

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Economics

2025

ESSAYS ON ECONOMETRICS: SPARSITY, DURATION, AND BINOMIAL PANEL DATA MODELS

Roelsgaard, Sebastian Toft

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp011544bs561>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

ESSAYS ON ECONOMETRICS: SPARSITY, DURATION,
AND BINOMIAL PANEL DATA MODELS

SEBASTIAN TOFT ROELSGAARD

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
ECONOMICS
ADVISERS: BO E. HONORÉ

SEPTEMBER 2025

© Copyright by Sebastian Toft Roelsgaard, 2025.

All rights reserved.

Abstract

This dissertation consists of three chapters on econometrics.

The first chapter, coauthored with Michal Kolesár and Ulrich K. Müller, documents that linear regression estimates which rely on an assumption of sparsity are fragile. First, we show that different choices of the regressor matrix that do not impact ordinary least squares (OLS) estimates, such as the choice of baseline category with categorical controls, can move sparsity-based estimates by two standard errors or more. Second, we develop two tests of the sparsity assumption based on comparing sparsity-based estimators with OLS. Across the three applications we find many rejections of the sparsity assumptions using these tests. Unless the number of regressors is comparable to or exceeds the sample size, OLS yields more robust inference at little efficiency cost.

The second chapter generalizes the canonical proportional hazards model to allow for two-way unobserved heterogeneity. The duration models of interest are those generated by a match between entities from two distinct groups, so the focus is on bipartite graphs. Identification of the coefficient on the explanatory variables and the difference of intra-group baseline hazards is proven when baseline hazard rates are “double-stratified.” In the special case where at least one group has time-invariant heterogeneity, we provide a consistent and asymptotically normal estimator of the coefficient vector when both groups grow at the same rate.

The final chapter, coauthored with Bo E. Honoré, generalizes recent developments for binary outcome panel data models with fixed effects to models with binomial outcomes. We first consider models where the outcome is the sum of conditionally independent and identically distributed “logit” Bernoulli random variables. We show that results from the

binary outcome literature concerning the existence of sufficient statistics carry over. In cases where a sufficient statistic does not exist, we find moment conditions. We then relax the assumption on identical distributions and show that certain results extend to the resulting “Poisson-Binomial” models. Finally, we connect the models analyzed to those in the literature on entry games.

Acknowledgements

I am deeply grateful to my adviser Bo for his guidance and support throughout the entire PhD program. I was fortunate enough to meet him before I even began at Princeton, and he helped me settle in before it was even clear (at least to me) that I was going to end up working on econometrics. He has been tremendously patient as I have stumbled my way through projects, and through countless meetings and numerous walks, he always helped me get unstuck. Besides the academic support that he provided, he would regularly check in on me during the pandemic, and he repeatedly convinced me to go volunteer at a forest restoration project with him, because “few things takes your mind off econometrics like putting your hands in dirt.” Bo is also a critical part of the social fabric not just at the Department of Economics, but also for the Danish community in Princeton. Together we have organized numerous social gatherings for the other Princeton Danes, and Bo has frequently taken on the role of host, providing a slice of home away from home.

I am likewise grateful to Ulrich for taking a shot at working with me. His enthusiasm for his craft is contagious, and few things in grad school have been as great as having a chat with him long into the afternoon, where he would frequently muse about the importance of being good at computing (I have surely disappointed him by not picking up Fortran), and concluding that “you learn to really respect hardware, when you code a flight simulator in assembly.” Ulrich was also extremely encouraging throughout my jobmarket, and took time to chat with me about his experiences in industry.

I am thankful for the the numerous other Princeton faculty who have helped me along the way, especially Michal Kolesár, Mikkel Plagborg-Møller, Mark Watson, and Richard Rogerson.

Laura Hedden deserves a shout-out not only for being an exceptional administrator who (almost) single-handedly keeps the graduate program together, but also for always taking the time to chat and lend an ear to the woes of me and numerous other graduate students.

Finally, my academic career would not have begun without John Kennes. He showed me the very best of what the life of a researcher has to offer, and he always had the utmost faith in me. Without his support and encouragement, I would not have ventured down this path. I am lucky to have had him first as a mentor and adviser, and now I am grateful to call him my friend.

Six years is a long time, but I was fortunate enough to spend them with so many fantastic people. Not all cohorts manage to have the cohesion that we had, and I additionally managed to call a good chunk of them my office mates. Impeccably little work was completed in their company. A special thank you to Raf for showing me the true meaning of work-life balance, running a film club with me, and showing me the beauty of America's national parks, to George and Narek for providing me with stoic advice and endless banter, and to Chansik for his impersonations. Thank you to Ruairidh and Edmund for being incredible roommates.

Thank you to my friends outside of the cohort, including Emily, Rachel, and Elettra. A very special thank you to Elaine for watching B-, nay, C-movies with me, and for the annual Christmas card.

Finally, in Princeton, I am grateful to all the members of McCosh Theatres. Especially to Allie for being a stellar movie buddy, and Francesco for his refined taste, thank you

to Suren, Jesse, Elif, Genevieve, and finally to Nicole, for helping me realize my inner potential.

I was lucky enough to have great friends from back home that stayed in touch despite my moving across the Atlantic. Thank you to Mikkell for the countless hours spent on the phone lamenting the current state of movies and music. To Jake for putting up with me as a roommate, our long chats, and for ensuring there is a kolonihave I can visit when I go home. To Artie for the many days we have spent together, for cutting my hair, and for our travels. To Nico for our many walks, Mathias for teaching me how to left join, and Hans for helping me realize that frog truly is stranger than fiction.

Thank you to my family for their support. To Mom and Dad for always believing in me. To my sisters Natascha and Kathinka for being the best sisters that I could have hoped for. To my brothers-in-law, Daniel and Henry, to my cousin Nicklas, to my nephews Arthur, Albert, and August and to my grandmother, Inge, for her wonderful stories.

And, to Carol, for the last five years and hopefully many more.

Til Mormor

Contents

Abstract	3
Acknowledgements	5
1 The Fragility of Sparsity	11
1.1 Empirical Illustrations	17
1.1.1 Normalizations of the Control Matrix	18
1.1.2 Alternative Normalizations of the Control Matrix	23
1.2 Sparse Representations Are Rare	26
1.2.1 Approximate Sparsity	26
1.2.2 Rotation	27
1.2.3 Categorical Data	28
1.2.4 Offset in Hermite Polynomial Regression	30
1.2.5 Discussion	32
1.3 Efficiency Gains Under Sparsity	35
1.4 Testing Sparsity	39
1.4.1 Hausman Test	39
1.4.2 Residual Test	41
1.4.3 Empirical Tests of Sparsity	44
1.5 Summary and Conclusions	45
References	49
Appendices	54
1.A Auxiliary Results	54
1.B Proofs	59
1.C Finite-Sample Size Properties of Sparsity Tests	78
2 Two-Way Duration Models	81
2.1 Identification	86
2.1.1 Comment on the Different Identification Proofs	90
2.2 Estimation and Inference	91

2.2.1	Sparsity Comments	92
2.2.2	Consistency	93
2.2.3	Asymptotic Normality	94
2.3	Conclusion	96
	References	98
	Appendices	101
2.A	Notes on Computation	101
2.B	Proofs	103
3	Binomial Models With Fixed Effects	117
3.1	Simplest Model, Propositions, and Conjectures	121
3.1.1	Conditional Likelihood Estimation of Models With No or One Lag	122
3.1.2	Pure AR Models With Additional Lags	124
3.1.3	Moment Conditions	127
3.1.4	Moment Conditions for Pure AR Models	130
3.2	Indicator AR Model	133
3.3	Poisson-Binomial Model	134
3.4	Models With Strategic Interactions	139
3.5	Conclusion	142
	References	143
	Appendices	147
3.A	Moment Conditions	147
3.B	Proofs	155
3.C	Moment Finding Procedure	163

Chapter 1

The Fragility of Sparsity*

The linear regression model is the most common workhorse for estimating causal effects. The key assumption allowing for a causal interpretation of the coefficient on the treatment or intervention of interest is that conditional on the set of controls included in the model, the remaining variation in the treatment is as good as random. Since the plausibility of this assumption increases with additional control variables, researchers commonly estimate regression models with many controls.

*This chapter is coauthored with Michal Kolesár and Ulrich K. Müller.

Michal Kolesár has presented this paper at invited seminars at LUM/TUM Munich, University of Warwick, University College London, University of California, Davis, the European Commission Joint Research Centre, Georgetown University, University of Wisconsin-Madison, the University of Calgary, Tulane University, University of Southern California, Columbia University, Northwestern University and the University of Groningen. Ulrich Müller has presented this paper at invited seminars at Stanford University, the Philadelphia Federal Reserve and at the University of California, San Diego as well as at the Causal Inference OCE Conference III in Chicago in 2024, the Econometric Society Summer Meetings in Nashville in 2024, and the Annual Meeting of the German Statistical Association in Regensburg in 2024. Sebastian Roelsgaard presented the paper at the ESIF Economics and AI+ML 2024 meeting at Cornell University as well as the DAEiNA 2023 annual meeting at Aarhus University.

To fix ideas, consider estimation of the coefficient β on treatment D_i in the regression model

$$Y_i = D_i\beta + W_i'\gamma + U_i, \quad E[U_i \mid D_i, W_i] = 0, \quad (1.1)$$

based on n observations indexed by i . Traditional estimation methods like ordinary least squares (OLS) limit the dimension p of the control vector W_i to be smaller than n . This limitation has fueled the development of alternative estimation methods that allow p to exceed n , provided one is willing to make the (approximate) sparsity assumption that the control function $W_i'\gamma$ can be well-approximated, in a mean-squared error sense, by a linear combination of a small number s of the controls, so that, up to log terms, $s^2/n \rightarrow 0$. In particular, several papers (see, among others, Javanmard & Montanari, 2014; van de Geer et al., 2014; Zhang & Zhang, 2014) have developed “debiased lasso” estimators of β that are able to purge bias present when eq. (1.1) is estimated directly using the lasso or related methods. Belloni et al. (2014) propose a corresponding post-double-selection procedure that first estimates the propensity score regression

$$D_i = W_i'\delta + \tilde{D}_i, \quad E[\tilde{D}_i \mid W_i] = 0, \quad (1.2)$$

by the lasso, and also runs a lasso regression of the outcome on the controls W_i . In a second step, β is estimated by regressing the outcome on D_i and a union of the controls selected by the two lasso regressions. A number of authors have developed variants of these methods that allow for treatment effect heterogeneity (see, among others, Athey et al., 2018; Belloni et al., 2017; Chernozhukov et al., 2018; Farrell, 2015). These developments have had a large influence on empirical practice in applied economics and related

fields, and researchers routinely use these sparsity-based estimators (SBEs) even in the regime $p < n$ as a complement to or replacement of OLS.

This paper argues that this practice may not lead to reliable inference for two reasons. First, SBEs, like the sparsity assumption underlying them, are not invariant to linear reparametrizations of the control matrix W . For instance, while normalization choices such as the choice of baseline category for a set of dummy variables, or how and whether to center the variables prior to taking powers or interactions, are immaterial when running OLS, they in general affect SBEs. We document that in three empirical applications of SBEs, which include the empirical illustration in Belloni et al. (2014), seemingly innocuous normalization choices of this type induce variation in SBEs of the same magnitude as their standard error. It is thus impossible to interpret the resulting inference without substantively engaging with the question whether the particular normalization choice used is appropriate. Yet, papers using SBEs do not even tell readers what this choice was, let alone argue that it is the right one to obtain a sparse representation. While it is well-known that SBEs are not invariant to linear reparametrizations in the theoretical literature, this practice implies that the extent of their fragility is not currently appreciated, even by researchers in the field.

Second, one might be skeptical of the plausibility of the sparsity assumption more generally. In social science applications, it is often difficult to point to theoretical or institutional arguments for why a small number of controls should be able to soak up most of the confounding. We thus develop two statistical tests of the null hypothesis that (approximate) sparsity holds. Both tests tend to reject in the baseline specification of the three empirical applications. What is more, we are often unable to find any normalization choice

for which the tests do not reject, suggesting that this second concern is not just a variation of the first.

Our conclusion is that applied researchers should be wary of using SBEs without a substantive defense of the sparsity assumption for the specific choice of the control matrix used in the analysis. Unless p is close to or exceeds n , a robust default is to simply run OLS.²

Our analysis relates to several strands of literature. Many authors mention the SBEs' lack of invariance to linear reparametrization as an undesirable feature. In the specific context of including categorical controls, alternatives to standard lasso regression have been developed that are invariant to the choice of baseline category (e.g. Bondell & Reich, 2009; Gertheiss & Tutz, 2010; Stokell et al., 2021). Similarly, one could perhaps incorporate the choice of a centering constant when taking powers as an additional tuning parameter when fitting SBEs. However, as we discuss further in Section 1.2.5 below, such solutions come with their own challenges. Using a Bayesian framework, Giannone et al. (2021) present empirical evidence suggesting sparsity may not be a compelling assumption in popular data sets used in economics. Wüthrich and Zhu (2023) give simulation evidence and theoretical arguments showing that even if sparsity arguably holds, SBEs may still display substantial bias in finite samples. Angrist and Frandsen (2022) explore robustness of SBEs to tuning parameter choices.

The remainder of the paper is organized as follows. In Section 1.1, we introduce the three empirical examples used to quantify fragility of SBEs: the empirical illustration in

²One only needs to be careful in standard error construction, as OLS no longer consistently estimates the residuals once $p \asymp n$. This renders the usual Eicker-Huber-White standard errors invalid; instead one needs to use a standard error formula that is robust to high-dimensional controls (see, e.g. Cattaneo et al., 2018b; Dobriban et al., 2024; D'Adamo, 2019; Kline et al., 2020; Jochmans, 2022).

Belloni et al. (2014) on the effect of abortion on crime, a Ferrara (2022) study of occupational upgrading by Black southerners, and a study of the effect of moral values on voting behavior by Enke (2020). Our analysis of fragility to linear reparametrizations focuses on two seemingly innocuous normalizations: first, we consider different ways of dropping collinear columns, such as which category to drop when including categorical variables. Second, when W includes powers and interactions of a set of baseline controls, we consider different ways of centering the baseline controls, such as demeaning, subtracting the median or no centering. We find that these normalizations can move the estimates by two standard errors to more. We also consider alternative normalizations that further expand the set of possible specifications of W : expressing a set of categorical controls as indicators for subsets of the categories, rather than just as indicators for each individual category; and continuously varying where a variable is centered prior to taking powers, rather than just considering discrete values such as the mean or the median. These normalizations lead to an even greater variation in the estimates.

In Section 1.2, we interpret these empirical results through the lens of a thought experiment: suppose a researcher picks one of the possible normalizations at random. How likely would such a random choice lead to an (approximately) sparse representation? We compute that probability in three stylized examples of normalizations and find that it rapidly decreases as function of n . Leaving default choices for how W is constructed to statistical software or happenstance is very unlikely to lead to sparsity.

Section 1.3 conducts an analysis of the relative efficiency of SBEs and OLS. We focus on the regime where p is smaller, but proportional to n . This is because if $p/n \rightarrow 0$, OLS achieves the semiparametric efficiency bound under homoskedastic errors, which

limits the arguments in favor of alternative estimators. On the other hand, if p exceeds n , the OLS benchmark is no longer available. We show that under homoskedasticity, the proportional variance reduction of SBEs relative to OLS is capped at $1 - p/n$ (effectively, we can at best avoid a degrees of freedom adjustment). This benchmark forms the basis for our recommendation to simply run OLS unless p is close to n : it yields robust estimates at little efficiency loss.

In Section 1.4, we develop two tests of the sparsity assumption, again in the regime $p \asymp n$.³ We circumvent the difficulty that the sparsity assumption only restricts rates, rather than the actual number of non-zero coefficients for a given n , by testing it indirectly. The first test compares OLS and lasso residuals: under (approximate) sparsity, the residual sum of squares of the lasso should exceed the OLS residual sum of squares only by a small amount. The second test is an application of the Hausman (1978) specification test: any difference between SBE and OLS estimates must be explained by differences in the relative efficiency of the estimators afforded by the sparsity assumption that SBE exploits, but OLS does not. Thus, the common practice of reporting SBE estimates alongside OLS should not be interpreted as a robustness check for the OLS specification; rather, divergence between the estimates indicates a failure of the sparsity assumption.

Section 1.5 concludes. All proofs are relegated to appendices.

³Carpentier and Verzelen (2021) develop a test in the high-dimensional regime $p/n \rightarrow \infty$. However, their test requires explicit specification of the sparsity level under the null, as does the test developed by He (2020) in the regime $p < n$. In the context of factor models, Beyhum and Striaukas (2024) consider testing the null of a zero vector against a sparse alternative.

1.1 Empirical Illustrations

In this section we revisit three applications that leveraged SBEs: the application in Belloni et al. (2014, BCH) that probes the investigation by Donohue and Levitt (2001) of the impact of abortion on crime, the Ferrara (2022) study of employment opportunities for Black southerners in the aftermath of World War II (WW2), and the examination of the relationship between moral values and voting behavior by Enke (2020). In each application, we apply the original SBE to eq. (1.1) after changing the specification of the control matrix W (where rows correspond to the control vectors W_i) in seemingly innocuous ways that do not impact OLS estimates. We show that the impact of these normalizations on SBEs is, on the other hand, substantial. To isolate the effect of the choice of control matrix from other implementation details, our analysis otherwise sticks to software defaults.⁴

In Section 1.1.1, we consider effects of two normalizations. First, we consider different ways of resolving multicollinearity in the control matrix by changing which columns are dropped to make the matrix W full rank. For example, if multicollinearity arises due to the inclusion of categorical variables, we change which category is dropped. Original implementations of SBEs in each of the three applications remove multicollinearity in W as a data-processing step, similar to standard implementations of least squares regression. Our exercise is therefore equivalent to changing the order of the columns of the control matrix, but otherwise conducting the analysis exactly as in the original. For lasso-based estimators, such a step is typically needed to ensure that so-called “compatibility conditions” or “restricted eigenvalue” assumptions hold; these assumptions ensure fast convergence rates

⁴For estimates using post-double selection, we use the `hdm` package and its defaults, including the choice of the tuning parameter and normalizing the columns of W to have unit variance.

for the lasso (see, e.g. Bickel et al., 2009).⁵ Second, when W includes powers and interactions of a baseline set of controls, we consider different normalizations of the baseline controls (such as demeaning vs subtracting the median) before taking powers and interactions. Appropriately centering the baseline variables puts them on the same scale, and may make the sparsity assumption more plausible and more easily interpretable. Relative to using raw polynomials, it typically also helps with numerical stability.

Section 1.1.2 considers alternative normalizations of W in the two scenarios above: expressing a set of categorical controls as indicators for subsets of the categories, rather than only as indicators for each individual category; and continuously varying where a variable is centered prior to taking powers and interactions, rather than just considering discrete values such as the mean or the median.

1.1.1 Normalizations of the Control Matrix

We begin by considering different ways of resolving multicollinearity issues that arise in each application.

The data in the first application consists of an annual panel of US states over the period 1985–97 originally analyzed by Donohue and Levitt (2001), who ran a two-way fixed effects regression of crime rates on effective abortion rates, state and year fixed effects, and 8 baseline covariates.⁶ BCH argue that this set of 8 controls may be insufficient to purge time-varying confounders. They consider a first-differences version of this specification,

⁵A necessary condition for these assumptions is that submatrices of W with $2s$ columns, where s is the sparsity index, are full rank. It implies that if, say, a categorical variable has fewer than $2s$ categories, we cannot include all categories as well as the intercept.

⁶These are: lags of the number of prisoners and police per capita, the unemployment rate, per-capita income and beer consumption, poverty rate, AFDC generosity lagged 15 years, and a dummy for a shall-issue concealed carry law.

with 12 time effects and first-differences of the 8 baseline controls, which they augment with 136 further variables obtained by squaring the baseline controls and interacting them with each other and with linear and quadratic trends.⁷ In addition, they include 49 variables that are time-invariant within each state, corresponding to initial values and averages of various transformations of the baseline controls, as well as interactions of these 49 variables with a time trend and its square. Since there are only 48 states in the data, these variables span the same column space as state fixed effects. The resulting control matrix has 303 columns, but rank of only 294: because time effects are included, 2 of the time-invariant variables are redundant, as are 2 of the interactions with a time trend and 2 with its square. In addition, one of the baseline controls, a shall-issue dummy, is binary and non-zero only 21 times, but it is interacted with 24 variables, so that 3 of the interactions are redundant. BCH first drop the collinear columns, and then estimate the model by the post-double lasso estimator that they develop.

Column 2 of Panel A in Table 1.1 replicates the BCH estimates for each of the three versions of the crime rate outcome variable considered by BCH.⁸ Comparing the standard error to the OLS standard error in column 1, we see that the post-double lasso estimator appears to be substantially more precise than the OLS estimate. It is also economically significant: the effective abortion rate is defined as the average legalized abortion rate among arrestee cohorts (the number of abortions per live birth in a cohort weighted by the

⁷In particular, BCH add squares of the first differences as well as lags and squared lags of the baseline controls, and they interact the first differences and their squares with a linear and quadratic trend. They also add interactions of the first-differences, and interact them with a linear and a quadratic trend. Since the time series of the shall-carry dummy in each state never changes from 1 to 0, its first difference is also binary. These transformations therefore only yield 136 unique columns.

⁸The replication differs slightly from the original (Table 2 in BCH) because the algorithm for dropping collinear columns in BCH had a coding error, yielding a matrix with 296 columns, and a rank of 293.

Table 1.1: Fragility of sparsity-based methods to normalizations of the control matrix.

		Sparsity-based estimation				
	OLS	Replication	Range (collinear)		Range (powers)	
Outcome	(1)	(2)	(3)	(4)	(5)	(6)
A: BCH						
$\Delta \ln(\text{violent crime/capita})$	0.006	−0.160	−0.216	−0.109	−0.160	−0.122
$(n = 576, p = 294)$	(0.755)	(0.112)	(0.118)	(0.093)	(0.112)	(0.097)
$\Delta \ln(\text{property crime/capita})$	−0.154	−0.110	−0.137	−0.054	−0.127	−0.078
$(n = 576, p = 294)$	(0.223)	(0.045)	(0.045)	(0.047)	(0.038)	(0.041)
$\Delta \ln(\text{murder/capita})$	2.240	−0.131	−0.225	−0.061	−0.149	−0.066
$(n = 576, p = 294)$	(2.819)	(0.146)	(0.140)	(0.149)	(0.151)	(0.161)
B: Ferrara (2022)						
% Semiskilled Black workers	0.118	0.548	0.242	0.657	0.482	0.548
$(n = 4,903, p = 2,252)$	(0.126)	(0.167)	(0.126)	(0.153)	(0.137)	(0.167)
C: Enke (2020)						
Trump − avg. GOP	−3.676	−1.921	−2.120	−1.840		
$(n = 2,452, p = 978)$	(1.416)	(0.941)	(0.950)	(0.934)		
Voted for Trump in 2016	−12.399	−12.335	−12.357	−11.955		
$(n = 2,852, p = 1,068)$	(1.357)	(1.050)	(1.049)	(1.028)		
Voted for Trump in primaries	−5.318	−7.778	−8.618	−7.716		
$(n = 1,544, p = 824)$	(2.660)	(1.544)	(1.534)	(1.541)		

Notes: Col. 2 replicates SBEs for the BCH, Ferrara (2022), and Enke (2020) studies discussed in the text. Col. 1 reports the unpenalized OLS estimate for the same specification. Cols. 3 and 4 report the range of estimates obtained under alternative ways of dropping collinear columns of the control matrix. Cols. 5 and 6 report the range of estimates obtained under alternative normalizations of the controls prior to taking powers and interactions: no normalization, demeaning, subtracting the median, and setting the range to $[-1, 1]$ and $[0, 1]$. The outcome variables in Panel C are multiplied by 100. “Trump − avg GOP” refers to the difference between the indicator for voting for Trump in the 2016 general election minus the vote share given to Republican candidates in the previous two presidential elections. Standard errors are given in parentheses. These are robust for Enke (2020), clustered by state for BCH and by county for Ferrara (2022).

cohort's share of arrestees, which is outcome-specific). For the violent crime outcome, its standard deviation across states in 1997 is about 1, so that the estimate implies a 16% reduction in crime rate per standard deviation increase in effective abortion rate. However, this result is very sensitive to how we resolve the collinearity. To resolve it, we may drop any 3 of the 24 interactions with the shall-issue dummy, and any 2 of the 49 time-invariant controls; the same holds when we interact these with a time trend and its square. This gives $\binom{24}{3} \binom{49}{2}^3 \approx 3 \times 10^{12}$ possible ways of obtaining a full rank control matrix. Columns 3 and 4 report a range of estimates we obtain by randomly choosing among these possibilities, showing that, depending on the outcome, the estimates move by 1.2 to 1.9 standard errors.

Similar collinearity issues arise in the second application that replicates the analysis in Ferrara (2022), who studies to what extent post WW2 occupational upgrading of Black workers from low-skilled to semi-skilled can be attributed to war casualties among semi-skilled white soldiers. Using a decennial 1920–1960 unbalanced panel of county-level observations in 16 predominantly Southern US states, the study runs a two-way fixed effects specification with county and time fixed effects, interactions between state and time fixed effects, the share of semi-skilled Black workers as the outcome, and treatment given by the white casualty rate interacted with a post-war indicator. To purge time-varying confounders, the study also includes 24 baseline controls (including the county draft rate, WW2 spending, demographic and socioeconomic controls), their squares and interactions, as well as interactions between the 24 baseline controls and state and time effects. In addition, two baseline controls (number of slaves in 1860 and the unemployment rate in 1937) are also included in triple interactions with other controls, time and state effects. After dropping a reference state and a reference year in each interaction, and dropping

zero and repeated columns, the resulting matrix has 2270 columns, but rank of only 2252, because the state of Delaware only contains 15 observations, but there are 33 Delaware-specific controls.⁹ Ferrara (2022) first forms a full-rank control matrix, and then uses double- t selection to estimate the model.¹⁰

We replicate this specification in column 2 of Panel B in Table 1.1.¹¹ However, similar to panel A, there are many ways of resolving multicollinearity in the control matrix, because there are many ways of specifying a reference state and a reference year in each interaction, and, if we keep 3 Delaware county effects, $\binom{30}{18}$ ways of dropping Delaware-specific controls. As shown in columns 3 and 4, depending on how the multicollinearity is resolved, the estimates vary by over three standard errors depending on how we order the state and time effects.¹²

Our final empirical example uses data from Enke (2020), who examines how voters' moral values affect their voting patterns. Enke uses survey data to construct an index of the relative importance of universalist moral values (individual rights, justice, fairness) vs communal values (loyalty, respect). He then runs a regression of three different measures

⁹In particular, since the two baseline controls entering triple interactions are time-invariant and the specification includes county effects, we need to exclude the main effects of these controls, their squares and state-interactions, which yields a control matrix with 2588 columns. 154 columns are collinear since we need to drop a reference state and year in each interaction, and 18 Delaware-specific controls are also collinear. In addition, 164 columns are repeated or zero, yielding a rank equal to 2252.

¹⁰Specifically, Ferrara (2022) first runs a selection step that regresses the outcome and the treatment on the control matrix using OLS, and selects controls with a t -statistic that is larger than 2.575 in absolute value in either regression. In a second step, the outcome is regressed on the treatment and selected controls using OLS.

¹¹The estimate is slightly higher than in the original because, to keep missing data and normalization issues separate, we restrict the sample to the 4,903 observations with no missing values in both estimation steps described in footnote 10. In addition, we include year fixed effects in both the first and second step, rather than exclude them from the first step.

