way to partial out fixed effects using a modification of the Gauss-Seidel algorithm proposed by Guimarães and Portugal (2010).

fixed effects. Correia, Guimarães, and Zylkin (2019) generalize this estimation procedure to arbitrary sets of fixed effects. Weidner and Zylkin (2018) investigate the incidental parameters problem in three-way fixed effects PPML models under fixed T asymptotics and suggest an appropriate jackknife bias correction. We contribute to this literature by characterizing and implementing analytical and jackknife bias corrections for our specific two- and three-way fixed effects in the context of binary choice models. This helps us mitigate the bias induced by estimating our theory-consistent model, requiring exporter-time (it), importer-time (jt), and in our preferred specification bilateral fixed effects (ij).[4]

The remainder of the paper is structured as follows. In section 5.2 we build a dynamic model of the extensive margin of international trade. The model yields aggregate predictions that can be structurally estimated using a probit model with high-dimensional fixed effects. In section 5.3 we describe the estimator and bias correction procedure. We show its performance in Monte Carlo simulations in section 5.4, before finally estimating the theoretical model in section 5.5. Section 5.6 concludes.

5.2 An Empirical Model of the Extensive Margin of Trade

As a theoretical foundation for our econometric specification, we consider a stylized dynamic Melitz (2003)-type heterogeneous firms model of international trade. Following Helpman, Melitz, and Rubinstein (2008, henceforth HMR) we assume a bounded productivity distribution, like a truncated Pareto in HMR's case. We deviate from HMR by explicitly stating a time dimension and, unlike in the standard Melitz setting, separate fixed exporting costs into costs of entering a new market and costs of selling in a given market (as in Alessandria and Choi 2007; Das, Roberts, and Tybout 2007).

There are N countries, indexed by i and j, each of which consumes and produces a continuum of products. The representative consumer in j receives utility according to a CES utility function:

$$u_{jt} = \left(\int_{\omega \in \Omega_{jt}} (\xi_{ijt})^{\frac{1}{\sigma}} q_{jt}(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right)^{\frac{\sigma}{\sigma-1}} \quad \text{with} \quad \sigma > 1.$$

where $q_{jt}(\omega)$ is j's consumption of product ω in period t, Ω_{jt} is the set of products available in j, σ is the elasticity of substitution across products, and ξ_{ijt} is a log-

4. An R implementation of the estimators developed in this paper will be provided on CRAN and is currently available from the authors upon request.

normally distributed idiosyncratic demand shock (with $\mu_\xi = 0$ and $\sigma_\xi = 1$) for goods from country i in country j and period t (similar to Eaton, Kortum, and Kramarz 2011). Demand in country j for good ω depends on this demand shock, j's overall expenditure E_{jt}, and the good price $p_{jt}(\omega)$ relative to the overall price level as captured by the price index P_{jt}:

$$q_{jt}(\omega) = \frac{p_{jt}(\omega)^{-\sigma}}{P_{jt}^{1-\sigma}} \xi_{ijt} E_{jt}.$$

$$\text{with} \qquad P_{jt} = \left(\int_{\omega \in \Omega_{jt}} \xi_{ijt} p_{jt}(\omega)^{1-\sigma} d\omega \right)^{\frac{1}{1-\sigma}}.$$

Each country has a fixed continuum of potentially active firms that have different productivities drawn from the distribution $G_{it}(\varphi)$, where $\varphi \in (0, \varphi^*_{it}]$. The productivity distribution evolves over time and firms' ranks within the productivity distribution can also change from period to period, though firms that in the last period did not export to a market already served by a domestic competitor are assumed not to directly jump to being the country's most productive firm in the next period.[5] Each period, a firm can decide to pay a fixed cost f_{it}^{prod} and start production of a differentiated variety using labour l as its only input, such that $l_t(\omega) = f_{it}^{prod} + q_t(\omega)/\varphi_t(\omega)$. A firm's marginal cost of providing one unit of its good to market j consists of iceberg trade costs τ_{ijt} and labour costs $w_{it}/\varphi_t(\omega)$. Firms compete with each other in monopolistic competition and charge a constant markup over marginal costs. Therefore, the price of a good ω produced in i and sold in j is

$$p_{ijt}(\omega) = \frac{\sigma}{\sigma - 1} \frac{\tau_{ijt} w_{it}}{\varphi_t(\omega)}.$$

A firm's *operating* profits in market j are hence given by

$$\tilde{\pi}_{ijt}(\omega) = \frac{1}{\sigma} \left(\frac{\sigma}{\sigma - 1} \frac{\tau_{ijt} w_{it}}{\varphi_t(\omega)} \right)^{1-\sigma} P_{jt}^{\sigma-1} \xi_{ijt} E_{jt}.$$

If a firm wants to export to a market j in period t, it has to pay a fixed exporting cost f_{ijt}^{exp}. The exporting fixed cost is higher by a market entry cost factor $f^{entry} \geq 1$ if the firm has not been active in the respective market in the previous period. For tractability, the entry cost factor is assumed to be constant across countries and time. Capturing the export decision by a binary variable $y_{ijt}(\omega)$, i.e. equal to one if the firm decides to serve market j in period t, we can formalize a firm's *realized* profits

5. Note that we could in principle also allow for new firm entry into the pool of potential producers without changing our final expression for the extensive margin as long as the new entrants cannot be the country's most productive firm right away.

in market j as follows:

$$\pi_{ijt}(\omega) = y_{ijt}(\omega)\left\{\tilde{\pi}_{ijt}(\omega) - f_{ijt}^{exp}(f^{entry})^{[1-y_{ij(t-1)}(\omega)]}\right\}.$$

In the absence of entry costs, a firm would simply compare its operating profits to the fixed exporting cost and decide to serve a market if the former are greater than the latter. With market entry costs, a firm might be willing to incur a loss in the current period if expected future profits from that same market outweigh the initial loss. Firms discount future profits at a rate δ per period. To keep things tractable and allow us to derive a theory-consistent estimation expression below, we assume that firms expect their future operating profits from and fixed costs of serving a given market to be equal to today's values, i.e. $\mathbb{E}_t[\tilde{\pi}_{ij(t+s)}] = \tilde{\pi}_{ijt}$ and $\mathbb{E}_t[f_{ij(t+s)}^{exp}] = f_{ijt}^{exp}$ $\forall s \in \mathbb{N}$.[6] The current value of today's and all future operating profits from market j is then given by $\sum_{s=0}^{\infty}(1-\delta)^s\tilde{\pi}_{ijt} = \tilde{\pi}_{ijt}/\delta$. A firm will decide to serve a destination market if these discounted expected profits exceed the sum of today's and discounted future fixed costs of entry and exporting, given by

$$f_{ijt}^{exp}(f^{entry})^{(1-y_{ij(t-1)}(\omega))} + \sum_{s=1}^{\infty}(1-\delta)^s f_{ijt}^{exp} = \frac{f_{ijt}^{exp}}{\delta}\left(1+\delta(f^{entry}-1)\right)^{(1-y_{ij(t-1)}(\omega))}.$$

Given this model setup, the question whether a country exports to another country *at all* can be considered by looking at the most productive firm (with φ_t^*) only. Denoting that firm's product by ω^*, we can capture the aggregate extensive margin by the binary variable y_{ijt} as follows:

$$y_{ijt} = y_{ijt}(\omega^*) = \begin{cases} 1 & \text{if} \quad \dfrac{\left(\frac{1}{\sigma}\left(\frac{\sigma}{\sigma-1}\frac{\tau_{ijt}w_{it}}{\varphi_{it}^*}\right)^{1-\sigma}P_{jt}^{\sigma-1}\xi_{ijt}E_{jt}\right)}{f_{ijt}^{exp}(1+\delta(f^{entry}-1))^{(1-y_{ij(t-1)})}} \geq 1, \\ 0 & \text{else.} \end{cases} \tag{5.1}$$

Country i is hence more likely to export to country j in period t if (i) bilateral variable trade costs are lower; (ii) wages in i, and hence production costs, are lower; (iii) the productivity of the most productive firm is higher, again reducing production costs; (iv) competitive pressure, inversely captured by the price index, in j is lower, corresponding to the idea of inward multilateral resistance coined by Anderson and Van Wincoop (2003) in the intensive margin context; (v) the market in j is larger; (vi) bilateral fixed costs of exporting are smaller; or (vii) i's most productive firm already served market j in the previous period and therefore does not have do pay the market entry cost. Note that (i) to (iv) all act via higher operating profits and

6. Note that our final expression for the extensive margin also holds if firms instead expect their operating profits from serving an export market to grow at a constant rate $\bar{g} < \delta$.

depend on the elasticity of substitution between goods. The higher this elasticity, the stronger the reaction of profits to changes in any of these factors. At the same time, a higher elasticity reduces the mark-up firms can charge and hence makes it generally harder to earn enough profits to mitigate the fixed costs of exporting. Further note that the importance of the entry costs depends on the discount factor. Intuitively, if agents are more patient, the one-time entry costs matter less compared to the repeatedly earned profits.

In order to turn equation (5.1) into the empirical expression that we will bring to the data, we take the natural logarithm and group all exporter-time and importer-time specific components and capture them with corresponding sets of fixed effects. Further, we need to specify the fixed and variable trade costs. In keeping with the existing literature, we model them as a linear combination of different observable bilateral variables, such as geographical distance, whether i and j are both WTO members, or whether i and j share a common currency. In our most general specification, we additionally include country pair fixed effects. Following Baier and Bergstrand (2007), this is common practice in the estimation of the determinants of the intensive margin of trade in order to avoid endogeneity due to unobserved heterogeneity. Further, these bilateral fixed effects may capture (part of) the strong persistence documented above.[7] We then end up with the following econometric model:

$$y_{ijt} = \begin{cases} 1 & \text{if} \quad \kappa + \lambda_{it} + \psi_{jt} + \beta_y y_{ij(t-1)} + \mathbf{x}'_{ijt}\boldsymbol{\beta}_x + \mu_{ij} \geq \zeta_{ijt}, \\ 0 & \text{else}, \end{cases} \tag{5.2}$$

where $\kappa = -\sigma\log(\sigma) - (1-\sigma)\log(\sigma-1) - \log(1+\delta(f^{entry}-1))$, $\lambda_{it} = (1-\sigma)(\log(w_{it}) - \log(\varphi^*_{it}))$, $\psi_{jt} = (\sigma-1)\log(P_{jt}) + \log(E_{jt})$, $\beta_y = \log(1+\delta(f^{entry}-1))$, $\mathbf{x}'_{ijt}\boldsymbol{\beta}_x + \mu_{ij} = (1-\sigma)\log(\tau_{ijt}) - \log(f^{exp}_{ijt})$, and $\zeta_{ijt} = -\log(\xi_{ijt}) \sim \mathcal{N}(0,1)$. The error term distribution implies that a probit estimator is the appropriate choice to estimate our model. Alternatively, we could deviate from Eaton, Kortum, and Kramarz (2011) and assume a log-logistic distribution for the idiosyncratic demand shocks, which would lead to a logit specification.

Our theoretical framework implies a flexible empirical specification that can reconcile the extensive margin estimation with the stylized fact presented in section 5.1. Note that we chose to make a number of simplifying assumptions in order to achieve the clear theory-consistent interpretation of specification (5.2). An alternative interpretation of equation (5.2) as a reduced-from representation of a more elaborate and realistic model (similar e.g. to how Roberts and Tybout 1997 motivate

7. If the trade costs further include any exporter(-time) or importer(-time) specific components, these are captured by the aforementioned corresponding sets of fixed effects.

their empirical consideration) is equally justifiable. At the same time, while our model is written along the lines of Helpman, Melitz, and Rubinstein (2008), which remains the benchmark for the empirical assessment of the (aggregate) extensive margin of trade, it is not decisive for our empirical specification that zero trade flows result from a truncated productivity distribution instead of a discrete number of firms (as in Eaton, Kortum, and Sotelo 2013) or from fixed exporting costs in a Krugman (1980)-type homogeneous firms setting (as in Egger and Larch 2011; Egger et al. 2011).

5.3 Binary Response Estimators with High-Dimensional Fixed Effects

Having set up the empirical framework, we now turn to the estimation procedure. As equation (5.2) demands two- or three-way fixed effects to capture unobservable characteristics, we describe how to implement suitable binary choice estimators. In a first step, we review a recent procedure for estimating probit and logit models with high-dimensional fixed effects proposed by Stammann (2018).[8] In a second step, we characterize appropriate bias correction techniques to address the induced incidental parameters problem.

5.3.1 Feasible Estimation

In this subsection, we sketch how to estimate structural parameters, average partial effects (APEs), and the corresponding standard errors in a binary response setting in the presence of high-dimensional fixed effects. Let $\mathbf{Z} = [\mathbf{D}, \mathbf{X}]$, where $\mathbf{D}$ is the dummy matrix corresponding to the fixed effects and $\mathbf{X}$ is a matrix of further regressors. Note that $\mathbf{X}$ may also include predetermined variables. Further, let $\boldsymbol{\alpha}$ denote the vector of fixed effects, $\boldsymbol{\beta}$ the vector of structural parameters, and $\boldsymbol{\theta} = [\boldsymbol{\alpha}', \boldsymbol{\beta}']'$.

The log-likelihood contribution of the ijt-th observation is

$$\ell_{ijt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_{ijt}) = y_{ijt} \log(F_{ijt}) + (1 - y_{ijt}) \log(1 - F_{ijt}),$$

where $\boldsymbol{\alpha}_{ijt} = [\lambda_{it}, \psi_{jt}]'$ in the case of two-way fixed effects and $\boldsymbol{\alpha}_{ijt} = [\lambda_{it}, \psi_{jt}, \mu_{ij}]'$ in the case of three-way fixed effects.[9] Further, F_{ijt} is either the logistic or the standard normal cumulative distribution function. See table 5.1 for the relevant expressions and derivatives.

8. We review the estimation procedure to reconcile the notation with the one used in Fernández-Val and Weidner (2016) and this article.

9. Note that we use for brevity notation for balanced data.

Table 5.1: *Expressions and Derivatives for Logit and Probit Models*

	Logit	Probit
F_{ijt}	$(1+\exp(-\eta_{ijt}))^{-1}$	$\Phi(\eta_{ijt})$
$\partial_\eta F_{ijt}$	$F_{ijt}(1-F_{ijt})$	$\phi(\eta_{ijt})$
$\partial_{\eta^2} F_{ijt}$	$\partial_\eta F_{ijt}(1-2F_{ijt})$	$-\eta_{ijt}\phi(\eta_{ijt})$
ν_{ijt}	$(y_{ijt}-F_{ijt})/\partial_\eta F_{ijt}$	$(y_{ijt}-F_{ijt})/\partial_\eta F_{ijt}$
H_{ijt}	1	$\partial_\eta F_{ijt}/(F_{ijt}(1-F_{ijt}))$
ω_{ijt}	$\partial_\eta F_{ijt}$	$H_{ijt}\partial_\eta F_{ijt}$
$\partial_\eta \ell_{ijt}$	$y_{ijt}-F_{ijt}$	$H_{ijt}(y_{ijt}-F_{ijt})$

Note: $\eta_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\beta} + \lambda_{it} + \psi_{jt}$ or $\eta_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\beta} + \lambda_{it} + \psi_{jt} + \mu_{ij}$ is the linear predictor.

The standard approach to estimate binary choice models is to maximize the following log-likelihood function:

$$\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} \ell_{ijt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_{ijt})$$

using Newton's method. The update in iteration $(r-1)$ is

$$(\boldsymbol{\theta}^r - \boldsymbol{\theta}^{r-1}) = (\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\mathbf{Z})^{-1}\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{\nu}}\,, \tag{5.3}$$

where $\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\mathbf{Z}$ and $\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{\nu}}$ denote the negative Hessian and gradient of the log-likelihood, respectively, and $\widehat{\boldsymbol{\Omega}}$ is a diagonal weighting matrix with $\text{diag}(\widehat{\boldsymbol{\Omega}}) = \hat{\boldsymbol{\omega}}$.

The brute-force computation of equation (5.3) quickly becomes computationally demanding, if not impossible.[10] Thus Stammann (2018) suggests a straightforward strategy called pseudo-demeaning that mimics the well known within transformation for linear regression models. The approach allows to update the structural parameters without having to explicitly update the incidental parameters, which leads to the following concentrated version of equation (5.3):

$$(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1}) = \left((\widehat{\mathbb{M}}\mathbf{X})'\widehat{\boldsymbol{\Omega}}(\widehat{\mathbb{M}}\mathbf{X})\right)^{-1}(\widehat{\mathbb{M}}\mathbf{X})'\widehat{\boldsymbol{\Omega}}(\widehat{\mathbb{M}}\hat{\boldsymbol{\nu}})\,, \tag{5.4}$$

where $(\widehat{\mathbb{M}}\mathbf{X})'\widehat{\boldsymbol{\Omega}}(\widehat{\mathbb{M}}\hat{\boldsymbol{\nu}})$ is the concentrated gradient, $(\widehat{\mathbb{M}}\mathbf{X})'\widehat{\boldsymbol{\Omega}}(\widehat{\mathbb{M}}\mathbf{X})$ is the concentrated negative Hessian, and $\widehat{\mathbb{M}} = \mathbf{I}_{IJT} - \widehat{\mathbb{P}} = \mathbf{I}_{IJT} - \mathbf{D}(\mathbf{D}'\widehat{\boldsymbol{\Omega}}\mathbf{D})^{-1}\mathbf{D}'\widehat{\boldsymbol{\Omega}}$ is known as the residual projection that partials out the fixed effects. After convergence of the optimization routine, the standard errors associated with the structural parameters can be computed from the inverse of the concentrated Hessian.

Since the computation of $\widehat{\mathbb{M}}$ itself is problematic even in moderately large data

10. In a balanced data set ($I = J = N$) with two-way fixed effects the routine requires to estimate $\approx 2NT$ fixed effects associated with a $2NT \times 2NT$ Hessian. In the case of three-way fixed effects, the number of parameters to be estimated is even $\approx N(N-1) \times 2NT$. In a trade panel data set with 200 countries and 50 years, the number of fixed effects in the latter case amounts to 59,800 parameters.

sets, Stammann (2018) proposes to calculate $\widehat{\mathbb{M}}\hat{\boldsymbol{\nu}}$ and $\widehat{\mathbb{M}}\mathbf{X}$ using the method of alternating projections (MAP), which only requires to repeatedly perform group-specific one-way weighted within transformations. This approach is feasible, since these within transformations translate into simple scalar transformations (see Stammann, Heiß, and McFadden 2016).[11] Note that all expressions that contain $\widehat{\mathbb{M}}$ or $\widehat{\mathbb{P}}$ can be calculated efficiently based on the MAP.

Next, we address the estimation of APEs. An estimator for the APEs is

$$\hat{\delta}_k = \frac{1}{IJT} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \widehat{\Delta}_{ijt}^{k} ,$$

where the partial effect of the k-th regressor $\widehat{\Delta}_{ijt}^{k}$ is either $\widehat{\Delta}_{ijt}^{k} = \partial \widehat{F}_{ijt} / \partial x_{ijtk}$ in the case of a continuous regressor or $\widehat{\Delta}_{ijt}^{k} = \widehat{F}_{ijt}|_{x_{ijtk}=1} - \widehat{F}_{ijt}|_{x_{ijtk}=0}$ in the case of a binary regressors.

Another question that arises in the context of APEs is how to calculate appropriate standard errors, even in the case of high-dimensional fixed effects. A possible candidate is the delta method, but in its standard form it requires the entire covariance matrix, which we do not obtain using the pseudo-demeaning approach. However, as outlined in Fernández-Val and Weidner (2016) and Czarnowske and Stammann (2019) in the context of individual and time fixed effects, it is possible to use a concentrated version of the delta method. In the following we present the feasible covariance estimators for our two-way and three-way error structure.[12] An appropriate covariance estimator for the APEs of the two-way fixed effects model is

$$\widehat{\mathbf{V}}^{\delta} = \frac{1}{I^2 J^2 T^2} \left(\underbrace{\left(\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \widehat{\boldsymbol{\Delta}}_{ijt} \right) \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \widehat{\boldsymbol{\Delta}}_{ijt} \right)'}_{v_1} + \underbrace{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \widehat{\boldsymbol{\Gamma}}_{ijt} \widehat{\boldsymbol{\Gamma}}_{ijt}'}_{v_2} \right) , \tag{5.5}$$

11. For further details, we refer the reader to appendix A.1, where we sketch the MAP for our application of two-way and three-way models, and provide the entire optimization routine corresponding to equation (5.4).

12. The corresponding asymptotic distribution of the estimators is provided in appendix A.3.

and of the three-way error component model

$$\widehat{\mathbf{V}}^{\delta} = \frac{1}{I^2J^2T^2}\Bigg(\underbrace{\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\widehat{\tilde{\mathbf{\Delta}}}_{ijt}\right)\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\widehat{\tilde{\mathbf{\Delta}}}_{ijt}\right)'}_{v_1} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\widehat{\mathbf{\Gamma}}_{ijt}\widehat{\mathbf{\Gamma}}'_{ijt}}_{v_2}$$
$$+\underbrace{2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{s>t}^{T}\widehat{\tilde{\mathbf{\Delta}}}_{ijt}\widehat{\mathbf{\Gamma}}'_{ijs}}_{v_3}\Bigg), \tag{5.6}$$

where in both cases $\widehat{\tilde{\mathbf{\Delta}}}_{ijt} = \widehat{\mathbf{\Delta}}_{ijt} - \hat{\boldsymbol{\delta}}$, $\widehat{\mathbf{\Delta}}_{ijt} = [\widehat{\Delta}^1_{ijt}, \dots, \widehat{\Delta}^m_{ijt}]'$, $\hat{\boldsymbol{\delta}} = [\hat{\delta}_1, \dots, \hat{\delta}_m]'$, and

$$\widehat{\mathbf{\Gamma}}_{ijt} = \left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\partial_\beta\widehat{\mathbf{\Delta}}_{ijt} - \left(\widehat{\mathbb{P}}\mathbf{X}\right)_{ijt}\partial_\eta\widehat{\mathbf{\Delta}}_{ijt}\right)'\widehat{\mathbf{A}}^{-1}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{ijt}\hat{\omega}_{ijt}\hat{\boldsymbol{\nu}}_{ijt} - \left(\widehat{\mathbb{P}}\widehat{\mathbf{\Psi}}\right)_{ijt}\partial_\eta\hat{\ell}_{ijt},$$

with $\widehat{\mathbf{A}} = (\widehat{\mathbb{M}}\mathbf{X})'\widehat{\mathbf{\Omega}}(\widehat{\mathbb{M}}\mathbf{X})$, $\widehat{\mathbf{\Psi}}_{ijt} = \partial_\eta\widehat{\mathbf{\Delta}}_{ijt}/\hat{\omega}_{ijt}$, and $\partial_\eta\hat{\ell}_{ijt}$ defined in table 5.1. To clarify notation, $\partial_\iota g(\cdot)$ denotes the first order partial derivative of an arbitrary function $g(\cdot)$ with respect to some parameter ι. Note that the term v_2 refers to the concentrated delta method. The terms v_1 and v_3 are in spirit of Fernández-Val and Weidner (2016) to improve the finite sample properties. These are on the one hand the variation induced by estimating sample instead of population means (v_1). On the other hand, if we are concerned about the strict exogeneity assumption (as we are in the case of dynamic three-way error structure models), the covariance between the estimation of sample means and parameters is another factor that should be incorporated (v_3). These computationally efficient covariance estimators can be readily applied not only to uncorrected APE estimators, but also to the bias-corrected APE estimators, which we will introduce below.

5.3.2 Incidental Parameter Bias Correction

As many nonlinear estimators, standard fixed effects versions of the logit and probit models suffer from the well-known incidental parameters problem first identified by Neyman and Scott (1948). The problem stems from the necessity to estimate many nuisance parameters which contaminate the estimator of the structural parameters and average partial effects. It can be further amplified by the inclusion of a lagged dependent variable. Note that this induces an incidental parameters problem even in the linear three-way fixed effects setting (see Nickell 1981) — and hence in our case also affects a linear probability model specification. Fernández-Val and Weidner (2018a) derive the order of the bias induced by incidental parameters to be given

by $bias \sim p/n$, where p and n are the numbers of parameters and observations, respectively. The literature suggests different types of bias corrections to reduce this incidental parameter bias. Jackknife corrections, like the leave-one-out jackknife proposed by Hahn and Newey (2004) or the split-panel jackknife (SPJ) introduced by Dhaene and Jochmans (2015), are the simplest approaches to obtain a bias correction, at the expense of being computationally costly. In contrast to analytical corrections, their application only requires knowledge of the order of the bias to form appropriate subpanels that are used to reestimate the model and to form an estimator of the bias terms. For analytical bias correction (ABC), it is necessary to derive the asymptotic distribution of the maximum likelihood estimator (MLE) in order to obtain an explicit expression of the asymptotic bias. This is then used to form a suitable estimator for the bias terms. Fernández-Val and Weidner (2016) propose analytical and split-panel jackknife bias corrections for structural parameters and APEs in the context of nonlinear models with individual and time fixed effects. In the following two subsections, we adapt and extend the bias corrections of Fernández-Val and Weidner (2016) to our two-way and three-way error component.[13]

Two-way fixed effects

The two-way fixed effects case with exporter-time and importer-time fixed effects is closely related to the two-way fixed effects models with a classical panel structure and individual and time fixed effects or with a pseudo-panel ij-structure and exporter and importer fixed effects as discussed by Fernández-Val and Weidner (2016) and Cruz-Gonzalez, Fernández-Val, and Weidner (2017), respectively. It is straightforward to see that in our case the overall bias consists of two components that are due to the inclusion of importer-time and exporter-time fixed effects, respectively, and takes the form $B_1/I + B_2/J$ (see appendix A.3).[14]

The form of the bias suggests to separately split the panel by I and J, leading to the following split-panel corrected estimator for the structural parameters:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{sp} &= 3\widehat{\boldsymbol{\beta}}_{I,J,T} - \widehat{\boldsymbol{\beta}}_{I/2,J,T} - \widehat{\boldsymbol{\beta}}_{I,J/2,T}, \quad \text{with} \\
\widehat{\boldsymbol{\beta}}_{I/2,J,T} &= \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_{\{i:i\leq\lceil I/2\rceil\},J,T} + \widehat{\boldsymbol{\beta}}_{\{i:i\geq\lfloor I/2+1\rfloor\},J,T}\right], \\
\widehat{\boldsymbol{\beta}}_{I,J/2,T} &= \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_{I,\{j:j\leq\lceil J/2\rceil,T\}} + \widehat{\boldsymbol{\beta}}_{I,\{j:j\geq\lfloor J/2+1\rfloor,T\}}\right],
\end{aligned} \tag{5.7}$$

where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the floor and ceiling functions. To clarify the notation, the subscript $\{i : i \leq \lceil I/2\rceil\}, J, T$ denotes that the estimator is based on a subsample,

13. We do not elaborate on the leave-one-out jackknife bias correction because the large number of fixed effects in our panel structure makes it unnecessary computationally demanding.

14. We also report the appropriate Neyman and Scott (1948) variance example in appendix A.2 as an illustration.

which contains all importers and time periods, but only the first half of all exporters.

In order to form the appropriate analytical bias correction, we need to specify the asymptotic distribution of the MLE, which we show in appendix A.3. The analytical bias-corrected estimator $\tilde{\boldsymbol{\beta}}^a$ is formed from estimators of the leading bias terms that are subtracted from the MLE of the full sample $\hat{\boldsymbol{\beta}}_{I,J,T}$. More precisely,

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}}^a &= \hat{\boldsymbol{\beta}}_{I,J,T} - \frac{\widehat{\mathbf{B}}_1^\beta}{I} - \frac{\widehat{\mathbf{B}}_2^\beta}{J}, \quad \text{with} \quad \widehat{\mathbf{B}}_1^\beta = \widehat{\mathbf{W}}^{-1}\widehat{\mathbf{B}}_1, \widehat{\mathbf{B}}_2^\beta = \widehat{\mathbf{W}}^{-1}\widehat{\mathbf{B}}_2, \quad \text{and} \\
\widehat{\mathbf{B}}_1 &= -\frac{1}{2JT}\sum_{j=1}^{J}\sum_{t=1}^{T}\frac{\sum_{i=1}^{I}\widehat{H}_{ijt}\partial_{\eta^2}\widehat{F}_{ijt}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{ijt}}{\sum_{i=1}^{I}\hat{\omega}_{ijt}}, \\
\widehat{\mathbf{B}}_2 &= -\frac{1}{2IT}\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\sum_{j=1}^{J}\widehat{H}_{ijt}\partial_{\eta^2}\widehat{F}_{ijt}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{ijt}}{\sum_{j=1}^{J}\hat{\omega}_{ijt}}, \\
\widehat{\mathbf{W}} &= \frac{1}{IJT}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\hat{\omega}_{ijt}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{ijt}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{ijt}',
\end{aligned}
$$

where $\partial_{\iota^2}g(\cdot)$ denotes the second order partial derivative of an arbitrary function $g(\cdot)$ with respect to some parameter ι. The explicit expressions of H_{ijt} and $\partial_{\eta^2}F_{ijt}$ are reported in table 5.1.

The split-panel jackknife estimator works similarly with APEs as with structural parameters. We simply replace in formula (5.7) the estimators for the structural parameters with estimators for the APEs. The following analytically bias-corrected estimator for the APEs is formed based on the asymptotic distribution presented in appendix A.3:

$$
\begin{aligned}
\tilde{\boldsymbol{\delta}}^a &= \hat{\boldsymbol{\delta}} - \frac{\widehat{\mathbf{B}}_1^\delta}{I} - \frac{\widehat{\mathbf{B}}_2^\delta}{J}, \quad \text{with} \\
\widehat{\mathbf{B}}_1^\delta &= \frac{1}{2JT}\sum_{j=1}^{J}\sum_{t=1}^{T}\frac{\sum_{i=1}^{I}-\widehat{H}_{ijt}\partial_{\eta^2}\widehat{F}_{ijt}\left(\widehat{\mathbb{P}}\widehat{\boldsymbol{\Psi}}\right)_{ijt}+\partial_{\eta^2}\widehat{\boldsymbol{\Delta}}_{ijt}}{\sum_{i=1}^{I}\hat{\omega}_{ijt}}, \\
\widehat{\mathbf{B}}_2^\delta &= \frac{1}{2IT}\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\sum_{j=1}^{J}-\widehat{H}_{ijt}\partial_{\eta^2}\widehat{F}_{ijt}\left(\widehat{\mathbb{P}}\widehat{\boldsymbol{\Psi}}\right)_{ijt}+\partial_{\eta^2}\widehat{\boldsymbol{\Delta}}_{ijt}}{\sum_{j=1}^{J}\hat{\omega}_{ijt}}.
\end{aligned}
$$

Note that all quantities are evaluated at bias-corrected structural parameters and the corresponding estimates of the fixed effects, where the latter can be obtained by reestimating the model using an offset algorithm as in Czarnowske and Stammann (2019). The covariance can be estimated according to equation (5.5).

Three-way fixed effects

Having adapted the two-way fixed effects bias correction of Fernández-Val and Weidner (2016) to the ijt-panel setting, we now move on to the more difficult case of

extending the consideration to three-way fixed effects. Fernández-Val and Weidner (2018a) conjecture based on their previously discussed formula, $bias \sim p/n$, that the bias is of order $(IT+JT+IJ)/(IJT)$ and of the form $B_1/I+B_2/J+B_3/T$. Intuitively, the inclusion of pair fixed effects induces another bias of order $1/T$ because there are only T informative observations per additionally included parameter. We support their conjecture by providing the appropriate Neyman and Scott (1948) variance example in appendix A.2 and propose novel analytical and jackknife bias corrections for three-way fixed effects models.

For the split-panel jackknife bias correction, this bias structure implies that we add an additional splitting dimension, leading to the following estimator for the structural parameters:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{sp} &= 4\widehat{\boldsymbol{\beta}}_{I,J,T} - \widehat{\boldsymbol{\beta}}_{I/2,J,T} - \widehat{\boldsymbol{\beta}}_{I,J/2,T} - \widehat{\boldsymbol{\beta}}_{I,J,T/2}, \quad \text{with} \\
\widehat{\boldsymbol{\beta}}_{I/2,J,T} &= \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_{\{i:i\leq\lfloor I/2\rfloor,J,T\}} + \widehat{\boldsymbol{\beta}}_{\{i:i\geq\lceil I/2+1\rceil,J,T\}}\right], \\
\widehat{\boldsymbol{\beta}}_{I,J/2,T} &= \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_{\{I,j:j\leq\lfloor J/2\rfloor,T\}} + \widehat{\boldsymbol{\beta}}_{\{I,j:j\geq\lceil J/2+1\rceil,T\}}\right], \\
\widehat{\boldsymbol{\beta}}_{I,J,T/2} &= \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_{\{I,J,t:t\leq\lfloor T/2\rfloor\}} + \widehat{\boldsymbol{\beta}}_{\{I,J,t:t\geq\lceil T/2+1\rceil\}}\right].
\end{aligned} \tag{5.8}$$

Combining insights from the classical panel structure in Fernández-Val and Weidner (2016), the pseudo-panel setting in Cruz-Gonzalez, Fernández-Val, and Weidner (2017), and the three-way fixed effects conjecture by Fernández-Val and Weidner (2018a), we formulate a conjecture for the asymptotic MLE distribution in the three-way setting (which we present in appendix A.3) and propose to extend the analytical two-way bias correction by a third part $\widehat{\mathbf{B}}_3$, such that

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}^{a} &= \hat{\boldsymbol{\beta}}_{I,J,T} - \frac{\widehat{\mathbf{B}}_1^{\beta}}{I} - \frac{\widehat{\mathbf{B}}_2^{\beta}}{J} - \frac{\widehat{\mathbf{B}}_3^{\beta}}{T}, \quad \text{with} \quad \widehat{\mathbf{B}}_3^{\beta} = \widehat{\mathbf{W}}^{-1}\widehat{\mathbf{B}}_3 \\
\widehat{\mathbf{B}}_3 &= -\frac{1}{2IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{t=1}^{T}\hat{\omega}_{ijt}\right)^{-1}\left(\sum_{t=1}^{T}\widehat{H}_{ijt}\partial_{\eta^2}\widehat{F}_{ijt}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{ijt}\right. \\
&\qquad \left. +2\sum_{l=1}^{L}(T/(T-L))\sum_{t=l+1}^{T}\partial_{\eta}\hat{\ell}_{ijt-l}\hat{\omega}_{ijt}\left(\widehat{\mathbb{M}}\mathbf{X}\right)_{ijt}\right).
\end{aligned}$$

L is a bandwidth parameter and is used for the estimation of spectral densities (Hahn and Kuersteiner 2007). In a model where all regressors are exogenous, L is set to zero, such that the second part in the numerator of $\widehat{\mathbf{B}}_3$ vanishes and all three estimators of the bias terms are symmetric. Otherwise, for instance in the dynamic model, Fernández-Val and Weidner (2016) suggest to conduct a sensitivity analysis with $L \in \{1,2,3,4\}$.

Again, for the APEs the split-panel jackknife estimator is formed by replacing the

estimators for the structural parameters with estimators for the APEs in formula (5.8). The analytically bias-corrected estimator, based on our conjecture for the asymptotic distribution provided in appendix A.3, is given by

$$\tilde{\boldsymbol{\delta}}^{a} = \hat{\boldsymbol{\delta}} - \frac{\widehat{\mathbf{B}}_1^{\delta}}{I} - \frac{\widehat{\mathbf{B}}_2^{\delta}}{J} - \frac{\widehat{\mathbf{B}}_3^{\delta}}{T}, \quad \text{with}$$

$$\widehat{\mathbf{B}}_3^{\delta} = \frac{1}{2IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\sum_{t=1}^{T} \hat{\omega}_{ijt} \right)^{-1} \left(\sum_{t=1}^{T} -\widehat{H}_{ijt} \partial_{\eta^2} \widehat{F}_{ijt} \left(\widehat{\mathbb{P}} \widehat{\boldsymbol{\Psi}} \right)_{ijt} + \partial_{\eta^2} \widehat{\boldsymbol{\Delta}}_{ijt} \right.$$
$$\left. + 2 \sum_{l=1}^{L} (T/(T-l)) \sum_{t=l+1}^{T} \partial_{\eta} \hat{\ell}_{ijt-l} \hat{\omega}_{ijt} \left(\widehat{\mathbb{M}} \widehat{\boldsymbol{\Psi}} \right)_{ijt} \right).$$

The last part of the numerator is again dropped if all regressors are assumed to be strictly exogenous. As previously mentioned, standard errors can still be obtained from equation (5.6).

5.4 Monte Carlo Simulations

In this section, we conduct extensive simulation experiments to investigate the properties of different estimators for both the structural parameters and the APEs. The estimators we study are MLE, ABC, SPJ and a (bias-corrected) ordinary least squares fixed effects estimator (LPM).[15] Our main focus are the biases and inference accuracies. To this end, we compute the relative bias and standard deviation (SD) in percent, the ratio between standard error and standard deviation (SE/SD), the relative root mean square error (RMSE) in percent, and the coverage probabilities (CPs) at a nominal level of 95 percent.

For the simulation experiments we adapt the design for a dynamic probit model of Fernández-Val and Weidner (2016) to our ijt-panel structure for the two cases with two- and three-way fixed effects.[16]

5.4.1 Two-Way Fixed Effects

The simulations in this section correspond to a theory-consistent estimation of the extensive margin outlined in section 5.2, taking into account unobserved time-varying exporter- and importer-specific terms as well as dynamics, but not allowing

15. Details on LPM and our suggested bias correction in this context are given in appendix A.4. We use the *R*-package *lfe* of Gaure (2013a) for the estimation of LPM and our own implementation for the estimation of the nonlinear estimators.

16. Further simulation experiments including static panel models are presented in appendix C.

for bilateral unobserved heterogeneity. Specifically, we generate data according to

$$y_{ijt} = \mathbf{1}\left[\beta_y y_{ijt-1} + \beta_x x_{ijt} + \lambda_{it} + \psi_{jt} \geq \epsilon_{ijt}\right],$$
$$y_{ij0} = \mathbf{1}\left[\beta_x x_{ij0} + \lambda_{i0} + \psi_{j0} \geq \epsilon_{ij0}\right],$$

where $i = 1,\dots,N$, $j = 1,\dots,N$, $t = 1,\dots,T$, $\lambda_{it} \sim$ iid. $\mathcal{N}(0,1/16)$, $\psi_{jt} \sim$ iid. $\mathcal{N}(0,1/16)$, and $\epsilon_{ijt} \sim$ iid. $\mathcal{N}(0,1)$.[17] Further, $x_{ijt} = 0.5x_{ijt-1} + \lambda_{it} + \psi_{jt} + \nu_{ijt}$, where $\nu_{ijt} \sim$ iid. $\mathcal{N}(0,0.5)$, $x_{ij0} \sim$ iid. $\mathcal{N}(0,1)$. To get an impression of how the different statistics evolve with changing panel dimensions, we consider all possible combinations of $N \in \{50,100,150\}$ and $T \in \{10,20,30,40,50\}$. For each of these combinations we generate $1,000$ samples.

Tables 5.8 – 5.13 in appendix B.1 report the extensive simulation results for the exogenous and predetermined regressors, respectively. The left panels contain the results of the structural parameters and the right panels the results of the APEs. In the following, we focus on the biases and coverage probabilities for $N \in \{50,150\}$, which we visualize in figures 5.2 and 5.3 for better comprehensibility.

First of all, we start with analyzing the properties of the different estimators for the structural parameters. MLE exhibits persistent biases which do not fade with increasing T but with increasing N. This result is as expected since MLE is fixed T consistent as shown in appendix A.3. Further, its CPs are too low and decreasing in T. The bias-corrected estimators perform clearly better than MLE. First, they reduce the bias considerably. ABC shows basically no bias for any considered sample size. SPJ performs slightly worse. Second, the bias corrections also dramatically improve the coverage probabilities. Whereas the CPs of ABC are close to the nominal value in all cases, the CPs of SPJ are somewhat too low for the exogenous regressor in the case of $N = 50$.

Next, we turn to the estimators of the APEs, where we now also consider LPM. It turns out that MLE as well as the two bias-corrected estimators are essentially unbiased. This is particularly noteworthy for MLE, since it exhibits a non-negligible bias for the structural parameters. Remarkably, LPM displays persistent biases which — different than for the nonlinear estimators — do not vanish with larger N. The bias is very small for the exogenous regressor but for the predetermined regressor it ranges between 5 and 6 percent.[18] This persistent biases also explain that LPM delivers too small CPs that decrease in T. Contrary, the CPs of the three nonlinear estimators are close to the nominal value in most cases.

17. Since $\{\lambda_{it}\}_{IT}$ and $\{\psi_{jt}\}_{JT}$ are independent sequences, and λ_{it} and ψ_{jt} are independent for all it, jt, we follow Fernández-Val and Weidner (2016) and incorporate this information in the covariance estimator for the APEs. The explicit expression is provided in the appendix A.3.

18. We found that the predicted probabilities of LPM exceed the boundaries of the unit interval considerably. This in turn affects the APEs for binary regressors, since they are based on differences of predicted probabilities.

Figure 5.2: *Dynamic: Two-Way Fixed Effects - Predetermined Regressor*

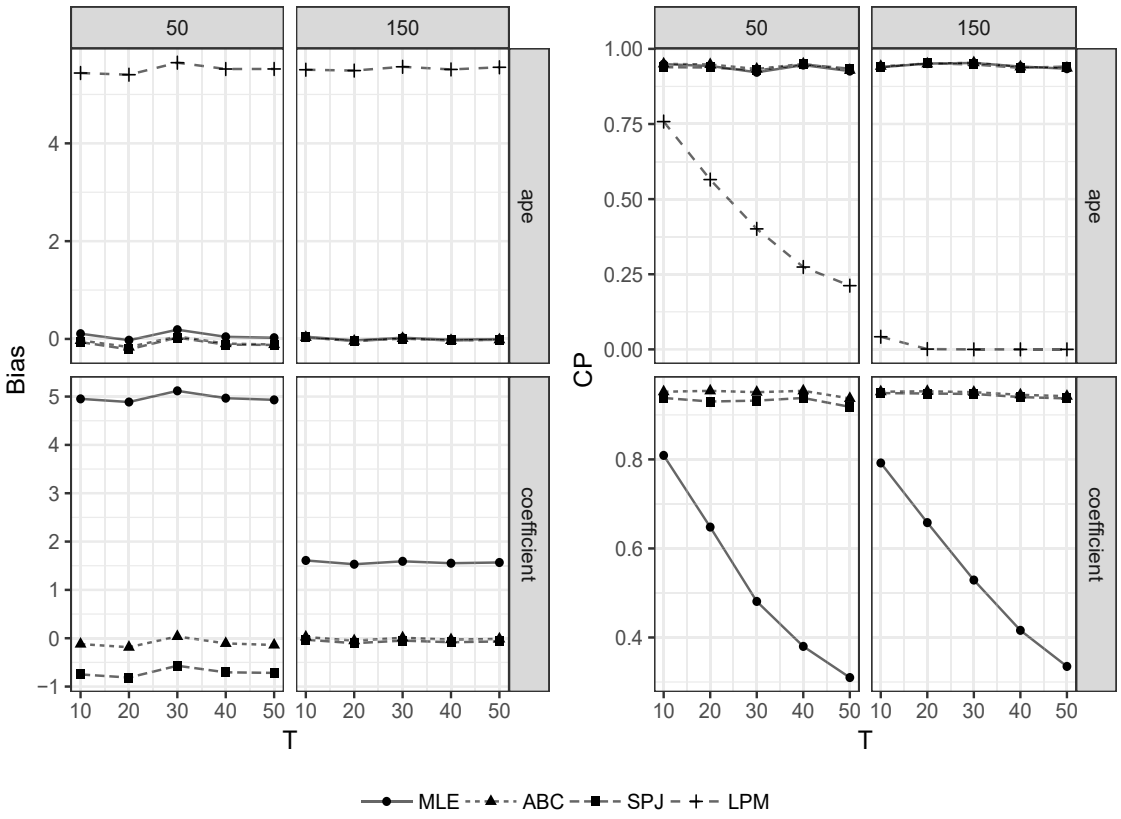
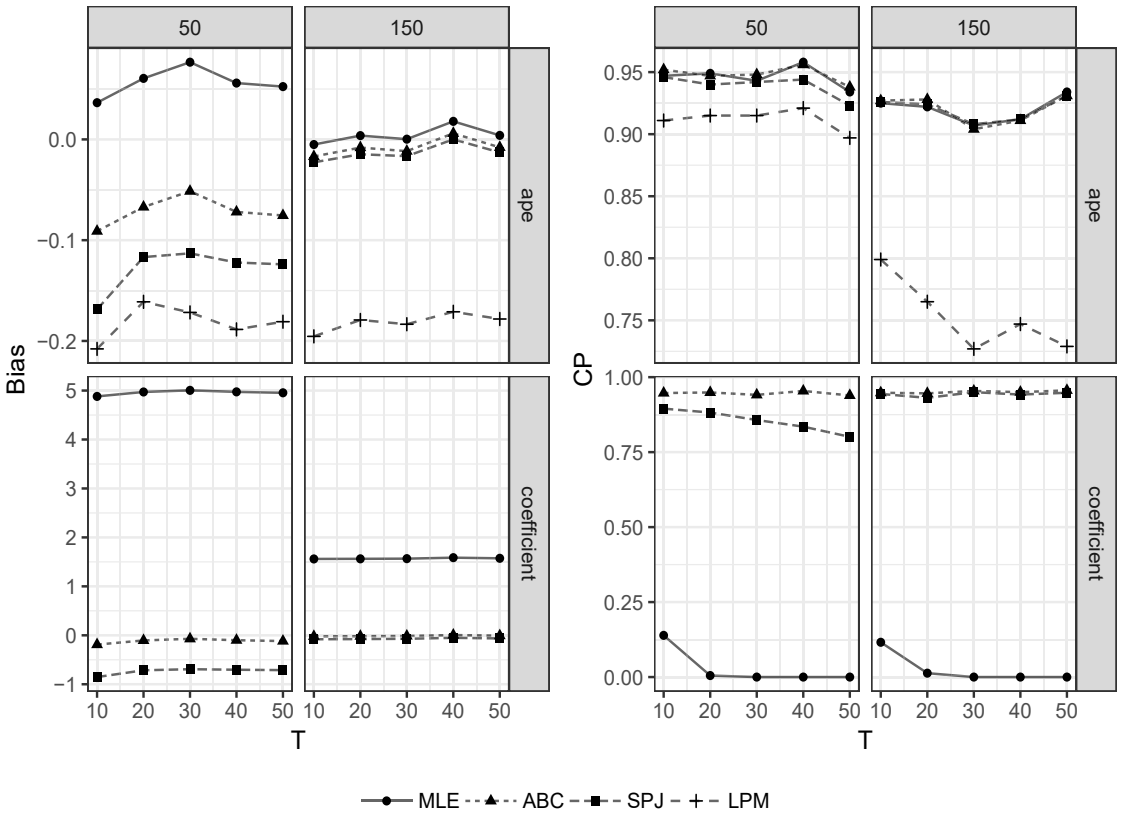


Figure 5.3: *Dynamic: Two-Way Fixed Effects - Exogenous Regressor*



All in all, our two-way fixed effects simulation results demonstrate that the bias-corrected estimators work extremely well in this context — for both structural parameters and APEs and both bias and coverage probabilities. Between the two, the analytical correction slightly outperforms the split-panel jackknife correction. If the interest lies only in APEs, the MLE works well, too, but for the structural parameters it shows bias and essentially useless coverage probabilities. LPM performs clearly worse than the probit estimators and should — given the availability of the nonlinear alternatives — only be used with great caution.

5.4.2 Three-Way Fixed Effects

The simulations in this section correspond to our preferred empirical specification for the extensive margin of international trade, in which we not only take into account the theoretically motivated it- and jt-fixed effects, but additionally allow for bilateral unobserved heterogeneity. In this three-way error structure environment, we generate data according to

$$y_{ijt} = \mathbf{1}\left[\beta_y y_{ijt-1} + \beta_x x_{ijt} + \lambda_{it} + \psi_{jt} + \mu_{ij} \geq \epsilon_{ijt}\right],$$
$$y_{ij0} = \mathbf{1}\left[\beta_x x_{ij0} + \lambda_{i0} + \psi_{j0} + \mu_{ij} \geq \epsilon_{ij0}\right],$$

where $i = 1,\dots,N$, $j = 1,\dots,N$, $t = 1,\dots,T$, $\beta_y = 0.5$, $\beta_x = 1$, $\lambda_{it} \sim$ iid. $\mathcal{N}(0,1/24)$, $\psi_{jt} \sim$ iid. $\mathcal{N}(0,1/24)$, $\mu_{ij} \sim$ iid. $\mathcal{N}(0,1/24)$, and $\epsilon_{ijt} \sim$ iid. $\mathcal{N}(0,1)$.[19] The exogenous regressor is modeled as an AR-1 process, $x_{ijt} = 0.5x_{ijt-1} + \lambda_{it} + \psi_{jt} + \mu_{ij} + \nu_{ijt}$, where $\nu_{ijt} \sim$ iid. $\mathcal{N}(0,0.5)$ and $x_{ij0} \sim$ iid. $\mathcal{N}(0,1)$. Again, we consider different sample sizes, specifically $N \in \{50,100,150\}$ and $T \in \{10,20,30,40,50\}$ and generate 1,000 data sets for each.

Tables 5.17 – 5.16 in appendix B.2 summarize the extensive simulation results for both regressors. For ABC and LPM we report two different choices of the bandwidth parameter, $L = 1$ and $L = 2$. Here, we again focus on the biases and coverage probabilities for $N \in \{50,150\}$ which are shown in figures 5.4 and 5.5.

We start by considering the different estimators for the structural parameters. For both kinds of regressors, MLE exhibits a severe bias that decreases with increasing T. However, even with $T = 50$, the estimator shows a distortion of 11 percent in the case of the predetermined regressor and 5 percent in the case of the exogenous regressor. We also find that the inference is not valid, since the CPs are zero or close to zero. The bias corrections bring a substantial improvement. First, they reduce the bias considerably. For example, the MLE of the predetermined regressor shows

19. We again follow Fernández-Val and Weidner (2016) and incorporate the information that $\{\lambda_{it}\}_{IT}$, $\{\psi_{jt}\}_{JT}$, and $\{\mu_{ij}\}_{IJ}$ are independent sequences, and λ_{it}, ψ_{jt}, and μ_{ij} are independent for all it, jt, ij in the covariance estimator for the APEs. The explicit expression is provided in appendix A.3.