¹²We also considered sensitivity of the results to ordering of the columns in the control matrix when the double- t estimator is replaced by the post double lasso estimator. This yielded a range of variation in the estimates equal to about two thirds of a standard error.

of voting behavior on this index, controlling for 10 continuous or binary controls, as well as 5 sets of categorical variables.¹³ Columns 1 and 2 in panel C of Table 1.1 replicate the OLS and post-double lasso estimates reported in the paper. Here the inclusion of multiple sets of categorical variables necessitates a specification of a reference category for each set. Columns 3 and 4 report the range of estimates obtained by changing these reference categories, which varies between a third and a half of standard error depending on the outcome. This is a narrower range than in panels A and B, albeit still large enough to affect the economic interpretation of the estimates.

Columns 5 and 6 of Table 1.1 show the range of estimates we obtain from considering different normalizations when taking powers and interactions between variables in the BCH and Ferrara (2022) applications. In addition to no normalization (the choice in the original analyses), we consider demeaning, centering at the median, and setting the range to $[-1, 1]$ and $[0, 1]$. The table shows that such normalizations can again substantially move the estimates, by up to 1.3 standard errors.

1.1.2 Alternative Normalizations of the Control Matrix

We now consider two alternatives to the construction of W in the presence of categorical variables and power and interactions.

In Section 1.1.1, a categorical variable with k categories was always specified as a set of $k - 1$ dummies for individual categories, with a reference dummy dropped to prevent collinearity. We now consider expressing such variables as $k - 1$ indicators for different

¹³The continuous or binary controls comprise: political liberalism, log of household income, education, log of population density of respondent's ZIP code, religiosity, gender, employment indicator, altruism, measure of trust, and the absolute value of the moral values index. The categorical variables are: county, year of birth, ethnicity, religious denomination and occupation fixed effects.

subsets of the k categories. As discussed in Section 1.2.3, if the subsets are carefully chosen, such a specification may be more likely to lead to sparsity in certain contexts. Here we pick the subsets at random, subject to the constraint that they span the same column space. For the Enke (2020) application, columns 3 and 4 of Table 1.1 correspond to a special case of this exercise, when the subsets are constrained to be of cardinality one. Correspondingly, the range of variation in the estimates over all possible subsets, reported in columns 1 and 2 of Table 1.2, is considerably greater, exceeding 0.9 standard errors in all regressions. For the Ferrara (2022) application, the range exceeds two standard errors.¹⁴

Similarly, the centering of baseline variables at the mean vs the median we considered in the previous subsection can be thought of as a special case of a more general problem: we could in principle choose the centering constant λ to be any number. At a theoretical level, we are not aware of arguments for why $\lambda = 0$ (no normalization), or setting it to the sample mean should more plausibly lead to sparsity than other choices. To further explore the sensitivity to the choice of λ , we standardize the baseline variables, and then vary the specification of W by picking λ from the range $[-1, 1]$ for each baseline variable. To align the exercise with the analytical calculations in Section 1.2.4, we use Hermite rather than raw polynomials. Columns 3 and 4 of Table 1.2 show that the resulting variation in the estimates is again substantial, exceeding 1.7 standard errors in all specifications. The sign of the estimated effect is also sensitive to the choice of λ in two of the BCH specifications.

¹⁴The range does not exceed that in columns 3 and 4 of Table 1.1 because we do not vary which Delaware-specific controls are dropped.

Table 1.2: Fragility of sparsity-based methods to further normalizations of the control matrix.

Outcome	Sparsity-based estimation			
	Range (category sums)		Range (offset)	
	(1)	(2)	(3)	(4)
A: BCH				
$\Delta \ln(\text{violent crime/capita})$ ($n = 576, p = 294$)			-0.239 (0.105)	0.107 (0.093)
$\Delta \ln(\text{property crime/capita})$ ($n = 576, p = 294$)			-0.193 (0.047)	-0.011 (0.053)
$\Delta \ln(\text{murder/capita})$ ($n = 576, p = 294$)			-0.338 (0.160)	0.128 (0.195)
B: Ferrara (2022)				
% Semiskilled Black workers ($n = 4,903, p = 2,252$)	0.401 (0.134)	0.685 (0.162)	0.397 (0.148)	0.646 (0.164)
C: Enke (2020)				
Trump – avg. GOP ($n = 2,452, p = 978$)	-2.333 (0.918)	-1.476 (0.906)		
Voted for Trump in 2016 ($n = 2,852, p = 1,068$)	-12.946 (1.029)	-11.202 (1.008)		
Voted for Trump in primaries ($n = 1,544, p = 824$)	-9.269 (1.526)	-7.083 (1.525)		

Notes: Cols. 1 and 2 report the range of estimates obtained under alternative ways of expressing a set of categorical variables. Cols. 3 and 4 report the range of estimates obtained when powers are constructed using Hermite polynomials, with an offset λ ranging between -1 and 1 . See notes to Table 1.1 for a description of outcome variables. Standard errors are given in parentheses. These are robust for Enke (2020), clustered by state for BCH and by county for Ferrara (2022).

1.2 Sparse Representations Are Rare

In this section we consider the thought experiment of picking a particular linear representation of the controls W at random. In particular, we consider (i) a full rotation of the column space; (ii) different ways of controlling for categorical variables; and (iii) an offset for the scalar variable in a polynomial series regression. We find that sparse representations are exceedingly rare in each example. This suggests that in a given application, undiscerning default choices for expressing the column space will typically not satisfy the sparsity assumptions. This helps to explain the findings of the previous section: most normalization choices cannot yield sparse representations. As we discuss in detail in Section 1.2.5 below, it also implies that the representation of the control matrix is a substantive choice and not just an implementation detail, like, say, the choice of the penalty parameter: it impacts not just the finite-sample estimate, but also the large-sample validity of SBE-based inference.

1.2.1 Approximate Sparsity

For simplicity, the calculations in this section focus on sparsity in the outcome regression (1.1). Let A be a full-rank $p \times p$ matrix. Then an alternative expression for the column space of W_i is $\tilde{W}_i = AW_i$ with coefficient $\tilde{\gamma} = A'^{-1}\gamma$, so that $W_i'\gamma = \tilde{W}_i'\tilde{\gamma}$. Note that any such transformation leaves the OLS coefficient on D_i numerically unaltered.

In our examples, γ is exactly sparse, with just one non-zero element, $\|\gamma\|_0 = 1$. The researcher uses the transformed regressors $\tilde{W}_i$ with coefficients $\tilde{\gamma}$, which typically is not exactly sparse. However, for sparsity-based inference to be asymptotically valid, it suffices for $\tilde{\gamma}$ to be *approximately sparse*. In particular, under our maintained assumption that

$p \asymp n$, it suffices that for a sparsity index

$$s = o(\sqrt{p}/\log p) \quad (1.3)$$

the mean square approximation error of $W_i'\gamma$ satisfies

$$\min_{\|v\|_0 \leq s} E[(W_i'\gamma - \tilde{W}_i'v)^2] = O(s/p) \quad (1.4)$$

(cf. Belloni et al., 2014, page 614). Here, and throughout the paper, all limits are taken as $p \rightarrow \infty$ (or, equivalently, as $n \rightarrow \infty$).

1.2.2 Rotation

By way of establishing a benchmark, we start with an extreme case, and consider the set of transformations

$$\tilde{W}_i = RW_i, \quad R \in \mathbf{R} = \{A \in \mathbb{R}^{p \times p} : A'A = I_p, \det(A) = 1\},$$

that is, $\tilde{W}_i$ is obtained by rotating W_i by R , and $\tilde{\gamma} = R\gamma$. For the purposes of studying sparsity, this is a very large set of transformations, as for any γ with $\|\gamma\|_2 > 0$, there exists R such that $\tilde{\gamma}$ has only one non-zero element. At the same time, there also exists an R such that all elements of $\tilde{\gamma}$ are equal (and equal to $\|\gamma\|_2/\sqrt{p}$).

Now suppose we take a random draw $\mathcal{R}$ from $\mathbf{R}$, with distribution equal to the Haar measure (so that for any $R \in \mathbf{R}$, $R\mathcal{R} \sim \mathcal{R}$). Recall that the multivariate standard normal distribution is spherically symmetrical, that is, for $\mathcal{Z} \sim \mathcal{N}(0, I_p)$ and any $R \in \mathbf{R}$, $R\mathcal{Z} \sim$

$\mathcal{Z}$. This implies that the induced distribution of $\tilde{\gamma} = \mathcal{R}\gamma$ satisfies

$$\mathcal{R}\gamma \sim \frac{\|\gamma\|_2}{\|\mathcal{Z}\|_2} \mathcal{Z}, \quad (1.5)$$

so the p elements of $\mathcal{R}\gamma$ are, up to a common rescaling, i.i.d. standard normal. The normal distribution is thin-tailed, making it very unlikely that a few largest absolute values of $\mathcal{Z}$ dominate. This suggests that with very high probability, $\mathcal{R}\gamma$ is not approximately sparse, as formalized in the following result.

Theorem 1.1. *Suppose that the eigenvalues of $E[W_i W_i']$ are bounded away from zero and infinity, and that $\|\gamma\|_2 \asymp 1$. Then the logarithm of the probability that the model with regressor $\tilde{W}_i = \mathcal{R}W_i$ satisfies eqs. (1.3) and (1.4) is of the order $-\frac{p}{4} \log p$.*

Given that a sparsity assumption allows for more informative inference, and that any coefficient vector can be rotated into an (extremely) sparse one, it is not surprising that a random rotation only rarely yields a sparse representation. The contribution of Theorem 1.1 is to quantify just how exceedingly rare this event is: for $p \geq 50$, say, $p^{-p/4} < 10^{-21}$. Of course, researchers rarely consider all rotations when specifying the regressor matrix. The relevance of this quantification lies in showing that even when the measure of plausible rotations is orders of magnitude smaller than the full set, the probability of arriving at a sparse representation is still minuscule.

1.2.3 Categorical Data

An empirically common form of controls is dummies for categorical variables. If there are p underlying categories, and a constant is included, then one must designate a reference

category by dropping one dummy to avoid perfect collinearity. Suppose the model is exactly sparse with sparsity index s when one chooses the right reference category. This means that $p - s$ categories have the same coefficient as the reference category. Thus, if one were to pick a reference category at random, one has a $1 - s/p$ chance of inducing sparsity, which is a probability close to one under (1.3).

However, an assumption that many categories have the same effect is not the only way to induce sparsity for a categorical variable. Suppose, for instance, that we want to non-parametrically control for a variable that measures age. Then a plausible form of sparsity arises under an assumption that the age effect is a step function, with some dividing line between young and old. If this threshold is not close to either very young or very old, then a reference category approach will not yield sparsity, but an appropriate re-expression of the column space will. Or maybe there are three distinct coefficients, one for the young, one for the middle-aged and one for the old, which again requires a different specification of the fixed effects to induce sparsity. Such specifications are less common in applied work. Arguably this is not because they are a priori less implausible, but rather, it is well understood that under OLS, they are all equivalent to the reference category specification. However, when imposing a sparsity assumption on the coefficients, this equivalence breaks down.

A general specification of the column space with categorical data takes the form

$$W_i = A_0 Z_i, \quad A_0 \in \mathbf{A} = \{A \in \mathbb{R}^{p \times p} : A_{ij} \in \{0, 1\}, \quad A \text{ is full rank}\}, \quad (1.6)$$

where the Z_i indicates the baseline categories (so each Z_i is a column of I_p). By construction, elements of W_i are all equal to zero or one in this specification, and since $A \in \mathbb{R}^{p \times p}$ is full rank, the constant vector is part of the column space spanned by W_i .

Suppose with $A_0 \in \mathbf{A}$, the coefficient vector on W_i is sparse. Now consider picking $\mathcal{A} \in \mathbf{A}$ at random by repeatedly drawing $p \times p$ matrices with i.i.d. Bernoulli(q) entries, $0 < q \leq 1/2$, until we find one that is full rank. By Theorem A of Tikhomirov (2020), the probability of discarding a matrix in this process is vanishingly small, as it is smaller than $(1 - q + \varepsilon)^p$ for all $\varepsilon > 0$ and large enough p . This allows us to essentially ignore rank-deficient matrices, yielding the following result.

Theorem 1.2. *Suppose a single coefficient on $W_i = A_0 Z_i$ is constant and non-zero, and the number of zeros K in the corresponding row of A_0 satisfies $0 < \lim_{n \rightarrow \infty} K/p < 1$. If all baseline categories have population fractions of the same order, then the probability that the model with $\tilde{W}_i = \mathcal{A} Z_i$ satisfies eqs. (1.3) and (1.4) is no larger than*

$$(1 - q + \varepsilon)^K$$

for all $\varepsilon > 0$ and large enough p .

1.2.4 Offset in Hermite Polynomial Regression

Our last example involves a large number of technical regressors. Specifically, suppose there is a scalar variable $z_i \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$ which we want to control for nonparametrically

by a series regression. We consider scaled Hermite polynomials

$$H_j(x) = (j!)^{-1/2}(-1)^j e^{x^2/2} \frac{d^j}{dx^j} e^{-x^2/2}, \quad j = 0, \dots, p-1$$

which are particularly convenient for an underlying standard Gaussian variable, since we have $E[\tilde{W}_i \tilde{W}_i'] = I_p$ for $\tilde{W}_{i,j} = H_{j-1}(z_i)$.

Suppose further that if z_i is shifted by a constant $\lambda \in \mathbb{R}$, the expansion is extremely sparse, with only the highest order term being non-zero, that is, $Y_i = D_i \beta + H_{p-1}(z_i + \lambda) + U_i$. If we were to draw λ uniformly on $[0, 1]$, what is the probability that a researcher ignoring the shift obtains an approximately sparse representation? The following result shows that the probability will be bounded above by $\varepsilon/\log(p)$, for all $\varepsilon > 0$ and large enough p .

Theorem 1.3. *Suppose $\lambda = L/\log(p)$, $L > 0$. If L is fixed, then for $1 \leq j \leq L\sqrt{p}/\log(p)$ and $p \geq \max\{(2L)^2, 6\}$, $\tilde{\gamma}_{p-j}^2 \geq Ce^{j/2}$, where C is an absolute constant, and eqs. (1.3) and (1.4) fail. In contrast, if $L \rightarrow 0$, eqs. (1.3) and (1.4) hold.*

While the rate of $\log(p)$ in Theorem 1.3 is quite slow, note that when sparsity fails, it does so dramatically in the sense that many coefficients diverge. This is not surprising. It is well-known that assuming that the first s (say) terms of the polynomial expansion provide a good approximation to the regression function amounts to imposing a smoothness assumption on it. In this case, the exact basis approximation would matter little, and sparsity-based methods would perform well regardless of the choice of λ . However, under smoothness SBEs are not needed, since a low-order series expansion (i.e. only controlling for the first s polynomial terms, with $s \ll p$) already performs well.

In the current context with technical regressors, the appeal of SBEs is that the assumption of sparsity they leverage is weaker than assuming smoothness, since we allow for the best s -term approximation to include high order polynomials. But if a good approximation to the regression function indeed requires high-order terms, it must be very non-smooth—and for non-smooth functions even small changes in the approximating basis imply wild swings in the coefficients. In other words, while sparsity is indeed weaker than smoothness, the particular class of non-smooth functions sparsity allows for is tightly linked to the basis used. If one suspects non-smoothness, and one wants to avoid running OLS with many polynomial terms, one therefore needs a substantive argument for why the particular basis chosen is likely to capture it.

1.2.5 Discussion

The fact that the sparsity assumption depends on normalization choices in the specification of W is well-known in the theoretical literature. The three stylized thought-experiments above quantify the extent of this dependence: the vast majority of normalization choices do not induce an approximately sparse representation. The results in Section 1.1 show that this lack of invariance induces variation in SBE estimates that is of the same order as sampling uncertainty. This fragility appears underappreciated in the literature. We reviewed the 50 most cited papers applying SBEs, and found that, with one exception, none made explicit what the normalization choices were, let alone discussed why they were appropriate to induce sparsity.¹⁵

¹⁵We ordered all papers citing one of the “debiased lasso” papers mentioned in the introduction by number of citations on Google Scholar. One paper mentioned that it included all dummy variables and did

A natural reaction to these results is to seek to modify SBEs in a way that they adapt to different forms of sparse representations. For instance, it may be possible to treat the offset, λ , as another parameter to be estimated. Similarly, a number of proposals have been developed that modify the lasso to make it invariant to the choice of reference category, including the group lasso (Yuan & Lin, 2006), or variants of fused lasso (Tibshirani et al., 2005), such as all-pairs penalties (Bondell & Reich, 2009; Gertheiss & Tutz, 2010) or the SCOPE estimator (Stokell et al., 2021). When combined with the debiasing techniques developed in the literature on debiased lasso (e.g. Javanmard & Montanari, 2014; van de Geer et al., 2014; Zhang & Zhang, 2014; Belloni et al., 2014), these approaches may yield algorithms that are less fragile.

However, it appears non-trivial to extend these modifications to the more complicated designs encountered in practice, such as when interactions are present, or to handle collinearity issues arising due to limited variability in the controls. Both the BCH and the Ferrara (2022) applications of Section 1.1 exhibit multicollinearity that goes beyond the issue of choosing baseline categories.

An alternative approach could be to exhaustively consider all plausible normalization choices, analogous to our analysis in Section 1.1, and report the union of the confidence intervals. However, this is practically difficult, as the number of possible normalizations is often very large. Moreover, such an approach would only yield valid inference if at least one of the considered choices induces a sparse representation.

To ensure that the latter condition holds, one needs a substantive argument. In some instances, such arguments may necessitate further expanding the set of normalization

not pick a reference category; the paper did not discuss the implications of this choice for the restricted eigenvalue condition, or whether other normalization choices were made.

choices. For instance, the grouped lasso approach in the references cited above require that the baseline categories can be partitioned into a small number of types, with all categories of the same type having (nearly) identical effect on the outcome variable. In the context of controlling for profession dummies, say, this amounts to an assumption that there are only a few profession types with heterogeneous effects, which is a strong assumption. It may be more reasonable to instead posit that professions are characterized by a combination of a small number of latent skills, and these skills affect the outcome in an additive manner. This then yields a different set of plausible normalizations that are not captured by grouped lasso-type approaches.

In absence of substantive arguments, the sparsity assumption may well fail to hold, even after considering a large set of potential normalization choices. We present evidence for this in our applications in Section 1.4 below. Thus, fully data-driven methods to modify SBEs such that they adapt to an appropriate sparse representation regardless of the context does not seem like a feasible solution.

Applications of SBEs require researchers to take a stand on why a particular representation (or a set of representations) admits a sparse representation based on domain-specific knowledge. This is analogous to using problem-specific arguments to defend other substantive assumptions, such as assuming selection on observables. But it is at odds with the purported appeal of SBEs to be fully automatic, with the aim of reducing the researchers' degrees of freedom in model choice.

While the sensitivity of SBEs to the control matrix specification may seem similar to the usual issue that with flexible methods, implementation issues, such as tuning parameter choice, can matter a lot in finite samples, it is fundamentally different. It also

differs from the well-known fact that the choice of basis matters in implementing series regression. Under smoothness, many basis choices for series regression are theoretically justified, as well as being first-order equivalent in particular settings, such as the partially linear model. Large sample validity of flexible methods can be justified under a wide range of tuning parameter choices, including the choice of penalty in implementing SBEs. In contrast, because the specification of the control matrix affects the validity of the sparsity assumption, and thus large-sample validity of inference, it is a substantive modeling choice, not merely an implementation detail. It needs to be defended as such.

1.3 Efficiency Gains Under Sparsity

One argument for using SBEs when p is large, but still smaller than the sample size, is that OLS estimates are too noisy. In this section, we quantify the potential efficiency gains of alternative estimators relative to OLS, and we argue that the potential gains are limited unless p is comparable to n .

We consider the model given in eqs. (1.1) and (1.2) in the introduction. We maintain the assumption that the residuals U_i and $\tilde{D}_i$ are conditionally mean zero to avoid technical complications; the assumption could be replaced by an assumption that restricts these conditional means to be small, so that the linear specifications in eqs. (1.1) and (1.2) could be thought of as approximating non-linear conditional means (see, e.g., Cattaneo et al., 2018b). Using the Frisch-Waugh-Lowell theorem, we can write the OLS estimator of β as

$$\hat{\beta}_{OLS} := \frac{\sum_{i=1}^n \ddot{D}_i Y_i}{\sum_{i=1}^n \ddot{D}_i^2},$$

where $\ddot{D}_i$ denotes the sample propensity score residual from estimating eq. (1.2) by OLS. Specifically, employing the usual matrix notation, $\ddot{D}_i$ corresponds to the i th element of $\ddot{D} := (I - P)D$, where $P = WW^+$ is the projection matrix onto the space spanned by the controls. Here W^+ denotes the pseudo-inverse (so that if W is full rank, $P = W(W'W)^{-1}W'$).

To analyze $\hat{\beta}_{OLS}$, we need to impose more structure on the model in eqs. (1.1) and (1.2).

Assumption 1.1. *For some constants $\eta > 0$ and $K \geq 1$: (i) $\{(U_i, \tilde{D}_i)\}_{i=1}^n$ are independent across i conditional on W ; (ii) $E[|U_i|^{2+\eta} | D, W] + E[|\tilde{D}_i|^4 | W] \leq K$ uniformly over i ; (iii) $1/E[\tilde{D}_i^2 | W] + 1/E[U_i^2 | D, W] \leq K$ uniformly over i ; (iv) $\limsup_{n \rightarrow \infty} p/n < 1$.*

Parts (i)–(iii) of assumption 1.1 are standard assumptions on sampling and the regression errors, ensuring that the errors are neither too thick-tailed nor degenerate. Part (iv) ensures that, when OLS estimates of (1.1) implicitly estimate the propensity score regression, we do not overfit so much that we eliminate any variation in the sample residuals $\ddot{D}_i$. This overfitting precludes using OLS in settings with $p > n$. However, as long as the limit of p/n is strictly smaller than 1, the OLS estimator remains asymptotically normal with the usual sandwich expression for its standard error:

Lemma 1.1. *Consider the model in eqs. (1.1) and (1.2). Under Assumption 1.1,*

$$\frac{\hat{\beta}_{OLS} - \beta}{s_{OLS}} \xrightarrow{d} \mathcal{N}(0, 1), \quad s_{OLS}^2 = \frac{1}{(\ddot{D}'\ddot{D})^2} \sum_{i=1}^n \ddot{D}_i^2 U_i^2.$$

The fact that OLS remains asymptotically normal in the regime $p \asymp n$ is not new—results similar to Lemma 1.1 appeared previously in, for instance, Cattaneo et al. (2018b)

who build on earlier work by Mammen (1993). We state it here under a simpler set of assumptions to allow us to consider its implications for potential efficiency gains. To this end, observe that the asymptotic variance is in general larger than the semiparametric efficiency bound unless $p/n \rightarrow 0$. In particular, when the errors U_i are homoskedastic, the efficient standard error is given by the square root of

$$s_*^2 = \frac{1}{(\tilde{D}'\tilde{D})^2} \sum_{i=1}^n \tilde{D}_i^2 U_i^2,$$

in the sense that the semiparametric efficiency bound is given by the probability limit of s_*^2/n (see, e.g., Robinson, 1988). Semiparametrically efficient estimators $\hat{\beta}^*$ of β thus need to satisfy

$$\hat{\beta}^* - \beta = (\tilde{D}'\tilde{D})^{-1} \tilde{D}'U + o_p(n^{-1/2}), \quad (1.7)$$

This lack of semiparametric efficiency is not a deficiency of OLS: under Gaussian errors, one-sided t -tests based on OLS are uniformly most powerful. Rather, it reflects the fact that when $p \asymp n$, achieving the semiparametric efficiency bound requires some type of restriction on γ . It turns out that restricting γ to be sparse (see Section 1.2.1) is sufficient; indeed a number of SBEs satisfy eq. (1.7) (e.g. Belloni et al., 2014; Javanmard & Montanari, 2014; van de Geer et al., 2014; Zhang & Zhang, 2014).

The standard error ratio s_*/s_{OLS} thus represents the potential efficiency gain from imposing sparsity. When U_i is homoskedastic and Assumption 1.1 holds, the ratio satisfies

$$\frac{s_*^2}{s_{OLS}^2} = (1 - p/n) \frac{\tilde{D}'\tilde{D}/(n-p)}{\tilde{D}'\tilde{D}/n} (1 + o_p(1)) = (1 - p/n) \kappa (1 + o_p(1)), \quad \kappa = \frac{E[\hat{\sigma}_{\tilde{D}}^2]}{\sigma_{\tilde{D}}^2},$$

where $\sigma_{\tilde{D}}^2$ is the variance of the error in the propensity score regression, and $\hat{\sigma}_{\tilde{D}}^2 = \tilde{D}'\tilde{D}/(n - p)$, the mean squared error in the propensity score regression, can be thought of as an estimator of $\sigma_{\tilde{D}}^2$. The ratio κ measures the relative bias of this estimator, with $\kappa = 1$ when the estimator is unbiased. This is the case when $\tilde{D}_i$ is homoskedastic, and the standard error ratio s_*/s_{OLS} then simplifies to a degrees-of-freedom correction, $\sqrt{1 - p/n}$, as noted previously in Cattaneo et al. (2018a). When $\tilde{D}_i$ is heteroskedastic, $\hat{\sigma}_{\tilde{D}}^2$ may display downward bias if the leverages P_{ii} are positively correlated with the conditional variances $E[\tilde{D}_i^2 \mid W]$; this can be seen from writing $\kappa = E[\sum_i (1 - P_{ii})E[\tilde{D}_i^2 \mid W]] / \sum_i (1 - P_{ii})E[\tilde{D}_i^2]$. However, the correlation would need to be substantial to induce downward bias that is sizable enough to allow for large efficiency gains. When $p/n = 0.2$ and $\kappa = 0.9$, for instance, corresponding to 10% downward bias, the standard error ratio implies a $1 - \sqrt{0.8 \cdot 0.9} = 15.1\%$ reduction in standard error. Consequently, the ratio p/n needs to be much larger than 0.2 to allow for sizable efficiency gains under sparsity: at $\kappa = 0.9$, we need $p/n \geq 6/16 \approx 0.38$ to reduce the standard errors by more than 25%.

While these efficiency calculations rely on homoskedastic errors U_i , similar conclusions are likely to hold under heteroskedasticity. The argument for this is that, in general,

$$\frac{s_*}{s_{OLS}} = \frac{s_*/s_{*,\text{hom}}}{s_{OLS}/s_{OLS,\text{hom}}} \sqrt{(1 - p/n)\kappa} \cdot (1 + o_p(1)),$$

where $s_{*,\text{hom}} = \sigma_U / \sqrt{\tilde{D}'\tilde{D}}$ and $s_{OLS,\text{hom}} = \sigma_U / \sqrt{\tilde{D}'\tilde{D}}$ are homoskedasticity-only standard error formulas. The ratios $s_*/s_{*,\text{hom}}$ and $s_{OLS}/s_{OLS,\text{hom}}$ measure the magnitude of the heteroskedasticity correction on the standard errors. Heteroskedasticity would thus need

to have a large *differential* impact on the standard errors for OLS versus $\hat{\beta}_*$ for the standard error ratio to deviate much from $\sqrt{(1 - p/n)\kappa}$. But magnitudes of heteroskedasticity corrections tend to be modest in practice, so that substantive efficiency gains are unlikely when the controls number 20% or less of the sample size. In such cases, simply running OLS offers a robust method of inference at little efficiency loss.

1.4 Testing Sparsity

In this section, we develop two tests of the sparsity assumption under our maintained assumption that p is smaller than, but proportional to the sample size. We then apply these tests to the empirical examples studied in Section 1.1.

1.4.1 Hausman Test

The first test we develop is a simple application of the idea popularized by Hausman (1978) that if we have two estimators, one of which is more efficient, but requires stronger assumptions for its validity, we can indirectly test these stronger assumptions by checking whether the estimates are statistically significantly different. The next lemma formalizes the result.

Lemma 1.2. *Consider the model in eqs. (1.1) and (1.2). Suppose Assumption 1.1 holds, and that $\text{tr}(P)/n \geq \tilde{K}$ a.s. for some $\tilde{K} > 0$. Then, for any estimator satisfying eq. (1.7),*

$$\frac{\hat{\beta}_{OLS} - \hat{\beta}_*}{s_H} \xrightarrow{d} \mathcal{N}(0, 1), \quad s_H^2 = \sum_{i=1}^n Z_i^2 U_i^2,$$

where $Z_i = \frac{\ddot{D}_i}{\ddot{D}'\ddot{D}} - \frac{\tilde{D}_i}{\tilde{D}'\tilde{D}}$.

Additionally, suppose the regression functions admit the decomposition $W_i'\gamma = f(W_i) + r_\gamma(W_i)$ and $W_i'\delta = g(W_i) + r_\delta(W_i)$, where the remainder terms satisfy (i) $\frac{1}{n} \sum_{i=1}^n E[r_\delta(W_i)^2 + r_\gamma(W_i)^2] \rightarrow 0$; (ii) $\frac{1}{n} \sum_{i=1}^n E[r_\gamma(W_i)^2 r_\delta(W_i)^2]$ is bounded; and (iii) $n \sum_{i=1}^n (Z_i^2 U_i^2 - \tilde{z}_i^2 (U_i + r_\gamma(W_i))^2) = o_p(1)$, with $\tilde{z}_i = \frac{\ddot{D}_i}{\ddot{D}'\ddot{D}} - \frac{\tilde{D}_i + r_\delta(W_i)}{\sum_{i=1}^n (\tilde{D}_i + r_\delta(W_i))^2}$. Suppose also that for some estimates $\hat{U} = \hat{U}(Y, D, W)$ and $\hat{D} = \hat{D}(W, D)$ (iv) $\max_i |\hat{U}_i - U_i - r_\gamma(W_i)| + \max_i |\hat{D}_i - D_i - r_\delta(W_i)| = o_p(1)$. Then the same conclusion holds with s_H^2 replaced by $\hat{s}_H^2 = \sum_{i=1}^n \hat{Z}_i^2 \hat{U}_i^2$, where $\hat{Z}_i = \frac{\ddot{D}_i}{\ddot{D}'\ddot{D}} - \frac{\hat{D}_i}{\hat{D}'\hat{D}}$.

The second part of Lemma 1.2 allows us to construct a simple plug-in estimator of the standard error s_H based on lasso or post-lasso residuals. In particular, when the regression functions admit a sparse approximation in the sense of eqs. (1.3) and (1.4), the additional condition (i) in the second part of Lemma 1.2 will hold with f and g given by the best sparse approximations to $W_i'\gamma$ and $W_i'\delta$, and condition (iv) will hold for the lasso or the post-lasso residuals. Conditions (ii) and (iii) are high-level conditions ensuring that if we include the approximation errors r_δ and r_γ in the definition of the residuals, replacing U_i with $U_i + r_\gamma(W_i)$, and $\tilde{D}_i$ with $\tilde{D}_i + r_\delta(W_i)$, this has negligible impact on the standard error s_H in large samples; it is similar to the condition ASTE (P) (v) in Belloni et al. (2014). We note that Lemma 1.2 only imposes very weak conditions on the control matrix W , allowing it to be reduced-rank, so long as the rank is proportional to p , and allowing the rows to be dependent, and not identically distributed.

Sometimes, SBEs are used as a “robustness check” alongside a main specification based on OLS. Lemma 1.2 shows that such practice is in fact the opposite of a robustness check: if the two estimates are not close to one another this indicates failure of the sparsity

assumption rather than lack of robustness in the OLS estimates. When U_i is homoskedastic, the Hausman standard error may be written as

$$s_H^2 = (s_{OLS}^2 - s_*^2)(1 + o_p(1)) = s_*^2 \left(\frac{1}{(1 - p/n)\kappa} - 1 \right) (1 + o_p(1)),$$

so that when the efficiency gain $\sqrt{(1 - p/n)\kappa}$ is small, the two estimates need to be tightly coupled, within a fraction of the SBE standard error.

1.4.2 Residual Test

Our second approach to testing sparsity is based on the idea that if the identities $\mathcal{S}^* \subseteq \{1, \dots, p\}$ of the controls that give the best sparse approximation to the regression function are known, testing sparsity is equivalent to testing that the coefficients on the remaining controls are small, which can be gauged using a conventional F -statistic. Although the identities $\mathcal{S}^*$ are unknown in practice, it turns out that under the null hypothesis of sparsity, the residuals from the infeasible short regression that only includes controls in $\mathcal{S}^*$ are sufficiently well approximated by residuals from a lasso regression, so that the hypothesis can be tested by comparing the lasso and OLS residuals.

To make the result precise, consider a linear regression

$$Y_i = X_i' \alpha + \epsilon_i, \quad E[\epsilon_i \mid X_i] = 0, \quad i = 1, \dots, n, \quad (1.8)$$

with ϵ_i independent across i , conditional on the regressors, and $p := \dim(X_i) < n$. In our setup, eq. (1.8) may correspond to one of three regressions. For testing sparsity of γ in eq. (1.1), which all SBEs require, we can set $Y_i = Y_i$ and $X_i = (D_i, W_i')'$. For testing

sparsity in the propensity score regression, which is needed, for instance, for the validity of the post-double selection estimator of Belloni et al. (2014), we can set $Y_i = D_i$ and $X_i = W_i$. Sparsity in both regressions can be tested jointly in the reduced form regression of $Y_i = Y_i$ onto $X_i = W_i$.

We wish to test the assumption that the regression function $X_i'\alpha$ admits a sparse approximation. To work out the implications of this hypothesis, we introduce some additional notation. For a subset $\mathcal{S} \subseteq \{1, \dots, p\}$ of the regressors, let $X_{\mathcal{S}}$ denote the submatrix of X that drops the columns corresponding to the complement of $\mathcal{S}$, let $P_{\mathcal{S}} = X_{\mathcal{S}}X_{\mathcal{S}}^+$ denote the projection matrix associated with $X_{\mathcal{S}}$, and let $P = XX^+$ denote the full projection matrix. In a slight departure from Section 1.2, we gauge the quality of the approximation conditional on X , so that the approximation error from only using the regressors $X_{\mathcal{S}}$ is given by $(I - P_{\mathcal{S}})X\alpha$, the residual from projecting $X\alpha$ onto $X_{\mathcal{S}}$. The assumption that $X\alpha$ is sparse can then be stated as:

Assumption 1.2. *There exists a subset $\mathcal{S}^* \subseteq \{1, \dots, p\}$ with cardinality s , such that $\|(I - P_{\mathcal{S}^*})X\alpha\|_2^2 = O_p(s)$ and $s \log(p)/\sqrt{p} \rightarrow 0$.*

If we knew the identity of the subset $\mathcal{S}^*$, and we also assumed that the sparsity was exact, so that the $O_p(s)$ term in Assumption 1.2 was zero, then a natural way of testing Assumption 1.2 would be to compare the restricted and unrestricted sum of squared residuals,

$$\mathcal{F} = Y'(I - P_{\mathcal{S}^*})Y - Y'(I - P)Y.$$

The statistic $\mathcal{F}$ is the numerator of the homoskedastic F -statistic. While the F -statistic critical values are only valid under homoskedasticity, we can leverage the fact that when

$p \rightarrow \infty$, as is the case under our asymptotics, $\mathcal{F}$ is asymptotically normal after centering and scaling to derive a critical value that is robust to heteroskedasticity. Furthermore, if we have an estimator $\hat{\alpha}$ such that $Y - X\hat{\alpha}$ approximates the residuals $(I - P_{S^*})Y$ from the infeasible short regression sufficiently well, we can replace the infeasible sum of squared residuals $Y'(I - P_{S^*})Y$ in $\mathcal{F}$ by $\|Y - X\hat{\alpha}\|_2^2$ to derive a feasible version of this test. Finally, it turns out that weakening exact sparsity to approximate sparsity in the sense of Assumption 1.2 does not impact the null rejection probability of the test in large samples. The next lemma formalizes these arguments.

Lemma 1.3. *Suppose Assumption 1.2 holds and that, for some $K \geq 1$, (a) $E[\epsilon_i^4 | X] \leq K$; (b) $\max_i P_{ii} < 1 - 1/K$; (c) $\text{tr}(P)/n \geq 1/K$; and (d) $E[\epsilon_i^2 | X] \geq 1/K$ a.s. Then, as $n \rightarrow \infty$,*

$$\frac{\mathcal{F} - \sum_i \epsilon_i^2 P_{ii}}{\sqrt{2 \sum_{i \neq j} \epsilon_i^2 \epsilon_j^2 P_{ij}^2}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (1.9)$$

Furthermore, suppose that for some estimator $\hat{\alpha}$, (i) $\|X\hat{\alpha} - X\alpha\|_2 \preceq_p \sqrt{s \log(p)}$; (ii) $\|\hat{\alpha} - (X'_{S^*} X_{S^*})^{-1} X'_{S^*} X \alpha\|_1 \preceq_p s \sqrt{\log(p)/n}$; and, in addition, (iii) $\max_{ij} |X_{ij}| s \sqrt{\log(p)/n} = o_p(1)$; and (iv) $\|\sum_i \epsilon_i P_{ii} X_i\|_\infty + \|\sum_i \epsilon_i X_i\|_\infty \preceq_p \sqrt{n \log(p)}$. Then, letting $\hat{\epsilon}_i = Y_i - X_i' \hat{\alpha}$,

$$\frac{\|\hat{\epsilon}\|_2^2 - Y'(I - P)Y - \sum_i \hat{\epsilon}_i^2 P_{ii}}{\sqrt{2 \sum_{i \neq j} \hat{\epsilon}_i^2 \hat{\epsilon}_j^2 P_{ij}^2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Conditions (a) and (d) are standard assumptions on the regression errors, while conditions (b) and (c) impose weak restrictions on the design matrix; condition (c) is analogous to the condition imposed in Lemma 1.2, and (b) bounds leverage away from one. Conditions (i) and (ii) in the second part of the lemma are standard mean-squared error and ℓ_1

rate conditions that hold for the lasso and post-lasso estimators (Bickel et al., 2009; Belloni & Chernozhukov, 2013). Conditions (iii) and (iv) are tail restrictions on the covariates and residuals analogous to those in Belloni et al. (2014).

Lemma 1.3 implies that we can test the sparsity assumption by calculating the lasso or post-lasso residuals $\hat{\epsilon}_i$, and then checking whether the residual sum of squares of the lasso is comparable to that of the OLS residual sum of squares. If the difference satisfies

$$\|\hat{\epsilon}\|_2^2 - Y'(I - P)Y \geq z_{1-\alpha} \sqrt{2 \sum_{i \neq j} \hat{\epsilon}_i^2 \hat{\epsilon}_j^2 P_{ij}^2 + \sum_i \hat{\epsilon}_i^2 P_{ii}}$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution (1.645 for $\alpha = 5\%$ level test), we reject Assumption 1.2.

In general, testing hypotheses about α when $\lim_{n \rightarrow \infty} p/n > 0$ can be quite involved, since plugging in regression residuals into asymptotic variance expressions leads to bias (see, e.g., Kline et al., 2020; Anatolyev & Sølvesten, 2023; Cattaneo et al., 2018b). As shown in the proof of Lemma 1.3 (see eq. (1.29)), we avoid such difficulties here by virtue of the fact that under the assumptions of the lemma, the lasso residuals $\hat{\epsilon}_i$ are sufficiently accurate so that the variance part of the test statistic, the denominator in eq. (1.9), can be consistently estimated under the null hypothesis.

1.4.3 Empirical Tests of Sparsity

We now apply the tests developed in Sections 1.4.1 and 1.4.2 to the empirical illustrations considered in Section 1.1. For each specification, we consider the Hausman test, which compares OLS with SBE, as well two versions of the residual test. The first version esti-

mates the outcome regression in eq. (1.1) using the post-lasso, and compares the residuals to those based on OLS. The second version compares post-lasso and OLS residuals from estimating the propensity score regression in eq. (1.2).

Table 1.3 reports the results. Column 1 applies the test to the original specification in each paper. We see that for 6 of the 7 outcomes, at least one of the tests rejects the assumption of sparsity.

To check that these results are not driven by finite-sample size distortions of these tests, Section 1.C conducts Monte Carlo simulations based on these applications; in these simulations, the size stays close to or below the nominal level.

One response to these findings is to seek alternative normalizations of the control matrix that are consistent with the sparsity assumptions. The remaining columns in Table 1.3 report the range of p -values under the four different normalizations we considered in Section 1.1. The table shows that in these applications, we were unable to find a normalization for five of the seven outcomes where at least one test did not reject. This is in spite of the large number of alternative specifications that these normalizations generate.

1.5 Summary and Conclusions

We have argued, using empirical evidence and theoretical arguments, that SBEs display a lack of robustness to the specification of the control matrix. In the three applications we have examined, the range of variation in the SBE estimates under equally plausible alternative specifications is of the same order of magnitude as sampling uncertainty. Two reasons underlie this fragility. First, whether a small number of covariates can account for the bulk

Table 1.3: p -values (in percentages) for tests of the sparsity assumption under different normalizations of the control matrix.

		Repl.	Collinearity		Powers		Category sums		Offset	
Outcome	Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A: BCH										
$\Delta \ln(\text{violent crime/capita})$	H	81.7	76.0	87.4	81.7	85.8			73.1	100.0
	OR	9.5	9.5	9.5	9.5	9.5			8.2	20.2
	PR	0.3	0.0	0.9	0.0	0.6			0.0	2.2
$\Delta \ln(\text{property crime/capita})$	H	82.6	61.8	93.4	70.9	89.7			47.5	100.0
	OR	12.0	12.0	12.0	12.0	12.0			10.4	45.4
	PR	28.8	12.6	35.0	0.0	29.9			0.0	39.9
$\Delta \ln(\text{murder/capita})$	H	21.0	19.7	22.5	20.4	22.6			17.3	26.8
	OR	43.3	43.3	43.3	43.3	43.3			43.3	45.4
	PR	0.4	0.2	1.1	0.4	1.2			0.0	26.0
B: Ferrara (2022)										
% Semiskilled Black workers	H	0.0	0.0	5.5	0.0	0.0	0.0	0.0	0.0	0.0
	OR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	PR	34.5	19.6	49.7	34.5	53.3	11.0	96.6	34.7	73.3
C: Enke (2020)										
Trump – avg. GOP	H	6.1	4.9	9.5			1.8	15.2		
	OR	13.0	5.1	13.0			5.0	30.3		
	PR	0.2	0.1	0.2			0.0	4.0		
Voted for Trump in 2016	H	94.6	63.1	96.4			19.0	100.0		
	OR	0.0	0.0	0.5			0.0	14.1		
	PR	0.1	0.0	0.1			0.0	1.4		
Voted for Trump in primaries	H	15.5	5.9	16.6			2.6	30.5		
	OR	9.6	2.8	9.6			0.7	40.6		
	PR	0.0	0.0	0.0			0.0	0.6		

Notes: See notes to Table 1.1 for description of the specifications. H: Hausman, OR: residual test based on outcome regression, PR: residual test based on propensity score regression. Col. 1 reports p -values under the original specification for the BCH, Ferrara (2022), and Enke (2020) studies. Cols. 2–3 report the range of p -values obtained under alternative ways of dropping collinear columns of the control matrix. Cols. 4–5, 6–7, and 8–9 report, respectively, the p -value range under alternative normalizations of the controls prior to taking powers and interactions, when categorical variables are expressed as indicators for different subsets, and when powers are constructed using Hermite polynomials, with an offset λ ranging between -1 and 1 .

of the confounding depends on the particular specification of the control matrix: even if the sparsity assumption holds under a particular way of expressing the column space of the controls, most alternative plausible normalizations do not admit a sparse approximation. Second, it may be the case no control matrix in a large class of normalizations admits a sparse approximation.

What should a practitioner take away from these results? We have argued that unless p is comparable to n , the potential efficiency gains of SBEs over OLS are limited. Simply reporting OLS with standard errors that are robust to the presence of many controls (e.g. Cattaneo et al., 2018a; Jochmans, 2022; Dobriban et al., 2024; D’Adamo, 2019) will deliver credible inference at little efficiency loss. When p is comparable to n , OLS becomes too noisy to be useful, and infeasible when the covariate dimension exceeds the sample size. Sparsity restrictions on the control vector, or other types of restrictions such as limiting the magnitude of the control coefficients (e.g. Li & Müller, 2021; Armstrong et al., 2023), are then necessary for informative inference. However, since these are substantive modeling restrictions, they need to be discussed and defended on substantive grounds, analogous to the discussion of other key assumptions, such as selection on observables. In particular, researchers who opt to leverage SBEs need to explain why sparsity should plausibly hold under the chosen specification of the control matrix, and not leave normalization choices to statistical software. The sparsity tests developed in this paper can serve as a complement to these arguments, provided they serve as a model specification check rather than a pretest.

We have focused on the sparsity assumption and SBEs because these estimators are used frequently, and their theory is well-developed. However, many other modern ma-

chine learning methods likewise lack invariance to linear reparametrization of the control matrix. When these methods are used for prediction, this lack of invariance is less important, for two reasons. First, one is typically interested in average performance over many predictions, and the overall prediction performance may be robust even if individual predictions are sensitive to normalizations. Second, one can gauge the performance of a given procedure directly using a test sample. When we incorporate these methods into econometric models, however, we are typically interested in inference on a single causal effect, and test sample benchmarking is unavailable. Understanding more generally when a lack of invariance leads to fragility is an interesting area for future research.

References

- Anatolyev, S., & Sølvesten, M. (2023). Testing many restrictions under heteroskedasticity. *Journal of Econometrics*, 236(1), Article 105473.
- Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, 40(S1), S97–S140.
- Armstrong, T., Kolesár, M., & Kwon, S. (2023, August). *Bias-aware inference in regularized regression models*. arXiv: 2012.14823.
- Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society*, 80(4), 597–623.
- Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. B. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1), 233–298.
- Belloni, A., Chernozhukov, V., & Hansen, C. B. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650.
- Beyhum, J., & Striaukas, J. (2024). Testing for sparse idiosyncratic components in factor-augmented regression models. *Journal of Econometrics*, 244(1), Article 105845.
- Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4), 1705–1732.

- Bondell, H. D., & Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65(1), 169–177.
- Bulinski, A. V. (2017). Conditional central limit theorem. *Theory of Probability & Its Applications*, 61(4), 613–631.
- Carpentier, A., & Verzelen, N. (2021). Optimal sparsity testing in linear regression model. *Bernoulli*, 27(2), 727–750.
- Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2), 277–301.
- Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018b). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523), 1350–1361.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.
- D’Adamo, R. (2019, April). *Cluster-robust standard errors for linear regression models with many controls*. arXiv: 1806.07314.
- Dobriban, E., Su, W. J., Yang, Y., & Zhang, Z. (2024, January). *Robust inference under heteroskedasticity via the Hadamard estimator*. arXiv: 1807.00347.
- Donohue, J. J., III, & Levitt, S. D. (2001). The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, 116(2), 379–420.
- Enke, B. (2020). Moral values and voting. *Journal of Political Economy*, 128(10), 3679–3729.

- Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1), 1–23.
- Ferrara, A. (2022). World War II and black economic progress. *Journal of Labor Economics*, 40(4), 1053–1091.
- Gertheiss, J., & Tutz, G. (2010). Sparse modeling of categorical explanatory variables. *The Annals of Applied Statistics*, 4(4), 2150–2180.
- Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5), 2409–2437.
- Gordon, L. (1994). A stochastic approach to the gamma function. *The American Mathematical Monthly*, 101(9), 858–865.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6), 1251–1271.
- He, J. (2020, March). *A test for sparsity* [Unpublished manuscript, Sciences Po].
- Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82), 2869–2909.
- Jochmans, K. (2022). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 117(538), 887–896.
- Kline, P., Saggio, R., & Sølvesten, M. (2020). Leave-out estimation of variance components. *Econometrica*, 88(5), 1859–1898.
- Li, C. (, & Müller, U. K. (2021). Linear regression with many controls of limited explanatory power. *Quantitative Economics*, 12(2), 405–442.

- Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1), 255–285.
- Robbins, H. (1955). A remark on Stirling’s formula. *The American Mathematical Monthly*, 62(1), 26–29.
- Robinson, P. M. (1988). Root- N -consistent semiparametric regression. *Econometrica*, 56(4), 931–954.
- Stokell, B. G., Shah, R. D., & Tibshirani, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society*, 83(3), 579–611.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, 67(1), 91–108.
- Tikhomirov, K. (2020). Singularity of random Bernoulli matrices. *Annals of Mathematics*, 191(2), 593–634.
- van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202.
- von Bahr, B., & Esseen, C.-G. (1965). Inequalities for the r th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1), 299–303.
- Wüthrich, K., & Zhu, Y. (2023). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *The Review of Economics and Statistics*, 105(4), 982–997.

- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1), 49–67.
- Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society*, 76(1), 217–242.

Appendix

1.A Auxiliary Results

For the next two results, we consider a quadratic form $\psi' H \psi$, where H is an orthogonal projection matrix with rank bounded by r . Conditional on some σ -algebra $\mathcal{Z}_n$, the elements of ψ are independent and mean zero, and H is non-random. We will prove a law of large numbers and a central limit theorem for $\psi' H \psi$.

Lemma 1.4. *Suppose that uniformly over i , $E[|\psi_i|^{2+\eta} \mid \mathcal{Z}_n] \leq K$ for some $K > 0$ and some $\eta \in [0, 2]$. Then*

$$\psi' H \psi = E[\psi' H \psi \mid \mathcal{Z}_n] + O_p(r^{1/2} + r^{2/(2+\eta)}).$$

Proof. Write

$$\psi' H \psi - E[\psi' H \psi \mid \mathcal{Z}_n] = \sum_i (\psi_i^2 - E[\psi_i^2 \mid \mathcal{Z}_n]) H_{ii} + 2 \sum_{i < j} \psi_i \psi_j H_{ij} =: T_1 + T_2.$$

By iterated expectations, and the inequality of von Bahr and Esseen (1965),

$$E|T_1|^{1+\eta/2} \leq 2E \sum_i |\psi_i^2 - E[\psi_i^2 \mid \mathcal{Z}_n]|^{1+\eta/2} |H_{ii}|^{1+\eta/2} \leq 2Kr,$$

so by Markov's inequality, $T_1 = O_p(r^{2/(2+\eta)})$. The term T_2 is mean zero with variance

$$4 \sum_{i < j} E[\psi_i^2 \psi_j^2 H_{ij}^2] \preceq r,$$

so by Markov's inequality, $T_2 = O_p(r^{1/2})$. $\square$

Lemma 1.5. *Let $\omega^2 = 2 \sum_{i \neq j} H_{ij}^2 E[\psi_i^2 \psi_j^2 \mid \mathcal{Z}_n]$. Suppose that for some constant $K \geq 1$, (a) $r/\omega^2 \leq K$, and (b) $E[\psi_i^4 \mid \mathcal{Z}_n] \leq K$ a.s. Then, as $r \rightarrow \infty$,*

$$\frac{\psi' H \psi - \sum_i H_{ii} \psi_i^2}{\omega} \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof. Write $(\psi' H \psi - \sum_i H_{ii} \psi_i^2)/\omega = \sum_i \mathcal{Y}_i$, where $\mathcal{Y}_i = \frac{2}{\omega} \psi_i \sum_{j=1}^{i-1} \psi_j H_{ij}$ is a martingale difference array with respect to the filtration $\mathcal{F}_i = \sigma(\psi_1, \dots, \psi_i, \mathcal{Z}_n)$. By the martingale central limit theorem, it suffices to verify the Lyapunov condition $\sum_i E[\mathcal{Y}_i^4] \rightarrow 0$, and convergence of the conditional variance, $\sum_i E[\mathcal{Y}_i^2 \mid \mathcal{F}_{i-1}] = 1 + o_p(1)$.

The Lyapunov condition follows from the bound

$$\begin{aligned} \sum_i E[\mathcal{Y}_i^4] &\leq \frac{2^4}{\omega^4} E \left[\psi_i^4 \cdot 3 \sum_{j=1}^{i-1} \psi_j^2 H_{ij}^2 \sum_{k=1}^{i-1} \psi_k^2 H_{ik}^2 \right] \\ &\leq \sum_i E \frac{48K^3}{\omega^4} \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} H_{ij}^2 H_{ik}^2 \\ &\leq \frac{48K^3}{\omega^4} r \\ &\leq \frac{48K^5}{r} \rightarrow 0, \end{aligned}$$

where the last inequality uses condition (a).

To show convergence of the conditional variance, decompose

$$\begin{aligned} \sum_i E[\mathcal{Y}_i^2 \mid \mathcal{F}_{i-1}] - 1 &= \frac{4}{\omega^2} \sum_{i=1}^n \sum_{j=1}^{i-1} E[\psi_i^2 \mid \mathcal{Z}_n] (\psi_j^2 - E[\psi_j^2 \mid \mathcal{Z}_n]) H_{ij}^2 \\ &\quad + \frac{8}{\omega^2} \sum_{i=1}^n E[\psi_i^2 \mid \mathcal{Z}_n] \sum_{j=1}^{i-1} \psi_j H_{ij} \sum_{k=1}^{j-1} \psi_k H_{ik} =: T_1 + T_2. \end{aligned}$$

We have

$$E[T_1^2 \mid \mathcal{Z}_n] \leq \frac{16}{\omega^4} \sum_{j=1}^n K^3 \left(\sum_{i=j+1}^n H_{ij}^2 \right)^2 \leq \frac{16K^5}{r},$$

so that $T_1 = o_p(1)$ as $r \rightarrow \infty$ by Markov's inequality. It remains to show that $T_2 = o_p(1)$.

Let $\|A\|_F$ denote the Frobenius norm of a matrix, and let $L_{ij} = \mathbf{I}\{i > j\} H_{ij}$. The second moment of T_2 can be bounded as

$$\begin{aligned} E[T_2^2 \mid \mathcal{Z}_n] &= \frac{64}{\omega^4} \sum_{k>j} E[\psi_j^2 \psi_k^2 \mid \mathcal{Z}_n] \left(\sum_{i=1}^n E[\psi_i^2 \mid \mathcal{Z}_n] L_{ij} H_{ik} \right)^2 \\ &\leq \frac{64K^4}{r^2} \sum_{j,k} \left(\sum_{i=1}^n E[\psi_i^2 \mid \mathcal{Z}_n] L_{ij} H_{ik} \right)^2 \\ &= \frac{64K^4}{r^2} \|H \operatorname{diag}(E[\psi_i^2 \mid \mathcal{Z}_n]) L\|_F^2 \\ &\preceq \frac{\|L\|_F^2}{r^2} \\ &\leq \frac{\|H\|_F^2}{r^2} \\ &= \frac{1}{r}, \end{aligned}$$

where the first inequality uses conditions (a) and (b), and the second inequality applies the inequality $\|AL\|_F^2 \leq \lambda_{\max}(A'A)\|L\|_F^2$ twice. The claim then follows by Markov's inequality. $\square$

Lemma 1.6. *Let $\mathcal{B}_p^K$ be a $K \times p$ matrix with i.i.d. Bernoulli(q) entries, $0 < q \leq 1/2$, and assume $0 < \lim_{p \rightarrow \infty} K/p \leq 1$. For any $\varepsilon > 0$, there exist $C_1, C_2 > 0$ only depending on ε and q such that*

$$P \left(\min_{\|v\|_2 \leq \lfloor C_1 p \rfloor, \|v\|_2=1} \|\mathcal{B}_p^K v\|_2^2 \leq C_2 p \right) < (1 - q + \varepsilon)^K$$

for all large enough p .