Figure 5.4: *Dynamic: Three-Way Fixed Effects - Predetermined Regressor*

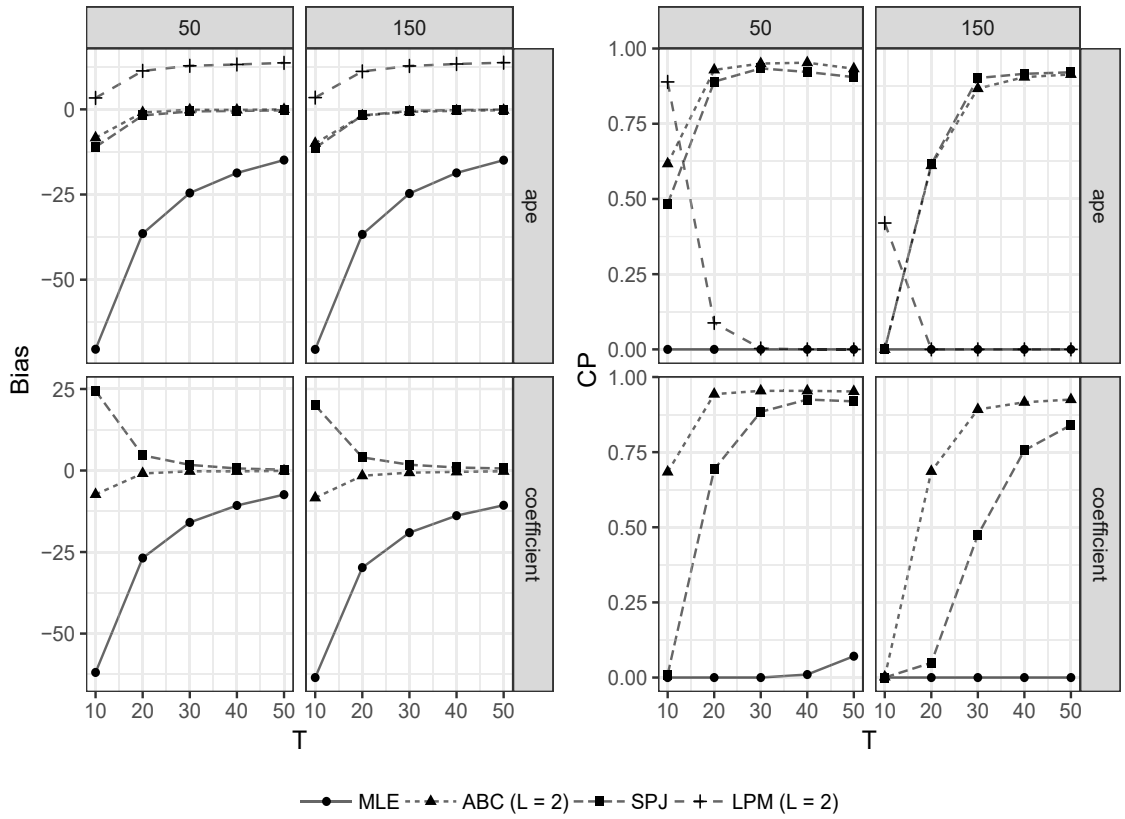
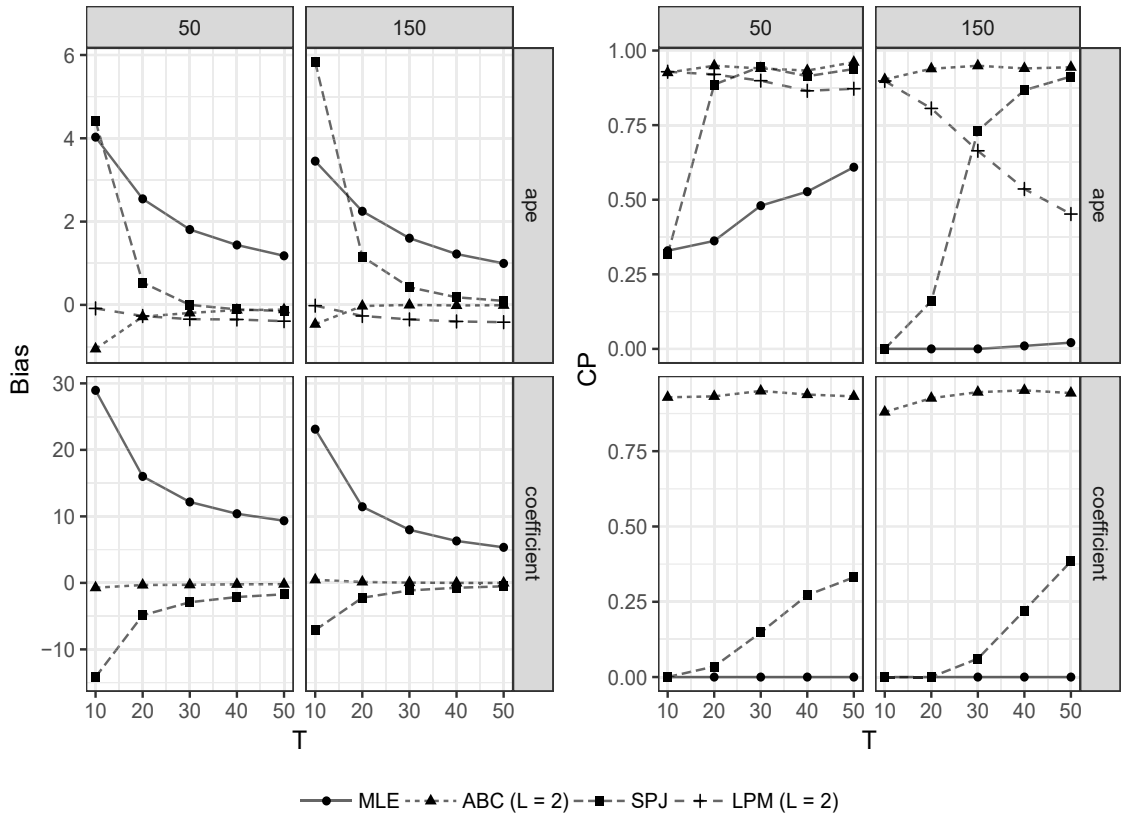


Figure 5.5: *Dynamic: Three-Way Fixed Effects - Exogenous Regressor*



a distortion of 63 percent for $T = 10$ and $N = 150$. ABC reduces the bias to 8 percent and SPJ to 20 percent. In the case of the exogenous regressor, MLE exhibits a bias of 23 percent, whereas ABC has a bias of 1 percent and SPJ of 7 percent. Irrespective of the type of the regressor, both bias-corrected estimators also converge quickly to the true parameter value with growing T. Second, the bias corrections improve the CPs. For the exogenous regressor the CPs of ABC are close to the desired level of 95 percent for all T, whereas SPJ remains far away from 95 percent even at $T = 50$. In the case of the predetermined regressor, the CPs of both corrections approach the nominal level when T rises. This happens faster for ABC.

We again proceed with the APEs, where we also consider LPM as an alternative estimator. Overall, we obtain similar findings as for the structural parameters. MLE is distorted over all settings, but the bias decreases as T increases. The distortion is especially severe in the case of the predetermined regressor. Even at $T = 50$, MLE suffers a bias of 15 percent. The bias corrections bring a substantial reduction in this case. Whereas ABC shows only a small distortion of 1 percent in the case of the exogenous regressor at $T = 10$, SPJ is even heavier distorted than MLE. However, with increasing T, both SPJ and ABC quickly converge to the true APE. Furthermore, unlike ABC, SPJ needs a sufficiently large number of time periods to get its CPs close to 95 percent. For the predetermined regressor, these convergence processes last longer for both bias corrections. Looking at LPM in the case of the exogenous regressors, it produces almost unbiased estimates irrespective of T, but its CPs fall dramatically with increasing T. Moreover, in the case of the predetermined regressor, we observe an increase in the bias up to 14 percent with increasing T.[20] These results illustrate the superiority of ABC and SPJ over LPM.

Overall, our three-way fixed effects simulation results confirm the conjecture of Fernández-Val and Weidner (2018a) about the general form and lend support to our conjecture for the specific structure of the bias terms in the three-way fixed effects specification. First, we find that the bias corrections indeed substantially mitigate the bias. Second, as already found in other studies, analytical bias corrections clearly outperform split-panel jackknife bias corrections (see among others Fernández-Val and Weidner 2016; Czarnowske and Stammann 2019). For samples with shorter time horizons, ABC is often less distorted and its dispersion is generally lower. This

20. A similar behaviour of LPM has been observed by Czarnowske and Stammann (2019) in the context of a dynamic probit model with individual and time fixed effects. To ensure that the bias correction presented in appendix A.4 in our three-way fixed effects specification is implemented correctly we have tested it in a data generation process for classical linear models, i.e. without binary dependent variables, and found that it works as intended. The undesirable behavior in our simulation design for the probit model is driven by the fact that due to the autoregressive process of $\mathbf{x}$, the predicted probabilities of LPM exceed the boundaries of the unit interval more and more frequently as T increases. This is particularly reflected in the APEs for binary regressors, since they are based on differences of predicted probabilities.

is also reflected by better CPs. Further, our three-way fixed effects simulation results suggest that estimates based on MLE or LPM should be treated with great caution. Generally, in the three-way fixed effects setting, a sufficiently large number of time periods appears to be crucial to obtain reliable results, even for the bias-corrected estimators.

5.5 Determinants of the Extensive Margin of Trade

Having described the estimation and bias correction procedures, we now turn to the estimation of the determinants of the extensive margin of international trade outlined in section 5.2.

Recall equation (5.2) that relates the incidence of nonzero aggregate trade flows to exporter-time and importer-time specific characteristics, as well as trade in the previous period, next to fixed and variable trade costs:

$$y_{ijt} = \begin{cases} 1 & \text{if} \quad \kappa + \lambda_{it} + \psi_{jt} + \beta_y y_{ij(t-1)} + \mathbf{x}'_{ijt}\boldsymbol{\beta}_x \geq \zeta_{ijt}, \\ 0 & \text{else}. \end{cases}$$

This yields the following probit model:

$$\Pr(y_{ijt} = 1 | \mathbf{x}_{ijt}, y_{ij(t-1)}, \lambda_{it}, \psi_{jt}) = F\left(\mathbf{x}'_{ijt}\boldsymbol{\beta}_x + \beta_y y_{ij(t-1)} + \lambda_{it} + \psi_{jt}\right), \tag{5.9}$$

in case we assume to capture bilateral variables and fixed trade costs entirely with observables, or

$$y_{ijt} = \begin{cases} 1 & \text{if} \quad \kappa + \lambda_{it} + \psi_{jt} + \beta_y y_{ij(t-1)} + \mathbf{x}'_{ijt}\boldsymbol{\beta}_x + \mu_{ij} \geq \zeta_{ijt}, \\ 0 & \text{else} \end{cases}$$

and

$$\Pr(y_{ijt} = 1 | \mathbf{x}_{ijt}, y_{ij(t-1)}, \lambda_{it}, \psi_{jt}, \mu_{ij}) = F\left(\mathbf{x}'_{ijt}\boldsymbol{\beta}_x + \beta_y y_{ij(t-1)} + \lambda_{it} + \psi_{jt} + \mu_{ij}\right), \tag{5.10}$$

in case we include a time-invariant bilateral fixed effect to capture unobservable country pair characteristics. $y_{ij(t-1)}$ is the lagged dependent variable, $\mathbf{x}$ is a vector of observable bilateral variables, β_y and β_x are the corresponding parameters.

We largely follow HMR and the wider literature on the determinants of the *intensive* margin of trade (compare Head and Mayer 2014) in the choice of these variables: Distance, a common land border, the same origin of the legal system, common language, previous colonial ties, a joint currency, an existing free trade agreement, or joint membership in the WTO. In terms of data, we turn to the

comprehensive gravity dataset provided alongside Head, Mayer, and Ries (2010) that encompasses information on trade flows and these variables of interest from 1948 – 2006.

Table 5.2: *Probit: Coefficients*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	-	-	1.664***	-	1.140***
	[-]	[-]	[1.719]	[-]	[1.057]
	(-)	(-)	(0.004)	(-)	(0.005)
log(Distance)	-	-0.800***	-0.528***	-	-
	[-0.656***]	[-0.821]	[-0.546]	[-]	[-]
	(0.003)	(0.003)	(0.004)	(-)	(-)
Land border	-	0.207***	0.118***	-	-
	[0.260***]	[0.214]	[0.124]	[-]	[-]
	(0.014)	(0.016)	(0.018)	(-)	(-)
Legal	-	0.137***	0.089***	-	-
	[0.090***]	[0.141]	[0.093]	[-]	[-]
	(0.004)	(0.004)	(0.005)	(-)	(-)
Language	-	0.426***	0.280***	-	-
	[0.380***]	[0.436]	[0.289]	[-]	[-]
	(0.005)	(0.006)	(0.007)	(-)	(-)
Colonial ties	-	0.657***	0.487***	-	-
	[0.190***]	[0.702]	[0.542]	[-]	[-]
	(0.02)	(0.031)	(0.036)	(-)	(-)
Currency Union	-	0.631***	0.424***	0.303***	0.214***
	[0.381***]	[0.649]	[0.443]	[0.335]	[0.255]
	(0.012)	(0.015)	(0.017)	(0.032)	(0.034)
FTA	-	0.543***	0.359***	0.073*	0.038
	[0.508***]	[0.552]	[0.364]	[0.072]	[0.033]
	(0.017)	(0.019)	(0.021)	(0.038)	(0.04)
WTO	-	0.152***	0.101***	0.052***	0.039**
	[0.286***]	[0.154]	[0.104]	[0.058]	[0.048]
	(0.005)	(0.008)	(0.009)	(0.016)	(0.017)
Fixed effects	i,j,t	it,jt	it,jt	it,jt,ij	it,jt,ij
Sample size	1,204,671	1,204,671	1,171,794	1,204,671	1,171,794
- perf. class.	12,298	147,760	141,537	370,617	374,067
Deviance	8.891×10^5	7.019×10^5	5.183×10^5	4.76×10^5	4.189×10^5

Note: Uncorrected coefficients in square brackets; standard errors in parentheses.

We report the bias-corrected coefficients in table 5.2 and the corresponding average partial effects in table 5.3.[21] For each uncorrected and (analytically) bias-corrected coefficients and average partial effects we also report the uncorrected one in square brackets, as well as the standard errors in parentheses below. In column (1) we first mimic the specification estimated by HMR.[22] Their specification includes

21. While the error term distribution assumed in section 5.2 suggests a probit estimator, we also estimate equations (5.9) and (5.10) with a logit estimator and show the corresponding results in tables 5.26 and 5.27 in appendix D. The coefficients and average partial effects are similar to those estimated with the probit model.

22. HMR use a dataset that ranges from 1970 to 1997. They also include dummy variables for

Table 5.3: *Probit: Average Partial Effects*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	-	-	0.346***	-	0.179***
	[-]	[-]	[0.344]	[-]	[0.138]
	(-)	(-)	(0.003)	(-)	(0.052)
log(Distance)	-	-0.135***	-0.066***	-	-
	[-0.136***]	[-0.135]	[-0.066]	[-]	[-]
	(0.005)	(0.005)	(0.001)	(-)	(-)
Land border	-	0.035***	0.015***	-	-
	[0.054***]	[0.035]	[0.015]	[-]	[-]
	(0.004)	(0.004)	(0.003)	(-)	(-)
Legal	-	0.023***	0.011***	-	-
	[0.019***]	[0.023]	[0.011]	[-]	[-]
	(0.001)	(0.001)	(0.001)	(-)	(-)
Language	-	0.071***	0.035***	-	-
	[0.078***]	[0.071]	[0.035]	[-]	[-]
	(0.003)	(0.001)	(0.001)	(-)	(-)
Colonial ties	-	0.107***	0.061***	-	-
	[0.039***]	[0.111]	[0.066]	[-]	[-]
	(0.004)	(0.007)	(0.005)	(-)	(-)
Currency Union	-	0.103***	0.053***	0.038***	0.024***
	[0.078***]	[0.103]	[0.054]	[0.037]	[0.025]
	(0.004)	(0.003)	(0.002)	(0.005)	(0.009)
FTA	-	0.090***	0.045***	0.009	0.004
	[0.103***]	[0.088]	[0.044]	[0.008]	[0.003]
	(0.005)	(0.004)	(0.003)	(0.007)	(0.006)
WTO	-	0.026***	0.013***	0.006**	0.004
	[0.061***]	[0.026]	[0.013]	[0.006]	[0.005]
	(0.002)	(0.002)	(0.001)	(0.003)	(0.003)
Fixed effects	i,j,t	it,jt	it,jt	it,jt,ij	it,jt,ij
Sample size	1,204,671	1,204,671	1,171,794	1,204,671	1,171,794
- perf. class.	12,298	147,760	141,537	370,617	374,067
Deviance	8.891×10^5	7.019×10^5	5.183×10^5	4.76×10^5	4.189×10^5

Note: Uncorrected average partial effects in square brackets; standard errors in parentheses.

exporter, importer and time fixed effects.[23] All coefficients have the expected sign, i.e. a negative impact of distance on the probability to trade, while all other variables are estimated to have a positive impact. Note the strong and highly significant impact of a common currency, free trade agreement or joint membership of the WTO. Ceteris paribus, each is estimated to increase the probability of nonzero flows by between 6 and 10 percentage points. Column (2) introduces a stricter set of fixed effects, namely at the exporter-time and importer-time level. Most coefficients and average partial effects are similar to those in column (1). This changes in column (3), which keeps the same fixed effects, but adds a lagged dependent variable. Assuming no unobservable bilateral heterogeneity, as in equation (5.9), this specification correctly estimates the model set up in section 5.2. The first thing to note is the highly significant coefficient for the lagged dependent variable, which reflects the strong impact of previous nonzero trade flows on current ones. Ceteris paribus, the average partial effect shows a 34 percentage points higher probability of nonzero trade, given the two countries were also engaged in trade in the previous year. The second observation is that essentially all coefficients are remarkably smaller than those in column (2), and average partial effects are reduced by about 50 percent across the board. This result underlines the need to explicitly take persistence into account. Note, however, that the APEs of the two specifications are not directly comparable, because the static model forces immediate effects and long-run dynamic adjustments into a single estimate.

Column (4) then takes one step back and one forward. While not including the lagged dependent variable in the estimation, it introduces a bilateral fixed effect that controls for bilateral unobserved heterogeneity. This follows the important insight by Baier and Bergstrand (2007), who show that controlling for unobserved bilateral heterogeneity produces considerably a different estimated impact of free trade agreements, among other variables, on the intensive margin of trade. While now an identification of many of the variables of interest is not possible anymore due to their time invariance, this specification reveals a much reduced estimated impact of the time-varying variables. The impact of a common currency on the probability of exporting is reduced to 3.8 percentage points, while those of a common free trade agreement and WTO are decreased to less than 1 percentage point. This result highlights the importance of controlling for unobserved country pair heterogeneity and possible endogeneity. Finally, in column (5) we present our preferred specification, estimating equation (5.10). The estimation now includes the

whether both countries are landlocked or islands, or follow the same religion. Hence our coefficients deviate somewhat from theirs, while remaining qualitatively similar.

23. Note that following Fernández-Val and Weidner (2018a) the incidental bias problem is small enough to ignore in this setting with i, j and t fixed effects, since the order of the bias is $1/IT + 1/JT + 1/IJ$, which in our case becomes negligible small since I, J and T are large.

"full set" of fixed effects, i.e. exporter-time, importer-time and bilateral fixed effect, in addition to the lagged dependent variable.[24] Again the coefficient on the latter is highly significant, entailing an average partial effect of about 18 percentage points. Importantly, the only remaining statistically significant average partial effect is estimated for a common currency at 2.4 percentage points. The impact of a free trade agreement or joint membership of the WTO are statistically insignificant.

Contrasting the results from column (5) to those of column (1), which currently constitutes the de-facto standard of estimating the determinants of the extensive margin of trade, underlines the importance of (i) appropriate exporter-time and importer-time fixed effects that capture all country-time specific variation; (ii) country pair fixed effects that capture all unobserved bilateral heterogeneity and address endogeneity concerns, analogous to Baier and Bergstrand (2007) on the intensive margin; (iii) dynamics, in that country pairs that have previously traded are significantly more likely to do so than otherwise comparable country pairs. This corroborates the stylized facts from section 5.1, which showed country pairs that had previously engaged in trade to be likely to do so again in the next year. Failing to observe any of these three insights produces widely different estimates.

Table 5.4: *Probit vs. LPM (Three-Way): Average Partial Effects*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	-	-	0.444***	0.474***	0.179***
	(-)	(-)	(0.001)	(0.001)	(0.052)
Currency Union	0.009***	0.038***	0.008***	0.008**	0.024***
	(0.003)	(0.005)	(0.003)	(0.003)	(0.009)
FTA	-0.121***	0.009	-0.065***	-0.062***	0.004
	(0.003)	(0.007)	(0.002)	(0.002)	(0.006)
WTO	0.017***	0.006**	0.008***	0.008***	0.004
	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)
Estimator	LPM	Probit	LPM	LPM	Probit
bias-corrected	false	true	false	true	true
Sample size	1,204,671	1,204,671	1,171,794	1,171,794	1,171,794

Note: All columns include exporter-time, importer-time, and pair fixed effects; standard errors in parentheses.

Another important insight is that the magnitude of the incidental parameters problem — at least in this specific setting — is not as severe as one might have feared. The most significant impact is observed on the coefficient for the lagged dependent variable, which in table 5.2 column (5) differs by about 10 percent, and even almost 24 percent in the respective average partial effect reported in table 5.3

24. Note that in the analytical bias correction we set the bandwidth parameter to $L = 2$. We report results for $L \in \{0,1,2,3,4\}$ in tables 5.28 to 5.33 in appendix D. The results remain robust with $L = 1-4$.

column (5). However, this does not carry through to other variables, in particular for average partial effects. As shown in simulations in section 5.4, this may not come as a big surprise. In this application we consider a panel that covers 57 years, meaning the relatively large T inhibits a strong bias (e.g. compare figure 5.5). As shown in the simulations, the bias is more severe in settings with fewer time periods and should be handled appropriately.

To show the superiority of using suitable binary choice estimators with high-dimensional fixed effects we also contrast the results to estimating equations (5.9) and (5.10) with a linear probability model (LPM). Table 5.4 shows that LPM with the same set of three-way fixed effects produces estimates that are far off the probit ones.[25] Columns (1) and (2) compare estimates without, columns (3) to (5) those with a lagged dependent variable.[26] Figure 5.6 underlines this impression: the LPM produces up to 28 percent of fitted probabilities < 0 or > 1. This result highlights that binary choice estimators with high-dimensional fixed effects cannot easily be mimicked by an OLS estimator.

5.6 Conclusion

In this paper we reexamined the determinants of the extensive margin of international trade. We set up a model that exhibits a dynamic component and allows for time-invariant unobserved bilateral trade cost factors, generating persistence — a feature in the data that has so far been paid little attention to. We estimated the model using a probit estimator with high-dimensional fixed effects. As fixed effects create an incidental parameters problem in binary choice settings, we characterized and implemented bias corrections for estimations with appropriate two- and three-way fixed effects. Finally, we showed that our estimates of the determinants of the extensive margin of trade differ significantly from previous ones. This highlights the importance of true state dependence and unobserved heterogeneity and therefore strongly supports the use of our bias-corrected dynamic fixed effects estimator.

The extensive margin of trade obviously extends beyond the aggregate level, warranting further research at lower levels of aggregation, in particular in the context of firms. While our model's prediction and its empirical specification rely on some abstractions, it provides a very tractable and flexible framework that can be estimated with recently established estimation procedures, when combined with the bias correction technique we introduce.

25. As for the probit estimates, we also report the bias-corrected LPM estimates with different bandwidth parameters in table 5.33. All in all, the results remain robust with $L = 1 - 4$. We also report estimates for two-way fixed effects in table 5.32 in appendix D.

26. In column (3) we ignore and in column (4) we apply the appropriate bias correction for the LPM with endogenous regressor, as detailed in appendix A.4.

References

Alessandria, George, and Horag Choi. 2007. “Do sunk costs of exporting matter for net export dynamics?” *The Quarterly Journal of Economics* 122 (1): 289–336.

Anderson, James E., and Eric Van Wincoop. 2003. “Gravity with gravitas: a solution to the border puzzle.” *The American Economic Review* 93 (1): 170–192.

Baier, Scott L., and Jeffrey H. Bergstrand. 2007. “Do free trade agreements actually increase members’ international trade?” *Journal of International Economics* 71 (1): 72–95.

Balazsi, Laszlo, Laszlo Matyas, and Tom Wansbeek. 2018. “The estimation of multidimensional fixed effects panel data models.” *Econometric Reviews* 37 (3): 212–227.

Berman, Nicolas, Vincent Rebeyrol, and Vincent Vicard. 2019. “Demand learning and firm dynamics: evidence from exporters.” *Review of Economics and Statistics* 101 (1): 91–106.

Bernard, Andrew B., Esther Ann Boler, Renzo Massari, Jose-Daniel Reyes, and Daria Taglioni. 2017. “Exporter dynamics and partial-year effects.” *The American Economic Review* 107 (10): 3211–28.

Correia, Sergio, Paulo Guimarães, and Thomas Zylkin. 2019. “PPMLHDFE: Fast poisson estimation with high-dimensional fixed effects.” *arXiv preprint:1903.01690*.

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.

Czarnowske, Daniel, and Amrei Stammann. 2019. “Binary Choice Models with High-Dimensional Individual and Time Fixed Effects.” *arXiv preprint:1904.04217*.

Das, Sanghamitra, Mark J. Roberts, and James R. Tybout. 2007. “Market entry costs, producer heterogeneity, and export dynamics.” *Econometrica* 75 (3): 837–873.

Dhaene, Geert, and Koen Jochmans. 2015. “Split-panel jackknife estimation of fixed-effect models.” *Review of Economic Studies* 82 (3): 991–1030.

Eaton, Jonathan, and Samuel Kortum. 2002. “Technology, geography, and trade.” *Econometrica* 70 (5): 1741–1779.

Eaton, Jonathan, Samuel Kortum, and Francis Kramarz. 2011. “An anatomy of international trade: Evidence from French firms.” *Econometrica* 79 (5): 1453–1498.

Eaton, Jonathan, Samuel Kortum, and Sebastian Sotelo. 2013. “International Trade: Linking Micro and Macro.” In *Advances in Economics and Econometrics: Tenth World Congress, Volume II: Applied Economics,* edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 329–370. Cambridge University Press.

Egger, Peter, and Mario Larch. 2011. “An assessment of the Europe agreements’ effects on bilateral trade, GDP, and welfare.” *European Economic Review* 55 (2): 263–279.

Egger, Peter, Mario Larch, Kevin E. Staub, and Rainer Winkelmann. 2011. “The trade effects of endogenous preferential trade agreements.” *American Economic Journal: Economic Policy* 3 (3): 113–43.

Feenstra, Robert C. 2015. *Advanced International Trade: Theory and Evidence.* Princeton University Press.

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.

Frisch, Ragnar, and Frederick V. Waugh. 1933. “Partial Time Regressions as Compared with Individual Trends.” *Econometrica* 1 (4): 387–401.

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

———. 2013b. “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, and Guido Kuersteiner. 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 (1-2): 96–99.

Head, Keith, and Thierry Mayer. 2014. “Chapter 3 - Gravity Equations: Workhorse, Toolkit, and Cookbook.” In *Handbook of International Economics,* edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 4:131–195. Handbook of International Economics. Elsevier.

Head, Keith, Thierry Mayer, and John Ries. 2010. “The erosion of colonial trade linkages after independence.” *Journal of International Economics* 81 (1): 1–14.

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, Alexander Hudlet, and Joschka Wanner. 2019. “Separating the Wheat from the Chaff: Fast Estimation of GLMs with High-Dimensional Fixed Effects.” *Unpublished Working Paper.*

Krugman, Paul. 1980. “Scale economies, product differentiation, and the pattern of trade.” *The American Economic Review* 70 (5): 950–959.

Larch, Mario, Joschka Wanner, Yoto V. Yotov, and Thomas Zylkin. 2019. “Currency Unions and Trade: A PPML Re-assessment with High-dimensional Fixed Effects.” *Oxford Bulletin of Economics and Statistics* 81 (3): 487–510.

Lovell, Michael C. 1963. “Seasonal adjustment of economic time series and multiple regression analysis.” *Journal of the American Statistical Association* 58 (304): 993–1010.