Proof. The proof follows from the same arguments as Proposition 3.6 of Tikhomirov (2020). There are only two instances in the proof where $p - K \neq 1$ matters: first, in the second to last displayed inequality on Tikhomirov's page 601, the second inequality now invokes Lemma 3.5 with $m = K$, so in his notation, the two $n - 1$ terms are replaced by K (where his n corresponds to our p). Second, in the last displayed inequality on page 601, the $n - 1$ term is again replaced by K . Since $\lim_{n \rightarrow \infty} K/p = c_1 > 0$, the second inequality now still holds for all $|\sum x_i|$ that are larger by a factor of $2/\sqrt{c_1}$, at least for all large enough n . This only affects the constant C in Tikhomirov's first displayed inequality on page 602, and the result follows as in Tikhomirov's proof. $\square$

Lemma 1.7. *Suppose that Assumption 1.2 holds, and that condition (a), and conditions (i), (ii), and (iv) of Lemma 1.3 hold. Then*

$$\|\mathbf{Y} - X\hat{\alpha}\|_2^2 - \mathbf{Y}'(I - P_{S^*})\mathbf{Y} = O_p(s \log(p)).$$

Proof of Lemma 1.7. Letting $r = (I - P_{\mathcal{S}^*})X\alpha$, and $\tilde{\alpha} = \hat{\alpha} - (X'_{\mathcal{S}^*}X_{\mathcal{S}^*})^{-1}X'_{\mathcal{S}^*}X\alpha$, we may write,

$$\begin{aligned} & \|\mathbf{Y} - X\hat{\alpha}\|_2^2 - \mathbf{Y}'(I - P_{\mathcal{S}^*})\mathbf{Y} \\ &= \|X\alpha - X\hat{\alpha}\|_2^2 + 2\epsilon'(X\alpha - X\hat{\alpha}) + \epsilon'\epsilon - \|(I - P_{\mathcal{S}^*})\epsilon + r\|_2^2 \\ &= \|X\alpha - X\hat{\alpha}\|_2^2 - 2\epsilon'X\tilde{\alpha} + \epsilon'P_{\mathcal{S}^*}\epsilon - \|r\|_2^2. \end{aligned}$$

Hence, by eq. (1.26), Assumption 1.2, and condition (i) of Lemma 1.3,

$$\left| \|\mathbf{Y} - X\hat{\alpha}\|_2^2 - \mathbf{Y}'(I - P_{\mathcal{S}^*})\mathbf{Y} \right| \preceq_p s \log(p) + \|\epsilon'X\|_{\infty} \|\tilde{\alpha}\|_1 \preceq_p s \log(p),$$

where the second inequality uses conditions (ii) and (iv) of Assumption 1.2. □

1.B Proofs

Proof of Theorem 1.1. Let A_p be the probability of the event that the model with regressor $\mathcal{R}W_i$ satisfies eqs. (1.3) and (1.4), and let

$$r_{s,\mathcal{R}} = \min_{\|v\|_0 \leq s} E[(W_i' \mathcal{R}' \mathcal{R} \gamma - W_i' \mathcal{R}' v)^2 \mid \mathcal{R}] = \min_{\|v\|_0 \leq s} (\mathcal{R} \gamma - v)' \mathcal{R} E[W_i W_i'] \mathcal{R}' (\mathcal{R} \gamma - v)$$

denote the mean square approximation error under the best s -sparse approximation in this model. Using the assumption that $E[W_i W_i']$ has eigenvalues bounded from below, eq. (1.5), and the assumption that $\|\gamma\|_2$ is bounded from below, it follows that for some $K > 0$ large enough,

$$r_{s,\mathcal{R}} \geq \frac{1}{K} \min_{\|v\|_0 \leq s} \|\mathcal{R} \gamma - v\|_2^2 \geq \frac{1}{K^2} \frac{\sum_{j=1}^{p-s} \mathcal{Z}_{j:p}^2}{\sum_{j=1}^p \mathcal{Z}_j^2},$$

where $\mathcal{Z}_{j:p}^2$ are the order statistics $\mathcal{Z}_{1:p}^2 \leq \mathcal{Z}_{2:p}^2 \leq \dots \leq \mathcal{Z}_{p:p}^2$ of the sample $\{\mathcal{Z}_j^2\}_{j=1}^p$, and $\mathcal{Z}_j$ are i.i.d. standard normal.

If eqs. (1.3) and (1.4) hold, then for p large enough, $r_{s,\mathcal{R}} \leq \eta := K^{-2}s/p$ for $s = \lfloor \sqrt{p}/\log(p) \rfloor$. Hence, by the union bound, for p large enough,

$$\begin{aligned} A_p &\leq P(r_{s,\mathcal{R}} \leq \eta) \\ &\leq \binom{p}{s} P\left(\frac{1}{K^2} \frac{\sum_{j=1}^{p-s} \mathcal{Z}_j^2}{\sum_{j=1}^p \mathcal{Z}_j^2} \leq \eta\right) \\ &= \binom{p}{s} P\left(\frac{\sum_{j=1}^{p-s} \mathcal{Z}_j^2 / (p-s)}{\sum_{j=p-s}^p \mathcal{Z}_j^2 / s} \leq \frac{s}{p-s} \frac{K^2 \eta}{1 - K^2 \eta}\right) \\ &= \binom{p}{s} I_{K^2 \eta}(p/2 - s/2, s/2), \end{aligned}$$

where the last line uses the fact that an F -distribution with d_1 and d_2 degrees of freedom has c.d.f. $I_{d_1 x / (d_1 x + d_2)}(d_1/2, d_2/2)$. Here $I_x(a; b) = \int_0^x t^{a-1}(1-t)^{b-1} dt / B(a, b)$ is the regularized incomplete beta function, and $B(a, b)$ is the beta function.

Since $\int_0^x t^{a-1}(1-t)^{b-1} dt \leq \int_0^x t^{a-1} dt = x^a/a$,

$$\begin{aligned}
A_p &\leq \binom{p}{s} \frac{1}{B(p/2 - s/2, s/2)} \frac{(K^2 \eta)^{p/2 - s/2}}{p/2 - s/2} \\
&\leq e^{3s/2} (p/s)^{3s/2 - 1} (K^2 \eta)^{p/2 - s/2} \\
&= e^{3s/2} (p/s)^{2s - 1 - p/2} \\
&\leq e^{3s/2} (\sqrt{p})^{2s - 1 - p/2},
\end{aligned} \tag{1.10}$$

where the second line uses the identity $\binom{p}{s} = \frac{p}{s(p-s)} \frac{1}{B(s, p-s)}$ and a double application of the beta function bound $\frac{p}{s(p-s)} \frac{1}{B(s, p-s)} \leq (ep/s)^s$ for all $s \geq 1$, and the last inequality uses the definition of s and holds for p large enough so that $2s - 1 - p/2$ is negative. The beta function bound follows from the generalization of Stirling's inequality to gamma functions due to Gordon (1994, Theorem 5),

$$\Gamma(t) = \sqrt{2\pi} t^{t-1/2} e^{-t} R_t, \quad R_t \in [1, \sqrt{2}] \tag{1.11}$$

for all $t \geq 1/2$.¹⁶ Taking logs of eq. (1.10) then yields the upper bound

$$\log(A_p) \leq \frac{3}{2}s - (p/4 - s + 1/2) \log(p) \asymp -\frac{p}{4} \log(p).$$

To derive a lower bound for A_p , observe that $r_{s,\mathcal{R}} \leq K^2 \sum_{j=1}^{p-s} \mathcal{Z}_{j:p}^2 / \sum_{j=1}^p \mathcal{Z}_j^2$ for a large enough K . Since the probability of an approximately sparse representation is lower bounded by the probability that $r_{s,\mathcal{R}} \leq K^2 s/p$ for $s = \lfloor \sqrt{p}/\log(p)^2 \rfloor$, we have

$$\begin{aligned} A_p &\geq P\left(\frac{\sum_{j=1}^{p-s} \mathcal{Z}_j^2}{\sum_{j=1}^p \mathcal{Z}_j^2} \leq s/p\right) \\ &= I_{s/p}(p/2 - s/2, s/2) \\ &\geq \frac{(1 - s/p)^{s/2-1} (s/p)^{p/2-s/2}}{(p/2 - s/2)B(p/2 - s/2, s/2)} \\ &\geq \frac{1}{2\sqrt{2\pi}} \frac{(s/p)^{p/2-s+1}}{(s/2)^{1/2} (1 - s/p)^{p/2-s+3/2}} \\ &\geq \frac{1}{2\sqrt{\pi}} \frac{(s/p)^{p/2-s+1}}{\sqrt{s}}, \end{aligned}$$

where the second line uses the bound $\int_0^x t^{a-1}(1-t)^{b-1}dt \geq (1-x)^{b-1} \int_0^x t^{a-1}dt = (1-x)^{b-1}x^a/a$, and the third line uses the inequality $\frac{p}{s(p-s)B(s,p-s)} \geq \frac{1}{2\sqrt{2\pi}} \frac{(p/s)^s}{s^{1/2}(1-s/p)^{p-s+1/2}}$ that follows from eq. (1.11). Hence, $\log(A_p) \gtrsim -\frac{p}{4} \log(p)$, as claimed. $\square$

¹⁶Specifically,

$$\begin{aligned} \frac{p}{s(p-s)} \frac{1}{B(s,p-s)} &= \frac{p\Gamma(p)}{s\Gamma(s)(p-s)\Gamma(p-s)} = \frac{R_p}{\sqrt{2\pi}R_sR_{p-s}} \frac{(p/s)^s}{s^{1/2}} \left(1 + \frac{s}{p-s}\right)^{p-s+1/2} \\ &\leq \frac{1}{\sqrt{\pi}} \sqrt{\frac{p}{s(p-s)}} (ep/s)^s \leq (ep/s)^s \end{aligned}$$

where the second equality uses eq. (1.11), the first inequality uses $1+x \leq e^x$ and the bounds on R_t in eq. (1.11), and the last inequality uses $p/(s(p-s)) \leq 2$.

Proof of Theorem 1.2. Let $s = \lfloor \sqrt{p}/\log(p) \rfloor$. If eqs. (1.3) and (1.4) hold, then for p large enough, $\min_{\|v\|_0 \leq s} E[(Z'_i A'_0 \gamma - Z'_i \mathcal{A}' v)^2 \mid \mathcal{A}] \leq s/p$. It hence suffices to show that given $\varepsilon > 0$, for all large enough p ,

$$P \left(\min_{\|v\|_0 \leq s} E[(Z'_i A'_0 \gamma - Z'_i \mathcal{A}' v)^2 \mid \mathcal{A}] \leq \frac{s}{p} \right) < (1 - q + \varepsilon)^K.$$

Let $\psi = A'_0 \gamma$, a $p \times 1$ vector with K elements equal to zero and $p - K$ elements equal to $\|\gamma\|_2$ (which is equal to the single non-zero element of γ). By assumption, $pE[Z_i Z'_i]$ is a diagonal matrix with all diagonal elements bounded below by some constant $C_0 > 0$. Thus, it suffices to show that

$$P \left(\min_{\|v\|_0 \leq s} \|\psi - \mathcal{A}' v\|_2^2 \leq s/C_0 \right) < (1 - q + \varepsilon)^K \quad (1.12)$$

for all p large enough.

Let $\mathcal{B}$ be a $p \times p$ matrix with i.i.d. Bernoulli(q) entries, and without loss of generality, assume $\mathcal{A} = \mathcal{B}$ if $\mathcal{B}$ is non-singular. By Theorem A of Tikhomirov (2020), given any $\varepsilon_0 > 0$, the probability of $\mathcal{B}$ being singular is smaller than $(1 - q + \varepsilon_0)^p$ for all large enough p . Applying this result with $\varepsilon_0 = \varepsilon/2$, and using the law of total probability, we can bound the left-hand side by

$$\begin{aligned} P \left(\min_{\|v\|_0 \leq s} \|\psi - \mathcal{A}' v\|_2^2 \leq s/C_0 \right) &\leq P(\mathcal{B} \text{ is singular}) + P \left(\min_{\|v\|_0 \leq s} \|\psi - \mathcal{B}' v\|_2^2 \leq s/C_0 \right) \\ &\leq (1 - q + \varepsilon/2)^p + P \left(\min_{\|v\|_0 \leq s} \|\psi - \mathcal{B}' v\|_2^2 \leq s/C_0 \right). \end{aligned}$$

Since $2(1 - q + \varepsilon/2)^p \leq (1 - q + \varepsilon)^p$ for $p \geq \log(2)/\log(\frac{1-q+\varepsilon}{1-q+\varepsilon/2})$, to show eq. (1.12), suffices to show that the second term in the above display is bounded by $(1 - q + \varepsilon/2)^p$.

With $\eta = \|\gamma\|_2/(2\sqrt{s})$, by the union bound,

$$P\left(\min_{\|v\|_0 \leq s} \|\psi - \mathcal{B}v\|_2^2 \leq s/C_0\right) \leq P\left(\min_{\|v\|_0 \leq s, \|v\|_2 < \eta} \|\psi - \mathcal{B}v\|_2^2 \leq s/C_0\right) + P\left(\min_{\|v\|_0 \leq s, \|v\|_2 \geq \eta} \|\psi - \mathcal{B}v\|_2^2 \leq s/C_0\right). \quad (1.13)$$

Note that all elements of $\mathcal{B}$ are either zero or one. By the Cauchy-Schwarz inequality, if $\|v\|_0 \leq s$ and $\|v\|_2 \leq \eta$, all elements of $\mathcal{B}v$ are bounded above by $\frac{1}{2}\|\gamma\|_2$, and hence $\|\psi - \mathcal{B}v\|_2^2 \geq \frac{1}{4}\|\gamma\|_2^2(p - K)$ almost surely. The first probability on the right-hand side of eq. (1.13) is thus equal to zero for all large enough p .

Let $\mathcal{B}_p^K$ be the $K \times p$ matrix that collects the K rows of $\mathcal{B}$ where the corresponding element in ψ is zero. Then for any $C_1 > 0$,

$$\min_{\|v\|_0 \leq s, \|v\|_2 > \eta} \|\psi - \mathcal{B}v\|_2^2 \geq \min_{\|v\|_0 \leq s, \|v\|_2 \geq \eta} \|\mathcal{B}_p^K v\|_2^2 \geq \eta^2 \min_{\|v\|_0 \leq \lfloor C_1 p \rfloor, \|v\|_2 = 1} \|\mathcal{B}_p^K v\|_2^2$$

almost surely, where the last inequality holds for all p large enough so that $\lfloor C_1 p \rfloor \geq s$.

Thus, for all $C_2 > 0$ and large enough p ,

$$\begin{aligned} P\left(\min_{\|v\|_0 \leq s, \|v\|_2 \geq \eta} \|\psi - \mathcal{B}v\|_2^2 \leq s/C_0\right) &\leq P\left(\min_{\|v\|_2 \leq \lfloor C_1 p \rfloor, \|v\|_2 = 1} \|\mathcal{B}_p^K v\|_2^2 \leq s/\eta^2 \cdot 1/C_0\right) \\ &\leq P\left(\min_{\|v\|_2 \leq \lfloor C_1 p \rfloor, \|v\|_2 = 1} \|\mathcal{B}_p^K v\|_2^2 \leq C_2 p\right) \end{aligned}$$

since $s/\eta^2 \asymp p/(\ln p)^2$. The result now follows from choosing C_1 and C_2 from Lemma 1.6, with $\varepsilon/2$ playing the role of ε . $\square$

Proof of Theorem 1.3. Let $p_0 = p - 1$. Using the identity

$$\sqrt{j!}H_j(x+y) = \sum_{k=0}^j \binom{j}{k} x^{j-k} \sqrt{k!}H_k(y),$$

we find

$$H_{p_0}(z_i + \lambda) = \sum_{j=0}^{p_0} \binom{p_0}{j} \sqrt{j!/p_0!} \lambda^{p_0-j} \tilde{W}_{i,j+1} = \sum_{k=0}^{p_0} \tilde{\gamma}_{p-k} \tilde{W}_{i,p-k},$$

where $\tilde{\gamma}_{p-k} = \binom{p_0}{p_0-k} \sqrt{(p_0-k)!/p_0!} \lambda^k$.

Applying the Stirling formula (Robbins, 1955)

$$\log(n!) = \frac{1}{2} \log(2\pi) + (n + 1/2) \log(n) - n + R_n, \quad R_n \in [1/(12n + 1), 1/12n]$$

we obtain, for $k \geq 1$,

$$\begin{aligned} \log(\tilde{\gamma}_{p-k}^2) &= \log\left(\frac{p_0! \lambda^{2k}}{k!^2 (p_0 - k)!}\right) \\ &= (p_0 - k + 1/2) \log\left(\frac{p_0}{p_0 - k}\right) + k - \log(k) + 2k \log\left(\frac{\sqrt{p_0} \lambda}{k}\right) + R'_{p_0,k}, \end{aligned} \tag{1.14}$$

where $R'_{p_0,k} = R_{p_0} - 2R_k - R_{p_0-k} - \log(2\pi)$ is bounded above and below by a constant, since $R_n \in [0, 1/12]$. Now, $\log(p_0/(p_0 - k)) = \log(1 + k/(p_0 - k)) \geq k/p_0$ since $\log(1 + x) \geq x/(1 + x)$ for $x \geq -1$. Hence, for some constant C , and for $1 \leq k \leq$

$$L\sqrt{p_0}/\log(p_0),$$

$$\begin{aligned}\log(\tilde{\gamma}_{p-k}^2) &\geq k \left[2 + \frac{1/2 - k}{p_0} - \log(k)/k \right] + 2k \log \left(\frac{\sqrt{p_0}L}{k \log(p_0)} \right) + C \\ &\geq k \left[2 + \frac{1/2 - L\sqrt{p_0}/\log(p_0)}{p_0} - \log(k)/k \right] + 2k \log(1) + C \\ &\geq k \left[1 - \frac{L}{\log(p_0)\sqrt{p_0}} \right] + C,\end{aligned}$$

where the last inequality uses $\log(k)/k \leq 1$. The expression in brackets is bounded below by $1/2$ for $p \geq \max\{6, (2L)^2\}$, which yields the result for L fixed.

To prove the result when $L \rightarrow 0$, eq. (1.14) implies, for some constant C ,

$$\begin{aligned}\log(\tilde{\gamma}_{p-k}^2) &\leq (p_0 - k + 1/2) \log \left(1 + \frac{k}{p_0 - k} \right) + k - \log(k) + 2k \log \left(\frac{\sqrt{p_0}\lambda}{k} \right) + C \\ &\leq 2k + \frac{1}{2} \frac{k}{p_0 - k} + 2k \log \left(\frac{\sqrt{p_0}\lambda}{k} \right) + C \\ &\leq k \left[3 + 2 \log \left(\frac{\sqrt{p_0}\lambda}{k} \right) \right] + C,\end{aligned}$$

where the second line uses $\log(1+x) \leq x$. Since for $k \geq \sqrt{p_0}\lambda e^2$ the expression in square brackets is smaller than -1 , if we approximate the regression function using the last $s = \lceil e^2 \max\{L/\log(p_0), 1/\log(p_0)^2\} \sqrt{p_0} \rceil$ regressors, we make an approximation error that is bounded by

$$\sum_{k \geq s} \tilde{\gamma}_{p-k}^2 \leq C \sum_{k \geq s} e^{-k} \preceq e^{-s} \preceq e^{-e^2 \sqrt{p}/\log(p)^2},$$

which is of lower order than s/p . □

Proof of Lemma 1.1. Write

$$\hat{\beta}_{OLS} - \beta = \sum_{i=1}^n \eta_i, \quad \eta_i = \frac{1}{\ddot{D}'\ddot{D}} \ddot{D}_i U_i.$$

Conditional on (D, W) , η_i are independent, with variance $\omega_{OLS}^2 := \text{var}(\sum_i \eta_i \mid D, W) = \frac{\sum_i \ddot{D}_i^2 E[U_i^2 \mid D, W]}{(\ddot{D}'\ddot{D})^2}$. By a conditional version of the Lindeberg-Feller theorem (Bulinski, 2017, Theorem 1), it therefore suffices to verify a conditional version of the Lyapunov condition

$$\sum_{i=1}^n \frac{E[|\eta_i|^{2+\eta} \mid D, W]}{\omega_{OLS}^{2+\eta}} = o_p(1),$$

and that

$$s_{OLS}^2 / \omega_{OLS}^2 = 1 + o_p(1). \quad (1.15)$$

To show the Lyapunov condition, substitute in the definition of η_i and ω_{OLS} to write the left-hand side as

$$\begin{aligned} \sum_{i=1}^n \frac{E[|U_i|^{2+\eta} \mid D, W] |\ddot{D}_i|^{2+\eta}}{(\sum_{j=1}^n \ddot{D}_j^2 E[U_j^2 \mid D, W])^{1+\eta/2}} &\leq K^{2+\eta/2} \frac{\sum_{i=1}^n |\ddot{D}_i|^{2+\eta}}{(\sum_{j=1}^n \ddot{D}_j^2)^{1+\eta/2}} \\ &\leq K^{2+\eta/2} \frac{\max_{i'} |\ddot{D}_{i'}|^\eta \sum_{i=1}^n \ddot{D}_i^2}{(\sum_{j=1}^n \ddot{D}_j^2)^{1+\eta/2}} \\ &= K^{2+\eta/2} \left(\frac{\max_i \ddot{D}_i^2 / n}{\frac{1}{n} \sum_{j=1}^n \ddot{D}_j^2} \right)^{\eta/2}, \end{aligned}$$

where the first inequality uses Assumption 1.1 (ii) and (iii).

Therefore, to verify the Lyapunov condition, it suffices to show that

$$\max_i |\ddot{D}_i| = O_p(n^{1/4}), \quad (1.16)$$

and that

$$\frac{1}{n} \sum_{i=1}^n \ddot{D}_i^2 \asymp_p 1. \quad (1.17)$$

Now, by Lemma 1.4 and Assumption 1.1 (ii) and (iii), $\frac{1}{n} \sum_{i=1}^n \ddot{D}_i^2 = \frac{1}{n} \sum_i M_{ii} E[\tilde{D}_i^2 | W] + o_p(1)$, $M_{ii} = (I - P)_{ii}$, with the first term bounded between $(1 - p/n)/K$ and K . Therefore, eq. (1.17) follows by Assumption 1.1 (iv). Furthermore, by the union bound and Markov's inequality

$$P(\max_i |\ddot{D}_i|/n^{1/4} \geq \epsilon) \leq \sum_{i=1}^n P(|\ddot{D}_i| \geq n^{1/4}\epsilon) \leq \sum_{i=1}^n \frac{E[\ddot{D}_i^4]}{n\epsilon^4} \leq \frac{1}{\epsilon^4} 4K,$$

which implies eq. (1.16). Here the last inequality follows from the bound

$$\begin{aligned} E[\ddot{D}_i^4 | W] &= \sum_j M_{ij}^4 E[\tilde{D}_j^4 | W] + 3 \sum_{j \neq k} M_{ij}^2 M_{ik}^2 E[\tilde{D}_j^2 \tilde{D}_k^2 | W] \\ &\leq 4K \sum_{j,k} M_{ij}^2 M_{ik}^2 \\ &= 4K M_{ii}^2. \end{aligned} \quad (1.18)$$

It remains to verify eq. (1.15). Write

$$\frac{s_{OLS}^2}{\omega_{OLS}^2} - 1 = \frac{\frac{1}{n} \sum_{i=1}^n \ddot{D}_i^2 (U_i^2 - E[U_i^2 | D, W])}{\frac{1}{n} \sum_{i=1}^n \ddot{D}_i^2 E[U_i^2 | D, W]}.$$

The denominator satisfies $\frac{1}{n} \sum_{i=1}^n \ddot{D}_i^2 E[U_i^2 \mid D, W] \geq \frac{1}{n} \sum_{i=1}^n \ddot{D}_i^2 / K \asymp_p 1$ by eq. (1.17).

The result therefore follows if we can show that the numerator is of the order $o_p(1)$. This follows from the fact that for any variable $\mathcal{U}_i$, that, conditional on (D, W) has mean zero and uniformly bounded $1 + \eta/2$ moments,

$$\frac{1}{n} \sum_i \ddot{D}_i^2 \mathcal{U}_i = o_p(1). \quad (1.19)$$

In turn, eq. (1.19) follows by Markov's inequality, and the bound

$$\begin{aligned} E \left| \frac{1}{n} \sum_{i=1}^n \ddot{D}_i^2 \mathcal{U}_i \right|^{1+\eta/2} &\preceq \frac{1}{n^{1+\eta/2}} \sum_{i=1}^n E |\ddot{D}_i|^{2+\eta} \\ &\leq \frac{1}{n^{1+\eta/2}} \sum_{i=1}^n E [\ddot{D}_i^4]^{(2+\eta)/4} \\ &\leq \frac{1}{n^{1+\eta/2}} E \sum_{i=1}^n (4K M_{ii}^2)^{(2+\eta)/4} \rightarrow 0. \end{aligned}$$

Here the first inequality uses iterated expectations, and the inequality of von Bahr and Esseen (1965), the second inequality uses Jensen's inequality, the third uses eq. (1.18), and the final limit uses $M_{ii}^2 \leq 1$. $\square$

Proof of Lemma 1.2. Observe first that

$$n \sum_{i=1}^n Z_i^2 = \frac{\tilde{D}' P \tilde{D}/n}{\tilde{D}' \tilde{D}/n \cdot \ddot{D}' \ddot{D}/n} = \frac{\frac{1}{n} \sum_i P_{ii} E[\tilde{D}_i^2 | W] + O_p(p^{1/2}/n)}{\tilde{D}' \tilde{D}/n \cdot \ddot{D}' \ddot{D}/n} \asymp_p \frac{\text{tr}(W)}{n}, \quad (1.20)$$

where the second equality follows by Lemma 1.4, and the last step follows by the law of large numbers, eq. (1.17), and Assumption 1.1 (ii) and (iii).

Hence, letting $\omega_H^2 = \sum_{i=1}^n Z_i^2 E[U_i^2 | D, W]$, it follows from Assumption 1.1 (iii), and the fact that $s_*^2 = O_p(1/n)$ that

$$\frac{s_*^2}{\omega_H^2} = \frac{O_p(1)}{n \sum_{i=1}^n Z_i^2 E[U_i^2 | D, W]} = O_p(n/\text{tr}(P)).$$

Therefore, under eq. (1.7), since $\liminf_{n \rightarrow \infty} n/\text{tr}(P)$ is bounded,

$$\frac{\hat{\beta}_{OLS} - \hat{\beta}^*}{\omega_H} = \frac{1}{\omega_H} \sum_{i=1}^n Z_i U_i + o_p(s_*/\omega_H) = \frac{1}{\omega_H} \sum_{i=1}^n Z_i U_i + o_p(1).$$

By a conditional version of the Lindeberg-Feller theorem (Bulinski, 2017, Theorem 1), the first term is asymptotically standard normal if a conditional version of the Lyapunov condition holds,

$$\sum_{i=1}^n \frac{|Z_i|^{2+\eta} E[|U_i|^{2+\eta} | D, W]}{\omega_H^{2+\eta}} = o_p(1). \quad (1.21)$$

By Assumption 1.1 (ii) and (iii) and eq. (1.20), the left-hand side is bounded above by

$$K^{2+\eta/2} \left(\frac{n \max_i Z_i^2}{n \sum_i Z_i^2} \right)^{\eta/2} \preceq_p (n^2/\text{tr}(P) \cdot \max_i Z_i^2)^{\eta/2}.$$

The claim in eq. (1.21) then follows from the bound

$$n \max_i Z_i^2 \leq 2 \frac{\max_i \ddot{D}_i^2/n}{(\ddot{D}' \ddot{D}/n)^2} + 2 \frac{\max_i \tilde{D}_i^2/n}{(\tilde{D}' \tilde{D}/n)^2}.$$

The first term is $O_p(n^{-1/2})$ by eqs. (1.16) and (1.17). The second term is also $O_p(n^{-1/2})$, since $\tilde{D}'_i \tilde{D}_i/n \asymp_p 1$ by the law of large numbers and Assumption 1.1 (ii) and (iii), and since

$$\max_i |\tilde{D}_i| = O_p(n^{1/4}). \quad (1.22)$$

Specifically, eq. (1.22) holds since by the union bound and Markov's inequality,

$$P(\max_i |\tilde{D}_i|/n^{1/4} > \epsilon) \leq \sum_{i=1}^n P(|\tilde{D}_i|/n^{1/4} > \epsilon) \leq \sum_{i=1}^n \frac{E[\tilde{D}_i^4]}{n\epsilon^4} \leq \frac{K}{\epsilon^4}.$$

To establish the first claim of Lemma 1.2, it now suffices to show that $\omega_H/s_H = 1 + o_p(1)$.

This follows from writing

$$\frac{s_H^2}{\omega_H^2} - 1 = \frac{n \sum_i Z_i^2 (U_i^2 - E[U_i^2 | D, W])}{n \sum_i Z_i^2 E[U_i^2 | D, W]}.$$

The denominator is of the order p/n by eq. (1.20) and Assumption 1.1 (ii) and (iii). To show that the numerator is $o_p(1)$, let $\mathcal{U}_i = U_i^2 - E[U_i^2 | D, W]$, and decompose it as

$$\begin{aligned} n \sum_i Z_i^2 \mathcal{U}_i &= \frac{\frac{1}{n} \sum_i \ddot{D}_i^2 \mathcal{U}_i}{(\ddot{D}' \ddot{D}/n)^2} + \frac{\frac{1}{n} \sum_i \tilde{D}_i^2 \mathcal{U}_i}{(\tilde{D}' \tilde{D}/n)^2} - 2 \frac{\frac{1}{n} \sum_i \ddot{D}_i \tilde{D}_i \mathcal{U}_i}{(\ddot{D}' \ddot{D}/n)(\tilde{D}' \tilde{D}/n)} \\ &= O_p(1) \frac{1}{n} \sum_i \ddot{D}_i^2 \mathcal{U}_i + O_p(1) \frac{1}{n} \sum_i \tilde{D}_i^2 \mathcal{U}_i + O_p(1) \frac{1}{n} \sum_i \ddot{D}_i \tilde{D}_i \mathcal{U}_i \\ &= o_p(1). \end{aligned}$$

Here the second equality uses eq. (1.17), and the last inequality uses eq. (1.19), the law of large numbers, and the result that

$$\frac{1}{n} \sum_i \ddot{D}_i \tilde{D}_i \mathcal{U}_i = o_p(1), \quad (1.23)$$

which follows by Markov's inequality and the bound

$$\begin{aligned} E \left| \frac{1}{n} \sum_i \ddot{D}_i \tilde{D}_i \mathcal{U}_i \right|^{1+\eta/2} &\leq \frac{1}{n^{1+\eta/2}} \sum_i E |\ddot{D}_i \tilde{D}_i|^{1+\eta/2} \\ &\leq \frac{1}{n^{1+\eta/2}} \sum_i (E \ddot{D}_i^2 \tilde{D}_i^2)^{(2+\eta)/4} \\ &= \frac{1}{n^{1+\eta/2}} \sum_i (E \sum_j M_{ij}^2 \tilde{D}_j^2 \tilde{D}_i^2)^{(2+\eta)/4} \\ &\leq \frac{1}{n^{1+\eta/2}} \sum_i (M_{ii} K)^{(2+\eta)/4} \rightarrow 0, \end{aligned}$$

where the first inequality uses iterated expectations, and the inequality of von Bahr and Esseen (1965), the second inequality uses Jensen's inequality, the third uses Assumption 1.1 (ii), and the final limit uses $M_{ii} \leq 1$.

To prove the second claim in Lemma 1.2, let $r_{\gamma,i} = r_\gamma(W_i)$, $r_{\delta,i} = r_\delta(W_i)$, $\tilde{u}_i = r_{\gamma,i} + U_i$ and $\tilde{d}_i = r_{\delta,i} + \tilde{D}_i$ so that we may write

$$\left| \frac{\hat{S}_H^2}{\omega_H^2} - 1 \right| \leq_p \left| n \sum_i (\hat{Z}_i^2 \hat{U}_i^2 - \tilde{z}_i^2 \tilde{u}_i^2) \right| + \left| n \sum_i (Z_i^2 U_i^2 - \tilde{z}_i^2 \tilde{u}_i^2) \right| + o_p(1). \quad (1.24)$$

The second term is of the order $o_p(1)$ by condition (iii), so it suffices to show that the first term is of the order $o_p(1)$.

Let $\check{U}_i = \hat{U}_i - \tilde{u}_i$ and $\check{D}_i = \hat{D}_i - \tilde{d}_i$, so we may write $n\hat{Z}_i\hat{U}_i = \frac{\check{D}_i\check{U}_i}{\hat{D}'\hat{D}/n} + \frac{\check{D}_i\tilde{u}_i}{\hat{D}'\hat{D}/n} - \frac{\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i}{\hat{D}'\hat{D}/n} - \frac{\tilde{d}_i\tilde{u}_i}{\hat{D}'\hat{D}/n}$ and $n\tilde{z}_i\tilde{u}_i = \frac{\tilde{u}_i\check{D}_i}{\check{D}'\check{D}/n} - \frac{\tilde{u}_i\tilde{d}_i}{\check{d}'\check{d}/n}$. Plugging these expressions into the first term in eq. (1.24) and expanding the term yields

$$\begin{aligned} n \sum_i (\hat{Z}_i^2 \hat{U}_i^2 - \tilde{z}_i^2 \tilde{u}_i^2) &= \frac{\frac{1}{n} \sum_i [(\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i)^2 + 2(\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i)\tilde{d}_i\tilde{u}_i]}{(\hat{D}'\hat{D}/n)^2} \\ &\quad + \frac{\frac{1}{n} \sum_i (\check{D}_i^2\check{U}_i^2 + 2\check{D}_i^2\check{U}_i\tilde{u}_i)}{(\check{D}'\check{D}/n)^2} \\ &\quad - \frac{2}{n} \sum_i \frac{(\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i)(\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i) + \check{D}_i\check{U}_i\tilde{d}_i\tilde{u}_i}{\check{D}'\check{D}/n \cdot \hat{D}'\hat{D}/n} \\ &\quad + \frac{1}{n} \sum_i \tilde{d}_i^2 \tilde{u}_i^2 \left[\frac{1}{(\hat{D}'\hat{D}/n)^2} - \frac{1}{(\check{d}'\check{d}/n)^2} \right] \\ &\quad + 2 \frac{\frac{1}{n} \sum_i \tilde{d}_i \check{D}_i \tilde{u}_i^2}{\check{D}'\check{D}/n} \left[\frac{1}{\check{d}'\check{d}/n} - \frac{1}{\hat{D}'\hat{D}/n} \right]. \end{aligned}$$

Using the inequality $|ab| \leq (a^2 + b^2)/2$, eq. (1.17), and the fact that by condition (iv), $\hat{D}'\hat{D}/n = \check{d}'\check{d}/n + o_p(1)$ and that $\check{d}'\check{d}/n \asymp_p 1$ by condition (i) and Assumption 1.1 (ii) and

(iii), it follows that the right-hand side is smaller than

$$\begin{aligned}
& \left| n \sum_i (\hat{Z}_i^2 \hat{U}_i^2 - \tilde{z}_i^2 \tilde{u}_i^2) \right| \\
& \leq O_p(1) \frac{1}{n} \sum_i [\check{D}_i^2 \check{U}_i^2 + \check{D}_i^2 \tilde{u}_i^2 + \tilde{d}_i^2 \check{U}_i^2 + |\check{D}_i \tilde{d}_i| \tilde{u}_i^2 + \tilde{d}_i^2 |\check{U}_i \tilde{u}_i|] \\
& + O_p(1) \frac{1}{n} \sum_i [\ddot{D}_i^2 \check{U}_i^2 + (\ddot{D}_i^4 + \tilde{u}_i^2) |\check{U}_i|] \\
& + O_p(1) \frac{1}{n} \sum_i [\check{U}_i^2 \check{D}_i^2 + \ddot{D}_i^2 \check{U}_i^2 + \check{D}_i^2 \tilde{u}_i^2 + \tilde{d}_i^2 \check{U}_i^2 + \tilde{u}_i^2 |\ddot{D}_i \check{D}_i| + |\check{U}_i| (\ddot{D}_i^2 + \tilde{d}_i^2 \tilde{u}_i^2)] \\
& + o_p(1) \frac{1}{n} \sum_i \tilde{d}_i^2 \tilde{u}_i^2 \\
& + o_p(1) \frac{1}{n} \sum_i (\ddot{D}_i^2 + \tilde{d}_i^2) \tilde{u}_i^2 \\
& \leq o_p(1) \frac{1}{n} \sum_i [1 + \tilde{u}_i^2 + \tilde{d}_i^2 + \tilde{d}_i^2 \tilde{u}_i^2 + \ddot{D}_i^4 + \ddot{D}_i^2 \tilde{u}_i^2],
\end{aligned}$$

where the second inequality uses condition (iv) and the inequality $2|a| \leq 1 + a^2$. We now show that each summand is bounded in probability. By condition (i), and Assumption 1.1 (ii), $\frac{1}{n} \sum_i (\tilde{u}_i^2 + \tilde{d}_i^2) = \frac{1}{n} \sum_i (U_i^2 + D_i^2) + o_p(1) = O_p(1)$. Next,

$$\frac{1}{n} \sum_i \tilde{d}_i^2 \tilde{u}_i^2 \leq \frac{4}{n} \sum_i (\tilde{D}_i^2 + r_{\delta,i}^2)(r_{\gamma,i}^2 + U_i^2),$$

which by Assumption 1.1 (ii) and conditions (i) and (ii) has expectation bounded by a constant times $\frac{1}{n} \sum_{i=1}^n E[1 + r_{\gamma,i}^2 + r_{\delta,i}^2 + r_{\delta,i}^2 r_{\gamma,i}^2] = O(1)$, so $\frac{1}{n} \sum_i \tilde{d}_i^2 \tilde{u}_i^2 = O_p(1)$ by Markov's inequality. Likewise, it follows from eq. (1.18) that $\frac{1}{n} \sum_i \ddot{D}_i^4 = O_p(1)$.

Finally, by Assumption 1.1 (ii), and the bound $E[\ddot{D}_i^2 \mid W] \leq KM_{ii} \leq K$, we have $\frac{1}{n} \sum_i E[\ddot{D}_i^2 \tilde{u}_i^2] \leq 2\frac{1}{n} \sum_i E[\ddot{D}_i^2(K + r_\gamma^2)] \leq 2E[K(K + r_\gamma^2)] = O(1)$, so that by Markov's inequality, the last term is also bounded in probability. $\square$

Proof of Lemma 1.3. Letting $r = (I - P_{S^*})X\alpha$, we may write

$$\mathcal{F} = \epsilon' P \epsilon - \epsilon' P_{S^*} \epsilon + 2\epsilon' r + r' r = \epsilon' P \epsilon + O_p(s + \|r\|_2 + \|r\|_2^2), \quad (1.25)$$

where the second equality follows since the second term satisfies

$$\epsilon' P_{S^*} \epsilon = E[\epsilon' P_{S^*} \epsilon] + O_p(s^{1/2}) \leq Ks + O_p(s^{1/2}) \quad (1.26)$$

by Lemma 1.4 and condition (a), and the third term is mean zero with variance bounded by $K\|r\|^2$ by condition (a). Furthermore, by conditions (b) and (d),

$$\omega^2 := 2 \sum_{i \neq j} E[\epsilon_i^2 \epsilon_j^2 \mid X] P_{ij}^2 \geq 2 \sum_{i \neq j} P_{ij}^2 / K^2 \geq \frac{2}{K^2} \text{tr}(P)(1 - \max_i P_{ii}) \geq 2 \text{tr}(P) / K^3.$$

Hence, by condition (c), condition (a) of Lemma 1.5 holds.

Since condition (b) of Lemma 1.5 holds by condition (a), we can apply Lemma 1.5 to $\epsilon' P \epsilon$ to yield $(\epsilon' P \epsilon - \sum_{i=1}^n \epsilon_i^2 P_{ii}) / \omega \xrightarrow{d} \mathcal{N}(0, 1)$. Combining this result with eq. (1.25) and Assumption 1.2 yields

$$\frac{\mathcal{F} - \sum_{i=1}^n \epsilon_i^2 P_{ii}}{\omega} = \frac{\epsilon' P \epsilon - \sum_{i=1}^n \epsilon_i^2 P_{ii}}{\omega} + o_p(1) \xrightarrow{d} \mathcal{N}(0, 1). \quad (1.27)$$

Furthermore, letting $\psi_i = \epsilon_i^2 - E[\epsilon_i^2 \mid X]$, and using the identity $\epsilon_i^2 \epsilon_j^2 - E[\epsilon_i^2 \epsilon_j^2 \mid X] = \psi_i \psi_j + \psi_i E[\epsilon_j^2 \mid X] + E[\epsilon_i^2 \mid X] \psi_j$, we have

$$\begin{aligned} \left| \frac{2 \sum_{i \neq j} \epsilon_i^2 \epsilon_j^2 P_{ij}^2 - \omega^2}{\omega^2} \right| &\leq \frac{K^3}{\text{tr}(P)} \left| \sum_{i \neq j} (\epsilon_i^2 \epsilon_j^2 - E[\epsilon_i^2 \epsilon_j^2 \mid X]) P_{ij}^2 \right| \\ &\leq \frac{2K^3}{\text{tr}(P)} \left| \sum_{i < j} \psi_i \psi_j P_{ij}^2 \right| + \frac{2K^3}{\text{tr}(P)} \left| \sum_{i \neq j} \psi_i E[\epsilon_j^2 \mid X] P_{ij}^2 \right|. \end{aligned}$$

By condition (a), conditional on X , the term inside the first absolute value function is mean zero with variance bounded by a constant times $\sum_{i,j} P_{ij}^4 \leq \text{tr}(P)$, and the term inside the second absolute value function is also mean zero with variance bounded by a constant times $\sum_{i=1}^n (\sum_{j=1}^n P_{ij}^2)^2 \leq \text{tr}(P)$, so that the above display is $O_p(\text{tr}(P)^{-1/2}) = o_p(1)$ by Markov's inequality. Combining this result with eq. (1.27) then yields the first claim.

To prove the second claim in Lemma 1.3, it suffices to show that

$$\|\mathbf{Y} - X \hat{\alpha}\|_2^2 - \mathbf{Y}'(I - P_{S^*})\mathbf{Y} = O_p(s \log(p)),$$

that

$$\sum_i P_{ii}(\hat{\epsilon}_i^2 - \epsilon_i^2) = o_p(p^{1/2}) \tag{1.28}$$

and that

$$\sum_{i \neq j} (\hat{\epsilon}_i^2 \hat{\epsilon}_j^2 - \epsilon_i^2 \epsilon_j^2) P_{ij}^2 = o_p(p). \tag{1.29}$$

The first assertion follows by Lemma 1.7.

To show eq. (1.28), decompose $\hat{\epsilon}_i = \epsilon_i - \tilde{f}_i = \epsilon_i + r_i - X'_i \tilde{\alpha}$, where $\tilde{f}_i = X'_i(\hat{\alpha} - \alpha)$, and $\tilde{\alpha} = \hat{\alpha} - (X'_{S^*} X_{S^*})^{-1} X'_{S^*} X \alpha$, so that we may write

$$\begin{aligned} \left| \sum_i P_{ii}(\hat{\epsilon}_i^2 - \epsilon_i^2) \right| &= \left| \sum_i P_{ii} \tilde{f}_i^2 + 2 \sum_i P_{ii} \epsilon_i r_i - 2 \sum_i P_{ii} \epsilon_i X'_i \tilde{\alpha} \right| \\ &\preceq_p s \log(p) + \|r\|_2 + 2 \left\| \sum_i P_{ii} \epsilon_i X_i \right\|_\infty \|\tilde{\alpha}\|_1 \preceq_p s \log(p) + \|r\|_2, \end{aligned}$$

where the first inequality follows by condition (i) and Markov's inequality, and since conditional on X , the second term is mean zero with standard deviation bounded by a constant times $\|r\|_2$. The second inequality applies conditions (ii) and (iv). Equation (1.28) then follows by Assumption 1.2.

To show eq. (1.29), write the left-hand side as

$$\begin{aligned} &\sum_{i \neq j} (\hat{\epsilon}_i^2 \hat{\epsilon}_j^2 - \epsilon_i^2 \epsilon_j^2) P_{ij}^2 \\ &= \sum_{i \neq j} \tilde{f}_i^2 \tilde{f}_j^2 P_{ij}^2 + 2 \sum_{i \neq j} \epsilon_i^2 \tilde{f}_j^2 P_{ij}^2 - 4 \sum_{i \neq j} \epsilon_j \tilde{f}_j \tilde{f}_i^2 P_{ij}^2 - 4 \sum_{i \neq j} \epsilon_i \tilde{f}_i \epsilon_j^2 P_{ij}^2 + 4 \sum_{i \neq j} \epsilon_i \tilde{f}_i \epsilon_j \tilde{f}_j P_{ij}^2 \\ &=: T_1 + 2T_2 - 4T_3 - 4T_4 + 4T_5 \\ &= O_p(s^2 \log(p)^2 + \|r\|_2^2) + o_p(p), \end{aligned}$$

where the bounds follow by bounding each term, as derived next. First, by condition (i), T_1 is bounded by a constant times $\|X(\hat{\alpha} - \alpha)\|_2^4 \preceq_p s^2 \log(p)^2$. Second,

$$T_2 \leq 2 \sum_{i \neq j} \epsilon_i^2 r_j^2 P_{ij}^2 + 2 \sum_{i \neq j} \epsilon_i^2 (X'_j \tilde{\alpha})^2 P_{ij}^2 = O_p(\|r\|_2^2) + o_p(p).$$

where the second equality follows by Markov's inequality, since the first term, conditional on X , has expectation bounded by a constant times $\sum_j r_j^2 P_{jj} \leq \|r\|_2^2$, and the second term is bounded by $\max_j (X'_j \tilde{\alpha})^2 \cdot \sum_{i \neq j} \epsilon_i^2 P_{ij}^2 = o_p(1) \cdot O_p(p)$, since $\sum_{i \neq j} E[\epsilon_i^2 P_{ij}^2]$ is bounded by a constant times $\sum_{i \neq j} P_{ij}^2 \leq p$, and by conditions (ii) and (iii),

$$\max_j |X'_j \tilde{\alpha}| \leq \max_j \|X_j\|_\infty \|\tilde{\alpha}\|_1 = o_p(1). \quad (1.30)$$

Third,

$$|T_3| \leq 2 \sum_{i \neq j} \epsilon_j^2 \tilde{f}_i^2 P_{ij}^2 + 2 \sum_{i \neq j} \tilde{f}_j^2 \tilde{f}_i^2 P_{ij}^2 = 2T_1 + 2T_2.$$

Fourth,

$$\begin{aligned} |T_4| &\leq \left| \sum_{i \neq j} \epsilon_i X'_i \tilde{\alpha} \epsilon_j^2 P_{ij}^2 \right| + 2 \left| \sum_{i \neq j} \epsilon_i r_i (\epsilon_j^2 - E[\epsilon_j^2 | X]) P_{ij}^2 \right| + \left| \sum_{i \neq j} \epsilon_i r_i E[\epsilon_j^2 | X] P_{ij}^2 \right| \\ &= o_p(p) + O_p(\|r\|_2). \end{aligned}$$

Here the conclusion follows since the first term is bounded by $\max_i |X'_i \tilde{\alpha}| \sum_{i \neq j} |\epsilon_i| \epsilon_j^2 P_{ij}^2 = o_p(p)$ by eq. (1.30) and the fact that $E \sum_{i \neq j} |\epsilon_i| \epsilon_j^2 P_{ij}^2 \preceq \sum_{i,j} P_{ij}^2 \leq p$, the second term is mean zero with variance bounded by a constant times $\sum_{i,j} r_i^2 P_{ij}^4 \leq \|r\|_2^2$, and the third term is mean zero with variance bounded by a constant times $\sum_i r_i^2 P_{ii}^2 \leq \|r\|_2^2$. Finally,

$$T_5 = 2 \sum_{i < j} \epsilon_i \epsilon_j r_j r_i P_{ij}^2 + 2 \sum_{i \neq j} \epsilon_i r_i \epsilon_j X'_j \tilde{\alpha} P_{ij}^2 + \sum_{i \neq j} \epsilon_i \epsilon_j (X'_j \tilde{\alpha}) (X'_i \tilde{\alpha}) P_{ij}^2.$$

The first term is mean zero with variance bounded by a constant times $\sum_{i < j} r_j^2 r_i^2 P_{ij}^4 \leq \|r\|_2^4$. The second term is bounded by a constant times $\max_j |X'_j \tilde{\alpha}| \cdot \sum_{i \neq j} r_i^2 |\epsilon_j| P_{ij}^2 +$

$\max_j |X_j' \tilde{\alpha}| \cdot \sum_{i \neq j} \epsilon_i^2 |\epsilon_j| P_{ij}^2 = o_p(\|r\|_2^2 + p)$, and the last term is bounded by $\max_j |X_j' \tilde{\alpha}|^2 \cdot \sum_{i \neq j} |\epsilon_i \epsilon_j| P_{ij}^2 = o_p(p)$. Hence, $T_5 = O_p(\|r\|^2) + o_p(p)$, concluding the proof. $\square$

1.C Finite-Sample Size Properties of Sparsity Tests

We now conduct a simple Monte Carlo simulation to assess the finite-sample performance of the Hausman and residual tests considered in Section 1.4.

The control matrix W in these simulations corresponds to one of seven designs, each corresponding to one of the seven specifications reported in the first two columns of Table 1.1. The treatment and outcome vectors D and Y are drawn from independent normal distributions with zero means and homoskedastic variances σ_Y^2 and σ_D^2 . For each control matrix, we consider two values for these variances: either the variance of the outcome and treatment, or else the variances of OLS residuals in each empirical specification. The first method yields standard deviation values (σ_Y, σ_D) given by $\{(0.082, 0.083), (0.052, 0.095), (0.264, 0.083)\}$ in the three BCH specifications, $(14.2, 2.1)$ in the Ferrara (2022) application, and $\{(37.5, 1.1), (49.71.0), (50.00.8)\}$ in the three Enke (2020) specifications. The second method gives standard deviation values $\{(0.045, 0.004), (0.030, 0.009), (0.197, 0.004)\}$ in the three BCH specifications, $(3.1, 0.7)$ in the Ferrara (2022) application, and $\{(27.2, 0.5), (29.8, 0.6), (31.4, 0.5)\}$ in the three Enke (2020) specifications.

In other words, each design corresponds exactly to the empirical specifications we studied in Section 1.4.3, except we replace the outcome and treatment with mean zero Gaussian variables. The parameter values β, γ and δ in the outcome and propensity score regressions in eqs. (1.1) and (1.2) therefore all equal 0, so that the sparsity assumption

holds trivially. For each design, we apply the same three tests as in Table 1.3, with the same implementation. In particular, the tests allow for heteroskedasticity in the Enke (2020) specification, and for clustering in the other two specifications. To benchmark the performance of the tests, we also compare the test to an oracle which replaces the SBE estimates with a constrained OLS estimate that sets all penalized coefficients to zero.¹⁷

We report the empirical null rejection rates of the sparsity assumption for each test in Table 1.4. The rejection rates do not exceed the nominal 5% level by more than one percentage point, and do not deviate from the oracle rejection rates by more than two percentage points in any specification. Both the feasible and the oracle version of the residual test is conservative in several specifications: this appears to be a consequence of the fact that the mean of the infeasible test statistic $\mathcal{F}$ is given by $E[\sum_i \epsilon_i^2 P_{ii} - \sum_i \epsilon_i^2 P_{S^*,ii}]$, but the recentering in eq. (1.9) only accounts for the first term, since the second term is asymptotically negligible. This may render the test conservative in finite samples when the true sparsity level is not zero, or when unpenalized covariates (that enter estimate of the sparsity index S^*) are present.

¹⁷The unpenalized controls comprise 12 year fixed effects in BCH, county fixed effects Ferrara (2022), and the intercept in Enke (2020).

Table 1.4: Empirical null rejection rates (in percentages) of sparsity tests with nominal level 5% in simulations.

		Large variance		Small variance	
		SBE	Oracle	SBE	Oracle
Outcome	Test	(1)	(2)	(3)	(4)
A: BCH					
$\Delta \ln(\text{violent crime/capita})$	H	5.5	5.6	5.1	5.1
	PR	1.8	1.8	1.5	1.5
	OR	1.7	1.7	1.6	1.6
$\Delta \ln(\text{property crime/capita})$	H	4.9	4.9	5.6	5.5
	PR	1.7	1.7	1.6	1.6
	OR	1.7	1.7	1.5	1.5
$\Delta \ln(\text{murder/capita})$	H	5.4	5.4	4.9	5.0
	PR	1.6	1.6	1.6	1.6
	OR	1.8	1.8	1.7	1.7
B: Ferrara (2022)					
% Semiskilled	H	4.9	4.8	5.1	5.1
Black workers	PR	3.9	4.2	3.9	4.2
	OR	4.1	4.4	4.1	4.4
C: Enke (2020)					
Trump – avg. GOP	H	5.1	5.0	5.2	5.2
	PR	3.0	4.6	3.3	5.0
	OR	3.2	4.8	3.3	4.9
Voted for Trump in 2016	H	5.1	5.1	4.9	4.8
	PR	3.1	4.8	3.2	4.9
	OR	3.3	4.8	3.3	4.8
Voted for Trump in primaries	H	5.1	5.0	5.0	5.1
	PR	3.1	4.5	3.3	5.0
	OR	3.2	4.7	3.2	4.8

Notes: H: Hausman, OR: residual test based on outcome regression, PR: residual test based on propensity score regression. Oracle corresponds to a constrained OLS estimator that sets all penalized coefficients to zero. The control matrix W corresponds to the control matrix in Cols. 1–2 of Table 1.1. In Cols. 1–2, the outcome and treatment variance correspond to the variance of the outcome and treatment in the corresponding specification in Cols. 1–2 of Table 1.1. In Cols. 3–4, the outcome and treatment variances correspond to variances of OLS residuals in these specifications. 10,000 simulation draws.