Melitz, Marc J. 2003. “The impact of trade on intra-industry reallocations and aggregate industry productivity.” *Econometrica* 71 (6): 1695–1725.

Neumann, John von. 1950. “Functional Operators. Vol. II. The geometry of orthogonal spaces, volume 22 (reprint of 1933 notes) of Annals of Math.” *Studies. Princeton University Press.*

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.

Piveteau, Paul. 2019. “An empirical dynamic model of trade with consumer accumulation.” *Working Paper.*

Redding, Stephen, and Anthony J. Venables. 2004. “Economic geography and international inequality.” *Journal of international Economics* 62 (1): 53–82.

Roberts, Mark J., and James R. Tybout. 1997. “The decision to export in Colombia: An empirical model of entry with sunk costs.” *The American Economic Review:* 545–564.

Ruhl, Kim J., and Jonathan L. Willis. 2017. “New exporter dynamics.” *International Economic Review* 58 (3): 703–726.

Stammann, Amrei. 2018. “Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects.” *arXiv preprint:1707.01815v3*.

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

Weidner, Martin, and Thomas Zylkin. 2018. “Bias and Consistency in Three-way Gravity Models.” *Working Paper*.

Appendix

A Computational and Econometric Details

A.1 Computational Details

In this section we briefly demonstrate how the method of alternating projections (MAP) works in the context of logit and probit models with a two- or three-way error component, and how it can be efficiently embedded into a standard Newton-Raphson optimization routine (see Stammann 2018 for further details).

First note that $\mathbb{M}\mathbf{v}$ is essentially a weighted within transformation, where $\mathbf{v}$ is an arbitrary $n \times 1$ vector, and $\mathbb{M} = \mathbf{I}_n - \mathbb{P} = \mathbf{I}_n - \mathbf{D}(\mathbf{D}'\mathbf{\Omega}\mathbf{D})^{-1}\mathbf{D}'\mathbf{\Omega}$. The computation of $\mathbb{M}$ is problematic even in moderately large data sets, and since $\mathbb{M}$ is non-sparse, there is also no general scalar expression to compute $\mathbb{M}\mathbf{v}$. Thus Stammann (2018) proposes to calculate $\mathbb{M}\mathbf{v}$ using a simple iterative approach based on the MAP tracing back to Neumann (1950) and Halperin (1962).[27] Let $\mathbf{D}_k$, denote the dummy variables corresponding to the k-th group, $k \in \{1,2,3\}$. Further, let $\mathbb{M}_{\mathbf{D}_k}\mathbf{v}$, with $\mathbb{M}_{\mathbf{D}_k} = \mathbf{I}_n - \mathbf{D}_k(\mathbf{D}_k'\mathbf{\Omega}\mathbf{D}_k)^{-1}\mathbf{D}_k'\mathbf{\Omega}$. The corresponding scalar expressions of $\mathbb{M}_{\mathbf{D}_k}\mathbf{v}$ are summarized in table (5.5).

Table 5.5: *Scalar Transformations*

group	$\mathbb{M}_{\mathbf{D}_k}\mathbf{v}$
importer-time ($k=1$)	$\mathbf{v}_{ijt} - \frac{\sum_{j=1}^{J}\omega_{ijt}v_{ijt}}{\sum_{j=1}^{J}\omega_{ijt}}$
exporter-time ($k=2$)	$\mathbf{v}_{ijt} - \frac{\sum_{i=1}^{I}\omega_{ijt}v_{ijt}}{\sum_{i=1}^{I}\omega_{ijt}}$
pair ($k=3$)	$\mathbf{v}_{ijt} - \frac{\sum_{t=1}^{T}\omega_{ijt}v_{ijt}}{\sum_{t=1}^{T}\omega_{ijt}}$

The MAP can be summarized by algorithm 7, where $K = 2$ in the case of two-way fixed effects and $K = 3$ in the case of three-way fixed effects. Thus, the MAP only requires to repeatedly apply weighted one-way within transformations (see Stammann 2018). The entire optimization routine is sketched by algorithm 8.

Algorithm 7 MAP: Neumann-Halperin

1: Initialize $\mathbb{M}\mathbf{v} = \mathbf{v}$.
2: **repeat**
3: **for** $k = 1,\dots,K$ **do**
4: Compute $\mathbb{M}_{\mathbf{D}_k}\mathbb{M}\mathbf{v}$ and update $\mathbb{M}\mathbf{v}$ such that $\mathbb{M}\mathbf{v} = \mathbb{M}_{\mathbf{D}_k}\mathbb{M}\mathbf{v}$
5: **until convergence**.

27. The MAP has been introduced to econometrics by Guimarães and Portugal (2010) and Gaure (2013b) in the context of linear models with multi-way fixed effects.

Algorithm 8 Efficient Newton-Raphson using the MAP

1: Initialize $\boldsymbol{\beta}^0$, $\boldsymbol{\eta}^0$, and $r = 0$.
2: **repeat**
3: Set $r = r + 1$.
4: Given $\hat{\boldsymbol{\eta}}^{r-1}$ compute $\hat{\boldsymbol{v}}$ and $\widehat{\boldsymbol{\Omega}}$.
5: Given $\hat{\boldsymbol{v}}$ and $\widehat{\boldsymbol{\Omega}}$ compute $\widehat{\mathbb{M}}\hat{\boldsymbol{v}}$ and $\widehat{\mathbb{M}}\mathbf{X}$ using the MAP
6: Compute $\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1} = \left((\widehat{\mathbb{M}}\mathbf{X})'\widehat{\boldsymbol{\Omega}}(\widehat{\mathbb{M}}\mathbf{X})\right)^{-1}(\widehat{\mathbb{M}}\mathbf{X})'\widehat{\boldsymbol{\Omega}}(\widehat{\mathbb{M}}\hat{\boldsymbol{v}})$
7: Compute $\hat{\boldsymbol{\eta}}^r = \hat{\boldsymbol{\eta}}^{r-1} + \hat{\boldsymbol{v}} - \widehat{\mathbb{M}}\hat{\boldsymbol{v}} + \widehat{\mathbb{M}}\mathbf{X}(\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r-1})$
8: **until convergence**.

A.2 Neyman-Scott Variance Example

In this section we study two variants of the classical Neyman and Scott (1948) variance example to support the form of the bias terms, and to illustrate the functionality of the bias corrections. To the best of our knowledge, the variance example of Neyman and Scott (1948) has not been investigated for our specific error components. We start with the more general three-way fixed effects case, which nests the two-way error structure.

Three-way Fixed Effects

Let $i = 1,\dots,I$, $j = 1,\dots,J$ and $t = 1,\dots,T$. Consider the following linear three-way fixed effects model:

$$y_{ijt} = \mathbf{x}_{ijt}'\boldsymbol{\beta} + \lambda_{it} + \psi_{jt} + \mu_{ij} + u_{ijt}\,. \tag{5.11}$$

According to Balazsi, Matyas, and Wansbeek (2018), the appropriate within transformation corresponding to equation (5.11) is given by

$$z_{ijt} - \bar{z}_{ij\cdot} - \bar{z}_{\cdot jt} - \bar{z}_{i\cdot t} + \bar{z}_{\cdot\cdot t} + \bar{z}_{\cdot j\cdot} + \bar{z}_{i\cdot\cdot} - \bar{z}_{\cdots}\,,$$

where

$$\bar{z}_{ij\cdot} = T^{-1}\sum_{t=1}^{T} z_{ijt}\,,\ \bar{z}_{\cdot jt} = I^{-1}\sum_{i=1}^{I} z_{ijt}\,,\ \bar{z}_{i\cdot t} = J^{-1}\sum_{j=1}^{J} z_{ijt}\,,$$
$$\bar{z}_{\cdot\cdot t} = (IJ)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ijt}\,,\ \bar{z}_{\cdot j\cdot} = (IT)^{-1}\sum_{i=1}^{I}\sum_{t=1}^{T} z_{ijt}\,,\ \bar{z}_{i\cdot\cdot} = (JT)^{-1}\sum_{j=1}^{J}\sum_{t=1}^{T} z_{ijt}\,,$$
$$\text{and } \bar{z}_{\cdots} = (IJT)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} z_{ijt}\,.$$

This result is helpful to study the following variant of the Neyman and Scott (1948) variance example

$$y_{ijt}|\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{\mu} \sim \mathcal{N}(\lambda_{it}+\psi_{jt}+\mu_{ij},\beta)\,,$$

where we now can easily form the uncorrected variance estimator

$$\hat{\beta}_{I,J,T} = \frac{1}{IJT}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}(y_{ijt}-\bar{y}_{ij\cdot}-\bar{y}_{\cdot jt}-\bar{y}_{i\cdot t}+\bar{y}_{\cdot\cdot t}+\bar{y}_{\cdot j\cdot}+\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdots})^2 \tag{5.12}$$

and the (degrees-of-freedom)-corrected counterpart

$$\hat{\beta}^{cor}_{I,J,T} = \frac{IJT}{(I-1)(J-1)(T-1)}\hat{\beta}_{I,J,T}\,.$$

Taking the expectation of (5.12) (conditional on the fixed effects) yields

$$\begin{aligned}\bar{\beta}_{I,J,T} = \mathbb{E}_{\alpha}[\hat{\beta}_{I,J,T}] &= \beta^0\left(\frac{(I-1)(J-1)(T-1)}{IJT}\right) \\ &= \beta^0\left(1-\frac{1}{I}-\frac{1}{J}-\frac{1}{T}+\frac{1}{IT}+\frac{1}{JT}+\frac{1}{IJ}-\frac{1}{IJT}\right),\end{aligned} \tag{5.13}$$

where β^0 is the true variance parameter. Thus, the three leading bias terms, which drive the main part of the asymptotic bias, are $\overline{\mathbf{B}}^{\beta}_{1,\infty} = -\beta^0$, $\overline{\mathbf{B}}^{\beta}_{2,\infty} = -\beta^0$, and $\overline{\mathbf{B}}^{\beta}_{3,\infty} = -\beta^0$.

Analytical Bias Correction

Using equation (5.13), we can form the analytically bias-corrected estimator

$$\tilde{\beta}^{a}_{I,J,T} = \hat{\beta}_{I,J,T} - \frac{\widehat{\mathbf{B}}^{\beta}_{1,I,J,T}}{I} - \frac{\widehat{\mathbf{B}}^{\beta}_{2,I,J,T}}{J} - \frac{\widehat{\mathbf{B}}^{\beta}_{3,I,J,T}}{T}\,, \tag{5.14}$$

where we set $\widehat{\mathbf{B}}^{\beta}_{1,I,J,T} = -\hat{\beta}_{I,J,T}$, $\widehat{\mathbf{B}}^{\beta}_{2,I,J,T} = -\hat{\beta}_{I,J,T}$, and $\widehat{\mathbf{B}}^{\beta}_{3,I,J,T} = -\hat{\beta}_{I,J,T}$ to reduce the order of the bias in equation (5.13) at costs of introducing higher order terms (see equation (5.16)). Thus, we can rewrite the analytically bias-corrected estimator (5.14)

$$\tilde{\beta}^{a}_{I,J,T} = \hat{\beta}_{I,J,T}\left(1+\frac{1}{I}+\frac{1}{J}+\frac{1}{T}\right). \tag{5.15}$$

Taking the expectation of (5.15) yields

$$\begin{aligned}\bar{\beta}^a_{I,J,T} &= \mathbb{E}_\alpha[\tilde{\beta}^a_{I,J,T}] \\ &= \beta^0\left(1-\frac{1}{I}-\frac{1}{J}-\frac{1}{T}+\frac{1}{IT}+\frac{1}{JT}+\frac{1}{IJ}-\frac{1}{IJT}\right)\left(1+\frac{1}{I}+\frac{1}{J}+\frac{1}{T}\right) \\ &= \beta^0\left(1-\frac{1}{IT}-\frac{1}{JT}-\frac{1}{T^2}-\frac{3}{IJ}+\frac{1}{I^3}+\frac{1}{J^3}+\frac{4}{IJT}+\frac{1}{IT^2}+\frac{1}{JT^2}\right. \\ &\quad \left.-\frac{1}{I^3T}-\frac{1}{J^3T}-\frac{1}{IJT^2}\right).\end{aligned} \tag{5.16}$$

Split-Panel Jackknife

As an alternative to equation (5.15) we can also form the following SPJ estimator:

$$\hat{\beta}^{spj}_{I,J,T} = 4\hat{\beta}_{I,J,T} - \hat{\beta}_{I/2,J,T} - \hat{\beta}_{I,J/2,T} - \hat{\beta}_{I,J,T/2}\,,$$

where $\hat{\beta}_{I/2,J,T}$ denotes the half panel estimator based on splitting the panel by exporters. This estimator also reduces the order of the bias in equation (5.13) as we see from its expected value:

$$\begin{aligned}\bar{\beta}^{spj}_{I,J,T} = E_\phi[\hat{\beta}^{spj}_{I,J,T}] &= 4\bar{\beta}_{I,J,T} - \bar{\beta}_{I/2,J,T} - \bar{\beta}_{I,J/2,T} - \bar{\beta}_{I,J,T/2} \\ &= \beta^0\left(1-\frac{1}{IT}-\frac{1}{JT}-\frac{1}{IJ}+\frac{2}{IJT}\right).\end{aligned} \tag{5.17}$$

Numerical Results

Table 5.6 shows numerical results for the uncorrected and the bias-corrected estimators in finite samples, where we assume symmetry, i.e. $I = J = N$. The results demonstrate that the bias corrections are effective in reducing the bias.

Table 5.6: *Numerical Results (Three-Way): Bias*

N	T	$(\bar{\beta}_{I,J,T}-\beta^0)/\beta^0$	$(\bar{\beta}^a_{I,J,T}-\beta^0)/\beta^0$	$(\bar{\beta}^{spj}_{I,J,T}-\beta^0)/\beta^0$
10	10	-0.271	-0.052	-0.028
25	10	-0.171	-0.021	-0.009
25	25	-0.115	-0.009	-0.005
50	10	-0.136	-0.015	-0.004
50	25	-0.078	-0.004	-0.002
50	50	-0.059	-0.002	-0.001

Two-way Fixed Effects

In the following we briefly review the example with two-way fixed effects:

$$y_{ijt}|\boldsymbol{\lambda},\boldsymbol{\psi} \sim \mathcal{N}(\lambda_{it}+\psi_{jt},\beta)\,.$$

Since it is a subcase of three-way fixed effects example, all previous results simplify by dropping the terms that exhibit T. The uncorrected variance estimator is[28]

$$\hat{\beta}_{I,J,T} = \frac{1}{IJT}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}(y_{ijt}-\bar{y}_{\cdot jt}-\bar{y}_{i\cdot t}+\bar{y}_{\cdot\cdot t})^2$$

and the (degrees-of-freedom)-corrected variance estimator is

$$\hat{\beta}^{cor}_{I,J,T} = \frac{IJ}{(I-1)(J-1)}\hat{\beta}_{I,J,T}\,.$$

Taking the expected value yields

$$\begin{aligned}\bar{\beta}_{I,J,T} = \mathbb{E}_\alpha[\hat{\beta}_{I,J,T}] &= \beta^0\left(\frac{(I-1)^2}{IJ}\right) \\ &= \beta^0\left(1-\frac{1}{I}-\frac{1}{J}+\frac{1}{IJ}\right).\end{aligned} \tag{5.18}$$

Analytical Bias Correction

Based on equation (5.18) we can form the following analytically bias-corrected estimator:

$$\tilde{\beta}^a_{I,J,T} = \hat{\beta}_{I,J,T}\left(1+\frac{1}{I}+\frac{1}{J}\right),$$

which has the expected value

$$\bar{\beta}^a_{I,J,T} = \mathbb{E}_\alpha[\tilde{\beta}^a_{I,J,T}] = \beta^0\left(1-\frac{3}{IJ}+\frac{1}{I^3}+\frac{1}{J^3}\right).$$

Split-Panel Jackknife

A suitable split-panel jackknife estimator is

$$\hat{\beta}^{spj}_{I,J,T} = 4\hat{\beta}_{I,J,T}-\hat{\beta}_{I/2,J,T}-\hat{\beta}_{I,J/2,T}\,,$$

28. We draw on the appropriate demeaning formula for the two-way fixed effects model $y_{ijt} = \boldsymbol{x}'_{ijt}\boldsymbol{\beta}+\lambda_{it}+\psi_{jt}+u_{ijt}$, which is given by $z_{ijt}-\bar{z}_{\cdot jt}-\bar{z}_{i\cdot t}+\bar{z}_{\cdot\cdot t}$.

which has the expected value

$$\begin{aligned}\bar{\beta}^{spj}_{I,J,T} = \mathbb{E}_\alpha[\hat{\beta}^{spj}_{I,J,T}] &= 3\bar{\beta}_{I,J,T} - \bar{\beta}_{I/2,J,T} - \bar{\beta}_{I,J/2,T} \\ &= \beta^0\left(1 - \frac{1}{IJ}\right).\end{aligned}$$

Numerical Results

The numerical results in table 5.7 demonstrate that the bias corrections work.

Table 5.7: *Numerical Results (Two-Way): Bias*

N	$(\bar{\beta}_{I,J,T} - \beta^0)/\beta^0$	$(\bar{\beta}^a_{I,J,T} - \beta^0)/\beta^0$	$(\bar{\beta}^{spj}_{I,J,T} - \beta^0)/\beta^0$
10	-0.190	-0.028	-0.010
25	-0.078	-0.005	-0.002
50	-0.040	-0.001	-0.000
100	-0.020	-0.000	-0.000

A.3 Asymptotic Bias Corrections

For the following expressions we draw on the results of Fernández-Val and Weidner (2016), who have already derived the asymptotic distributions of the MLE for structural parameters and APEs in classical two-way fixed effects models based on it-panels. As outlined in Cruz-Gonzalez, Fernández-Val, and Weidner (2017) the bias corrections of Fernández-Val and Weidner (2016) can be easily adjusted to two-way fixed effects models based on pseudo-panels with an ij-structure (i corresponds to importer and j to exporter), and importer and exporter fixed effects. We give an intuitive explanation. Since only J observations are informative per exporter fixed effects, we get a bias of order J for including exporter fixed effects, and vice versa a bias of order I for including importer fixed effects. Further, since there are no predetermined regressors in an ij-structure, we get two symmetric bias terms

$$\overline{\mathbf{B}}_{1,\infty} = \text{plim}_{I,J\to\infty}\left[-\frac{1}{2J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I}\mathbb{E}_\alpha[H_{ij}\partial_{\eta^2}F_{ij}(\mathbb{M}\mathbf{X})_{ij}]}{\sum_{i=1}^{I}\mathbb{E}_\alpha[\omega_{ij}]}\right], \tag{5.19}$$

$$\overline{\mathbf{B}}_{2,\infty} = \text{plim}_{I,J\to\infty}\left[-\frac{1}{2I}\sum_{i=1}^{I}\frac{\sum_{j=1}^{J}\mathbb{E}_\alpha[H_{ij}\partial_{\eta^2}F_{ij}(\mathbb{M}\mathbf{X})_{ij}]}{\sum_{j=1}^{J}\mathbb{E}_\alpha[\omega_{ij}]}\right], \tag{5.20}$$

where ω_{ij} is the ij-th diagonal entry of $\mathbf{\Omega}$, and $\mathbb{M} = \mathbf{I}_{IJ} - \mathbf{D}(\mathbf{D}'\mathbf{\Omega}\mathbf{D})^{-1}\mathbf{D}'\mathbf{\Omega}$. $\partial_{\iota^2}g(\cdot)$ denotes the second order partial derivative of an arbitrary function $g(\cdot)$ with respect to some parameter ι. The explicit expressions of H_{ijt} and $\partial_{\eta^2}F_{ijt}$ are reported in table 5.1. Equations (5.19) and (5.20) are essentially $\overline{\mathbf{D}}_\infty$ from Fernández-Val and

Weidner (2016) with adjusted indices. The same adjustment can be transferred to the APEs.

In the following we apply the same logic to derive the asymptotic bias terms in our two- and three-way error structure.

Two-way fixed effects

We get a bias of order J for including exporter-time fixed effects, since J observations are informative per exporter-time fixed effect. In the same way we get a bias of order I for including importer-time fixed effects. Similar to the case of the ij-structure of Cruz-Gonzalez, Fernández-Val, and Weidner (2017) we get two symmetric bias terms in the distributions of the structural parameters and the APEs, respectively, because including predetermined regressors does not violate the strict exogeneity assumption.

Asymptotic distribution of $\hat{\boldsymbol{\beta}}$

$$\sqrt{IJ}(\widehat{\boldsymbol{\beta}}_{I,J,T} - \boldsymbol{\beta}^0) \xrightarrow{d} \overline{\mathbf{W}}_{\infty}^{-1} \mathcal{N}(\kappa \overline{\mathbf{B}}_{1,\infty} + \kappa^{-1} \overline{\mathbf{B}}_{2,\infty}, \overline{\mathbf{W}}_{\infty}), \quad \text{with} \tag{5.21}$$

$$\overline{\mathbf{B}}_{1,\infty} = \operatorname{plim}_{I,J\to\infty} \left[-\frac{1}{2J} \sum_{t=1}^{T} \sum_{j=1}^{J} \frac{\sum_{i=1}^{I} \mathbb{E}_{\alpha}[H_{ijt} \partial_{\eta^2} F_{ijt} (\mathbb{M}\mathbf{X})_{ijt}]}{\sum_{i=1}^{I} \mathbb{E}_{\alpha}[\omega_{ijt}]} \right],$$

$$\overline{\mathbf{B}}_{2,\infty} = \operatorname{plim}_{I,J\to\infty} \left[-\frac{1}{2I} \sum_{t=1}^{T} \sum_{i=1}^{I} \frac{\sum_{j=1}^{J} \mathbb{E}_{\alpha}[H_{ijt} \partial_{\eta^2} F_{ijt} (\mathbb{M}\mathbf{X})_{ijt}]}{\sum_{j=1}^{J} \mathbb{E}_{\alpha}[\omega_{ijt}]} \right],$$

$$\overline{\mathbf{W}}_{\infty} = \operatorname{plim}_{I,J\to\infty} \left[\frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \mathbb{E}_{\alpha}[\omega_{ijt} (\mathbb{M}\mathbf{X})_{ijt} (\mathbb{M}\mathbf{X})_{ijt}'] \right],$$

where $\sqrt{J/I} \to \kappa$ as $I, J \to \infty$ and $0 < \kappa < \infty$.

Asymptotic distribution of $\hat{\boldsymbol{\delta}}$

$$r(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} - I^{-1} \overline{\mathbf{B}}_{1,\infty}^{\delta} - J^{-1} \overline{\mathbf{B}}_{2,\infty}^{\delta}) \xrightarrow{d} \mathcal{N}(0, \overline{\mathbf{V}}_{\infty}), \quad \text{with} \tag{5.22}$$

$$\overline{\mathbf{B}}_{1,\infty}^{\delta} = \operatorname{plim}_{I,J\to\infty} \left[\frac{1}{2JT} \sum_{t=1}^{T} \sum_{j=1}^{J} \frac{\sum_{i=1}^{I} -\mathbb{E}_{\alpha}[H_{ijt} \partial_{\eta^2} F_{ijt}] \mathbb{E}_{\alpha}[(\mathbb{P}\boldsymbol{\Psi})_{ijt}] + \mathbb{E}_{\alpha}[\partial_{\eta^2} \boldsymbol{\Delta}_{ijt}]}{\sum_{i=1}^{I} \mathbb{E}_{\alpha}[\omega_{ijt}]} \right],$$

$$\overline{\mathbf{B}}_{2,\infty}^{\delta} = \operatorname{plim}_{I,J\to\infty} \left[\frac{1}{2IT} \sum_{t=1}^{T} \sum_{i=1}^{I} \frac{\sum_{j=1}^{J} -\mathbb{E}_{\alpha}[H_{ijt} \partial_{\eta^2} F_{ijt}] \mathbb{E}_{\alpha}[(\mathbb{P}\boldsymbol{\Psi})_{ijt}] + \mathbb{E}_{\alpha}[\partial_{\eta^2} \boldsymbol{\Delta}_{ijt}]}{\sum_{j=1}^{J} \mathbb{E}_{\alpha}[\omega_{ijt}]} \right],$$

$$\overline{\mathbf{V}}_{\infty}^{\delta} = \operatorname{plim}_{I,J\to\infty} \frac{r^2}{I^2 J^2 T^2} \mathbb{E}_{\alpha} \left[\left(\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \boldsymbol{\Delta}_{ijt} \right) \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \boldsymbol{\Delta}_{ijt} \right)' + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \boldsymbol{\Gamma}_{ijt} \boldsymbol{\Gamma}_{ijt}' \right],$$

where $\bar{\mathbf{\Delta}}_{ijt} = \mathbf{\Delta}_{ijt} - \boldsymbol{\delta}$, $\mathbf{\Delta}_{ijt} = [\Delta^1_{ijt},\dots,\Delta^m_{ijt}]'$, $\boldsymbol{\delta} = [\delta_1,\dots,\delta_m]'$, $\delta_k = (IJT)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\Delta^k_{ijt}$, $\mathbf{\Psi}_{ijt} = \partial_\eta \mathbf{\Delta}_{ijt}/\omega_{ijt}$, r is a convergence rate, and

$$\begin{aligned}\mathbf{\Gamma}_{ijt} = \mathbb{E}_\alpha\left[(IJ)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\partial_\beta \mathbf{\Delta}_{ijt} - (\mathbb{P}\mathbf{X})_{ijt}\,\partial_\eta \mathbf{\Delta}_{ijt}\right]' \overline{\mathbf{W}}_\infty^{-1}\,\mathbb{E}_\alpha\left[(\mathbb{M}\mathbf{X})_{ijt}\,\omega_{ijt}\boldsymbol{\nu}_{ijt}\right] \\ - \mathbb{E}_\alpha\left[(\mathbb{P}\mathbf{\Psi})_{ijt}\,\partial_\eta \ell_{ijt}\right].\end{aligned}$$

$\partial_\iota g(\cdot)$ denotes the first order partial derivative of an arbitrary function $g(\cdot)$ with respect to some parameter ι. The expression $\overline{\mathbf{V}}^{\delta}_{\infty}$ can be modified by assuming that $\{\lambda_{it}\}_{IT}$ and $\{\psi_{jt}\}_{JT}$ are independent sequences, and λ_{it} and ψ_{jt} are independent for all it, jt:

$$\begin{aligned}\overline{\mathbf{V}}^{\delta}_{\infty} = \operatorname{plim}_{I,J\to\infty}\frac{r^2}{I^2J^2T^2}\,\mathbb{E}_\alpha\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{r=1}^{J}\bar{\mathbf{\Delta}}_{ijt}\bar{\mathbf{\Delta}}'_{irt} + \sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{i\neq p}^{I}\bar{\mathbf{\Delta}}_{ijt}\bar{\mathbf{\Delta}}'_{pjt}\right. \\ \left. + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\mathbf{\Gamma}_{ijt}\mathbf{\Gamma}'_{ijt}\right].\end{aligned}$$

Three-way fixed effects

With the inclusion of pair fixed effects, we introduce an additional bias of order T, since only T observations are informative per pair fixed effect. Another difference that occurs in contrast to the two-way fixed effects case is that predetermined regressors lead to a violation of the strict exogeneity assumption. To deal with this issue we adapt the asymptotic bias terms $\overline{\mathbf{B}}_\infty$ and $\overline{\mathbf{B}}^{\delta}_\infty$ of Fernández-Val and Weidner (2016) to the new structure.