Chapter 2

Two-Way Duration Models

This paper provides results pertaining to the identification and estimation of the coefficient vector in proportional hazard models for matches, allowing for the presence of two-way unobserved heterogeneity. The most general hazard rate for the duration of a match between entities i and a is assumed to have functional form

$$\lambda_{ia}(t \mid x_{ia}) = \lambda_{0i}(t)\lambda_{0a}(t)\varphi(x_{ia}), \quad (2.1)$$

where λ_{0i} , λ_{0a} are entity-specific baseline hazards and x_{ia} are match-specific covariates. The researcher observes (T_{ia}, x_{ia}) for matched entities with indices $i = 1, 2, \dots, n$ and $a = 1, 2, \dots, m$. We show that identification of φ is possible up to scale in this general model. When we further make the common assumption that $\varphi(x_{ia}) = \exp(x'_{ia}\beta_0)$ and (at least) one of the two entities have constant unobserved heterogeneity, i.e. a “fixed effect,” we prove consistency and asymptotic normality of an estimator of β_0 under two-sided asymptotics.

The hazard rate is in this case given by

$$\lambda_{ia}(t \mid x_{ia}) = \lambda_{0i}(t) \exp(x'_{ia}\beta_0 + \psi_a), \quad (2.2)$$

which naturally nests a “two-way fixed effects” version of the proportional hazards model, where the hazard rate is

$$\lambda_{ia}(t \mid x_{ia}) = \lambda_0(t) \exp(x'_{ia}\beta_0 + \alpha_i + \psi_a). \quad (2.3)$$

The hazard rates mentioned could pertain to the duration of worker i ’s job spell at firm a , where the observed match-specific covariates, x_{ia} , may be the wage, weeks of vacation and the amount of paid sick-leave the worker has at their disposal. The duration could alternatively be the time until recidivism for person i after serving time at correctional facility a , where x_{ia} could contain the length of the sentence that individual i served at correctional facility a .

In the applications mentioned i and a refers to entities coming from two distinct groups, where inter-group matches may exist but intra-group matches do not, commonly referred to in the literature on the econometrics of networks as a bipartite setting. While the identification results in this paper hold in settings where all entities come from the same group with intra-group matches, the subsequent estimation and inference assumes a setting with two distinct groups, although the arguments can be modified to a non-bipartite setting. For the remainder of this paper, we will occasionally refer to i and a as worker and firm respectively.

This paper sits at the intersection of the econometric literature on duration models and the econometric literature on networks. Duration models were introduced in econometrics by Lancaster (1979), who applied the Cox (1972) proportional hazard model to unemployment data to evaluate the then new literature on search models. The large-sample properties of the coefficient estimator for the canonical proportional hazards model was investigated by Tsiatis (1981). A more recent paper by Alvarez et al. (2019) presents a class of optimal stopping models that generate durations, these include search models.

Applications of duration models to *employment* data likewise exist in the literature, a recent example of which is found in Baydur and Mukoyama (2020), who consider the changing job characteristics over the course of the business cycle. They consider two restricted versions of the hazard rates considered here.¹ In one of them, they allow for the duration to be one-way stratified, so that there is a worker-specific baseline hazard,

$$\lambda_{ia}(t \mid x_{ia}) = \lambda_{0i}(t) \exp(x'_{ia}\beta_0),$$

for which a stratified partial likelihood estimator was proposed by Ridder and Tunalı (1999). An alternative specification considered by Baydur and Mukoyama (2020), is one where the individual specific heterogeneity is time-invariant

$$\lambda_{ia}(t \mid x_{ia}) = \lambda_0(t) \exp(x'_{ia}\beta_0 + \alpha_i).$$

¹The authors consider these with competing risks, so their analysis is not nested by the framework considered in this paper.

Time invariant unobserved heterogeneity is often referred to as the *frailty* term in the literature on duration models, the distribution of which is sometimes specified akin to a “random effect.”

An alternative specification for heterogeneity, which has seen popular use in economics, is the mixed proportional hazards model, where the unobserved heterogeneity does not decompose into entity-specific effects but is time-invariant

$$\lambda_{ia}(t \mid x_{ia}) = \lambda_0(t) \exp(x'_{ia}\beta)\xi_{ia}.$$

The mixed proportional hazards model has been considered in Heckman and Singer (1984a, 1984b) and Honoré (1990). The handbook chapter by Van Den Berg (2001) covers the model extensively.

Beyond the duration model literature, the models considered in this paper are also closely related to the burgeoning literature on two-way unobserved heterogeneity and networks.

Two-way unobserved heterogeneity was pioneered in labor economics by Abowd et al. (1999), where the authors sought to decompose earnings into match-specific explanatory variables, a worker fixed effect and a firm fixed effect. For the two-way fixed effects model there is likewise an interest in estimating the agent-specific effects, but the statistical properties of the resulting estimator depends on the density of the graph, i.e. the number of connections for each node, as shown by Jochmans and Weidner (2019). Naturally, one can reduce the parameter space by assuming a distribution for the unobserved heterogeneity, as done by Allman et al. (2009), thereby alleviating some of the concerns arising from sparsity. In labor markets, firms tend to be more well-connected than workers, so one may

wish to impose a distributional assumption for the less well-connected side of the market as in Bonhomme et al. (2019).

Due to the covariates entering through an exponential single-index function, the resulting estimator will take a logit form, which is common in the literature on network formation. An excellent treatment of network models is found in the book by B. Graham and de Paula (2020). Specifically, the estimator will look similar to that of Charbonneau (2017), whose asymptotic properties were analyzed by Jochmans (2018). The identification argument developed by Charbonneau (2017), and in this paper, borrow ideas from a long literature on panel data logit models, stretching back to Rasch (1960) and Rasch (1961). They showed that fixed effects can be eliminated by conditioning the likelihood function on a sufficient statistic. A similar logic is employed by Cox (1972) to estimate the common parameter β_0 in the canonical proportional hazards model. A more thorough overview of the literature on logit panel data models can be found in the introduction to Chapter 3.

The remainder of the paper is structured as follows. Section 2.1 discusses identification of the models. Section 2.2 introduces an estimator for the coefficient β_0 in the model with hazard rate given by (2.2) and proves consistency and asymptotic normality. Section 2.3 concludes.

Notation

This subsection serves to collect some of the notation that we will use throughout. The outcome is a duration for a match between entities i and a denoted by T_{ia} , the observable covariates are x_{ia} and, when relevant, the fixed effects are denoted by α_i and ψ_a . The

hazard rates will be denoted by λ , with subscripts indicating either that it is the baseline hazard rate, λ_0 , entity specific, e.g. λ_{0i} , or match specific λ_{ia} .

From the identification argument it becomes clear that a lot of the analysis hinges on quadruples of two workers and two firms. For notation it will be convenient to define $\tilde{x}_{ijab} \equiv x_{ib} + x_{ja} - x_{ia} - x_{jb}$, and to minimize the number of summation signs, we will occasionally refer to a generic quadruple as κ and write $\tilde{x}_\kappa$, or in cases where we refer to a pair (tuple) we will occasionally denote these by τ instead of (i, a) .

Furthermore, for the proofs we will be using the σ -algebra generated by the covariates and fixed effects $\mathcal{F}_{mn} = \sigma(\{\{x_{ia}\}_{mn}, \{\alpha_i\}_n, \{\psi_a\}_m, \lambda_0\})$ and $\mathcal{F} = \sigma(\cup_{m \in \mathbb{N}} \cup_{n \geq m} \mathcal{F}_{mn})$. The proofs hinges on projections of random variables and vectors onto the aforementioned σ -algebras, these projections will be denoted by a superscript calligraphic $\mathcal{F}$.

2.1 Identification

Our data consists of observations (T_{ia}, x_{ia}) where $i = 1, 2, \dots, n$ and $a = 1, 2, \dots, m$. We will show that identification of φ is possible up to scale as well as identification of the ratio of intra-group hazard rates if the data-generating process has hazard rate given by (2.1). As corollaries to this, identification of β_0 and the intra-group differences in fixed effects are identified if the data-generating process has hazard rate given by (2.3).

To gain intuition for the identification arguments, consider the model with two-way fixed effects in (2.3), and proceed in a similar fashion as in Cox (1972). In that paper, they condition on durations still at risk that all have a common baseline hazard rate. The probability of a certain duration being the first to terminate does not depend on the baseline

hazard. As the models in this paper have entity-specific unobserved heterogeneity, one should intuitively restrict attention to only the at-risk durations that share workers or firms in common. To see this, consider a worker, i , with employment spells at two firms, a and b . Assume it is known that one of the two spells terminated at duration t . That is, consider the events

$$\begin{aligned}\tilde{\mathcal{W}}_{1,iab}(t, \Delta) &= \{T_{iA} \in [t, t + \Delta), T_{iB} > t + \Delta\}, \\ \tilde{\mathcal{W}}_{2,iab}(t, \Delta) &= \{T_{iA} > t + \Delta, T_{iB} \in [t, t + \Delta)\},\end{aligned}$$

Then the probability that it is the employment spell with firm a that terminated first is given by

$$\begin{aligned}\lim_{\Delta \rightarrow 0^+} P(\tilde{\mathcal{W}}_{1,iab}(t, \Delta) \mid \tilde{\mathcal{W}}_{1,iab}(t, \Delta) \cup \tilde{\mathcal{W}}_{2,iab}(t, \Delta), \lambda_{0i}, \lambda_{0a}, \lambda_{0b}, x_{ia}, x_{ib}) \\ &= \frac{\lambda_{ia}(t|x_{ia})}{\lambda_{ia}(t|x_{ia}) + \lambda_{ib}(t|x_{ib})} \\ &= \frac{\lambda_{0i}(t)\lambda_{0a}(t)\varphi(x_{ia})}{\lambda_{0i}(t)\lambda_{0a}(t)\varphi(x_{ia}) + \lambda_{0i}(t)\lambda_{0b}(t)\varphi(x_{ib})} \\ &= \frac{\lambda_{0a}(t)\varphi(x_{ia})}{\lambda_{0a}(t)\varphi(x_{ia}) + \lambda_{0b}(t)\varphi(x_{ib})} \\ &= \frac{1}{1 + \frac{\lambda_{0b}(t)\varphi(x_{ib})}{\lambda_{0a}(t)\varphi(x_{ia})}}.\end{aligned}$$

Thereby eliminating the entity specific hazard rate for i .

A similar derivation could then be made for a worker j , who also has spells at the same two firms. Specifically, consider worker j and switch around the two events we get

$$\begin{aligned} \lim_{\Delta \rightarrow 0^+} P(\tilde{\mathcal{W}}_{2,jab}(t, \Delta) \mid \tilde{\mathcal{W}}_{1,jab}(t, \Delta) \cup \tilde{\mathcal{W}}_{2,jab}(t, \Delta), \lambda_{0j}, \lambda_{0a}, \lambda_{0b}, x_{ja}, x_{jb}) \\ = \frac{1}{1 + \frac{\lambda_{0a}(t) \varphi(x_{ja})}{\lambda_{0b}(t) \varphi(x_{jb})}}. \end{aligned}$$

Taking intersections of the events across the two workers, we derive at the conditioning sets that we use to prove identification

$$\begin{aligned} \mathcal{W}_{1,ijab}(t, \Delta) &= \{T_{ia} \in [t, t + \Delta), T_{ib} > t + \Delta\} \cap \{T_{ja} > t + \Delta, T_{jb} \in [t, t + \Delta)\}, \\ \mathcal{W}_{2,ijab}(t, \Delta) &= \{T_{ia} > t + \Delta, T_{ib} \in [t, t + \Delta)\} \cap \{T_{ja} \in [t, t + \Delta), T_{jb} > t + \Delta\}. \end{aligned} \tag{2.4}$$

Hence, under a support condition for the covariates, φ is identified up to scale.

Identification therefore rests on the existence of a quadruple $\{i, j; a, b\}$, such that i and j have both had employment spells at firms a and b . In words, the shortest spell for both workers have duration t , but the firm with which each worker has their shortest spell is not the same. Given this knowledge, which of the pairs, T_{ib}, T_{ja} or T_{ia}, T_{jb} , that terminates first depends only on the observables.

Theorem 2.1. *Assume the data-generating process has hazard rate given by (2.1) Consider i, j both having positive durations at a, b . If $x_{ia} \mid x_{ja}, x_{ib}, x_{jb}$ has full support, then φ is identified up to scale.*

An identical argument for the more restricted data-generating processes, where the functional form of φ is specified to be an exponential single-index function, yields the following corollary, where the support condition has been substituted instead by a rank condition analogous to one that yields identification in the linear regression context, showing that β_0 is identified.

For notational economy, we define $\tilde{x}_{ijab} \equiv x_{ib} + x_{ja} - x_{ia} - x_{jb}$.

Corollary 2.1. *Assume the data-generating process has hazard rate given by (2.2) or (2.3). Consider i, j both having positive duration at a, b . If $E[\tilde{x}_{ijab}\tilde{x}'_{ijab}]$ is non-singular then β_0 is identified.*

Finally, with either φ identified or if there exists a worker with $x_{ia} = x_{ib}$, we can identify the ratio of intra-group hazards. If e.g. worker i has worked at both firms a and b we can identify $\lambda_{0b}/\lambda_{0a}$. Note that for any function $g(t) > 0$ substituting the baselines in equation (2.1) with

$$\tilde{\lambda}_{0i}(t) := \frac{\lambda_{0i}(t)}{g(t)}, \quad \tilde{\lambda}_{0a}(t) := \lambda_{0a}(t) \cdot g(t)$$

would yield the exact same hazard rate for worker i at firm a , so we can at best identify the baselines up to scale.

Theorem 2.2. *Assume the data-generating process has hazard rate given by (2.1). Consider firms a and b that both have had positive duration with worker i . If either 1) φ is identified (up to scale), or 2) $x_{ia} = x_{ib}$, then the ratio $\frac{\lambda_{0b}(t)}{\lambda_{0a}(t)}$ is identified.*

Naturally, this also allows the researcher to identify the difference of fixed effects.

Corollary 2.2. *Assume the data-generating process has hazard rate given by (2.2). Consider firms a and b that both have had positive duration with worker i . If either 1) β_0 is identified, or 2) $x_{ia} = x_{ib}$, then the difference $\psi_b - \psi_a$ is identified.*

Note that if worker i has worked for firms a and b and worker j has worked for firms b and c we immediately get identification of $\lambda_{0c}/\lambda_{0a}$ or, in the two-way fixed effects model, $\psi_c - \psi_a$ as well. In the terminology of networks econometrics, the preceding theorem and corollary allows us to identify the ratio of the intra-group unobserved heterogeneity for all entities in a maximally connected component.

2.1.1 Comment on the Different Identification Proofs

In the restricted model where the unobserved heterogeneity for at least one side is time invariant, (2.2) or (2.3), the identification argument is simpler, as the termination time of the pairs, T_{ia} and T_{jb} in $\mathcal{W}_{1,ijab}(t, \Delta)$, and T_{ib} and T_{ja} in $\mathcal{W}_{2,ijab}(t, \Delta)$, need not be close to each other. Rather the conditioning argument can be expressed by the indicator variables²

$$\begin{aligned} W_{1,ijab} &= \mathbb{1}\{T_{ia} < T_{ib}, T_{ja} > T_{jb}\}, \\ W_{2,ijab} &= \mathbb{1}\{T_{ia} > T_{ib}, T_{ja} < T_{jb}\}. \end{aligned} \tag{2.5}$$

The different identification arguments lead to estimators that condition on sets of “effective quadruples,” which are quadruples where both workers have positive duration of

²Analogously, if we instead want the stratified baseline hazard to be on the firm side, so the hazard rate for the duration is given by $\lambda_{ia}(t | x_{ia}) = \lambda_{0a}(t) \exp(x'_{ia}\beta_0 + \alpha_i)$, we would use the indicator variables

$$\begin{aligned} \tilde{W}_{1,ijab} &= \mathbb{1}\{T_{ia} > T_{ja}, T_{ib} < T_{jb}\}, \\ \tilde{W}_{2,ijab} &= \mathbb{1}\{T_{ia} < T_{ja}, T_{ib} > T_{jb}\}, \end{aligned}$$

redefining $\tilde{x}_{ijab}$ as necessary. Both work for the case with two-way fixed effects as in (2.3).

work at both firms that satisfy the conditioning requirements. However, since the identification argument for model (2.1) requires the durations T_{ia} and T_{jb} to be close to each other there is a (potentially much) smaller number of effective quadruples, and it furthermore necessitates the use of kernels as the durations are continuous random variables, which in turn requires careful consideration of bandwidth choice. In contrast, in the simpler models (2.2) and (2.3), effective quadruples are all those that satisfy that $W_{1,ijab} + W_{2,ijab} = 1$, which does not necessitate the use of kernels and hence does not require the econometrician to choose a bandwidth.

The focus will be on the latter case where, under the assumption of two-sided asymptotics, we prove consistency and asymptotic normality by adapting the proofs in Jochmans (2018).

2.2 Estimation and Inference

In this section we consider estimation and inference for the coefficient vector on the observables. The focus is on the case where the unobserved heterogeneity for one group, or both, is time-invariant. In this instance, we allow for two-sided asymptotics and by arguments adapted from the network formation literature, specifically B. S. Graham (2017) and Jochmans (2018), we prove consistency and asymptotic normality of the estimator.

Unlike the network formation papers, the outcome variable in our setting is not whether a link is formed, but rather a duration conditional on a link. Trying to endogenize the network formation is beyond the scope of this paper. We can therefore consider either that all inter-group matches exist or that the network formation is exogenous. One may imple-

ment the latter by adding indicator functions for whether a link exists and all calculations of convergence rates therefore need to account for this as well. We proceed with the former.

The estimator is the maximizer of the partial likelihood function given by

$$Q_{mn}(\beta) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{a=1}^{m-1} \sum_{b=a+1}^m \underbrace{(W_{1,ijab} \log(F(\tilde{x}'_{ijab}\beta)) + W_{2,ijab} \log(F(-\tilde{x}'_{ijab}\beta)))}_{q_{ijab}(\beta)}, \quad (2.6)$$

where $F(z) = 1/(1+\exp(-z))$ the (standard) logistic CDF. We define $\kappa : \{1, 2, \dots, m\}^2 \times \{1, 2, \dots, n\}^2 \rightarrow \{1, 2, \dots, k_{mn}\}$, where $k_{mn} = \binom{m}{2} \binom{n}{2}$, which maps quadruple indices into the natural numbers, and abusing notation we will write $q_{\kappa}(\beta)$ to denote the objective function contribution associated with a generic set of indices, and we will replace the four summations with a sum just over κ instead.

Furthermore, we denote by $k_{mn}^* = \sum_{\kappa} (W_{1,\kappa} + W_{2,\kappa}) \equiv \sum_{\kappa} W_{\kappa}$, the quadruples that contribute to the objective function (the effective sample size). Finally, define the expected fraction of quadruples that contribute to the objective function

$$p_{mn} = \frac{E[k_{mn}^*]}{k_{mn}} = \frac{\sum_{\kappa} E[W_{\kappa}]}{k_{mn}}. \quad (2.7)$$

2.2.1 Sparsity Comments

Throughout the paper we maintain the (unrealistic) assumption that all inter-group matches exist. The theory readily extends to the case where only a subset of links exists, as long as the presence of a link is exogenous to the subsequent duration itself conditional on the covariates and fixed effects. The objective function would then have to be modified

by “tagging on” an indicator for the presence of a link, diminishing the effective sample size, hence care has to be taken to ensure that the number of effective quadruples are still growing at an appropriate rate to ensure convergence.

For example, consider that each of the n workers get matched randomly to 2 of the m firms in their career. Any pair of firms $l = (a, b)$ will be matched with $x_{m,l} \sim \text{Binom}(n, \rho_{m,l})$ workers, where $\sum_l \rho_{m,l} = 1$, giving rise to $\binom{x_{m,l}}{2}$ quadruples involving firm pair l . This implies that the expected number of quadruples in the economy will be

$$\sum_l \binom{m}{2} E \left[\binom{x_{m,l}}{2} \right] = \binom{n}{2} \sum_l \rho_{m,l}^2.$$

If all firms are equally sized in expectation, so that $\rho_{m,l} = \binom{m}{2}^{-1}$ for all l , which yields $\binom{n}{2} \binom{m}{2}^{-1}$ expected quadruple links. In this instance, we would need $m = o(n)$ to have the number of quadruples growing (and likely even stronger requirements as not all quadruples are useful). If on the other hand, $\rho_{m,l} = 1$ and $\rho_{m,l'} = 0$ for $l' \neq l$, there will be $\binom{n}{2}$ matched quadruples in the economy.

2.2.2 Consistency

The proof for consistency follow those of Jochmans (2018) closely and we make similar assumptions. The main differences are that the estimator considered here is a two-sample estimator,³ and that the indicator function corresponding to the conditioning argument

³Since the arguments are analogous to those in Jochmans (2018), this implies that the arguments used in this paper also prove that the estimator in Charbonneau (2017) readily extends to a bipartite setting.

differ from those used in Jochmans (2018), who has a binary outcome variable as it is a network formation model.

We make the following assumptions

Assumption 2.1. *We assume that the nodes in the network are sampled independently.*

Assumption 2.2. *We assume that β_0 is interior to $\Theta_0 \subset \mathbb{R}^p$, which is assumed compact.*

Assumption 2.3. *We assume that $E[\|x_{ia}\|^2] < \infty$ for all i, a .*

Assumption 2.4. *We assume that $mp_{mn} \rightarrow \infty$ and that the Hessian,*

$$\lim_{m,n \rightarrow \infty} \frac{1}{k_{mn}p_{mn}} \sum_{\kappa} E [W_{\kappa} \tilde{x}_{\kappa} \tilde{x}_{\kappa}' F(\tilde{x}_{\kappa}' \beta_0) (1 - F(\tilde{x}_{\kappa}' \beta_0))],$$

is invertible.

Theorem 2.3. *Assume the data-generating process has hazard rate given by (2.2) and that Assumptions 2.1-2.4 hold. Then $\hat{\beta} \xrightarrow{p} \beta_0$.*

2.2.3 Asymptotic Normality

In line with the proof of asymptotic normality in Jochmans (2018), we will show that evaluating the score conditioning on the information-set generated by all covariates and fixed effects leads to an asymptotically normal random vector. We then establish asymptotic equivalence between this and the score without any conditioning.

The sample score vector is given by

$$\hat{G}_{mn}(\beta) = \sum_{i=1}^n \sum_{a=1}^m \sum_{j \neq i} \sum_{b \neq a} \underbrace{\frac{\partial}{\partial \beta} q_{ijab}(\beta)}_{g_{ijab}(\beta)}.$$

In the proofs we show that the covariance of the terms in the score vector is zero unless they share a dyad, not just a node, and it is useful therefore to define the following object, which we will refer to as the sample covariance of the score

$$\hat{\Gamma}_{mn}(\beta) = \sum_{i=1}^n \sum_{a=1}^m \sum_{j \neq i} \sum_{k \neq i} \sum_{b \neq a} \sum_{c \neq a} g_{ijab}(\beta) g_{ikac}(\beta)'. \quad (2.8)$$

The sample Hessian is given by

$$\hat{H}_{mn}(\beta) \equiv \frac{\partial^2}{\partial \beta \partial \beta'} Q_{mn}(\beta) = - \sum_{\kappa} W_{\kappa} \tilde{x}_{\kappa} \tilde{x}_{\kappa}' F(\tilde{x}_{\kappa}' \beta_0) (1 - F(\tilde{x}_{\kappa}' \beta_0)). \quad (2.9)$$

Use these to define

$$\hat{\Omega}_{mn}(\beta) = \hat{H}_{mn}(\beta)^{-1} \hat{\Gamma}_{mn}(\beta) \hat{H}_{mn}(\beta)^{-1}. \quad (2.10)$$

We make the following additional assumption

Assumption 2.5. *We assume that $E[\|x_{ia}\|^6] < \infty$ for all i, a .*

Finally, the proofs rely on showing asymptotic equivalence between the score vector and the score vector projected onto the covariates and a duration. We make the following definition and technical assumption

Definition 2.1. *We define the projected score vector as*

$$V_{mn} = \sum_{i=1}^n \sum_{a=1}^m \underbrace{\sum_{j \neq i} \sum_{b \neq a} E[s_{ijab}(\beta_0) \mid T_{ia}, \mathcal{F}]}_{v_{ia}}.$$

Assumption 2.6. *We assume that the conditional variance of the projected score vector,*

$$\Gamma_{mn}^{\mathcal{F}} = \sum_{i=1}^n \sum_{a=1}^m E[v_{ia} v'_{ia} \mid \mathcal{F}],$$

satisfies a rank-condition

$$\lambda_{\min}(\Gamma_{mn}^{\mathcal{F}}) = \omega_p(m^2 n^2).$$

That is, the smallest eigenvalue of $\Gamma_{mn}^{\mathcal{F}}$ is of order strictly greater than $m^2 n^2$ in probability.

This yields the following theorem

Theorem 2.4. *Assume the data-generating process has hazard rate given by (2.2) or (2.3).*

If Assumptions 2.1-2.6 hold, then

$$\hat{\Omega}_{mn}(\hat{\beta})^{-1/2} \left(\hat{\beta} - \beta_0 \right) \xrightarrow{d} N(0, I)$$

as $m, n \rightarrow \infty$.

The proof for the theorem follows that of Jochmans (2018), except we invoke a different conditional Central Limit Theorem, which requires us to make Assumption 2.6.

2.3 Conclusion

This paper has focused on extending the proportional hazard model by allowing two-way unobserved heterogeneity. We did so by first proving that the unobserved heterogeneity can be purged with a conditioning argument. In a simpler model, where the heterogeneity is time-invariant, we proved that the partial likelihood estimator stemming from the

conditioning argument is consistent and asymptotically normal, rendering estimation and inference possible.

While we view these as important steps in the direction of making duration models more useful for two-sided networks, such as labor markets, there are further steps needed to make the models presented here practically applicable. At least four points warrant further investigation, and they are mentioned here in no particular order.

In most practical examples censoring plays a key role when looking at duration data, as data usually do not cover enough years for all relevant durations to have terminated with probability 1.

In many economic settings, such as the labor market, competing risks are likewise important, if one want to treat the transition from employment to further employment as different from the transition to unemployment or non-employment.

Furthermore, the analysis so far does not allow for time-varying covariates, which is unrealistic in many labor market scenarios, where the covariates, such as wages, typically vary over time. Specifically, wages tend to increase with tenure, which may impact the amount of time a worker spends at a firm.

Finally, the analysis here pre-supposes that the duration data is continuous, though in practice the data is sampled at a much lower frequency.

In addition to the aforementioned points for future research, that are commonplace in other duration models, an interesting extension would be to incorporate peer effects, so that the duration of an employment match depends on the composition of a worker's coworkers.

References

- Abowd, J. M., Kramarz, F., & Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, 67(2), 251–333.
- Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A), 3099–3132.
- Alvarez, F., Borovičková, K., & Shimer, R. (2019, October). *Decomposing Duration Dependence in a Stopping Time Model* (tech. rep. No. w22188). National Bureau of Economic Research. Cambridge, MA.
- Baydur, I., & Mukoyama, T. (2020). Job duration and match characteristics over the business cycle. *Review of Economic Dynamics*, 37, 33–53.
- Bonhomme, S., Lamadon, T., & Manresa, E. (2019). A Distributional Framework for Matched Employer Employee Data. *Econometrica*, 87(3), 699–739.
- Bulinski, A. V. (2017). Conditional Central Limit Theorem. *Theory of Probability & Its Applications*, 61(4), 613–631.
- Charbonneau, K. B. (2017). Multiple fixed effects in binary response panel data models. *The Econometrics Journal*, 20(3), S1–S13.
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220.
- Graham, B., & de Paula, Á. (2020). *The econometric analysis of network data*. Academic Press is an imprint of Elsevier.
- Graham, B. S. (2017). An Econometric Model of Network Formation With Degree Heterogeneity. *Econometrica*, 85(4), 1033–1063.

- Heckman, J., & Singer, B. (1984a). The Identifiability of the Proportional Hazard Model. *The Review of Economic Studies*, 51(2), 231–241.
- Heckman, J., & Singer, B. (1984b). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, 52(2), 271–320.
- Honoré, B. E. (1990). Simple Estimation of a Duration Model with Unobserved Heterogeneity. *Econometrica*, 58(2), 453–73.
- Jochmans, K. (2018). Semiparametric Analysis of Network Formation. *Journal of Business & Economic Statistics*, 36(4), 705–713.
- Jochmans, K., & Weidner, M. (2019). Fixed-Effect Regressions on Network Data. *Econometrica*, 87(5), 1543–1560.
- Lancaster, T. (1979). Econometric Methods for the Duration of Unemployment. *Econometrica*, 47(4), 939–56.
- Newey, W. K., & McFadden, D. (1994). Chapter 36 Large sample estimation and hypothesis testing. Elsevier.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Denmark: Pædagogiske Institut.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, 4, 321–333.
- Ridder, G., & Tunalı, İ. (1999). Stratified partial likelihood estimation. *Journal of Econometrics*, 92(2), 193–232.

Tsiatis, A. A. (1981). A Large Sample Study of Cox's Regression Model. *The Annals of Statistics*, 9(1), 93–108.

Van Den Berg, G. J. (2001). Chapter 55 - Duration Models: Specification, Identification and Multiple Durations. In J. J. Heckman & E. Leamer (Eds.). Elsevier.

Appendix

2.A Notes on Computation

This subsection provides a few comments about computation. The objective function, gradient and Hessian are computationally expensive to evaluate as they consist of summing over quadruples, to which we apply the exponential function. Massive speedups can be gained from storing the covariates in contiguous memory as requesting data from RAM is a major bottleneck on modern computers. We therefore define $\check{x}_{\kappa^*} = (W_{1,\kappa^*} - W_{2,\kappa^*})\tilde{x}_{\kappa^*}$, where $\kappa^* = \{\kappa : W_{1,\kappa} + W_{2,\kappa} = 1\}$. If the covariate is scalar, instead of having it stored in an $n \times m$ matrix, and accessing the elements from this that satisfy $W_{1,\kappa} + W_{2,\kappa} = 1$, we can instead store only the relevant ones in a vector, which is significantly faster to access. Additionally, modern computers substantially speed up computation from vectorized operations, so the storage format allows quicker function evaluation as well.

Furthermore, notice that

$$\log(F(z)) = \log\left(\frac{1}{1 + \exp(-z)}\right) = -\log(1 + \exp(-z)),$$

which is cheaper than having to do the division. Then

$$\begin{aligned}\check{Q}_{mn}(\beta) &= -\sum_{\kappa^*=1}^{k^*} \log(1 + \exp(-\check{x}'_{\kappa^*}\beta)), \\ \check{G}_{mn}(\beta) &= \sum_{\kappa^*=1}^{k^*} \frac{\check{x}_{\kappa^*}}{1 + \exp(\check{x}'_{\kappa^*}\beta)}, \\ \check{H}_{mn}(\beta) &= -\sum_{\kappa^*=1}^{k^*} \left(\frac{\check{x}_{\kappa^*}}{1 + \exp(\check{x}'_{\kappa^*}\beta)}\right)^2 \exp(\check{x}'_{\kappa^*}\beta).\end{aligned}$$

Optimization by e.g. Newton's method requires evaluating the first and second derivative of the objective function numerous times. For a given value of β , it is therefore possible to calculate $\exp(\tilde{x}'_{\kappa^*}\beta)$ once for each relevant quadruple, κ^* , use that in the calculation of $\frac{\tilde{x}_{\kappa^*}}{1+\exp(\tilde{x}'_{\kappa^*}\beta)}$, which provides “building blocks” for constructing the contribution of κ^* to both $\tilde{G}_{mn}$ and $\tilde{H}_{mn}$ in a more efficient manner, by just “plugging in” the calculated value to the score and Hessian.

2.B Proofs

Proof of Theorem 2.1: Let events $\mathcal{W}_{1,ijab}(t, \Delta)$ and $\mathcal{W}_{2,ijab}(t, \Delta)$ be as defined in (2.4).

Let $\mathcal{F}_{ijab} = \sigma(\{x_{ia}, x_{ib}, x_{ja}, x_{jb}, \lambda_{0i}, \lambda_{0j}, \lambda_{0a}, \lambda_{0b}\})$ denote the information set pertaining to the corresponding observations. Then

$$\begin{aligned} & \lim_{\Delta \rightarrow 0^+} \frac{P(\mathcal{W}_{1,ijab}(t, \Delta) \mid \mathcal{F}_{ijab})}{\Delta} \\ &= \lambda_{0i}(t)\lambda_{0a}(t)\varphi(x_{ia})S_{ia}(t \mid x_{ia})S_{ib}(t \mid x_{ib})\lambda_{0j}(t)\lambda_{0b}(t)\varphi(x_{jb})S_{jb}(t \mid x_{jb})S_{ja}(t \mid x_{ja}), \end{aligned}$$

and

$$\begin{aligned} & \lim_{\Delta \rightarrow 0^+} \frac{P(\mathcal{W}_{2,ijab}(t, \Delta) \mid \mathcal{F}_{ijab})}{\Delta} \\ &= \lambda_{0i}(t)\lambda_{0b}(t)\varphi(x_{ib})S_{ib}(t \mid x_{ib})S_{ia}(t \mid x_{ia})\lambda_{0j}(t)\lambda_{0a}(t)\varphi(x_{ja})S_{ja}(t \mid x_{ja})S_{jb}(t \mid x_{jb}), \end{aligned}$$

where $S_{ia}(t \mid x_{ia}) = \exp\left(-\int_0^t \lambda_{ia}(u \mid x_{ia}) du\right)$ is the survivor for the duration T_{ia} . Then the probability of one event, conditioning on the union of the two, we get

$$\begin{aligned} & \lim_{\Delta \rightarrow 0^+} P(\mathcal{W}_{1,ijab}(t, \Delta) \mid \mathcal{W}_{1,ijab}(t, \Delta) \cup \mathcal{W}_{2,ijab}(t, \Delta), \mathcal{F}_{ijab}) \\ &= \lim_{\Delta \rightarrow 0^+} \frac{P(\mathcal{W}_{1,ijab}(t, \Delta) \mid \mathcal{F}_{ijab})}{P(\mathcal{W}_{1,ijab}(t, \Delta) \mid \mathcal{F}_{ijab}) + P(\mathcal{W}_{2,ijab}(t, \Delta) \mid \mathcal{F}_{ijab})} \\ &= \frac{\varphi(x_{ia})\varphi(x_{jb})}{\varphi(x_{ia})\varphi(x_{jb}) + \varphi(x_{ib})\varphi(x_{ja})} \\ &= \frac{1}{1 + \frac{\varphi(x_{ib})\varphi(x_{ja})}{\varphi(x_{ia})\varphi(x_{jb})}}. \end{aligned}$$

which does not depend on the λ 's nor on the time period t . Proving that φ is identified up to scale under the support condition specified. $\square$

Proof of Theorem 2.2: Consider the events $\tilde{\mathcal{W}}_{1,iab}(t, \Delta) = \{T_{ia} \in [t, t + \Delta), T_{ib} > t + \Delta\}$ and $\tilde{\mathcal{W}}_{2,iab}(t, \Delta) = \{T_{ia} > t + \Delta, T_{ib} \in [t, t + \Delta)\}$ and $\tilde{\mathcal{F}}_{iab} = \sigma(\{x_{ia}, x_{ib}, \lambda_{0i}, \lambda_{0a}, \lambda_{0b}\})$. Then

$$\lim_{\Delta \rightarrow 0^+} P(\tilde{\mathcal{W}}_{1,iab}(t, \Delta) | \tilde{\mathcal{W}}_{1,iab}(t, \Delta) \cup \tilde{\mathcal{W}}_{2,iab}(t, \Delta), \tilde{\mathcal{F}}_{iab}) = \frac{1}{1 + \frac{\lambda_{0b}(t)\varphi(x_{ib})}{\lambda_{0a}(t)\varphi(x_{ia})}}.$$

$\square$

Comment on the Order of the Number of Terms

The purpose of this subsection is to derive certain calculations for the number of terms that arise in different calculations. The number of terms play a key role in the proofs that follow.

We assume, without loss of generality, that $n \geq m$. In variance calculations we consider products such as $E[(\sum_{\kappa} q_{\kappa}(\beta))(\sum_{\kappa'} q_{\kappa'}(\beta))]$, which contain $O(n^4 m^4)$ terms. By assumption they will be independent whenever they share no indices in common. Consider therefore $\kappa = \kappa(\{i, j; a, b\})$ and $\kappa' = \kappa(\{i', j'; a', b'\})$, having indices in common and therefore potentially exhibiting dependence. If they share the index $a = a'$, there are $\binom{m-1}{2} = O(m^2)$ choices for b, b' to be distinct, and $\binom{n}{4} = O(n^4)$ choices for i, i', j, j' to be distinct. As there are m different combinations for $a = a'$, the number of quadruples that have at least one node in common are $O(n^4 m^3)$ (assuming that $n \geq m$). This also implies, that for a quadruple $\kappa = \kappa(\{i, j; a, b\})$, there is $O(n^2 m)$ other quadruples sharing

at least one index with it. Therefore, further abusing notation we will write $\mathbb{1}\{\kappa \cap \kappa' \neq \emptyset\}$ to indicate the quadruples that share indices.

When proving asymptotic normality, we condition on the σ -algebra generated by the covariances and the fixed effects, which we will show implies that functions of quadruples will be independent unless they share at least one worker-firm dyad, e.g. $i = i'$ and $a = a'$. There are $O(n^3 m^3)$ such terms. Note however, that the order is in fact smaller as not all quadruples are “useful” and therefore do not contribute to the objective function. We will return to this in the proofs.

Finally, when discussing limits, we consider sequences of m and n such that they are of the same order. Hence we consider $m, n \rightarrow \infty$, where $m/(m+n) \rightarrow c \in (0, 1)$. A more formal write-up would index the variables by some $\nu \in \mathbb{N}$, so $m = m_\nu \rightarrow \infty$ and $n = n_\nu \rightarrow \infty$ as $\nu \rightarrow \infty$ in such a way that $m_\nu/(m_\nu + n_\nu) \rightarrow c \in (0, 1)$, however we will omit such an index for notational economy.

Proof of Theorem 2.3: Note that the partial-likelihood contributions are bounded by

$$\begin{aligned}
|q_\kappa(\beta)| &\leq |\log(F(\tilde{x}'_\kappa\beta))| + |\log(F(-\tilde{x}'_\kappa\beta))| \\
&\leq |\log(1 + \exp(\tilde{x}'_\kappa\beta))| + |\log(1 + \exp(-\tilde{x}'_\kappa\beta))| \\
&\leq 2|\tilde{x}'_\kappa\beta| + 2\log(2) \\
&\leq 2(\|\tilde{x}_\kappa\| \|\beta\| + \log(2)),
\end{aligned}$$

where the last inequality follows from an application of Cauchy-Schwarz inequality. Let $Q_{mn}(\beta) = \sum_\kappa q_\kappa(\beta)$ and let $\bar{q}_\kappa(\beta) = q_\kappa(\beta) - E[q_\kappa(\beta)]$, the demeaned contribution to the partial likelihood. Then

$$\frac{Q_{mn}(\beta)}{k_{mn}^*} - \frac{E[Q_{mn}(\beta)]}{E[k_{mn}^*]} = \frac{\sum_\kappa \bar{q}_\kappa(\beta)}{E[k_{mn}^*]} + \frac{\sum_\kappa q_\kappa(\beta)}{E[k_{mn}^*]} \left(\frac{E[k_{mn}^*]}{k_{mn}^*} - 1 \right).$$

For the first right hand term, note that by assumption 2.1 and the bound derived above, the variance of $q_\kappa(\beta)$ exists. Therefore, Chebyshev's inequality holds,

$$P\left(\left|\frac{\sum_\kappa \bar{q}_\kappa(\beta)}{E[k_{mn}^*]}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2} \frac{E[(\sum_\kappa \bar{q}_\kappa(\beta))^2]}{E[k_{mn}^*]^2},$$

for any given β . Using the notation above, we have for the numerator

$$\begin{aligned}
E \left[\left(\sum_{\kappa} \bar{q}_{\kappa}(\beta) \right) \left(\sum_{\kappa'} \bar{q}_{\kappa'}(\beta) \right) \right] &= \sum_{\kappa} \sum_{\kappa'} E [\bar{q}_{\kappa}(\beta) \bar{q}_{\kappa'}(\beta)] \\
&= \sum_{\kappa} \sum_{\kappa'} \mathbb{1}_{\{\kappa \cap \kappa' \neq \emptyset\}} E [\bar{q}_{\kappa}(\beta) \bar{q}_{\kappa'}(\beta)] \\
(\text{Cauchy-Schwarz}) &\leq \sum_{\kappa} \sum_{\kappa'} \mathbb{1}_{\{\kappa \cap \kappa' \neq \emptyset\}} \sqrt{V [\bar{q}_{\kappa}(\beta)]} \sqrt{V [\bar{q}_{\kappa'}(\beta)]} \\
&= O(n^2 m k_{mn} p_{mn}),
\end{aligned}$$

where the last inequality follows from the variance of each $\bar{q}_{\kappa}(\beta)$ being bounded, and since in expectation there are $k_{mn} p_{mn}$ effective quadruples and $O(n^2 m)$ quadruples sharing at least one index with it. Since $E[k_{mn}^*] = k_{mn} p_{mn}$ and $k_{mn} = O(n^2 m^2)$, the upper bound from Chebyshev's inequality is

$$P \left(\left| \frac{\sum_{\kappa} \bar{q}_{\kappa}(\beta)}{E[k_{mn}^*]} \right| > \varepsilon \right) \leq \frac{1}{\varepsilon^2} \frac{E[(\sum_{\kappa} \bar{q}_{\kappa}(\beta))^2]}{E[k_{mn}^*]^2} = O \left(\frac{1}{m p_{mn}} \right),$$

which converges to zero by assumption 2.4. Note then that for the second term we want to show that

$$\frac{E[k_{mn}^*]}{k_{mn}^*} - 1 \xrightarrow{p} 0,$$

which is equivalent, by the definition of p_{mn} in (2.7), to showing that

$$\frac{k_{mn}^*}{k_{mn}} - p_{mn} = \frac{\sum_{\kappa} (W_{\kappa} - E[W_{\kappa}])}{k_{mn}} \xrightarrow{p} 0.$$

To do this, we mimic the proof above, denoting $\bar{W}_\kappa = W_\kappa - E[W_\kappa]$ and note that by Chebyshev's inequality, for $\varepsilon > 0$

$$P\left(\left|\frac{\sum_\kappa \bar{W}_\kappa}{k_{mn}}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2} \frac{E[(\sum_\kappa \bar{W}_\kappa)^2]}{k_{mn}^2} = O\left(\frac{n^2 m k_{mn} p_{mn}}{k_{mn}^2}\right) = O\left(\frac{p_{mn}}{m}\right),$$

which goes to zero as $p_{mn} \in (0, 1)$ and $m \rightarrow \infty$. $\square$

Proof of Theorem 2.4. In the proof we first show asymptotic equivalence between the sample score vector and the projected score vector. We then show that the Hessian converges uniformly to its expectation. We then prove that the projected score vector obeys a central limit theorem. Finally, we tie it all together to show the stated result.

Throughout the proofs we will be conditioning on the information sets consisting of the following σ -algebras generated by the covariates and fixed effects $\mathcal{F}_{mn} = \sigma(\{\{x_{ia}\}_{mn}, \{\alpha_i\}_n, \{\psi_a\}_m, \lambda_0\})$ as well as $\mathcal{F} = \sigma(\cup_{m \in \mathbb{N}} \cup_{n \geq m} \mathcal{F}_{mn})$.

Furthermore, for the score vector, $\hat{G}_{mn}$ and the projected score vector, V_{mn} , we will not make explicit their dependence on the true parameter vector β_0 , when this is redundant.

Note that from the identification argument we see

$$\begin{aligned} E[W_{1,ijab} \mid \mathcal{F}_{mn}] &= P(W_{1,ijab} = 1 \mid \mathcal{F}_{mn}) \\ &= F(\tilde{x}'_{ijab} \beta_0) P(W_{1,ijab} + W_{2,ijab} = 1 \mid \mathcal{F}_{mn}), \\ E[W_{2,ijab} \mid \mathcal{F}_{mn}] &= P(W_{2,ijab} = 1 \mid \mathcal{F}_{mn}) \\ &= F(-\tilde{x}'_{ijab} \beta_0) P(W_{1,ijab} + W_{2,ijab} = 1 \mid \mathcal{F}_{mn}). \end{aligned}$$

Therefore, since the derivative of the likelihood contributions is given by

$$g_{ijab}(\beta_0) = W_{1,ijab}F(-\tilde{x}'_{ijab}\beta_0)\tilde{x}'_{ijab} - W_{2,ijab}F(\tilde{x}'_{ijab}\beta_0)\tilde{x}'_{ijab},$$

it follows that

$$\begin{aligned} E[g_{ijab}(\beta_0) \mid \mathcal{F}_{mn}] \\ &= E[W_{1,ijab} \mid \mathcal{F}_{mn}]F(-\tilde{x}'_{ijab}\beta_0)\tilde{x}'_{ijab} - E[W_{2,ijab} \mid \mathcal{F}_{mn}]F(\tilde{x}'_{ijab}\beta_0)\tilde{x}'_{ijab} \\ &= 0 \end{aligned}$$

almost surely, and therefore $E[g_{ijab}(\beta_0)] = 0$ by the law of iterated expectations. Furthermore, note that since $T_{ia} \perp\!\!\!\perp T_{jb}$ conditional on $\mathcal{F}_{mn}$ whenever $i \neq j$ or $a \neq b$, the covariance of the score contribution associated with two quadruples is non-zero only if they share a dyad, not just a single agent. That is

$$\begin{aligned} E[g_{ijab}(\beta_0)g_{ikcd}(\beta_0)] &= E[E[g_{ijab}(\beta_0)g_{ikcd}(\beta_0) \mid \mathcal{F}_{mn}]] \\ &= E[E[g_{ijab}(\beta_0) \mid \mathcal{F}_{mn}]E[g_{ikcd}(\beta_0) \mid \mathcal{F}_{mn}]] \\ &= 0 \end{aligned}$$

if $j \neq k, a \neq c$, and $b \neq d$.

Therefore, we can write the variance of the sample score as

$$E[\hat{G}_{mn}\hat{G}'_{mn}] = \sum_{i=1}^n \sum_{a=1}^m \sum_{j \neq i} \sum_{k \neq i} \sum_{b \neq a} \sum_{c \neq a} E[g_{ijab}(\beta_0)g_{ikac}(\beta_0)'],$$

and in particular we note that $E[\hat{G}_{mn}\hat{G}'_{mn}] = E[\hat{\Gamma}_{mn}]$.

Recall the projected score vector, V_{mn} , is constructed by projecting the score vector onto the information set and each dyad's duration

$$V_{mn} = \sum_{i=1}^n \sum_{a=1}^m \underbrace{\sum_{j \neq i} \sum_{b \neq a} E[g_{ijab}(\beta_0) \mid T_{ia}, \mathcal{F}_{mn}]}_{v_{ia}} = \sum_{i=1}^n \sum_{a=1}^m v_{ia}.$$

Hence, $E[v_{ia}] = 0$ and $E[V_{mn}] = 0$. Furthermore, note well, that

$$E[v_{ia}v'_{jb}] = E[E[v_{ia} \mid \mathcal{F}_{mn}]E[v'_{jb} \mid \mathcal{F}_{mn}]] = 0,$$

whenever $i \neq j$ or $a \neq b$. Hence, consider the matrix

$$\begin{aligned} \Gamma_{mn} &= E[V_{mn}V'_{mn}] \\ &= \sum_{i=1}^n \sum_{a=1}^m E[v_{ia}v'_{ia}] \\ &= \sum_{i=1}^n \sum_{a=1}^m \sum_{j \neq i} \sum_{k \neq i} \sum_{b \neq a} \sum_{c \neq a} E[E[g_{ijab}(\beta_0) \mid T_{ia}, \mathcal{F}_{mn}]E[g_{ikac}(\beta_0)' \mid T_{ia}, \mathcal{F}_{mn}]] \\ &= E[\hat{G}_{mn}\hat{G}'_{mn}] + o_p(n^3m^3), \end{aligned}$$

where the last equality stems from the fact that when the two quadruples share exactly one dyad, i.e. $k \neq j$ and $c \neq b$, we get

$$\begin{aligned} E [E [g_{ijab}(\beta_0) \mid T_{ia}, \mathcal{F}_{mn}] E [g_{ikac}(\beta_0)' \mid T_{ia}, \mathcal{F}_{mn}]] \\ = E [E [g_{ijab}(\beta_0) g_{ikac}(\beta_0)' \mid T_{ia}, \mathcal{F}_{mn}]] \\ = E [g_{ijab}(\beta_0) g_{ikac}(\beta_0)'] , \end{aligned}$$

which is exactly the same as the contribution they would make to $E[\hat{G}_{mn} \hat{G}_{mn}']$, and there are $o(n^3 m^3)$ terms with more than one dyad in common (see subsection 2.B), where the contribution of these are each $O_p(1)$ by Assumption 2.3.

Asymptotic Equivalence Between Score and Projection

We will show asymptotic equivalence between V_{mn} and $\hat{G}_{mn}$, that is

$$\lim_{m,n \rightarrow \infty} \Gamma_{mn}^{-1/2} E \left[(V_{mn} - \hat{G}_{mn})(V_{mn} - \hat{G}_{mn})' \right] \Gamma_{mn}^{-1/2} = 0.$$

Note that $E[(V_{mn} - \hat{G}_{mn})(V_{mn} - \hat{G}_{mn})'] = E[V_{mn} V_{mn}' - \hat{G}_{mn} V_{mn}' - V_{mn} \hat{G}_{mn}' + \hat{G}_{mn} \hat{G}_{mn}']$.

From the preceding calculations, we get

$$\lim_{m,n \rightarrow \infty} \Gamma_{mn}^{-1/2} E[\hat{G}_{mn} \hat{G}_{mn}'] \Gamma_{mn}^{-1/2} = I + o(1),$$

and similar calculation shows the same is true for $E[V_{mn} \hat{G}_{mn}']$ proving the asymptotic equivalence.

Uniform Convergence of Hessian

We next prove uniform convergence of the sample Hessian to its expectation, and an analogous argument then gives uniform convergence of $\hat{\Gamma}_{mn}(\hat{\beta})$ to Γ_{mn} .

We want to show that

$$\sup_{\beta \in \Theta} \left\| \frac{H_{mn}(\beta)}{k_{mn}^*} - \frac{E[H_{mn}(\beta)]}{E[k_{mn}^*]} \right\|_F \xrightarrow{p} 0$$

as $m, n \rightarrow \infty$. Note that, from the proof of consistency, $k_{mn}^*/k_{mn} - p_{mn} \xrightarrow{p} 0$, so we can rewrite the expression as

$$\frac{\sup_{\beta} \|H_{mn}(\beta) - E[H_{mn}(\beta)]\|_F}{k_{mn}p_{mn}} \xrightarrow{p} 0,$$

where the result follows from verifying the conditions of Lemma 2.9 in Newey and McFadden (1994). We first show that $\|H_{mn}(\beta_1) - H_{mn}(\beta_2)\|_F = O_p(1)\|\beta_1 - \beta_2\|$ for all β_1, β_2 thereby proving stochastic equicontinuity of the Hessian matrix. Using the mean value theorem and the fact that $|F(z)| \leq 1$, we get

$$\begin{aligned} & \frac{1}{k_{mn}p_{mn}} \|H_{mn}(\beta_1) - H_{mn}(\beta_2)\|_F \\ & \leq \frac{1}{k_{mn}p_{mn}} \sum_{\tau} |F(\tilde{x}'_{\tau}\beta_1)(1 - F(\tilde{x}'_{\tau}\beta_1)) - F(\tilde{x}'_{\tau}\beta_2)(1 - F(\tilde{x}'_{\tau}\beta_2))| \|\tilde{x}_{\tau}\|^2 \\ & \leq \|\beta_1 - \beta_2\| \frac{1}{k_{mn}p_{mn}} \sum_{\tau} \|\tilde{x}_{\tau}\|^3 \\ & = O_p(1)\|\beta_1 - \beta_2\|. \end{aligned}$$

The first inequality follows from the compatibility of the Frobenius norm with the Euclidean norm and the second inequality follows from the fact that the derivative of $F(z)(1 - F(z))$ evaluated at its supremum is less than 1 in absolute value. The final equality follows from Assumption 2.5 after applying Chebyshev's inequality.

The other condition we need to verify, point-wise convergence of the Hessian to its expectation, can be established using similar arguments as theorem 1, by applying Chebyshev's inequality to

$$\frac{1}{k_{mn}p_{mn}} \|H_{mn}(\beta) - E[H_{mn}(\beta)]\|_F,$$

and noting that the variance of this is less than

$$\frac{1}{k_{mn}p_{mn}} \sum_{\tau} \sum_{\tau'} \mathbb{1}\{\tau \cap \tau' \neq \emptyset\} \sqrt{E[\|\tilde{x}_{\tau}\|^4]E[\|\tilde{x}_{\tau'}\|^4]} = O\left(\frac{1}{mp_{mn}}\right),$$

which converges to zero by Assumption 2.4.

Asymptotic Distribution of the Projected Score

We want to show that $(\Gamma_{mn}^{\mathcal{F}})^{-1/2}V_{mn} \xrightarrow{d} N(0, I)$. To do so, we rely on a conditional Central Limit Theorem. Here we notably depart from Jochmans (2018), as the conditional Central Limit Theorem cited in their paper assumes identical conditional distribution which does not hold. Instead we rely on the conditional Central Limit Theorem in Bulinski (2017), which applies to conditionally independent non-identically distributed random variables. We have to show that the $\mathcal{F}$ -Lindeberg condition in the paper is satisfied.

We will be using a Cramer-Wold device. Let $\eta \in \mathbb{R}^p \setminus \{0\}$. Define $\tilde{v}_{ia} = \eta'v_{ia}$ and $\tilde{\Gamma}_{mn}^{\mathcal{F}} = \sum_{\tau} E[\tilde{v}_{\tau}^2 | \mathcal{F}] = \sum_{\tau} \eta' E[v_{\tau}v'_{\tau} | \mathcal{F}] \eta = \eta' \Gamma_{mn}^{\mathcal{F}} \eta$.