Conjectured asymptotic distribution of $\hat{\boldsymbol{\beta}}$

$$\sqrt{IJT}(\widehat{\boldsymbol{\beta}}_{I,J,T} - \boldsymbol{\beta}^0) \xrightarrow{d} \overline{\mathbf{W}}_\infty^{-1}\mathcal{N}(\kappa_1\overline{\mathbf{B}}_{1,\infty} + \kappa_2\overline{\mathbf{B}}_{2,\infty} + \kappa_3\overline{\mathbf{B}}_{3,\infty}, \overline{\mathbf{W}}_\infty), \quad \text{with}$$

$$\overline{\mathbf{B}}_{1,\infty} = \operatorname{plim}_{I,J,T\to\infty}\left[-\frac{1}{2JT}\sum_{t=1}^{T}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I}\mathbb{E}_\alpha[H_{ijt}\partial_{\eta^2}F_{ijt}(\mathbb{M}\mathbf{X})_{ijt}]}{\sum_{i=1}^{I}\mathbb{E}_\alpha[\omega_{ijt}]}\right],$$

$$\overline{\mathbf{B}}_{2,\infty} = \operatorname{plim}_{I,J,T\to\infty}\left[-\frac{1}{2IT}\sum_{t=1}^{T}\sum_{i=1}^{I}\frac{\sum_{j=1}^{J}\mathbb{E}_\alpha[H_{ijt}\partial_{\eta^2}F_{ijt}(\mathbb{M}\mathbf{X})_{ijt}]}{\sum_{j=1}^{J}\mathbb{E}_\alpha[\omega_{ijt}]}\right],$$

$$\begin{aligned}\overline{\mathbf{B}}_{3,\infty} = \operatorname{plim}_{I,J,T\to\infty}\left[-\frac{1}{2IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{t=1}^{T}\mathbb{E}_\alpha[\omega_{ijt}]\right)^{-1}\left(\sum_{t=1}^{T}\mathbb{E}_\alpha[H_{ijt}\partial_{\eta^2}F_{ijt}(\mathbb{M}\mathbf{X})_{ijt}]\right.\right. \\ \left.\left. +2\sum_{\tau=t+1}^{T}\mathbb{E}_\alpha[H_{ijt}(Y_{ijt}-F_{ijt})\omega_{ijt}(\mathbb{M}\mathbf{X})_{ijt}]\right)\right],\end{aligned}$$

$$\overline{\mathbf{W}}_{\infty} = \operatorname{plim}_{I,J,T\to\infty}\left[\frac{1}{IJT}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\mathbb{E}_{\alpha}[\omega_{ijt}(\mathbb{M}\mathbf{X})_{ijt}(\mathbb{M}\mathbf{X})'_{ijt}]\right],$$

where $\sqrt{(JT)/I}\to\kappa_1$, $\sqrt{(IT)/J}\to\kappa_2$, $\sqrt{(IJ)/T}\to\kappa_3$ as $I,J,T\to\infty$, and $0<\kappa_{\iota}<\infty$ for $\iota=1,2,3$. The second term in the numerator of $\overline{\mathbf{B}}_{3,\infty}$ is dropped if all regressors are assumed to be strictly exogenous.

Conjectured asymptotic distribution of $\hat{\boldsymbol{\delta}}$

$$r(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}-I^{-1}\overline{\mathbf{B}}^{\delta}_{1,\infty}-J^{-1}\overline{\mathbf{B}}^{\delta}_{2,\infty}-T^{-1}\overline{\mathbf{B}}^{\delta}_{3,\infty})\xrightarrow{d}\mathcal{N}(0,\overline{\mathbf{V}}^{\delta}_{\infty}),\quad\text{with}$$

$$\overline{\mathbf{B}}^{\delta}_{1,\infty}=\operatorname{plim}_{I,J,T\to\infty}\left[\frac{1}{2JT}\sum_{t=1}^{T}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I}-\mathbb{E}_{\alpha}[H_{ijt}\partial_{\eta^2}F_{ijt}]\mathbb{E}_{\alpha}[(\mathbb{P}\boldsymbol{\Psi})_{ijt}]+\mathbb{E}_{\alpha}[\partial_{\eta^2}\boldsymbol{\Delta}_{ijt}]}{\sum_{i=1}^{I}\mathbb{E}_{\alpha}[\omega_{ijt}]}\right],$$

$$\overline{\mathbf{B}}^{\delta}_{2,\infty}=\operatorname{plim}_{I,J,T\to\infty}\left[\frac{1}{2IT}\sum_{t=1}^{T}\sum_{i=1}^{I}\frac{\sum_{j=1}^{J}-\mathbb{E}_{\alpha}[H_{ijt}\partial_{\eta^2}F_{ijt}]\mathbb{E}_{\alpha}[(\mathbb{P}\boldsymbol{\Psi})_{ijt}]+\mathbb{E}_{\alpha}[\partial_{\eta^2}\boldsymbol{\Delta}_{ijt}]}{\sum_{j=1}^{J}\mathbb{E}_{\alpha}[\omega_{ijt}]}\right],$$

$$\begin{aligned}\overline{\mathbf{B}}^{\delta}_{3,\infty}=\operatorname{plim}_{I,J,T\to\infty}\Bigg[\frac{1}{2IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{t=1}^{T}\mathbb{E}_{\alpha}[\omega_{ijt}]\right)^{-1}\Bigg(\sum_{t=1}^{T}&-\mathbb{E}_{\alpha}[H_{ijt}\partial_{\eta^2}F_{ijt}]\mathbb{E}_{\alpha}[(\mathbb{P}\boldsymbol{\Psi})_{ijt}]\\&+\mathbb{E}_{\alpha}[\partial_{\eta^2}\boldsymbol{\Delta}_{ijt}]+2\sum_{\tau=t+1}^{T}\mathbb{E}_{\alpha}[\partial_{\eta}\ell_{ijt-l}\omega_{ijt}(\mathbb{M}\boldsymbol{\Psi})_{ijt}]\Bigg)\Bigg].\end{aligned}$$

$$\begin{aligned}\overline{\mathbf{V}}^{\delta}_{\infty}=\operatorname{plim}_{I,J,T\to\infty}\frac{r^2}{I^2J^2T^2}\mathbb{E}_{\alpha}\Bigg[&\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\bar{\boldsymbol{\Delta}}_{ijt}\right)\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\bar{\boldsymbol{\Delta}}_{ijt}\right)'\\&+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\boldsymbol{\Gamma}_{ijt}\boldsymbol{\Gamma}'_{ijt}+2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{s>t}^{T}\bar{\boldsymbol{\Delta}}_{ijt}\boldsymbol{\Gamma}'_{ijs}\Bigg],\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Gamma}_{ijt}=\mathbb{E}_{\alpha}\left[(IJT)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\partial_{\beta}\boldsymbol{\Delta}_{ijt}-(\mathbb{P}\mathbf{X})_{ijt}\partial_{\eta}\boldsymbol{\Delta}_{ijt}\right]'&\overline{\mathbf{W}}^{-1}_{\infty}\mathbb{E}_{\alpha}\left[(\mathbb{M}\mathbf{X})_{ijt}\omega_{ijt}\boldsymbol{\nu}_{ijt}\right]\\&-\mathbb{E}_{\alpha}\left[(\mathbb{P}\boldsymbol{\Psi})_{ijt}\partial_{\eta}\ell_{ijt}\right],\end{aligned}$$

and r is a convergence rate. The second term in the numerator of $\overline{\mathbf{B}}_{3,\infty}$ and the last term in $\overline{\mathbf{V}}^{\delta}_{\infty}$ are dropped if all regressors are assumed to be strictly exogenous. The expression $\overline{\mathbf{V}}^{\delta}_{\infty}$ can be further modified by assuming that $\{\lambda_{it}\}_{IT}$, $\{\psi_{jt}\}_{JT}$ and $\{\mu_{ij}\}_{IJ}$ are independent sequences, and λ_{it}, ψ_{jt} and μ_{ij} are independent for all it, jt, ij:

$$\begin{aligned}\widehat{\mathbf{V}}^{\delta}=\operatorname{plim}_{I,J,T\to\infty}\frac{r^2}{I^2J^2T^2}\mathbb{E}_{\alpha}\Bigg(&\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{r=1}^{J}\bar{\boldsymbol{\Delta}}_{ijt}\bar{\boldsymbol{\Delta}}'_{irt}+\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{i\neq p}^{I}\bar{\boldsymbol{\Delta}}_{ijt}\bar{\boldsymbol{\Delta}}'_{pjt}\\&+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{s\neq t}^{T}\bar{\boldsymbol{\Delta}}_{ijt}\bar{\boldsymbol{\Delta}}'_{ijs}+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\boldsymbol{\Gamma}_{ijt}\boldsymbol{\Gamma}'_{ijt}+2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{s>t}^{T}\bar{\boldsymbol{\Delta}}_{ijt}\boldsymbol{\Gamma}'_{ijs}\Bigg).\end{aligned}$$

A.4 Bias-Corrected Ordinary Least Squares

Consider the three-way fixed effects linear probability model

$$y_{ijt} = \lambda_{it} + \psi_{jt} + \mu_{ij} + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \epsilon_{ijt}\,,$$

which can also be rewritten in matrix notation:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,. \tag{5.23}$$

We first deal with the computational burden. Applying the three-way fixed effects residual projection $\mathbb{M} = \mathbf{I}_{IJT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ to (5.23), leads to the following concentrated regression:

$$\mathbb{M}\mathbf{y} = \mathbb{M}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,. \tag{5.24}$$

The demeaning can be efficiently carried out by using the method of alternating projections (see Gaure 2013b).

Hahn and Moon (2006) have derived the bias of dynamic linear models with individual and time fixed effects. They show that there is only a bias of order $1/T$ stemming from the inclusion of individual effects in combination with predetermined regressors. Transferring their result to our problem with the three-way error component suggests that the inclusion of pair fixed effects in combination with predetermined regressors leads to the same order of the bias. Thus, the linear probability model needs only to be bias-corrected if not all regressors are strictly exogenous. This is for example the case in a dynamic model, where we include $\mathbf{y}_{t-1}$ to our set of regressors.

An estimator of the bias is given by

$$\widehat{\mathbf{B}} = \left(\frac{1}{IJT}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}(\mathbb{M}\mathbf{X})_{ijt}(\mathbb{M}\mathbf{X})'_{ijt}\right)^{-1}\left(-\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{L}\frac{1}{T-l}\sum_{t=l+1}^{T}\mathbf{X}_{ijt}\hat{\boldsymbol{\epsilon}}_{ijt-l}\right),$$

where $\hat{\boldsymbol{\epsilon}}$ is the residual of (5.24) and L is a bandwidth parameter.[29] This yields the bias-corrected estimator

$$\hat{\boldsymbol{\beta}} - \frac{\widehat{\mathbf{B}}}{IJT}\,, \tag{5.25}$$

where $\hat{\boldsymbol{\beta}} = \left((\mathbb{M}\mathbf{X})'(\mathbb{M}\mathbf{X})\right)^{-1}(\mathbb{M}\mathbf{X})'\mathbb{M}\mathbf{y}$.

29. The residuals of equation (5.23) and equation (5.24) are identical (see Gaure 2013b).

B Detailed Monte Carlo Results

B.1 Two-Way Fixed Effects

Table 5.8: *Properties: Dynamic (Two-Way) - x_{ijt} - $N = 50$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 50; T = 10									
MLE	5	2	5	0.95	0.14	0	1	1	0.97	0.95
ABC	-0	2	2	0.99	0.95	-0	1	1	0.98	0.95
SPJ	-1	2	2	0.96	0.90	-0	1	1	0.96	0.95
LPM						-0	1	1	0.89	0.91
	N = 50; T = 20									
MLE	5	1	5	0.97	0.00	0	1	1	0.97	0.95
ABC	-0	1	1	1.01	0.95	-0	1	1	0.98	0.95
SPJ	-1	1	1	0.97	0.88	-0	1	1	0.96	0.94
LPM						-0	1	1	0.88	0.92
	N = 50; T = 30									
MLE	5	1	5	0.93	0.00	0	1	1	0.97	0.94
ABC	-0	1	1	0.97	0.94	-0	1	1	0.98	0.95
SPJ	-1	1	1	0.93	0.86	-0	1	1	0.96	0.94
LPM						-0	1	1	0.90	0.92
	N = 50; T = 40									
MLE	5	1	5	0.98	0.00	0	1	1	1.00	0.96
ABC	-0	1	1	1.03	0.95	-0	1	1	1.01	0.96
SPJ	-1	1	1	0.98	0.83	-0	1	1	0.98	0.94
LPM						-0	1	1	0.92	0.92
	N = 50; T = 50									
MLE	5	1	5	0.92	0.00	0	1	1	0.95	0.93
ABC	-0	1	1	0.96	0.94	-0	1	1	0.95	0.94
SPJ	-1	1	1	0.94	0.80	-0	1	1	0.93	0.92
LPM						-0	1	1	0.86	0.90

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.9: *Properties: Dynamic (Two-Way) - x_{ijt} - $N = 100$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 100; T = 10									
MLE	2	1	3	0.97	0.12	0	1	1	0.95	0.94
ABC	-0	1	1	0.99	0.94	-0	1	1	0.95	0.94
SPJ	-0	1	1	0.97	0.94	-0	1	1	0.94	0.93
LPM						-0	1	1	0.79	0.87
	N = 100; T = 20									
MLE	2	1	2	0.96	0.01	0	1	1	0.90	0.92
ABC	-0	1	1	0.98	0.94	-0	1	1	0.90	0.92
SPJ	-0	1	1	0.96	0.93	-0	1	1	0.89	0.91
LPM						-0	1	1	0.73	0.82
	N = 100; T = 30									
MLE	2	0	2	0.97	0.00	0	0	0	0.92	0.93
ABC	-0	0	0	0.99	0.95	-0	0	0	0.92	0.93
SPJ	-0	0	0	0.98	0.93	-0	0	0	0.91	0.92
LPM						-0	0	1	0.75	0.83
	N = 100; T = 40									
MLE	2	0	2	0.97	0.00	0	0	0	0.89	0.92
ABC	-0	0	0	0.99	0.95	-0	0	0	0.89	0.92
SPJ	-0	0	0	0.99	0.92	-0	0	0	0.88	0.92
LPM						-0	0	0	0.73	0.81
	N = 100; T = 50									
MLE	2	0	2	0.99	0.00	0	0	0	0.92	0.93
ABC	-0	0	0	1.00	0.95	-0	0	0	0.92	0.94
SPJ	-0	0	0	0.99	0.93	-0	0	0	0.91	0.93
LPM						-0	0	0	0.74	0.83

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.10: *Properties: Dynamic (Two-Way) - x_{ijt} - N = 150*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 150; T = 10									
MLE	2	1	2	0.98	0.12	-0	1	1	0.91	0.92
ABC	-0	1	1	0.99	0.95	-0	1	1	0.91	0.93
SPJ	-0	1	1	0.99	0.94	-0	1	1	0.91	0.93
LPM						-0	1	1	0.67	0.80
	N = 150; T = 20									
MLE	2	0	2	0.99	0.01	0	0	0	0.91	0.92
ABC	-0	0	0	1.00	0.95	-0	0	0	0.90	0.93
SPJ	-0	0	0	0.98	0.93	-0	0	0	0.90	0.92
LPM						-0	0	0	0.67	0.76
	N = 150; T = 30									
MLE	2	0	2	1.01	0.00	0	0	0	0.86	0.91
ABC	-0	0	0	1.02	0.95	-0	0	0	0.86	0.90
SPJ	-0	0	0	1.01	0.95	-0	0	0	0.86	0.91
LPM						-0	0	0	0.63	0.73
	N = 150; T = 40									
MLE	2	0	2	0.99	0.00	0	0	0	0.88	0.91
ABC	0	0	0	1.00	0.95	0	0	0	0.88	0.91
SPJ	-0	0	0	0.98	0.94	0	0	0	0.88	0.91
LPM						-0	0	0	0.66	0.75
	N = 150; T = 50									
MLE	2	0	2	1.02	0.00	0	0	0	0.90	0.93
ABC	-0	0	0	1.03	0.96	-0	0	0	0.90	0.93
SPJ	-0	0	0	1.02	0.95	-0	0	0	0.90	0.93
LPM						-0	0	0	0.67	0.73

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.11: *Properties: Dynamic (Two-Way) - $y_{ij(t-1)}$ - $N = 50$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 50; T = 10									
MLE	5	4	7	0.99	0.81	0	4	4	0.99	0.95
ABC	-0	4	4	1.03	0.95	-0	4	4	1.01	0.95
SPJ	-1	4	4	1.00	0.94	-0	4	4	0.98	0.94
LPM						5	4	7	0.97	0.76
	N = 50; T = 20									
MLE	5	3	6	0.96	0.65	-0	3	3	0.96	0.94
ABC	-0	3	3	1.00	0.95	-0	3	3	0.97	0.95
SPJ	-1	3	3	0.97	0.93	-0	3	3	0.94	0.94
LPM						5	3	6	0.96	0.56
	N = 50; T = 30									
MLE	5	3	6	0.95	0.48	0	3	3	0.94	0.92
ABC	0	3	3	0.99	0.95	0	3	3	0.96	0.93
SPJ	-1	3	3	0.97	0.93	0	3	3	0.94	0.93
LPM						6	3	6	0.94	0.40
	N = 50; T = 40									
MLE	5	2	5	0.98	0.38	0	2	2	0.99	0.95
ABC	-0	2	2	1.02	0.95	-0	2	2	1.01	0.95
SPJ	-1	2	2	1.01	0.94	-0	2	2	0.99	0.95
LPM						6	2	6	0.97	0.27
	N = 50; T = 50									
MLE	5	2	5	0.92	0.31	0	2	2	0.93	0.93
ABC	-0	2	2	0.96	0.94	-0	2	2	0.95	0.93
SPJ	-1	2	2	0.94	0.92	-0	2	2	0.92	0.93
LPM						6	2	6	0.93	0.21

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.12: *Properties: Dynamic (Two-Way) - $y_{ij(t-1)}$ - $N = 100$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 100; T = 10									
MLE	2	2	3	0.96	0.80	0	2	2	0.94	0.94
ABC	0	2	2	0.98	0.94	0	2	2	0.95	0.94
SPJ	-0	2	2	0.97	0.94	0	2	2	0.95	0.94
LPM						5	2	6	0.91	0.30
	N = 100; T = 20									
MLE	2	2	3	0.99	0.63	0	2	2	0.99	0.94
ABC	-0	1	1	1.01	0.95	-0	2	2	1.00	0.94
SPJ	-0	2	2	0.99	0.94	-0	2	2	0.98	0.94
LPM						6	2	6	0.96	0.06
	N = 100; T = 30									
MLE	2	1	3	0.97	0.52	0	1	1	0.97	0.94
ABC	-0	1	1	0.99	0.94	-0	1	1	0.98	0.94
SPJ	-0	1	1	0.96	0.94	-0	1	1	0.96	0.93
LPM						6	1	6	0.94	0.01
	N = 100; T = 40									
MLE	2	1	3	0.99	0.42	0	1	1	0.97	0.94
ABC	-0	1	1	1.01	0.95	-0	1	1	0.98	0.94
SPJ	-0	1	1	0.99	0.94	-0	1	1	0.96	0.94
LPM						6	1	6	0.94	0.00
	N = 100; T = 50									
MLE	2	1	3	0.94	0.31	0	1	1	0.92	0.93
ABC	-0	1	1	0.96	0.93	-0	1	1	0.92	0.93
SPJ	-0	1	1	0.95	0.93	-0	1	1	0.91	0.92
LPM						6	1	6	0.90	0.00

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.13: *Properties: Dynamic (Two-Way) - $y_{ij(t-1)}$ - $N = 150$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 150; T = 10									
MLE	2	1	2	0.98	0.79	0	2	2	0.96	0.94
ABC	0	1	1	0.99	0.95	0	2	2	0.97	0.94
SPJ	-0	1	1	0.98	0.95	0	2	2	0.95	0.94
LPM						6	2	6	0.92	0.04
	N = 150; T = 20									
MLE	2	1	2	0.98	0.66	-0	1	1	1.00	0.95
ABC	-0	1	1	1.00	0.95	-0	1	1	1.00	0.95
SPJ	-0	1	1	0.99	0.95	-0	1	1	0.99	0.95
LPM						5	1	6	0.96	0.00
	N = 150; T = 30									
MLE	2	1	2	0.98	0.53	0	1	1	0.99	0.95
ABC	0	1	1	1.00	0.95	0	1	1	0.99	0.95
SPJ	-0	1	1	0.98	0.95	0	1	1	0.98	0.95
LPM						6	1	6	0.94	0.00
	N = 150; T = 40									
MLE	2	1	2	0.96	0.42	-0	1	1	0.96	0.94
ABC	-0	1	1	0.97	0.94	-0	1	1	0.96	0.94
SPJ	-0	1	1	0.96	0.94	-0	1	1	0.95	0.94
LPM						6	1	6	0.91	0.00
	N = 150; T = 50									
MLE	2	1	2	0.94	0.34	-0	1	1	0.93	0.93
ABC	-0	1	1	0.95	0.94	-0	1	1	0.94	0.94
SPJ	-0	1	1	0.94	0.94	-0	1	1	0.93	0.94
LPM						6	1	6	0.90	0.00

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

B.2 Three-Way Fixed Effects

Table 5.14: *Properties: Dynamic (Three-Way) - x_{ijt} - $N = 50$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 50; T = 10									
MLE	29	3	29	0.82	0.00	4	2	4	1.01	0.33
ABC (1)	-1	2	2	1.02	0.94	-1	2	2	1.09	0.94
ABC (2)	-1	2	2	1.01	0.93	-1	2	2	1.08	0.93
SPJ	-14	3	14	0.62	0.00	4	2	5	0.87	0.32
LPM (1)						0	2	2	0.94	0.93
LPM (2)						-0	2	2	0.94	0.93
	N = 50; T = 20									
MLE	16	1	16	0.87	0.00	3	1	3	0.97	0.36
ABC (1)	-0	1	1	0.98	0.94	-0	1	1	1.00	0.95
ABC (2)	-0	1	1	0.97	0.93	-0	1	1	0.99	0.95
SPJ	-5	1	5	0.86	0.04	1	1	1	0.91	0.89
LPM (1)						-0	1	1	0.90	0.93
LPM (2)						-0	1	1	0.90	0.92
	N = 50; T = 30									
MLE	12	1	12	0.92	0.00	2	1	2	1.00	0.48
ABC (1)	-0	1	1	1.01	0.95	-0	1	1	1.01	0.95
ABC (2)	-0	1	1	1.01	0.95	-0	1	1	1.01	0.94
SPJ	-3	1	3	0.93	0.15	-0	1	1	0.96	0.95
LPM (1)						-0	1	1	0.89	0.92
LPM (2)						-0	1	1	0.89	0.90
	N = 50; T = 40									
MLE	10	1	10	0.89	0.00	1	1	2	0.97	0.53
ABC (1)	-0	1	1	0.97	0.94	-0	1	1	0.98	0.93
ABC (2)	-0	1	1	0.97	0.94	-0	1	1	0.97	0.93
SPJ	-2	1	2	0.88	0.27	-0	1	1	0.91	0.91
LPM (1)						-0	1	1	0.84	0.89
LPM (2)						-0	1	1	0.84	0.86
	N = 50; T = 50									
MLE	9	1	9	0.90	0.00	1	1	1	1.01	0.61
ABC (1)	-0	1	1	0.97	0.94	-0	1	1	1.01	0.96
ABC (2)	-0	1	1	0.97	0.93	-0	1	1	1.01	0.96
SPJ	-2	1	2	0.90	0.33	-0	1	1	0.94	0.94
LPM (1)						-0	1	1	0.86	0.88
LPM (2)						-0	1	1	0.86	0.87

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; values in parentheses after ABC and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 5.15: *Properties: Dynamic (Three-Way) - x_{ijt} - $N = 100$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 100; T = 10									
MLE	24	1	24	0.89	0.00	4	1	4	1.04	0.02
ABC (1)	0	1	1	1.05	0.95	-0	1	1	1.08	0.94
ABC (2)	0	1	1	1.05	0.96	-1	1	1	1.08	0.91
SPJ	-9	1	9	0.70	0.00	6	1	6	0.89	0.00
LPM (1)						0	1	1	0.88	0.91
LPM (2)						-0	1	1	0.87	0.91
	N = 100; T = 20									
MLE	13	1	13	0.89	0.00	2	1	2	0.96	0.02
ABC (1)	0	1	1	0.98	0.93	0	1	1	0.98	0.95
ABC (2)	0	1	1	0.98	0.94	-0	1	1	0.97	0.94
SPJ	-3	1	3	0.86	0.01	1	1	1	0.89	0.54
LPM (1)						-0	1	1	0.85	0.89
LPM (2)						-0	1	1	0.85	0.87
	N = 100; T = 30									
MLE	9	1	9	0.91	0.00	2	0	2	0.96	0.05
ABC (1)	0	0	0	0.97	0.95	0	0	0	0.96	0.94
ABC (2)	-0	0	0	0.97	0.94	-0	0	0	0.96	0.94
SPJ	-1	1	2	0.91	0.14	0	0	1	0.93	0.86
LPM (1)						-0	0	1	0.82	0.86
LPM (2)						-0	0	1	0.82	0.81
	N = 100; T = 40									
MLE	7	0	7	0.91	0.00	1	0	1	0.94	0.12
ABC (1)	0	0	0	0.96	0.94	0	0	0	0.94	0.93
ABC (2)	-0	0	0	0.96	0.94	-0	0	0	0.94	0.92
SPJ	-1	0	1	0.92	0.32	0	0	0	0.91	0.91
LPM (1)						-0	0	1	0.79	0.81
LPM (2)						-0	0	1	0.79	0.73
	N = 100; T = 50									
MLE	6	0	6	0.94	0.00	1	0	1	1.00	0.17
ABC (1)	0	0	0	0.99	0.94	0	0	0	1.00	0.95
ABC (2)	-0	0	0	0.98	0.94	-0	0	0	1.00	0.95
SPJ	-1	0	1	0.95	0.48	0	0	0	0.96	0.94
LPM (1)						-0	0	0	0.80	0.76
LPM (2)						-0	0	1	0.80	0.69

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; values in parentheses after ABC and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 5.16: *Properties: Dynamic (Three-Way) - x_{ijt} - N = 150*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 150; T = 10									
MLE	23	1	23	0.86	0.00	3	1	4	1.06	0.00
ABC (1)	1	1	1	1.01	0.82	-0	1	1	1.09	0.94
ABC (2)	0	1	1	1.00	0.88	-0	1	1	1.07	0.90
SPJ	-7	1	7	0.67	0.00	6	1	6	0.89	0.00
LPM (1)						0	1	1	0.84	0.88
LPM (2)						-0	1	1	0.83	0.90
	N = 150; T = 20									
MLE	11	0	11	0.94	0.00	2	0	2	0.97	0.00
ABC (1)	0	0	0	1.02	0.89	0	0	0	0.97	0.94
ABC (2)	0	0	0	1.01	0.93	-0	0	0	0.97	0.94
SPJ	-2	0	2	0.89	0.00	1	0	1	0.90	0.16
LPM (1)						-0	0	0	0.81	0.88
LPM (2)						-0	0	0	0.81	0.81
	N = 150; T = 30									
MLE	8	0	8	0.92	0.00	2	0	2	0.96	0.00
ABC (1)	0	0	0	0.98	0.91	0	0	0	0.97	0.93
ABC (2)	0	0	0	0.98	0.95	-0	0	0	0.97	0.95
SPJ	-1	0	1	0.91	0.06	0	0	1	0.92	0.73
LPM (1)						-0	0	0	0.79	0.80
LPM (2)						-0	0	0	0.79	0.66
	N = 150; T = 40									
MLE	6	0	6	0.95	0.00	1	0	1	0.95	0.01
ABC (1)	0	0	0	1.00	0.94	0	0	0	0.95	0.92
ABC (2)	-0	0	0	1.00	0.95	-0	0	0	0.95	0.94
SPJ	-1	0	1	0.94	0.22	0	0	0	0.92	0.87
LPM (1)						-0	0	0	0.75	0.68
LPM (2)						-0	0	0	0.75	0.54
	N = 150; T = 50									
MLE	5	0	5	0.95	0.00	1	0	1	0.97	0.02
ABC (1)	0	0	0	0.99	0.93	0	0	0	0.97	0.93
ABC (2)	-0	0	0	0.99	0.94	-0	0	0	0.97	0.94
SPJ	-1	0	1	0.95	0.38	0	0	0	0.95	0.91
LPM (1)						-0	0	0	0.76	0.61
LPM (2)						-0	0	0	0.76	0.45

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; values in parentheses after ABC and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 5.17: *Properties: Dynamic (Three-Way) - $y_{ij(t-1)}$ - $N = 50$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 50; T = 10									
MLE	-62	5	62	0.95	0.00	-70	4	71	1.02	0.00
ABC (1)	-6	4	7	1.14	0.81	-7	5	8	1.11	0.76
ABC (2)	-7	5	9	1.05	0.68	-8	5	10	1.02	0.62
SPJ	24	6	25	0.77	0.01	-11	6	12	0.94	0.48
LPM (1)						2	5	5	1.02	0.95
LPM (2)						3	5	6	0.94	0.89
	N = 50; T = 20									
MLE	-27	4	27	0.94	0.00	-36	3	37	0.95	0.00
ABC (1)	-3	3	4	1.05	0.87	-3	3	5	1.00	0.85
ABC (2)	-1	3	3	1.00	0.94	-1	3	4	0.96	0.93
SPJ	5	4	6	0.89	0.69	-2	4	4	0.89	0.89
LPM (1)						8	3	9	0.95	0.28
LPM (2)						11	4	12	0.91	0.09
	N = 50; T = 30									
MLE	-16	3	16	0.97	0.00	-25	3	25	0.97	0.00
ABC (1)	-2	3	3	1.06	0.88	-2	3	3	1.01	0.87
ABC (2)	-0	3	3	1.03	0.95	-0	3	3	0.98	0.95
SPJ	2	3	3	0.95	0.88	-1	3	3	0.92	0.93
LPM (1)						10	3	11	0.96	0.03
LPM (2)						13	3	13	0.94	0.00
	N = 50; T = 40									
MLE	-11	2	11	0.96	0.01	-19	2	19	0.95	0.00
ABC (1)	-2	2	3	1.03	0.86	-2	2	3	0.99	0.85
ABC (2)	-0	2	2	1.01	0.95	-0	2	2	0.97	0.95
SPJ	1	2	3	0.93	0.92	-0	3	3	0.90	0.92
LPM (1)						11	2	12	0.95	0.01
LPM (2)						13	2	13	0.93	0.00
	N = 50; T = 50									
MLE	-7	2	8	0.94	0.07	-15	2	15	0.92	0.00
ABC (1)	-2	2	3	1.01	0.89	-2	2	3	0.95	0.87
ABC (2)	-0	2	2	0.99	0.95	-0	2	2	0.93	0.93
SPJ	0	2	2	0.92	0.92	-0	2	2	0.87	0.90
LPM (1)						12	2	12	0.92	0.00
LPM (2)						14	2	14	0.91	0.00

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; values in parentheses after ABC and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 5.18: *Properties: Dynamic (Three-Way) - $y_{ij(t-1)}$ - $N = 100$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 100; T = 10									
MLE	-63	3	63	0.98	0.00	-70	2	70	1.04	0.00
ABC (1)	-6	2	7	1.13	0.22	-8	2	8	1.10	0.09
ABC (2)	-8	2	8	1.04	0.08	-9	2	10	1.01	0.03
SPJ	21	3	21	0.80	0.00	-11	3	11	0.94	0.02
LPM (1)						2	2	3	1.00	0.84
LPM (2)						4	3	4	0.92	0.66
	N = 100; T = 20									
MLE	-29	2	29	0.96	0.00	-37	2	37	0.96	0.00
ABC (1)	-3	2	4	1.03	0.42	-4	2	4	0.99	0.37
ABC (2)	-1	2	2	0.99	0.86	-2	2	2	0.95	0.83
SPJ	4	2	5	0.91	0.26	-2	2	3	0.90	0.80
LPM (1)						8	2	9	0.95	0.00
LPM (2)						11	2	11	0.91	0.00
	N = 100; T = 30									
MLE	-18	1	18	0.97	0.00	-25	1	25	0.96	0.00
ABC (1)	-3	1	3	1.03	0.50	-3	1	3	0.98	0.49
ABC (2)	-1	1	1	1.00	0.93	-1	1	2	0.95	0.92
SPJ	2	1	2	0.94	0.72	-1	1	2	0.90	0.90
LPM (1)						10	1	10	0.95	0.00
LPM (2)						13	1	13	0.92	0.00
	N = 100; T = 40									
MLE	-13	1	13	1.01	0.00	-19	1	19	1.01	0.00
ABC (1)	-2	1	2	1.06	0.57	-2	1	2	1.04	0.56
ABC (2)	-0	1	1	1.04	0.94	-0	1	1	1.02	0.94
SPJ	1	1	2	0.98	0.86	-0	1	1	0.96	0.93
LPM (1)						11	1	11	0.98	0.00
LPM (2)						13	1	13	0.96	0.00
	N = 100; T = 50									
MLE	-10	1	10	0.98	0.00	-15	1	15	0.97	0.00
ABC (1)	-2	1	2	1.03	0.61	-2	1	2	0.99	0.62
ABC (2)	-0	1	1	1.01	0.95	-0	1	1	0.98	0.94
SPJ	1	1	1	0.98	0.91	-0	1	1	0.95	0.93
LPM (1)						12	1	12	0.94	0.00
LPM (2)						14	1	14	0.93	0.00

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; values in parentheses after ABC and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

Table 5.19: *Properties: Dynamic (Three-Way) - $y_{ij(t-1)}$ - $N = 150$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 150; T = 10									
MLE	-63	2	64	0.95	0.00	-70	1	70	1.02	0.00
ABC (1)	-7	1	7	1.09	0.01	-8	2	9	1.08	0.00
ABC (2)	-8	2	9	1.01	0.00	-10	2	10	1.00	0.00
SPJ	20	2	20	0.78	0.00	-11	2	11	0.92	0.00
LPM (1)						2	2	3	0.98	0.71
LPM (2)						3	2	4	0.90	0.42
	N = 150; T = 20									
MLE	-30	1	30	0.99	0.00	-37	1	37	1.00	0.00
ABC (1)	-4	1	4	1.07	0.05	-4	1	4	1.03	0.03
ABC (2)	-2	1	2	1.02	0.69	-2	1	2	0.99	0.61
SPJ	4	1	4	0.92	0.05	-2	1	2	0.90	0.61
LPM (1)						8	1	8	0.96	0.00
LPM (2)						11	1	11	0.92	0.00
	N = 150; T = 30									
MLE	-19	1	19	0.98	0.00	-25	1	25	0.97	0.00
ABC (1)	-3	1	3	1.04	0.15	-3	1	3	0.99	0.13
ABC (2)	-1	1	1	1.01	0.89	-1	1	1	0.97	0.87
SPJ	2	1	2	0.96	0.47	-0	1	1	0.92	0.90
LPM (1)						10	1	10	0.93	0.00
LPM (2)						13	1	13	0.91	0.00
	N = 150; T = 40									
MLE	-14	1	14	1.01	0.00	-19	1	19	0.99	0.00
ABC (1)	-2	1	2	1.06	0.20	-2	1	2	1.01	0.19
ABC (2)	-0	1	1	1.03	0.92	-0	1	1	0.99	0.90
SPJ	1	1	1	0.96	0.76	-0	1	1	0.93	0.92
LPM (1)						11	1	11	0.96	0.00
LPM (2)						13	1	13	0.94	0.00
	N = 150; T = 50									
MLE	-11	1	11	0.97	0.00	-15	1	15	0.95	0.00
ABC (1)	-2	1	2	1.01	0.30	-2	1	2	0.97	0.30
ABC (2)	-0	1	1	0.99	0.92	-0	1	1	0.95	0.91
SPJ	1	1	1	0.96	0.84	-0	1	1	0.92	0.92
LPM (1)						12	1	12	0.92	0.00
LPM (2)						14	1	14	0.90	0.00

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; values in parentheses after ABC and LPM indicate the chosen bandwidth parameter; results based on 1,000 repetitions.

C Further Monte Carlo Simulations

Although the main focus of our article is on the dynamic two- and three-way fixed effects model, the static counterparts are also highly relevant for applied work. For this reason, we study the finite sample properties of MLE, ABC, SPJ and LPM for these model specifications, too. In the following we briefly sketch the designs. Let $i = 1,\dots,N$, $j = 1,\dots,N$, $t = 1,\dots,T$, $\beta_y = 0.5$, $\beta = 1$.

Design - Two-way fixed effects

$$y_{ijt} = \mathbf{1}\left[\beta x_{ijt} + \lambda_{it} + \psi_{jt} \geq \epsilon_{ijt}\right],$$

where $\lambda_{it} \sim$ iid. $\mathcal{N}(0, 1/16)$, $\psi_{jt} \sim$ iid. $\mathcal{N}(0, 1/16)$, and $\epsilon_{ijt} \sim$ iid. $\mathcal{N}(0,1)$. Further, $x_{ijt} = 0.5x_{ijt-1} + \lambda_{it} + \psi_{jt} + \nu_{ijt}$, where $\nu_{ijt} \sim$ iid. $\mathcal{N}(0, 0.5)$, $x_{ij0} \sim$ iid. $\mathcal{N}(0,1)$.

Design - Three-way fixed effects

$$y_{ijt} = \mathbf{1}\left[\beta x_{ijt} + \lambda_{it} + \psi_{jt} + \mu_{ij} \geq \epsilon_{ijt}\right],$$

where $\lambda_{it} \sim$ iid. $\mathcal{N}(0, 1/24)$, $\psi_{jt} \sim$ iid. $\mathcal{N}(0, 1/24)$, $\mu_{ij} \sim$ iid. $\mathcal{N}(0, 1/24)$, and $\epsilon_{ijt} \sim$ iid. $\mathcal{N}(0,1)$. Further, $x_{ijt} = 0.5x_{ijt-1} + \lambda_{it} + \psi_{jt} + \mu_{ij} + \nu_{ijt}$, where $\nu_{ijt} \sim$ iid. $\mathcal{N}(0, 0.5)$, $x_{ij0} \sim$ iid. $\mathcal{N}(0,1)$.

Note that unlike in the dynamic three-way fixed effects model, the OLS estimator of the linear probability model (LPM) does not require a bias correction for the specifications considered in this section.

We now review the key results of the simulation experiments.

Results - Two-way fixed effects

Static (see tables 5.20, 5.21, 5.22): Although MLE shows a distortion in the structural parameter estimates, the bias does not carry over to the estimates of APEs. The bias corrections ABC and SPJ work well. They reduce the biases of the structural parameters and APEs to 1 or zero percent, and bring the CPs close to the nominal level. Overall, ABC, SPJ and MLE work similarly well if APEs are of interest. In terms of structural parameters, ABC exhibits a lower bias and better CPs than SPJ in samples with smaller N. LPM shows no distortion of the APEs in all settings, but we observe that with increasing N, the standard errors are underestimated, resulting in too low CPs.

Note that MLE is consistent under fixed T asymptotics. This is also evident from the simulation results, where the properties of the estimator do not change with T.

Results - Three-way fixed effects

Static (see tables 5.23, 5.24, 5.25): We find a considerable distortion in the MLE estimates of the structural parameters, which decreases with rising T, but is not negligibly small even at $T = 50$. ABC and SPJ both reduce this bias considerably, but ABC works better in samples with smaller T. While the CPs of ABC quickly converge to the nominal level, the CPs of SPJ are still far away from 95 percent even at $T = 50$. If we look at the APEs, we see that all estimators have either a very small bias of 1 percent or none at all. With increasing T, their CPs are also getting closer to 95 percent.

Table 5.20: *Properties: Static (Two-Way) - x_{ijt} - $N = 50$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 50; T = 10									
MLE	5	2	5	0.97	0.10	0	1	1	0.98	0.94
ABC	-0	1	1	1.01	0.94	-0	1	1	0.99	0.94
SPJ	-1	1	2	0.98	0.93	-0	1	1	0.96	0.93
LPM						0	1	1	0.96	0.93
	N = 50; T = 20									
MLE	5	1	5	0.99	0.01	0	1	1	1.06	0.96
ABC	-0	1	1	1.03	0.96	-0	1	1	1.07	0.97
SPJ	-1	1	1	0.98	0.91	-0	1	1	1.04	0.95
LPM						-0	1	1	1.05	0.96
	N = 50; T = 30									
MLE	5	1	5	0.98	0.00	0	1	1	1.01	0.95
ABC	-0	1	1	1.02	0.95	-0	1	1	1.03	0.95
SPJ	-1	1	1	1.00	0.89	-0	1	1	1.00	0.95
LPM						0	1	1	0.99	0.94
	N = 50; T = 40									
MLE	5	1	5	0.94	0.00	0	1	1	0.98	0.95
ABC	-0	1	1	0.97	0.94	-0	1	1	0.99	0.95
SPJ	-1	1	1	0.95	0.84	-0	1	1	0.97	0.94
LPM						-0	1	1	0.97	0.94
	N = 50; T = 50									
MLE	5	1	5	0.97	0.00	0	1	1	1.02	0.95
ABC	-0	1	1	1.01	0.96	-0	1	1	1.04	0.96
SPJ	-1	1	1	0.98	0.83	-0	1	1	1.00	0.95
LPM						0	1	1	1.00	0.95

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.21: *Properties: Static (Two-Way) - x_{ijt} - $N = 100$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 100; T = 10									
MLE	2	1	2	0.95	0.13	0	1	1	0.96	0.94
ABC	-0	1	1	0.97	0.94	-0	1	1	0.96	0.93
SPJ	-0	1	1	0.95	0.93	-0	1	1	0.95	0.93
LPM						0	1	1	0.85	0.90
	N = 100; T = 20									
MLE	2	1	2	0.98	0.00	0	0	0	0.99	0.96
ABC	-0	1	1	1.00	0.95	-0	0	0	1.00	0.95
SPJ	-0	1	1	0.99	0.94	-0	0	0	0.99	0.95
LPM						-0	0	0	0.89	0.92
	N = 100; T = 30									
MLE	2	0	2	1.00	0.00	0	0	0	1.03	0.95
ABC	0	0	0	1.02	0.96	-0	0	0	1.03	0.95
SPJ	-0	0	0	1.00	0.95	-0	0	0	1.03	0.96
LPM						0	0	0	0.92	0.93
	N = 100; T = 40									
MLE	2	0	2	0.98	0.00	0	0	0	0.97	0.94
ABC	-0	0	0	1.00	0.94	-0	0	0	0.97	0.94
SPJ	-0	0	0	0.98	0.93	-0	0	0	0.96	0.94
LPM						-0	0	0	0.87	0.91
	N = 100; T = 50									
MLE	2	0	2	1.00	0.00	0	0	0	0.99	0.95
ABC	-0	0	0	1.02	0.96	-0	0	0	0.99	0.95
SPJ	-0	0	0	1.02	0.94	-0	0	0	0.99	0.95
LPM						-0	0	0	0.88	0.92

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.22: *Properties: Static (Two-Way) - x_{ijt} - $N = 150$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 150; T = 10									
MLE	1	0	2	0.99	0.12	0	0	0	1.02	0.96
ABC	-0	0	0	1.01	0.96	-0	0	0	1.02	0.96
SPJ	-0	0	0	1.00	0.95	-0	0	0	1.01	0.95
LPM						-0	0	0	0.84	0.90
	N = 150; T = 20									
MLE	1	0	2	0.95	0.01	0	0	0	0.95	0.94
ABC	0	0	0	0.96	0.94	0	0	0	0.95	0.94
SPJ	-0	0	0	0.95	0.93	0	0	0	0.95	0.94
LPM						0	0	0	0.79	0.86
	N = 150; T = 30									
MLE	1	0	2	1.01	0.00	0	0	0	0.96	0.95
ABC	-0	0	0	1.03	0.95	-0	0	0	0.97	0.94
SPJ	-0	0	0	1.02	0.94	-0	0	0	0.96	0.95
LPM						-0	0	0	0.79	0.88
	N = 150; T = 40									
MLE	1	0	2	0.99	0.00	0	0	0	0.97	0.94
ABC	-0	0	0	1.00	0.95	-0	0	0	0.97	0.94
SPJ	-0	0	0	0.99	0.94	-0	0	0	0.96	0.94
LPM						-0	0	0	0.80	0.88
	N = 150; T = 50									
MLE	1	0	2	0.99	0.00	0	0	0	0.95	0.94
ABC	-0	0	0	1.00	0.94	-0	0	0	0.95	0.94
SPJ	-0	0	0	0.99	0.94	-0	0	0	0.95	0.94
LPM						0	0	0	0.78	0.88

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.23: *Properties: Static (Three-Way) - x_{ijt} - $N = 50$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 50; T = 10									
MLE	22	2	22	0.85	0.00	1	1	2	1.00	0.89
ABC	-1	2	2	1.03	0.88	-1	1	2	1.07	0.86
SPJ	-12	2	12	0.72	0.00	-0	2	2	0.88	0.91
LPM						0	1	1	1.04	0.96
	N = 50; T = 20									
MLE	12	1	12	0.92	0.00	0	1	1	1.00	0.94
ABC	-1	1	1	1.02	0.92	-0	1	1	1.03	0.93
SPJ	-4	1	4	0.88	0.08	-1	1	1	0.92	0.89
LPM						-0	1	1	1.04	0.96
	N = 50; T = 30									
MLE	10	1	10	0.94	0.00	0	1	1	1.02	0.94
ABC	-0	1	1	1.02	0.94	-0	1	1	1.04	0.94
SPJ	-2	1	2	0.93	0.28	-0	1	1	0.96	0.89
LPM						0	1	1	1.01	0.95
	N = 50; T = 40									
MLE	8	1	8	0.93	0.00	0	1	1	1.02	0.95
ABC	-0	1	1	0.99	0.92	-0	1	1	1.03	0.94
SPJ	-2	1	2	0.94	0.40	-0	1	1	0.99	0.90
LPM						-0	1	1	0.98	0.94
	N = 50; T = 50									
MLE	8	1	8	0.96	0.00	0	1	1	1.04	0.94
ABC	-0	1	1	1.03	0.93	-0	1	1	1.06	0.95
SPJ	-1	1	2	0.95	0.46	-0	1	1	0.99	0.91
LPM						0	1	1	0.99	0.94

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.24: *Properties: Static (Three-Way) - x_{ijt} - $N = 100$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 100; T = 10									
MLE	18	1	18	0.89	0.00	1	1	1	1.05	0.90
ABC	-1	1	1	1.05	0.80	-1	1	1	1.09	0.70
SPJ	-8	1	8	0.74	0.00	0	1	1	0.89	0.87
LPM						-0	1	1	1.04	0.96
	N = 100; T = 20									
MLE	9	1	9	0.93	0.00	0	0	1	1.01	0.92
ABC	-0	1	1	1.00	0.92	-0	0	1	1.03	0.92
SPJ	-2	1	2	0.94	0.01	-0	1	1	0.96	0.91
LPM						0	0	0	0.96	0.93
	N = 100; T = 30									
MLE	7	0	7	0.95	0.00	0	0	0	1.05	0.95
ABC	-0	0	0	1.01	0.93	-0	0	0	1.06	0.95
SPJ	-1	0	1	0.93	0.21	-0	0	0	0.98	0.92
LPM						-0	0	0	0.97	0.95
	N = 100; T = 40									
MLE	6	0	6	0.96	0.00	0	0	0	1.00	0.94
ABC	-0	0	0	1.00	0.94	-0	0	0	1.01	0.94
SPJ	-1	0	1	0.95	0.44	-0	0	0	0.95	0.92
LPM						-0	0	0	0.93	0.93
	N = 100; T = 50									
MLE	5	0	5	0.94	0.00	0	0	0	0.99	0.94
ABC	-0	0	0	0.98	0.94	-0	0	0	1.00	0.94
SPJ	-1	0	1	0.94	0.57	-0	0	0	0.97	0.92
LPM						-0	0	0	0.91	0.93

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

Table 5.25: *Properties: Static (Three-Way) - x_{ijt} - $N = 150$*

	Coefficients					APE				
	Bias	SD	RMSE	SE/SD	CP .95	Bias	SD	RMSE	SE/SD	CP .95
	N = 150; T = 10									
MLE	16	1	16	0.87	0.00	0	0	1	1.04	0.87
ABC	-1	1	1	1.02	0.77	-1	0	1	1.07	0.51
SPJ	-7	1	7	0.76	0.00	1	1	1	0.91	0.73
LPM						-0	0	0	0.95	0.94
	N = 150; T = 20									
MLE	8	0	8	0.92	0.00	0	0	0	1.00	0.91
ABC	-0	0	0	0.99	0.91	-0	0	0	1.01	0.89
SPJ	-2	0	2	0.89	0.00	-0	0	0	0.93	0.91
LPM						-0	0	0	0.93	0.93
	N = 150; T = 30									
MLE	6	0	6	0.93	0.00	0	0	0	0.97	0.93
ABC	-0	0	0	0.97	0.93	-0	0	0	0.98	0.92
SPJ	-1	0	1	0.93	0.08	-0	0	0	0.92	0.90
LPM						-0	0	0	0.88	0.92
	N = 150; T = 40									
MLE	5	0	5	0.95	0.00	0	0	0	1.01	0.93
ABC	-0	0	0	0.99	0.94	-0	0	0	1.02	0.94
SPJ	-1	0	1	0.93	0.33	-0	0	0	0.98	0.93
LPM						-0	0	0	0.90	0.93
	N = 150; T = 50									
MLE	4	0	4	0.98	0.00	0	0	0	1.05	0.95
ABC	-0	0	0	1.01	0.94	-0	0	0	1.05	0.95
SPJ	-0	0	0	0.97	0.51	-0	0	0	1.00	0.94
LPM						-0	0	0	0.92	0.92

Note: Bias, SD, and RMSE denote biases, standard deviations, and root mean squared errors in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; MLE, ABC, SPJ, and LPM denote the (bias-corrected) estimators; results based on 1,000 repetitions.

D Application

Table 5.26: *Logit: Coefficients*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	-	-	2.869***	-	1.929***
	[-]	[-]	[2.985]	[-]	[1.798]
	(-)	(-)	(0.008)	(-)	(0.009)
log(Distance)	-	-1.454***	-0.980***	-	-
	[-1.181***]	[-1.494]	[-1.012]	[-]	[-]
	(0.005)	(0.006)	(0.007)	(-)	(-)
Land border	-	0.621***	0.231***	-	-
	[0.660***]	[0.643]	[0.244]	[-]	[-]
	(0.026)	(0.029)	(0.033)	(-)	(-)
Legal	-	0.262***	0.169***	-	-
	[0.172***]	[0.269]	[0.176]	[-]	[-]
	(0.007)	(0.008)	(0.009)	(-)	(-)
Language	-	0.737***	0.514***	-	-
	[0.663***]	[0.757]	[0.529]	[-]	[-]
	(0.009)	(0.01)	(0.012)	(-)	(-)
Colonial ties	-	1.345***	1.002***	-	-
	[0.342***]	[1.443]	[1.102]	[-]	[-]
	(0.036)	(0.061)	(0.07)	(-)	(-)
Currency Union	-	1.137***	0.775***	0.578***	0.421***
	[0.660***]	[1.173]	[0.807]	[0.64]	[0.497]
	(0.021)	(0.027)	(0.031)	(0.06)	(0.064)
FTA	-	1.059***	0.664***	0.130*	0.072
	[0.955***]	[1.077]	[0.674]	[0.123]	[0.054]
	(0.031)	(0.036)	(0.04)	(0.07)	(0.075)
WTO	-	0.228***	0.187***	0.095***	0.087***
	[0.462***]	[0.232]	[0.191]	[0.105]	[0.102]
	(0.009)	(0.014)	(0.016)	(0.028)	(0.031)
Fixed effects	i,j,t	it,jt	it,jt	it,jt,ij	it,jt,ij
Sample size	1,204,671	1,204,671	1,171,794	1,204,671	1,171,794
- perf. class.	12,298	147,760	141,537	370,617	374,067
Deviance	8.857×10^5	6.976×10^5	5.2×10^5	4.728×10^5	4.184×10^5

Note: Uncorrected coefficients in square brackets; standard errors in parentheses.

Table 5.27: *Logit: Average Partial Effects*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	-	-	0.331^{***}	-	0.168^{***}
	[-]	[-]	[0.332]	[-]	[0.13]
	(-)	(-)	(0.002)	(-)	(0.049)
log(Distance)	-	-0.138^{***}	-0.067^{***}	-	-
	$[-0.140^{***}]$	[-0.137]	[-0.067]	[-]	[-]
	(0.005)	(0.005)	(0.001)	(-)	(-)
Land border	-	0.058^{***}	0.016^{***}	-	-
	$[0.077^{***}]$	[0.059]	[0.016]	[-]	[-]
	(0.004)	(0.004)	(0.003)	(-)	(-)
Legal	-	0.025^{***}	0.012^{***}	-	-
	$[0.020^{***}]$	[0.025]	[0.012]	[-]	[-]
	(0.001)	(0.001)	(0.001)	(-)	(-)
Language	-	0.069^{***}	0.035^{***}	-	-
	$[0.078^{***}]$	[0.069]	[0.035]	[-]	[-]
	(0.003)	(0.001)	(0.001)	(-)	(-)
Colonial ties	-	0.122^{***}	0.069^{***}	-	-
	$[0.040^{***}]$	[0.127]	[0.074]	[-]	[-]
	(0.004)	(0.006)	(0.006)	(-)	(-)
Currency Union	-	0.104^{***}	0.053^{***}	0.041^{***}	0.027^{***}
	$[0.077^{***}]$	[0.104]	[0.054]	[0.04]	[0.028]
	(0.004)	(0.003)	(0.002)	(0.006)	(0.009)
FTA	-	0.098^{***}	0.046^{***}	0.009	0.004
	$[0.110^{***}]$	[0.097]	[0.045]	[0.008]	[0.003]
	(0.005)	(0.004)	(0.003)	(0.006)	(0.006)
WTO	-	0.022^{***}	0.013^{***}	0.007^{**}	0.005^{*}
	$[0.056^{***}]$	[0.021]	[0.013]	[0.006]	[0.006]
	(0.002)	(0.002)	(0.001)	(0.003)	(0.003)
Fixed effects	i,j,t	it,jt	it,jt	it,jt,ij	it,jt,ij
Sample size	1,204,671	1,204,671	1,171,794	1,204,671	1,171,794
- perf. class.	12,298	147,760	141,537	370,617	374,067
Deviance	8.857×10^5	6.976×10^5	5.2×10^5	4.728×10^5	4.184×10^5

Notes: Uncorrected average partial effects in square brackets; standard errors in parentheses.

Table 5.28: *Probit with Different Bandwidths: Coefficients*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	0.961^{***}	1.112^{***}	1.140^{***}	1.154^{***}	1.161^{***}
	(0.036)	(0.037)	(0.039)	(0.04)	(0.04)
Currency Union	0.228^{***}	0.217^{***}	0.214^{***}	0.214^{***}	0.216^{***}
	(0.05)	(0.048)	(0.048)	(0.048)	(0.047)
FTA	0.035	0.037	0.038	0.042	0.043
	(0.056)	(0.054)	(0.053)	(0.053)	(0.053)
WTO	0.041	0.039	0.039	0.040	0.042^{*}
	(0.026)	(0.025)	(0.025)	(0.025)	(0.025)
Trim	$L = 0$	$L = 1$	$L = 2$	$L = 3$	$L = 4$

Note: All columns include exporter-time, importer-time, and pair fixed effects; standard errors in parentheses.

Table 5.29: *Probit with Different Bandwidths: Average Partial Effects*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	0.144^{***}	0.173^{***}	0.179^{***}	0.182^{***}	0.183^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Currency Union	0.026^{***}	0.025^{***}	0.024^{***}	0.024^{***}	0.025^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
FTA	0.004	0.004	0.004	0.005	0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
WTO	0.005^{**}	0.004^{**}	0.004^{**}	0.005^{**}	0.005^{**}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Trim	$L = 0$	$L = 1$	$L = 2$	$L = 3$	$L = 4$

Note: All columns include exporter-time, importer-time, and pair fixed effects; standard errors in parentheses.

Table 5.30: *Logit with Different Bandwidths: Coefficients*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	1.606***	1.879***	1.929***	1.953***	1.965***
	(0.037)	(0.038)	(0.039)	(0.04)	(0.04)
Currency Union	0.448***	0.426***	0.421***	0.421***	0.425***
	(0.057)	(0.054)	(0.054)	(0.054)	(0.053)
FTA	0.065	0.069	0.072	0.077	0.080
	(0.063)	(0.061)	(0.06)	(0.06)	(0.06)
WTO	0.091***	0.087***	0.087***	0.088***	0.091***
	(0.028)	(0.027)	(0.027)	(0.027)	(0.027)
Trim	$L = 0$	$L = 1$	$L = 2$	$L = 3$	$L = 4$

Note: All columns include exporter-time, importer-time, and pair fixed effects; standard errors in parentheses.

Table 5.31: *Logit with Different Bandwidths: Average Partial Effects*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	0.133***	0.162***	0.168***	0.170***	0.172***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Currency Union	0.028***	0.027***	0.027***	0.027***	0.027***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
FTA	0.004	0.004	0.004	0.005	0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
WTO	0.006***	0.005***	0.005***	0.006***	0.006***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Trim	$L = 0$	$L = 1$	$L = 2$	$L = 3$	$L = 4$

Note: All columns include exporter-time, importer-time, and pair fixed effects; standard errors in parentheses.

Table 5.32: *Probit vs. LPM (Two-Way): Average Partial Effects*

	Dependent variable: y_{ijt}			
	(1)	(2)	(3)	(5)
$y_{ij(t-1)}$	-	-	0.599***	0.346***
	(-)	(-)	(0.001)	(0.003)
log(Distance)	-0.133***	-0.135***	-0.053***	-0.066***
	(0.001)	(0.005)	(0)	(0.001)
Land border	0.014***	0.035***	0.003*	0.015***
	(0.002)	(0.004)	(0.002)	(0.003)
Legal	0.008***	0.023***	0.002***	0.011***
	(0.001)	(0.001)	(0.001)	(0.001)
Language	0.098***	0.071***	0.040***	0.035***
	(0.001)	(0.001)	(0.001)	(0.001)
Colonial ties	0.021***	0.107***	0.008***	0.061***
	(0.003)	(0.007)	(0.002)	(0.005)
Currency Union	0.107***	0.103***	0.046***	0.053***
	(0.003)	(0.003)	(0.002)	(0.002)
FTA	-0.155***	0.090***	-0.063***	0.045***
	(0.002)	(0.004)	(0.002)	(0.003)
WTO	-0.010***	0.026***	-0.008***	0.013***
	(0.001)	(0.002)	(0.001)	(0.001)
Estimator	LPM	Probit	LPM	Probit
bias-corrected	false	true	false	true
Sample size	1204671	1204671	1171794	1171794

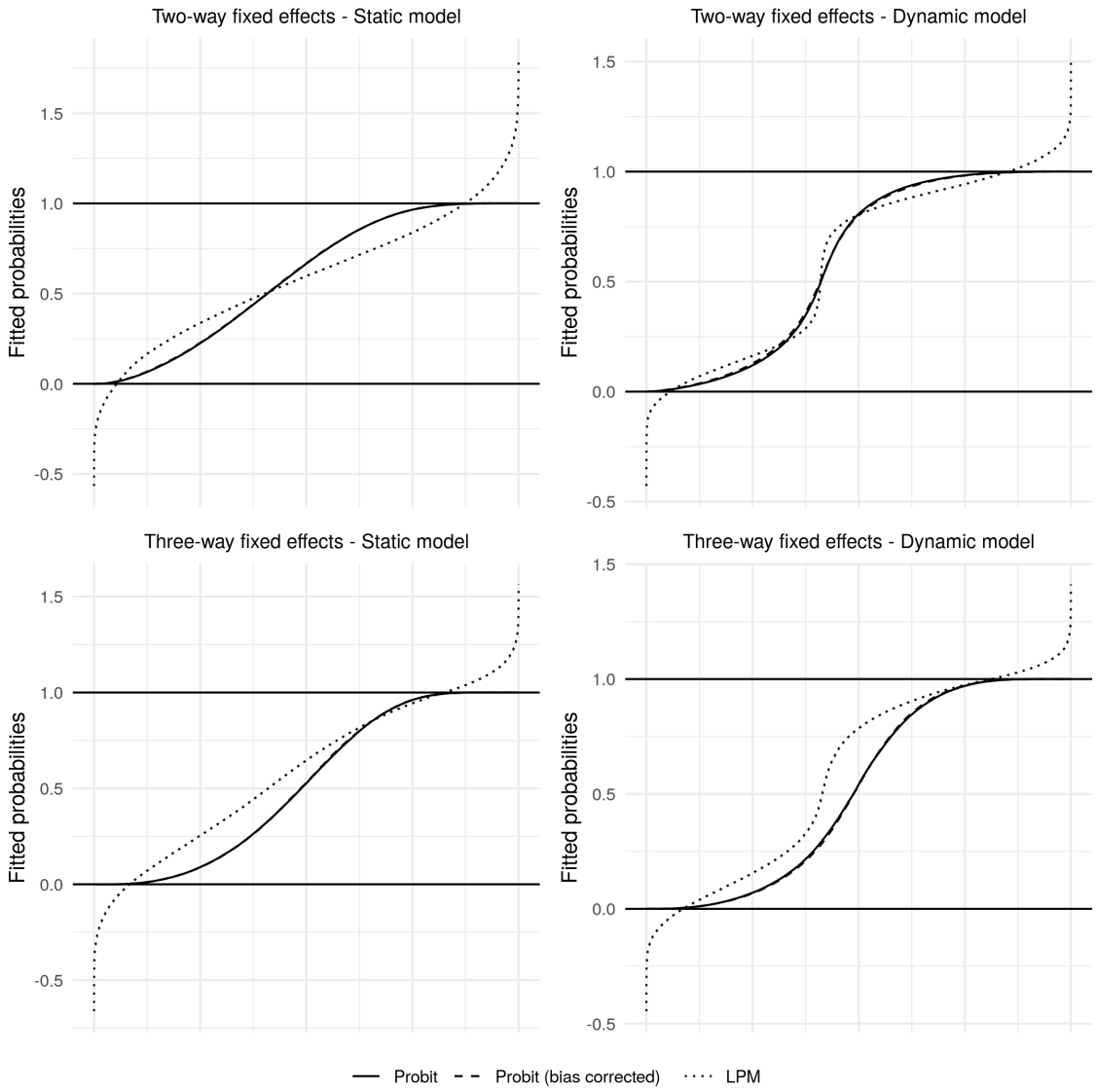
Note: All columns include exporter-time and importer-time fixed effects; standard errors in parentheses.

Table 5.33: *LPM with Different Bandwidths: Average Partial Effects*

	Dependent variable: y_{ijt}				
	(1)	(2)	(3)	(4)	(5)
$y_{ij(t-1)}$	0.444***	0.466***	0.474***	0.480***	0.485***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Currency Union	0.008***	0.008**	0.008**	0.008**	0.008**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
FTA	-0.065***	-0.062***	-0.062***	-0.061***	-0.061***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
WTO	0.008***	0.008***	0.008***	0.008***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Trim	$L = 0$	$L = 1$	$L = 2$	$L = 3$	$L = 4$

Note: All columns include exporter-time, importer-time, and pair fixed effects; standard errors in parentheses.

Figure 5.6: *Fitted Probabilities*



Chapter 6

Conclusion

This thesis focused on solving computational and statistical challenges arising from the application of nonlinear fixed effects models. First, computationally efficient algorithms were derived that allow to estimate these models even in the presence of high-dimensional fixed effects. Additionally it was shown how these algorithms can be combined with and adapted to bias corrections to mitigate the incidental parameters problem. Second, this thesis provided deeper insights on the finite sample properties of various bias corrections. Third, novel three-way fixed effects bias corrections were proposed that are particularly relevant in international trade.

The algorithms presented in this thesis are directly applicable to every generalized linear model. However, it would be beneficial to derive computationally efficient fixed effects algorithms for other popular nonlinear models like tobit or ordered logit. Further, a theoretical proof for the three-way bias corrections proposed in chapter 5 would still be useful, although their validity was credibly demonstrated through extensive simulation experiments. Overall, a major finding of this thesis is that especially analytical bias corrections are promising candidates for empirical research. This encourages the development of novel bias corrections for more complex error components that have not been studied so far.

Appendix A

R-Package *bife*

The *R*-package *bife* was developed in the context of chapter 2. It estimates fixed effects binary choice models (logit and probit) with potentially many individual fixed effects and computes average partial effects. The incidental parameter bias can be reduced with an asymptotic bias correction proposed by Fernández-Val (2009).

The corresponding user manual is provided in the following.

Package 'bife'

May 25, 2019

Type Package

Title Binary Choice Models with Fixed Effects

Version 0.6

Description Estimates fixed effects binary choice models (logit and probit) with potentially many individual fixed effects and computes average partial effects. Incidental parameter bias can be reduced with an asymptotic bias-correction proposed by Fernandez-Val (2009) <doi:10.1016/j.jeconom.2009.02.007>.

License GPL (>= 2)

Depends R (>= 3.1.0)

Imports data.table, Formula, Rcpp, stats

LinkingTo Rcpp, RcppArmadillo

URL https://github.com/amrei-stammann/bife

BugReports https://github.com/amrei-stammann/bife/issues

RoxygenNote 6.1.1

LazyData true

Suggests alpaca, knitr

VignetteBuilder knitr

NeedsCompilation yes

Author Amrei Stammann [aut, cre],
Daniel Czarnowske [aut] (<https://orcid.org/0000-0002-0030-929X>),
Florian Heiss [aut],
Daniel McFadden [ctb]

Maintainer Amrei Stammann <amrei.stammann@hhu.de>

Repository CRAN

Date/Publication 2019-05-24 23:50:17 UTC

R topics documented:

`bias_corr` *Asymptotic bias-correction for binary choice Models with fixed effects*

Description

`bias_corr` is a post-estimation routine that can be used to substantially reduce the incidental parameter bias problem (Neyman and Scott (1948)) present in non-linear fixed effects models (see Fernandez-Val and Weidner (2018) for an overview). The command applies the analytical bias-correction derived by Fernandez-Val (2009) to obtain bias-corrected estimates of the structural parameters.

Remark: Fernandez-Val (2009) further refined the bias-correction of Hahn and Newey (2004). The correction is now also applicable to dynamic models.

Usage

```
bias_corr(object, L = 0L)
```

Arguments

`object`	an object of class `"bife"`.
`L`	unsigned integer indicating a bandwidth for the estimation of spectral densities proposed by Hahn and Kuersteiner (2011). Default is zero, which should be used if all regressors are assumed to be strictly exogenous. In the presence of weakly exogenous or predetermined regressors, Fernandez-Val and Weidner (2018) suggest to choose a bandwidth not higher than four.

Value

The function bias_corr returns a named list of class "bife".

References

Fernandez-Val, I. (2009). "Fixed effects estimation of structural parameters and marginal effects in panel probit models". Journal of Econometrics 150(1), 71-85.

Fernandez-Val, I. and Weidner, M. (2018). "Fixed effects estimation of large-t panel data models". Annual Review of Economics, 10, 109-138.

Hahn, J. and Kuersteiner, G. (2011). "Bias reduction for dynamic nonlinear panel models with fixed effects". Econometric Theory, 27(6), 1152-1191.

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

Neyman, J. and Scott, E. L. (1948). "Consistent estimates based on partially consistent observations". Econometrica, 16(1), 1-32.

Stammann, A., Heiss, F., and and McFadden, D. (2016). "Estimating Fixed Effects Logit Models with Large Panel Data". Working paper.

See Also

bife

Examples

```
# Load 'psid' dataset
library(bife)
dataset <- psid

# Fit a static logit model
mod <- bife(LFP ~ I(AGE^2) + log(INCH) + KID1 + KID2 + KID3 + factor(TIME) | ID, dataset)
summary(mod)

# Apply analytical bias-correction
mod_bc <- bias_corr(mod)
summary(mod_bc)
```

---

bife *Efficiently fit binary choice models with fixed effects*

Description

bife can be used to fit fixed effects binary choice models (logit and probit) based on an unconditional maximum likelihood approach. It is tailored for the fast estimation of binary choice models with potentially many individual fixed effects. The routine is based on a special pseudo demeaning algorithm derived by Stammann, Heiss, and McFadden (2016). The estimates obtained are identical to the ones of glm, but the computation time of bife is much lower.

Remark: The term fixed effect is used in econometrician's sense of having a full set of individual specific intercepts. All other parameters in the model are referred to as structural parameters.

Usage

```
bife(formula, data = list(), model = c("logit", "probit"),
  beta_start = NULL, control = list(), bias_corr = NULL,
  tol_demeaning = NULL, iter_demeaning = NULL, tol_offset = NULL,
  iter_offset = NULL)
```

Arguments

formula	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. formula must be of type $y\ x\|id$ where the id refers to an individual identifier (fixed effect category).
data	an object of class "data.frame" containing the variables in the model.
model	the description of the error distribution and link function to be used in the model. For bife this has to be a character string naming the model function. Default is "logit".
beta_start	an optional vector of starting values used for the structural parameters in the optimization algorithm. Default is zero for all structural parameters.
control	a named list of parameters for controlling the fitting process. See bife_control for details.
bias_corr	deprecated; see bias_corr.
tol_demeaning, iter_demeaning, tol_offset, iter_offset	deprecated; see bife_control.

Details

bife drops all observations of cross-sectional units (individuals) with non-varying response. This can de done because these observations do not contribute to the identification of the structural parameters (perfect classification).

If bife does not converge this is usually a sign of linear dependence between one or more regressors and the fixed effects. In this case, you should carefully inspect your model specification.

Value

The function bife returns a named list of class "bife".

References

Stammann, A., Heiss, F., and and McFadden, D. (2016). "Estimating Fixed Effects Logit Models with Large Panel Data". Working paper.

Examples

```
# Load 'psid' dataset
library(bife)
dataset <- psid

# Fit a static logit model
mod <- bife(LFP ~ I(AGE^2) + log(INCH) + KID1 + KID2 + KID3 + factor(TIME) | ID, dataset)
summary(mod)
```

`bife_control` *Set* `bife` *Control Parameters*

Description

Set and change parameters used for fitting `bife`.

Usage

```
bife_control(dev_tol = 1e-08, rho_tol = 1e-04, conv_tol = 1e-06,
  iter_max = 100L, trace = FALSE)
```

Arguments

`dev_tol`	tolerance level for the first stopping condition of the maximization routine. The stopping condition is based on the relative change of the deviance in iteration r and can be expressed as follows: $(dev_{r-1} - dev_r)/(0.1 + dev_r) < tol$. Default is `1.0e-08`.
`rho_tol`	tolerance level for the stephalving in the maximization routine. Stephalving only takes place if the deviance in iteration r is larger than the one of the previous iteration. If this is the case, $\|\|\boldsymbol{\beta}_r - \boldsymbol{\beta}_{r-1}\|\|_2$ is halfed until the deviance is less or numerically equal compared to the deviance of the previous iteration. Stephalving fails if the the following condition holds: $\rho < tol$, where ρ is the stepcorrection factor. If stephalving fails the maximization routine is canceled. Default is `1.0e-04`.
`conv_tol`	tolerance level that accounts for rounding errors inside the stephalving routine when comparing the deviance with the one of the previous iteration. Default is `1.0e-06`.
`iter_max`	unsigned integer indicating the maximum number of iterations in the maximization routine. Default is `100L`.
`trace`	logical indicating if output should be produced in each iteration. Default is `FALSE`.

Value

The function bife_control returns a named list of control parameters.

See Also

bife

---

coef.bife *Extract estimates of structural parameters or fixed effects*

Description

coef.bife is a generic function which extracts estimates of the structural parameters or fixed effects from objects returned by bife.

Usage

```
## S3 method for class 'bife'
coef(object, type = c("sp", "fe"), corrected = NULL,
  fixed = NULL, ...)
```

Arguments

object	an object of class "bife".
type	the type of parameter estimates that should be returned; structural parameters or fixed effects. Default is "sp" referring to the structural parameters.
corrected, fixed	deprecated.
...	other arguments.

Value

The function coef.bife returns a named vector of estimates of the requested parameters.

See Also

bife

coef.bifeAPEs *Extract estimates of average partial effects*

Description

coef.bifeAPEs is a generic function which extracts estimates of the average partial effects from objects returned by get_APEs.

Usage

```
## S3 method for class 'bifeAPEs'
coef(object, ...)
```

Arguments

object	an object of class "APEs".
...	other arguments.

Value

The function coef.bifeAPEs returns a named vector of estimates of the average partial effects.

See Also

get_APEs

fitted.bife *Extract* bife *fitted values*

Description

fitted.bife is a generic function which extracts fitted values from an object returned by bife.

Usage

```
## S3 method for class 'bife'
fitted(object, ...)
```

Arguments

object	an object of class "bife".
...	other arguments.

Value

The function fitted.bife returns a vector of fitted values.

See Also

bife

---

get_APEs *Compute average partial effects for binary choice models with fixed effects*

Description

get_APEs is a post-estimation routine that can be used to estimate average partial effects with respect to all covariates in the model and the corresponding covariance matrix. The estimation of the covariance is based on a linear approximation (delta method). Note that the command automatically determines which of the regressors are continuous or binary.

Remark: The routine currently does not allow to compute average partial effects based on functional forms like interactions and polynomials.

Note: apeff_bife is deprecated and will be removed soon.

Usage

```
get_APEs(object, n_pop = NULL, weak_exo = FALSE)

apeff_bife(...)
```

Arguments

object	an object of class "bife".
n_pop	unsigned integer indicating a finite population correction for the estimation of the covariance matrix of the average partial effects proposed by Cruz-Gonzalez, Fernandez-Val, and Weidner (2017). The correction factor is computed as follows: $(n^* - n)/(n^* - 1)$, where n^* and n are the size of the entire population and the full sample size. Default is NULL, which refers to a factor of one and is equal to an infinitely large population.
weak_exo	logical indicating if some of the regressors are assumed to be weakly exogenous (e.g. predetermined). If object is returned by bias_corr, the option will be automatically set to TRUE if the choosen bandwidth parameter is larger than zero. Note that this option only affects the estimation of the covariance matrix. Default is FALSE, which assumes that all regressors are strictly exogenous.
...	arguments passed to the deprecated function apeff_bife.

Value

The function get_APEs returns a named list of class "bifeAPEs".

References

Cruz-Gonzalez, M., Fernandez-Val, I., and Weidner, M. (2017). "Bias corrections for probit and logit models with two-way fixed effects". The Stata Journal, 17(3), 517-545.

Fernandez-Val, I. (2009). "Fixed effects estimation of structural parameters and marginal effects in panel probit models". Journal of Econometrics 150(1), 71-85.

Fernandez-Val, I. and Weidner M. (2018). "Fixed effects estimation of large-t panel data models". Annual Review of Economics, 10, 109-138.

Neyman, J. and Scott, E. L. (1948). "Consistent estimates based on partially consistent observations". Econometrica, 16(1), 1-32.

Stammann, A., Heiss, F., and and McFadden, D. (2016). "Estimating Fixed Effects Logit Models with Large Panel Data". Working paper.

See Also

bias_corr, bife

Examples

```
# Load 'psid' dataset
library(bife)
dataset <- psid

# Fit a static logit model
mod <- bife(LFP ~ I(AGE^2) + log(INCH) + KID1 + KID2 + KID3 + factor(TIME) | ID, dataset)
summary(mod)

# Compute average partial effects
mod_ape <- get_APEs(mod)
summary(mod_ape)

# Apply analytical bias-correction
mod_bc <- bias_corr(mod)
summary(mod_bc)

# Compute bias-corrected average partial effects
mod_ape_bc <- get_APEs(mod_bc)
summary(mod_ape_bc)
```

---

predict.bife *Predict method for* bife *fits*

Description

predict.bife is a generic function which obtains predictions from an object returned by bife.

Usage

```
## S3 method for class 'bife'
predict(object, type = c("link", "response"),
  X_new = NULL, alpha_new = NULL, corrected = NULL, ...)
```

Arguments

object	an object of class "bife".
type	the type of prediction required. "link" is on the scale of the linear predictor whereas "response" is on the scale of the response variable. Default is "link".
X_new	a regressor matrix for predictions. If not supplied predictions are based on the regressor matrix returned by the object bife. See Details.
alpha_new	a scalar or vector of fixed effects. If not supplied predictions are based on the vector of fixed effects returned by bife. See Details.
corrected	deprecated.
...	other arguments

Details

The model frame returned by the object bife only includes individuals that were not dropped before the fitting process (due to perfect classification). The linear predictors of perfectly classified observations are equal to - Inf or Inf whereas the predicted probabilities are equal to their response. In-sample predictions are only based on non-perfectly classified observations.

If alpha_new is supplied as a scalar the linear predictor is computed using the same value of the fixed effect for each observation. If alpha_new is supplied as a vector it has to be of same length as the rows of the corresponding regressor matrix.

Value

The function predict.bife returns a vector of predictions.

See Also

bife

---

print.bife *Print* bife

Description

print.bife is a generic function which displays some minimal information from objects returned by bife.

Usage

```
## S3 method for class 'bife'
print(x, digits = max(3L, getOption("digits") - 3L), ...)
```

Arguments

x	an object of class "bife".
digits	unsigned integer indicating the number of decimal places. Default is max(3L, getOption("digits") -
...	other arguments.

See Also

bife

---

print.bifeAPEs *Print* bifeAPEs

Description

print.bifeAPEs is a generic function which displays some minimal information from objects returned by get_APEs.

Usage

```
## S3 method for class 'bifeAPEs'
print(x, digits = max(3L, getOption("digits") - 3L),
  ...)
```

Arguments

x	an object of class "bifeAPEs".
digits	unsigned integer indicating the number of decimal places. Default is max(3L, getOption("digits") -
...	other arguments.

See Also

get_APEs

`print.summary.bife` *Print* `summary.bife`

Description

`print.summary.bife` is a generic function which displays summary statistics from objects returned by `summary.bife`.

Usage

```
## S3 method for class 'summary.bife'
print(x, digits = max(3L, getOption("digits") -
  3L), ...)
```

Arguments

`x`	an object of class `"summary.bife"`.
`digits`	unsigned integer indicating the number of decimal places. Default is `max(3L, getOption("digits") -`
`...`	other arguments.

See Also

`bife`

`print.summary.bifeAPEs`
Print `summary.bifeAPEs`

Description

`print.summary.bifeAPEs` is a generic function which displays summary statistics from objects returned by `summary.bifeAPEs`.

Usage

```
## S3 method for class 'summary.bifeAPEs'
print(x, digits = max(3L, getOption("digits")
  - 3L), ...)
```

Arguments

`x`	an object of class `"summary.bifeAPEs"`.
`digits`	unsigned integer indicating the number of decimal places. Default is `max(3L, getOption("digits") -`
`...`	other arguments.

See Also

get_APEs

---

psid *Female labor force participation*

Description

The sample was obtained from the "Panel Study of Income Dynamics" and contains information about $N = 1461$ women that were observed over $T = 9$ years.

Usage

```
psid
```

Format

A data frame with 13,149 rows:

ID individual identifier

LFP labor force participation

KID1 # of kids aged between 0 and 2

KID2 # of kids aged between 3 and 5

KID3 # of kids aged between 6 and 17

INCH income husband

AGE age of woman

TIME time identifier

References

Hyslop, D. (1999). "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women". Econometrica 67(6), 1255-1294.

Fernandez-Val, I. (2009). "Fixed effects estimation of structural parameters and marginal effects in panel probit models". Journal of Econometrics 150(1), 71-85.

See Also

bife

summary.bife *Summarizing models of class* bife

Description

Summary statistics for objects of class "bife".

Usage

```
## S3 method for class 'bife'
summary(object, type = c("sp", "fe"), corrected = NULL,
  fixed = NULL, ...)
```

Arguments

object	an object of class "bife".
type	the type of parameter estimates the summary statistics are related to: structural parameters or fixed effects. Default is "sp" referring to the structural parameters.
corrected, fixed	deprecated.
...	other arguments.

Value

Returns an object of class "summary.bife" which is a list of summary statistics of object.

See Also

bife

summary.bifeAPEs *Summarizing models of class* bifeAPEs

Description

Summary statistics for objects of class "bifeAPEs".

Usage

```
## S3 method for class 'bifeAPEs'
summary(object, ...)
```

Arguments

`object`	an object of class `"bifeAPEs"`.
`...`	other arguments.

Value

Returns an object of class `"summary.bifeAPEs"` which is a list of summary statistics of `object`.

See Also

`get_APEs`

---

vcov.bife *Extract estimates of the covariance matrix*

Description

`vcov.bife` computes an estimate of the covariance matrix of the estimator of the structural parameters from objects returned by `bife`. The estimate is obtained using the inverse of the negative Hessian after convergence.

Usage

```
## S3 method for class 'bife'
vcov(object, ...)
```

Arguments

`object`	an object of class `"bife"`.
`...`	other arguments.

Value

The function `vcov.bife` returns a named matrix of covariance estimates.

See Also

`bife`

vcov.bifeAPEs *Extract estimates of the covariance matrix*

Description

vcov.bifeAPEs computes an estimate of the covariance matrix of the estimator of the average partial parameters from objects returned by get_APEs.

Usage

```
## S3 method for class 'bifeAPEs'
vcov(object, ...)
```

Arguments

object	an object of class "bifeAPEs".
...	other arguments.

Value

The function vcov.bifeAPEs returns a named matrix of covariance estimates.

See Also

get_APEs

Index

Appendix B

R-Package *alpaca*

The *R*-package *alpaca* was developed in the context of chapter 3 and 4. It provides a routine to concentrate out factors with many levels during the optimization of the log-likelihood function of the corresponding generalized linear model (GLM). The package is based on the algorithm proposed by Stammann (2018) and is restricted to GLMs that are based on maximum likelihood estimation and nonlinear. It also offers an efficient algorithm to recover estimates of the fixed effects in a post-estimation routine and includes robust and multi-way clustered standard errors. Further the package provides an analytical bias correction for two-way fixed effects binary choice models (logit and probit) derived by Fernández-Val and Weidner (2016).

The corresponding user manual is provided in the following.

Package 'alpaca'

May 24, 2019

Type Package

Title Fit GLM's with High-Dimensional k-Way Fixed Effects

Version 0.3.1

Description Provides a routine to concentrate out factors with many levels during the optimization of the log-likelihood function of the corresponding generalized linear model (glm). The package is based on the algorithm proposed by Stammann (2018) <arXiv:1707.01815> and is restricted to glm's that are based on maximum likelihood estimation and non-linear. It also offers an efficient algorithm to recover estimates of the fixed effects in a post-estimation routine and includes robust and multi-way clustered standard errors. Further the package provides an analytical bias-correction for binary choice models (logit and probit) derived by Fernandez-Val and Weidner (2016) <doi:10.1016/j.jeconom.2015.12.014>.

License GPL-3

Depends R (>= 3.1.0)

Imports data.table, Formula, MASS, Rcpp, stats, utils

LinkingTo Rcpp, RcppArmadillo

URL https://github.com/amrei-stammann/alpaca

BugReports https://github.com/amrei-stammann/alpaca/issues

RoxygenNote 6.1.1

Suggests bife, car, knitr, lfe

VignetteBuilder knitr

NeedsCompilation yes

Author Amrei Stammann [aut, cre],
Daniel Czarnowske [aut] (<https://orcid.org/0000-0002-0030-929X>)

Maintainer Amrei Stammann <amrei.stammann@hhu.de>

Repository CRAN

Date/Publication 2019-05-24 15:50:02 UTC

R topics documented:

`alpaca-package` *alpaca: A package for fitting glm's with high-dimensional k-way fixed effects*

Description

Concentrates out factors with many levels during the optimization of the log-likelihood function of the corresponding generalized linear model (glm). The package is restricted to glm's that are based on maximum likelihood estimation. This excludes all quasi-variants of glm. The package also offers an efficient algorithm to recover estimates of the fixed effects in a post-estimation routine and includes robust and multi-way clustered standard errors. Further the package provides an analytical bias-correction for binary choice models (logit and probit) derived by Fernandez-Val and Weidner (2016).

Note: Linear models are also beyond the scope of this package since there is already a comprehensive procedure available felm.

biasCorr *Asymptotic bias-correction after fitting binary choice models with two-way error component*

Description

biasCorr is a post-estimation routine that can be used to substantially reduce the incidental parameter bias problem (Neyman and Scott (1948)) present in non-linear fixed effects models (see Fernandez-Val and Weidner (2018) for an overview). The command applies the analytical bias-correction derived by Fernandez-Val and Weinder (2016) to obtain bias-corrected estimates of the structural parameters and is currently restricted to logit and probit models.

Usage

```
biasCorr(object = NULL, L = 0L)
```

Arguments

object	an object of class "feglm"; currently restricted to binomial with "logit" or "probit" link function.
L	unsigned integer indicating a bandwidth for the estimation of spectral densities proposed by Hahn and Kuersteiner (2011). Default is zero, which should be used if all regressors are assumed to be strictly exogenous. In the presence of weakly exogenous or predetermined regressors, Fernandez-Val and Weidner (2016, 2018) suggest to choose a bandwidth not higher than four.

Value

The function biasCorr returns a named list of classes "biasCorr" and "feglm".

References

Czarnowske, D. and Stammann, A. (2019). "Binary Choice Models with High-Dimensional Individual and Time Fixed Effects". ArXiv e-prints.

Fernandez-Val, I. and Weidner, M. (2016). "Individual and time effects in nonlinear panel models with large N, T". Journal of Econometrics, 192(1), 291-312.

Fernandez-Val, I. and Weidner, M. (2018). "Fixed effects estimation of large-t panel data models". Annual Review of Economics, 10, 109-138.

Hahn, J. and Kuersteiner, G. (2011). "Bias reduction for dynamic nonlinear panel models with fixed effects". Econometric Theory, 27(6), 1152-1191.

Neyman, J. and Scott, E. L. (1948). "Consistent estimates based on partially consistent observations". Econometrica, 16(1), 1-32.

See Also

feglm

Examples

```
# Generate an artificial data set for logit models
library(alpaca)
data <- simGLM(1000L, 20L, 1805L, model = "logit")

# Fit 'feglm()'
mod <- feglm(y ~ x1 + x2 + x3 | i + t, data)

# Apply analytical bias-correction
mod.bc <- biasCorr(mod)
summary(mod.bc)
```

`coef.APEs` *Extract estimates of average partial effects*

Description

`coef.APEs` is a generic function which extracts estimates of the average partial effects from objects returned by `getAPEs`.

Usage

```
## S3 method for class 'APEs'
coef(object, ...)
```

Arguments

`object`	an object of class `"APEs"`.
`...`	other arguments.

Value

The function `coef.APEs` returns a named vector of estimates of the average partial effects.

See Also

`getAPEs`

coef.feglm *Extract estimates of structural parameters*

Description

coef.feglm is a generic function which extracts estimates of the structural parameters from objects returned by feglm.

Usage

```
## S3 method for class 'feglm'
coef(object, ...)
```

Arguments

object	an object of class "feglm".
...	other arguments.

Value

The function coef.feglm returns a named vector of estimates of the structural parameters.

See Also

feglm

coef.summary.feglm *Extract coefficient matrix of structural parameters*

Description

coef.summary.feglm is a generic function which extracts a coefficient matrix of structural parameters from objects returned by feglm.

Usage

```
## S3 method for class 'summary.feglm'
coef(object, ...)
```

Arguments

object	an object of class "summary.feglm".
...	other arguments.

Value

The function coef.summary.feglm returns a named matrix of estimates related to the structural parameters.

See Also

feglm

---

feglm *Efficiently fit glm's with high-dimensional k-way fixed effects*

Description

feglm can be used to fit generalized linear models with many high-dimensional fixed effects. The estimation procedure is based on unconditional maximum likelihood and can be interpreted as a "pseudo demeaning" approach that combines the work of Gaure (2013) and Stammann et. al. (2016). For technical details see Stammann (2018). The routine is well suited for large data sets that would be otherwise infeasible to use due to memory limitations.

Remark: The term fixed effect is used in econometrician's sense of having intercepts for each level in each category.

Usage

```
feglm(formula = NULL, data = NULL, family = binomial(),
  beta.start = NULL, eta.start = NULL, control = NULL)
```

Arguments

formula	an object of class "formula": a symbolic description of the model to be fitted. formula must be of type y ~ x | k, where the second part of the formula refers to factors to be concentrated out. It is also possible to pass additional variables to feglm (e.g. to cluster standard errors). This can be done by specifying the third part of the formula: y ~ x | k | add.
data	an object of class "data.frame" containing the variables in the model.
family	a description of the error distribution and link function to be used in the model. Similiar to glm.fit this has to be the result of a call to a family function. Default is binomial(). See family for details of family functions.
beta.start	an optional vector of starting values for the structural parameters in the linear predictor. Default is $\boldsymbol{\beta} = \mathbf{0}$.
eta.start	an optional vector of starting values for the linear predictor.
control	a named list of parameters for controlling the fitting process. See feglmControl for details.

Details

If feglm does not converge this is usually a sign of linear dependence between one or more regressors and a fixed effects category. In this case, you should carefully inspect your model specification.

Value

The function feglm returns a named list of class "feglm".

References

Gaure, S. (2013). "OLS with Multiple High Dimensional Category Variables". Computational Statistics and Data Analysis, 66.

Stammann, A., Heiss, F., and McFadden, D. (2016). "Estimating Fixed Effects Logit Models with Large Panel Data". Working paper.

Stammann, A. (2018). "Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-Way Fixed Effects". ArXiv e-prints.

Examples

```
# Generate an artificial data set for logit models
library(alpaca)
data <- simGLM(1000L, 20L, 1805L, model = "logit")

# Fit 'feglm()'
mod <- feglm(y ~ x1 + x2 + x3 | i + t, data)
summary(mod)
```

`feglm.nb` *Efficiently fit negative binomial glm's with high-dimensional k-way fixed effects*

Description

feglm.nb can be used to fit negative binomial generalized linear models with many high-dimensional fixed effects (see feglm).

Usage

```
feglm.nb(formula = NULL, data = NULL, beta.start = NULL,
  eta.start = NULL, init.theta = NULL, link = c("log", "identity",
  "sqrt"), control = NULL)
```

Arguments

formula, data, beta.start, eta.start, control	see feglm.
init.theta	an optional initial value for the theta parameter (see glm.nb).
link	the link function. Must be one of "log", "sqrt", or "identity".

Details

If feglm.nb does not converge this is usually a sign of linear dependence between one or more regressors and a fixed effects category. In this case, you should carefully inspect your model specification.

Value

The function feglm.nb returns a named list of class "feglm".

References

Gaure, S. (2013). "OLS with Multiple High Dimensional Category Variables". Computational Statistics and Data Analysis. 66.

Stammann, A., F. Heiss, and D. McFadden (2016). "Estimating Fixed Effects Logit Models with Large Panel Data". Working paper.

Stammann, A. (2018). "Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-Way Fixed Effects". ArXiv e-prints.

See Also

glm.nb, feglm

---

feglmControl *Set* feglm *Control Parameters*

Description

Set and change parameters used for fitting feglm.

Note: feglm.control is deprecated and will be removed soon.

Usage

```
feglmControl(dev.tol = 1e-08, center.tol = 1e-05, rho.tol = 1e-04,
  conv.tol = 1e-06, iter.max = 100L, limit = 10L, trace = FALSE,
  drop.pc = TRUE, pseudo.tol = NULL, step.tol = NULL)

feglm.control(...)
```

Arguments

`dev.tol`	tolerance level for the first stopping condition of the maximization routine. The stopping condition is based on the relative change of the deviance in iteration r and can be expressed as follows: $(dev_{r-1} - dev_r)/(0.1 + dev_r) < tol$. Default is `1.0e-08`.
`center.tol`	tolerance level for the stopping condition of the centering algorithm. The stopping condition is based on the relative change of euclidean norm in iteration i and can be expressed as follows: $\|\|\mathbf{v}_i - \mathbf{v}_{i-1}\|\|_2 < tol\|\|\mathbf{v}_{i-1}\|\|$. Default is `1.0e-05`.
`rho.tol`	tolerance level for the stephalving in the maximization routine. Stephalving only takes place if the deviance in iteration r is larger than the one of the previous iteration. If this is the case, $\|\|\boldsymbol{\beta}_r - \boldsymbol{\beta}_{r-1}\|\|_2$ is halfed until the deviance is less or numerically equal compared to the deviance of the previous iteration. Stephalving fails if the the following condition holds: $\rho < tol$, where ρ is the stepcorrection factor. If stephalving fails the maximization routine is canceled. Default is `1.0e-04`.
`conv.tol`	tolerance level that accounts for rounding errors inside the stephalving routine when comparing the deviance with the one of the previous iteration. Default is `1.0e-06`.
`iter.max`	unsigned integer indicating the maximum number of iterations in the maximization routine. Default is `100L`.
`limit`	unsigned integer indicating the maximum number of iterations of `theta.ml`. Default is `10L`.
`trace`	logical indicating if output should be produced in each iteration. Default is `FALSE`.
`drop.pc`	logical indicating to drop observations that are perfectly classified (perfectly seperated) and hence do not contribute to the log-likelihood. This option is useful to reduce the computational costs of the maximization problem, since it reduces the number of observations and does not affect the estimates. Default is `TRUE`.
`pseudo.tol`	deprecated; use `center.tol` instead.
`step.tol`	depreacted; termination conditions is now similar to `glm`.
`...`	arguments passed to the deprecated function `feglm.control`.

Value

The function `feglmControl` returns a named list of control parameters.

See Also

`feglm`

fitted.feglm *Extract* feglm *fitted values*

Description

fitted.feglm is a generic function which extracts fitted values from an object returned by feglm.

Usage

```
## S3 method for class 'feglm'
fitted(object, ...)
```

Arguments

object	an object of class "feglm".
...	other arguments.

Value

The function fitted.feglm returns a vector of fitted values.

See Also

feglm

getAPEs *Compute average partial effects after fitting binary choice models with two-way error component*

Description

getAPEs is a post-estimation routine that can be used to estimate average partial effects with respect to all covariates in the model and the corresponding covariance matrix. The estimation of the covariance is based on a linear approximation (delta method). Note that the command automatically determines which of the regressors are continuous or binary.

Remark: The routine currently does not allow to compute average partial effects based on functional forms like interactions and polynomials.

Usage

```
getAPEs(object = NULL, n.pop = NULL, weak.exo = FALSE)
```

Arguments

object	an object of class "biasCorr" or "feglm"; currently restricted to binomial with "logit" or "probit" link function.
n.pop	unsigned integer indicating a finite population correction for the estimation of the covariance matrix of the average partial effects proposed by Cruz-Gonzalez, Fernandez-Val, and Weidner (2017). The correction factor is computed as follows: $(n^* - n)/(n^* - 1)$, where n^* and n are the size of the entire population and the full sample size. Default is NULL, which refers to a factor of one and is equal to an infinitely large population.
weak.exo	logical indicating if some of the regressors are assumed to be weakly exogenous (e.g. predetermined). If object is of class "biasCorr", the option will be automatically set to TRUE if the choosen bandwidth parameter is larger than zero. Note that this option only affects the estimation of the covariance matrix. Default is FALSE, which assumes that all regressors are strictly exogenous.

Value

The function getAPEs returns a named list of class "APEs".

References

Cruz-Gonzalez, M., Fernandez-Val, I., and Weidner, M. (2017). "Bias corrections for probit and logit models with two-way fixed effects". The Stata Journal, 17(3), 517-545.

Czarnowske, D. and Stammann, A. (2019). "Binary Choice Models with High-Dimensional Individual and Time Fixed Effects". ArXiv e-prints.

Fernandez-Val, I. and Weidner, M. (2016). "Individual and time effects in nonlinear panel models with large N, T". Journal of Econometrics, 192(1), 291-312.

Fernandez-Val, I. and Weidner, M. (2018). "Fixed effects estimation of large-t panel data models". Annual Review of Economics, 10, 109-138.

Neyman, J. and Scott, E. L. (1948). "Consistent estimates based on partially consistent observations". Econometrica, 16(1), 1-32.

See Also

biasCorr, feglm

Examples

```
# Generate an artificial data set for logit models
library(alpaca)
data <- simGLM(1000L, 20L, 1805L, model = "logit")

# Fit 'feglm()'
mod <- feglm(y ~ x1 + x2 + x3 | i + t, data)

# Compute average partial effects
mod.ape <- getAPEs(mod)
```

```
summary(mod.ape)

# Apply analytical bias-correction
mod.bc <- biasCorr(mod)
summary(mod.bc)

# Compute bias-corrected average partial effects
mod.ape.bc <- getAPEs(mod.bc)
summary(mod.ape.bc)
```

`getFEs` *Efficiently recover estimates of the fixed effects after fitting* `feglm`

Description

Recover estimates of the fixed effects by alternating between the normal equations of the fixed effects as shown by Stammann (2018).

Remark: The system might not have a unique solution since we do not take collinearity into account. If the solution is not unique, an estimable function has to be applied to our solution to get meaningful estimates of the fixed effects. See Gaure (n. d.) for an extensive treatment of this issue.

Usage

```
getFEs(object = NULL, alpha.tol = 1e-08)
```

Arguments

`object`	an object of class `"feglm"`.
`alpha.tol`	tolerance level for the stopping condition. The algorithm is stopped in iteration i if $\|\|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{i-1}\|\|_2 < tol\|\|\boldsymbol{\alpha}_{i-1}\|\|_2$. Default is `1.0e-08`.

Value

The function `getFEs` returns a named list containing named vectors of estimated fixed effects.

References

Gaure, S. (n. d.). "Multicollinearity, identification, and estimable functions". Unpublished.

Stammann, A. (2018). "Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects". ArXiv e-prints.

See Also

`feglm`

predict.feglm *Predict method for* feglm *fits*

Description

predict.feglm is a generic function which obtains predictions from an object returned by feglm.

Usage

```
## S3 method for class 'feglm'
predict(object, type = c("link", "response"), ...)
```

Arguments

object	an object of class "feglm".
type	the type of prediction required. "link" is on the scale of the linear predictor whereas "response" is on the scale of the response variable. Default is "link".
...	other arguments.

Value

The function predict.feglm returns a vector of predictions.

See Also

feglm

print.APEs *Print* APEs

Description

print.APEs is a generic function which displays some minimal information from objects returned by getAPEs.

Usage

```
## S3 method for class 'APEs'
print(x, digits = max(3L, getOption("digits") - 3L), ...)
```

Arguments

x	an object of class "APEs".
digits	unsigned integer indicating the number of decimal places. Default is max(3L, getOption("digits") -
...	other arguments.

See Also

getAPEs

---

print.feglm *Print* feglm

Description

print.feglm is a generic function which displays some minimal information from objects returned by feglm.

Usage

```
## S3 method for class 'feglm'
print(x, digits = max(3L, getOption("digits") - 3L), ...)
```

Arguments

x	an object of class "feglm".
digits	unsigned integer indicating the number of decimal places. Default is max(3L, getOption("digits") -
...	other arguments.

See Also

feglm

---

print.summary.APEs *Print* summary.APEs

Description

print.summary.APEs is a generic function which displays summary statistics from objects returned by summary.APEs.

Usage

```
## S3 method for class 'summary.APEs'
print(x, digits = max(3L, getOption("digits") -
  3L), ...)
```

Arguments

x	an object of class "summary.APEs".
digits	unsigned integer indicating the number of decimal places. Default is max(3L, getOption("digits") -
...	other arguments.

See Also

getAPEs

---

print.summary.feglm *Print* summary.feglm

Description

print.summary.feglm is a generic function which displays summary statistics from objects returned by summary.feglm.

Usage

```
## S3 method for class 'summary.feglm'
print(x, digits = max(3L, getOption("digits") -
  3L), ...)
```

Arguments

x	an object of class "summary.feglm".
digits	unsigned integer indicating the number of decimal places. Default is max(3L, getOption("digits") -
...	other arguments.

See Also

feglm

---

simGLM *Generate an artificial data set for some GLM's with two-way fixed effects*

Description

Constructs an artificial data set with n cross-sectional units observed for t time periods for logit, poisson, or gamma models. The "true" linear predictor ($\boldsymbol{\eta}$) is generated as follows:

$$\eta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma_t \,,$$

where $\mathbf{X}$ consists of three independent standard normally distributed regressors. Both parameter refering to the unobserved heterogeneity (α_i and γ_t) are generated as iid. standard normal and the structural parameters are set to $\boldsymbol{\beta} = [1, -1, 1]'$.

Note: The poisson and gamma model are based on the logarithmic link function.

Usage

```
simGLM(n = NULL, t = NULL, seed = NULL, model = c("logit",
  "poisson", "gamma"))
```

Arguments

n	a strictly positive integer equal to the number of cross-sectional units.
t	a strictly positive integer equal to the number of time periods.
seed	a seed to ensure reproducibility.
model	a string equal to "logit", "poisson", or "gamma".

Value

The function simGLM returns a data.frame with 6 variables.

See Also

feglm

---

summary.APEs *Summarizing models of class* APEs

Description

Summary statistics for objects of class "APEs".

Usage

```
## S3 method for class 'APEs'
summary(object, ...)
```

Arguments

object	an object of class "APEs".
...	other arguments.

Value

Returns an object of class "summary.APEs" which is a list of summary statistics of object.

See Also

getAPEs

summary.feglm *Summarizing models of class* feglm

Description

Summary statistics for objects of class "feglm".

Usage

```
## S3 method for class 'feglm'
summary(object, type = c("hessian", "outer.product",
  "sandwich", "clustered"), cluster = NULL, cluster.vars = NULL, ...)
```

Arguments

object	an object of class "feglm".
type	the type of covariance estimate required. "hessian" refers to the inverse of the negative expected Hessian after convergence and is the default option. "outer.product" is the outer-product-of-the-gradient estimator, "sandwich" is the sandwich estimator (sometimes also refered as robust estimator), and "clustered" computes a clustered covariance matrix given some cluster variables.
cluster	a symbolic description indicating the clustering of observations.
cluster.vars	deprecated; use cluster instead.
...	other arguments.

Details

Multi-way clustering is done using the algorithm of Cameron, Gelbach, and Miller (2011). An example is provided in the vignette "Replicating an Empirical Example of International Trade".

Value

Returns an object of class "summary.feglm" which is a list of summary statistics of object.

References

Cameron, C., J. Gelbach, and D. Miller (2011). "Robust Inference With Multiway Clustering". Journal of Business & Economic Statistics 29(2).

See Also

feglm

vcov.feglm *Extract estimates of the covariance matrix*

Description

vcov.feglm computes an estimate of the covariance matrix of the estimator of the structural parameters from objects returned by feglm. The estimate is obtained using the Hessian, the scores, or a combination of boths after convergence.

Usage

```
## S3 method for class 'feglm'
vcov(object, type = c("hessian", "outer.product",
  "sandwich", "clustered"), cluster = NULL, cluster.vars = NULL, ...)
```

Arguments

object	an object of class "feglm".
type	the type of covariance estimate required. "hessian" refers to the inverse of the negative expected Hessian after convergence and is the default option. "outer.product" is the outer-product-of-the-gradient estimator, "sandwich" is the sandwich estimator (sometimes also refered as robust estimator), and "clustered" computes a clustered covariance matrix given some cluster variables.
cluster	a symbolic description indicating the clustering of observations.
cluster.vars	deprecated; use cluster instead.
...	other arguments.

Details

Multi-way clustering is done using the algorithm of Cameron, Gelbach, and Miller (2011). An example is provided in the vignette "Replicating an Empirical Example of International Trade".

Value

The function vcov.feglm returns a named matrix of covariance estimates.

References

Cameron, C., J. Gelbach, and D. Miller (2011). "Robust Inference With Multiway Clustering". Journal of Business & Economic Statistics 29(2).

See Also

feglm

Index