Then, since $E[\tilde{v}_{ia} \mid \mathcal{F}] = 0$, we need for all $\varepsilon > 0$

$$\frac{1}{\tilde{\Gamma}_{mn}^{\mathcal{F}}} \sum_{i=1}^n \sum_{a=1}^m E[\tilde{v}_{ia}^2 \mathbb{1}\{|\tilde{v}_{ia}| > \varepsilon \sqrt{\tilde{\Gamma}_{mn}^{\mathcal{F}}}\} \mid \mathcal{F}] \xrightarrow{p} 0, \quad (2.11)$$

as $m, n \rightarrow \infty$. Note that this condition is implied by an analogous $\mathcal{F}$ –Lyapunov condition, since for $\delta > 0$

$$\frac{1}{\tilde{\Gamma}_{mn}^{\mathcal{F}}} \sum_{i=1}^n \sum_{a=1}^m E[\tilde{v}_{ia}^2 \mathbb{1}\{|\tilde{v}_{ia}| > \varepsilon \sqrt{\tilde{\Gamma}_{mn}^{\mathcal{F}}}\} \mid \mathcal{F}] \leq \frac{1}{\varepsilon^\delta} \frac{1}{(\tilde{\Gamma}_{mn}^{\mathcal{F}})^{1+\delta/2}} \sum_{i=1}^n \sum_{a=1}^m E[|\tilde{v}_{ia}|^{2+\delta} \mid \mathcal{F}].$$

Therefore, (2.11) is satisfied if

$$\frac{1}{(\tilde{\Gamma}_{mn}^{\mathcal{F}})^{1+\delta/2}} \sum_{i=1}^n \sum_{a=1}^m E[|\tilde{v}_{ia}|^{2+\delta} \mid \mathcal{F}] \xrightarrow{p} 0, \quad (2.12)$$

as $m, n \rightarrow \infty$. Note therefore that for the numerator

$$\begin{aligned} \sum_{i=1}^n \sum_{a=1}^m E[|\tilde{v}_{ia}|^{2+\delta} \mid \mathcal{F}] &= \sum_{i=1}^n \sum_{a=1}^m E[|\tilde{v}_{ia}|^\delta \tilde{v}_{ia}^2 \mid \mathcal{F}] \\ &\leq \sum_{i=1}^n \sum_{a=1}^m \|\eta\|^\delta E[\|\tilde{v}_{ia}\|^\delta \tilde{v}_{ia}^2 \mid \mathcal{F}] \\ &\leq \|\eta\|^\delta m^\delta n^\delta \max_{\kappa} \|\tilde{x}_{\kappa}\|^\delta \sum_{i=1}^n \sum_{a=1}^m E[\tilde{v}_{ia}^2 \mid \mathcal{F}] \\ &= \|\eta\|^\delta m^\delta n^\delta \max_{\kappa} \|\tilde{x}_{\kappa}\|^\delta \tilde{\Gamma}_{mn}^{\mathcal{F}}, \end{aligned}$$

which implies that

$$\begin{aligned}
\frac{1}{(\tilde{\Gamma}_{mn}^{\mathcal{F}})^{1+\delta/2}} \sum_{i=1}^n \sum_{a=1}^m E[|\tilde{v}_{ia}|^{2+\delta} \mid \mathcal{F}] &\leq \frac{1}{(\tilde{\Gamma}_{mn}^{\mathcal{F}})^{\delta/2}} \|\eta\|^\delta m^\delta n^\delta \max_{\kappa} \|\tilde{x}_{\kappa}\|^\delta \\
&= \left(\frac{\|\eta\|^2 m^2 n^2 \max_{\kappa} \|\tilde{x}_{\kappa}\|^2}{\tilde{\Gamma}_{mn}^{\mathcal{F}}} \right)^{\delta/2} \\
&\leq \left(\frac{m^2 n^2 \max_{\kappa} \|\tilde{x}_{\kappa}\|^2}{\lambda_{\min}(\Gamma_{mn}^{\mathcal{F}})} \right)^{\delta/2},
\end{aligned}$$

where $\lambda_{\min}(\Gamma_{mn}^{\mathcal{F}})$ is the minimum eigenvalue of the covariance matrix for the projected score vector, which is of order strictly greater than $m^2 n^2$ by Assumption 2.6. Assumption 2.5 implies that $\max_{\kappa} \|\tilde{x}_{\kappa}\|^2 = O_p(1)$, so the right hand side goes to zero in probability. This proves that $(\Gamma_{mn}^{\mathcal{F}})^{-1/2} V_{mn} \xrightarrow{d} N(0, I)$.

Asymptotic Distribution of the Estimator

We have now established the asymptotic distribution of the projected score vector, V_{mn} . By its equivalence to the sample score vector, $\hat{G}_{mn}$, we can finally derive the asymptotic distribution of the estimator.

Note that since $E[\Gamma_{mn}^{\mathcal{F}}] = \Gamma_{mn}$, we have $\|\Gamma_{mn}^{\mathcal{F}} - \Gamma_{mn}\|_F \xrightarrow{p} 0$ under Assumption 2.5. By the previously established convergence of $\hat{\Gamma}_{mn}$ to Γ_{mn} the continuous mapping theorem yields

$$\hat{\Gamma}_{mn}^{-1/2} \hat{G}_{mn}(\beta_0) \xrightarrow{d} N(0, I).$$

Hence

$$0 = \hat{G}_{mn}(\hat{\beta}) = \hat{G}_{mn}(\beta_0) + \hat{H}_{mn}(\beta_*) (\hat{\beta} - \beta_0),$$

where β_* is a vector that is coordinate wise between $\hat{\beta}$ and β_0 . Therefore

$$(\hat{\beta} - \beta_0) = -\hat{H}_{mn}(\beta_*)^{-1} \hat{G}_{mn}(\beta_0) = -(\hat{H}_{mn}^{-1}(\beta_*) \hat{\Gamma}_{mn}^{1/2}) \hat{\Gamma}_{mn}^{-1/2} \hat{G}_{mn}(\beta_0),$$

where $\hat{H}_{mn}(\beta_*)^{-1} \hat{\Gamma}_{mn}^{1/2}$ converges to $\hat{\Omega}_{mn}(\hat{\beta})^{-1/2}$ and the last term converges to a standard Gaussian. Combining these, we get

$$\hat{\Omega}_{mn}(\hat{\beta})^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, I),$$

as desired. □

Chapter 3

Binomial Models With Fixed Effects*

In this paper, we consider panel data models where the outcome in a given period is the sum of binary outcomes. In the simplest framework, the explanatory variables are identical across the outcomes, and these outcomes are independent conditional on a common fixed effect and those explanatory variables. In this case, the observed outcome is binomially distributed, and one can consider our model to be the simplest possible extension of the dynamic binary panel data model studied by Heckman (1978, 1981a, 1981b, 1981c) and many others since him. We extend results from the binary outcome panel data literature to this setting. Afterwards, we let the explanatory variables and fixed effects differ between outcomes and show that some results carry over. Finally, we consider letting the outcomes depend on each other in a “strategic way.” This allows us to make a connection to simple econometric models for entry.

This paper is part of a sizable literature that is concerned with estimating the common parameters of nonlinear panel data models with individual specific fixed effects. Within

*This chapter is coauthored with Bo E. Honoré.

this literature, the paper is most closely connected to models with binary outcomes, and it is this literature that we wish to expand. Research in this area dates back to (at least) Rasch (1960) and Rasch (1961), who consider static panel data logit models with fixed effects and strictly exogenous explanatory variables. In this case, it is possible to find a sufficient statistic for the fixed effect, and the likelihood function that conditions on this sufficient statistic can be used to estimate the common parameters of the model with as little as two time periods. This result has been generalized in a number of ways. Manski (1987) showed that the conditioning argument used in Rasch (1960) can be used to construct consistent maximum score estimators of the common parameters in static “threshold crossing” binary response models for which the underlying (time-invariant) error distribution is unknown. The resulting estimator converges at the $n^{-1/3}$ and Chamberlain (1993) and Chamberlain (2010) showed that the regular root- n consistent estimation is generally not possible with two time periods outside the logit model even if one knows the error distribution.¹

The literature has also considered dynamic versions of the fixed effects binary panel data logit model. For example, Chamberlain (1978) derived the sufficient statistic for the fixed effect in models with lagged dependent variables, but no other explanatory variables (see also Cox, 1958; Magnac, 2000; d’Addio & Honoré, 2010). Using the conditional likelihood to estimate the common parameters of this model requires additional time periods. For example, models with one lag requires a total of four time periods (the first of which provides an initial condition, for which the logit model does not need to apply), and a model with two lags requires six periods (of which the first two provide the initial con-

¹Building on insight from Horowitz (1992), Kyriazidou (1997) and Charlier et al. (1995) showed that one can achieve rates of convergence close to root- n if the underlying error distribution is sufficiently smooth.

ditions). Honoré and Kyriazidou (2000) and d’Addio and Honoré (2010) noted that it is sometimes possible to extend the logic in the papers above to settings where the explanatory variables include strictly exogenous variables as well as lagged dependent variables. However, the usefulness of this approach is limited by the fact that they need to condition on particular sequences of the vector of strictly exogenous explanatory variables, which will lead to estimators that converge at rate slower than $\text{root-}n$ when the explanatory variables are continuous.² Inspired by this, Kitazawa (2022), Honoré and Weidner (2024), and Dano (2023) constructed moment conditions that must be satisfied at the true parameters and discussed how these might be used for estimation. See Kruiniger (2023) and Dobronyi et al. (2024) for additional results about moment conditions in dynamic logit models.³

The literature has a long history of also using either conditional likelihood or moment conditions to estimate the common parameters in fixed effects versions of other “textbook” nonlinear models. For example, Hausman et al. (1984) considered conditional likelihood estimation of panel data versions of Poisson regression models, while Honoré (1992, 1993) and Hu (2002) constructed estimators for static and dynamic censored and truncated regression models based on moment functions, and Chamberlain (1992) and Wooldridge (1997) use moment conditions in a class of static and dynamic multiplicative models. Kyriazidou (1997, 2001) combine the two approaches to construct estimators of static and dynamic sample selection models.

²Ouyang and Yang (2024) note that the argument in Honoré and Kyriazidou (2000) and d’Addio and Honoré (2010) can be modified to allow one to condition on the combined effect of the explanatory variables rather than on the full vector. This can improve the rate of convergence of the estimator.

³The paper by Khan et al. (2023) discusses identification in a version of the model that relaxes the logit assumption. We will not consider generalizations of this since the results in Chamberlain (1993) and Chamberlain (2010) suggest that one will generally not be able to achieve regular $\text{root-}n$ consistent estimation outside the logit model.

The rest of the paper is organized as follows. In Section 3.1 we consider the simple extension of results from the binary outcome panel data literature to the case where the outcome is the sum of independent and identically distributed Bernoulli random variables, yielding a binomial distribution for the outcome. We prove that sufficient statistics analogous to those found in the binary outcome literature exist for the fixed effects, rendering estimation of the coefficients on the explanatory variables possible through a conditional likelihood approach, when the model is either static or when the only explanatory variable is a lagged dependent variable. In the cases where a conditional likelihood approach is not possible, we find moment conditions making it possible to nonetheless estimate the coefficients on the explanatory variables. The subsequent sections considers extensions to the baseline model. Specifically, in Section 3.2 we allow the coefficient on the lagged dependent variable in an AR model to be non-linear, through the use of indicator functions. We find moment conditions for this model. In Section 3.3 we allow the Bernoulli random variables the outcome is summed over to have non-identical distribution. We show that for the autoregressive version of this model, the sufficient statistics from 3.1 carry over. We further find moment conditions for the static version of the model, and using numerical calculations we conjecture the existence of moments for a model with both exogenous and lagged dependent variables. Section 3.4 discusses the relation between the models considered in this paper and of models with strategic interactions, such as is found in a simple static entry game. Using numerical procedures, we establish conjectures about the presence of moment conditions under different parameter restrictions. Section 3.5 concludes.

3.1 Simplest Model, Propositions, and Conjectures

In this section, we consider simple extensions of the results for the binary outcome logit model to the binomial distributions. In subsection 3.1.1, we review and extend results regarding conditional likelihood estimation when there is (at most) one lag. The main conclusion from this section is that the results for binary panel data models extend to the binomial case, namely that a conditional likelihood approach is possible when the model is static or when it is autoregressive with no other explanatory variables. In subsection 3.1.2, we then turn to conditional likelihood estimation of the common parameters in binomial fixed effects logit models with additional lags as the only explanatory variables. The main conclusion from this section is that only a subset of the common parameters are identified using this approach, as is the case in the binary models. Subsection 3.1.3 describes the general approach to constructing moment functions that can be used to estimate the common parameters in panel data models with individual specific fixed effects when the outcome is discrete. This approach is then used to derive moment conditions for a fixed effects binomial logit model with strictly exogenous explanatory variables and lagged dependent variables. Finally, subsection 3.1.4 continues the discussion of moment conditions by developing moment conditions for a pure AR(2) binomial fixed effects logit model, highlighting some of the differences between conditional likelihood and moment-based estimation.

3.1.1 Conditional Likelihood Estimation of Models With No or One Lag

We consider a panel data model, where the outcome in each period is the sum of M Bernoulli random variables that are independent and identically distributed conditional on common explanatory variables and fixed effects.

The distribution of $Y = (Y_1, Y_2, \dots, Y_T)$ conditional on the explanatory variables $X = (X_1, X_2, \dots, X_T)$, the initial condition Y_0 , and a fixed effect A is characterized by

$$Y_t | (\{Y_s\}_{s < t}, X, A) \sim \text{binom} \left(M, \frac{\exp(X'_t \beta + Y_{t-1} \gamma + A)}{1 + \exp(X'_t \beta + Y_{t-1} \gamma + A)} \right), \quad (3.1)$$

where the distribution of (Y_0, A) given X is left unspecified. The econometrician observes (Y_t, X_t) for $t = 1, \dots, T$ as well as the initial condition Y_0 .

When $M = 1$, this model corresponds to the dynamic logit model studied by, for example, Honoré and Kyriazidou (2000), Kitazawa (2022), Honoré and Weidner (2024), and Dano (2023). The propositions and conjectures below demonstrate that the results from this special case generalize to the setting with $M > 1$.

Propositions 3.1 and 3.2 below generalize Rasch (1960)'s and Chamberlain (1978) for binary logit outcomes to the binomial model in (3.1). Proposition 3.1 below is concerned with the static case in which the coefficient on Y_{t-1} , γ , is restricted to be 0, in which case the minimal sufficient statistic is identical to that in Rasch (1960).

Proposition 3.1. *Consider the special case of (3.1) in which $\gamma = 0$. The statistic*

$$\sum_{t=1}^T y_t$$

is minimally sufficient for A .

Proposition 3.2 below is concerned with the pure autoregressive model without other explanatory variables. In other words, β is restricted to be 0.

Proposition 3.2. *Consider the special case of (3.1) in which $\beta = 0$. The statistic*

$$\left(Y_0, \sum_{t=1}^T 1 \{Y_t = 0\}, \sum_{t=1}^T 1 \{Y_t = 1\}, \sum_{t=1}^T 1 \{Y_t = 2\}, \dots, \sum_{t=1}^T 1 \{Y_t = M\}, Y_T \right)$$

is minimally sufficient for A .

Note that in the case where $M = 1$, so there is a binary outcome, the statistic in Proposition 3.2 collapses to $(Y_0, \sum_{t=1}^T Y_t, Y_T)$, which is the well-known sufficient statistic for the fixed effect in an AR(1) binary logit model.

The following proposition establishes that it is generally not possible to estimate the parameters β and γ in (3.1) by conditional likelihood. Specifically, the proposition says that, generically, the only sufficient statistic for A is the sequence of all the Y 's.

Proposition 3.3. *Consider (3.1) and let S be a function of $(Y_0, Y_1, \dots, Y_T)$ such that for all $(X'_0\beta, X'_1\beta, \dots, X'_T\beta)$ in some open ball, the distribution of $(Y_0, Y_1, \dots, Y_T)$ conditional on $(S, X_0, X_1, \dots, X_T, A)$ does not depend on A . Then if $\gamma \neq 0$, $S = (Y_0, Y_1, \dots, Y_T)$.*

Proposition 3.3 generalizes the common knowledge that no (non-trivial) sufficient statistic exists for the fixed effects in a panel data logit model with lagged dependent

variables and strictly exogenous explanatory variables, if these have a non-discrete distribution.⁴ In subsection 3.1.3 we outline a procedure to find moment functions that nevertheless allow us to estimate both β and γ .

3.1.2 Pure AR Models With Additional Lags

In this subsection we consider an autoregressive model with a greater number of lags. It is possible to find a sufficient statistic for the fixed effects, but, as is the case in the binary panel data literature, only the coefficient on the last lag term remains and is estimable via a conditional likelihood approach.

Consider the binary outcome model,

$$P(Y_t = 1 \mid \{Y_s\}_{s < t}, A) = \frac{\exp(Y_{t-1}\gamma_1 + Y_{it-2}\gamma_2 + A)}{1 + \exp(Y_{t-1}\gamma_1 + Y_{it-2}\gamma_2 + A)}. \quad (3.2)$$

Let $S_1 = \sum_{t=1}^T Y_t$ and $S_{11} = \sum_{t=1}^T Y_t Y_{t-1}$. Chamberlain (1984, 1985) showed that conditional on $(Y_{-1}, Y_0, S_1, S_{11}, Y_{T-1}, Y_T)$, the distribution of $(Y_1, \dots, Y_T)$ does not depend on (A, γ_1) , but it does depend on γ_2 provided that there are four time periods in addition to the two that deliver the initial conditions. This implies that one can estimate γ_2 on a panel data set with at least six time periods, even if A and γ_1 are individual specific. On the other hand, the conditional likelihood approach does not identify γ_1 even if one has a panel of length larger than six.

Models like (3.2) can be useful for studying discrete time duration models. For example, Y_t could be an indicator for whether an individual is employed in period t , and

⁴We refer to this as “common knowledge,” because we are not aware of a formal statement to this effect.

(3.2) would then be a model for the transition in and out of employment. When γ_2 is 0 and A and γ_1 are individual specific, (3.2) describes a Markov chain (with states $Y = 0$ and $Y = 1$) in which the transition probabilities are unrestricted across individuals. On the other hand, $\gamma_2 \neq 0$ implies duration dependence in each of the two states. In other words, (3.2) constrains the duration dependence in the state $Y = 0$ to be similar to the state dependence in the state $Y = 1$, because both depend on the same parameter γ_2 . This motivates a generalization of (3.2) in which γ_2 depends on Y_{t-1} :

$$P(Y_t = 1 \mid \{Y_s\}_{s < t}, A) = \frac{\exp(Y_{t-1}\gamma_1 + Y_{it-2}\gamma_{2,Y_{t-1}} + A)}{1 + \exp(Y_{t-1}\gamma_1 + Y_{it-2}\gamma_{2,Y_{t-1}} + A)}.$$

Magnac (2000) show that conditional on $(Y_{-1}, Y_0, S_1, S_{11}, Y_{T-1}, Y_T)$, the distribution of $(Y_1, \dots, Y_T)$ does not depend on (A, γ_1) , but it does depend on $(\gamma_{2,0}, \gamma_{2,1})$ (when $T \geq 4$). In other words, one can allow (A, γ_1) to be individual specific and estimate coefficients on Y_{t-2} that depends on Y_{t-1} . This allows one to consider state dependence in the state $Y = 0$ separately from state dependence in the state $Y = 1$.

If one is only interested in state dependence in one of the two states, say $Y = 0$, then it might be desirable to allow γ_2 for the other state to be individual specific. This motivates a model in which $\gamma_{2,1}$ is also individual-specific. d'Addio and Honoré (2010) showed that in this case, we would need to also condition on

$$S_{111} = \sum_{t=3}^T Y_t Y_{t-1} Y_{t-2}$$

in order to get an expression that depends on $\gamma_{2,0}$ but not on the individual-specific effects.

Many of the ideas in the discussion above carry over to binomial versions of the binary outcome models above. Consider, for example, a generalization of (3.2) with J lags rather than 2:

$$Y_t \mid (\{Y_s\}_{s < t}, A) \sim \text{binom} \left(M, \frac{\exp \left(\sum_{j=1}^J Y_{t-j} \gamma_j + A \right)}{1 + \exp \left(\sum_{j=1}^J Y_{t-j} \gamma_j + A \right)} \right), \quad (3.3)$$

It is useful to define $\gamma = (\gamma_1, \dots, \gamma_J)'$ and $Y_{t-} = (Y_{t-1}, Y_{t-2}, \dots, Y_{t-J})'$, which allows one to write $\sum_{j=1}^J Y_{t-j} \gamma_j = Y_{t-}' \gamma$.

Proposition 3.4. *Let L_k (for $k = 1, \dots, (M+1)^J$), denote the J -tuples of integers between 0 and M . Then*

$$S = \left(Y_{1-j}, Y_{2-j}, \dots, Y_0, \sum_{t=1}^T Y_t, \sum_{t=1}^T 1 \{Y_{t-} = L_1\}, \sum_{t=1}^T 1 \{Y_{t-} = L_2\}, \dots, \sum_{t=1}^T 1 \{Y_{t-} = L_{(M+1)^J}\} \right)$$

is sufficient for A and the distribution of Y given S is

$$p(Y = y \mid S) = \frac{\prod_{t=1}^T \binom{M}{y_t} \exp(y_{t-}' \gamma)^{y_t}}{\sum_{w \in \mathcal{W}} \prod_{t=1}^T \binom{M}{w_t} \exp(w_{t-}' \gamma)^{w_t}} = \frac{\prod_{t=1}^T \binom{M}{y_t} \exp(y_{t-J} y_t \gamma_J)}{\sum_{w \in \mathcal{W}} \prod_{t=1}^T \binom{M}{w_t} \exp(w_{t-J} w_t \gamma_J)}, \quad (3.4)$$

where

$$\mathcal{W} = \left\{ (w_1, w_2, \dots, w_T)' : \sum_{t=1}^T w_t = \sum_{t=1}^T y_t, \right. \\ \left. \sum_{t=1}^T 1 \{w_{t-} = L_k\} = \sum_{t=1}^T 1 \{y_{t-} = L_k\} \text{ for } k = 1, \dots, (M+1)^J \right\}.$$

As is the case for the binary outcome model, the conditional likelihood based on (3.4) only identifies the coefficient on the J^{th} lag of Y . In subsection 3.1.3 we discuss the procedure for finding moment conditions that render identification (and estimation) of the other lag coefficients possible.

3.1.3 Moment Conditions

Proposition 3.3 established that it is not possible to estimate the parameters of (3.3) by conditional likelihood after conditioning on a sufficient statistic. In the same spirit, Proposition 3.4 makes it clear that the conditional likelihood approach will only identify the parameter of the last lag, γ_J in an $\text{AR}(J)$ model. Below, we discuss how it is nonetheless possible to construct moment conditions that do not depend on the fixed effects, A , but are informative about the common parameters. This generalizes the finding in Kitazawa (2022), Honoré and Weidner (2024), and Dano (2023) for the binary logit model to binomials.

The moments are obtained through a procedure similar to the one found in Honoré and Weidner (2024), by recognizing that since the moment is a function of the parameter that is equal to zero at the true value and since the outcome variable is discrete, it amounts to

finding a vector orthogonal to the probability vector of the outcome sequence at the true parameter values.

Concretely, consider model the model in equation (3.1). Let $\theta = (\beta, \gamma)$ and let θ_0 denote the true parameter value. If $M = 2$ and $T = 3$, we can then write the probability of the different outcomes as a vector given by

$$p(X, Y_0, A; \theta_0) = \begin{pmatrix} P((Y_1, Y_2, Y_3) = (0, 0, 0) \mid Y_0, X, A; \theta_0) \\ P((Y_1, Y_2, Y_3) = (0, 0, 1) \mid Y_0, X, A; \theta_0) \\ \vdots \\ P((Y_1, Y_2, Y_3) = (2, 2, 2) \mid Y_0, X, A; \theta_0), \end{pmatrix}$$

and hence the expected value of a function of the data and parameters, $m(Y, X, Y_0, A; \theta)$, can be written as

$$E[m(Y, X, Y_0, A; \theta)] = m(Y, X, Y_0, A; \theta)' p(X, Y_0, A; \theta_0).$$

A valid moment function must satisfy that it is orthogonal to the probability vector when evaluated at θ_0 . Furthermore, the moment vector is only useful if we can evaluate it based only on observable data, so it should not depend on A .

Such a moment function would therefore reside in the null space of the matrix whose rows consist of (the transposes of) the probability vectors above, where each row differs only in the value of A . Finding elements in the null space of a symbolic matrix is a problem that computer algebraic systems such as Mathematica is in principle well-suited for. In reality, the problem scales poorly as there are $(M + 1)^T$ outcome sequences, so it

becomes too computationally expensive to find the elements in the null space of the full matrix symbolically.

The dimensionality of the null space, which coincides with the number of potential moment functions, can be evaluated numerically at much lower cost. Inputting numerical values for the data, parameters, and the fixed effects, one can calculate the rank of the matrix and the difference between the number of columns and the rank is equal to the dimensionality of the null space. While this exercise will not provide a definite proof of whether it is possible to construct moment functions, it will provide a strong indication. Moreover, if more than one moment exists, any linear combinations of them will also be a moment function. This then implies that one may be able to construct moment functions that do not use all the possible outcomes (all the possible columns of the matrix).

Our aim is therefore to find a sparse vector in the nullspace. This is equivalent to finding a set of k outcomes, with k small, where the corresponding sub-matrix of the probability matrix is rank-deficient. We can then use Mathematica to find the symbolic expression for the nullspace vector and verify that it is indeed a moment.

Even though numerically evaluating the nullity is a lot less computationally costly, care is still required as there are $(M + 1)^T$ choose k sub-matrices of k columns. Algorithms exist for finding sparse vectors in a linear subspace under certain conditions (see, e.g., Qu et al., 2014), and while the conditions that guarantee success are not satisfied the algorithm is reasonably fast to run and the output is verifiable. In the appendix we detail our alternative numerical procedure allowing us to evaluate the rank deficiency of sub-matrices of k columns quickly. The main insight is that many of the sub-matrices share everything but one column, projecting the column differentiating them onto the shared

columns is informative about rank-deficiency and calculating the projection matrix for the shared columns need only be done once.

When $M = 2$ and $T = 3$, the nullity of the numerical probability matrix for the model in equation (3.1), with $\beta \neq 0$ and $\gamma \neq 0$, is equal to 12, suggesting that one can construct 12 moment conditions. It turns out that eight of them can be expressed in terms of seven of the 27 outcomes. Appendix 3.A gives the expression for these.

3.1.4 Moment Conditions for Pure AR Models

Proposition 3.4 established that although one can find a sufficient statistic for the fixed effect in a pure $\text{AR}(J)$ binomial distribution, the resulting conditional likelihood function only depends on the coefficient on the J^{th} lag, so it cannot be used to estimate the other common parameters. The proposition also implies that in a model with two lags, one needs a total of six time periods (two of which provide the initial conditions) to identify γ_2 from a conditional likelihood approach.

In this subsection we show that despite this it is possible to construct moment conditions that identify both γ_1 and γ_2 in an $\text{AR}(2)$ binomial model with $M = 2$, even if one only has a total of five time periods.

In order to make progress, we first establish numerically that for each set of initial values of Y , (y_{-1}, y_0) , there are four moment conditions, and that when $y_{-1} \neq y_0$ they all depend on both γ_1 and γ_2 . When $y_{-1} = y_0$, the moment functions only depend on γ_1 . We also establish numerically that two of the four moment conditions for a set of initial conditions are especially simple and share a somewhat similar structure across the initial conditions. These moment functions are given below.

For $y_0 = 0$ and $y_{-1} = \ell$ (for $\ell = 0, 1, 2$), and with $y = (y_1, y_2, y_3)$, we have the moment functions

$$m_{\ell 0}^a(y; \gamma_1, \gamma_2) = \begin{cases} -\exp(\ell \gamma_2), & y = (0, 1, 0) \\ -\exp(-\gamma_1 + \ell \gamma_2), & y = (0, 1, 1) \\ -\exp(-2\gamma_1 + \ell \gamma_2), & y = (0, 1, 2) \\ 1, & y = (1, 0, \cdot) \end{cases}$$

and

$$m_{\ell 0}^b(y; \gamma_1, \gamma_2) = \begin{cases} -\exp(2\ell \gamma_2), & y = (0, 2, 0) \\ -\exp(-2\gamma_1 + 2\ell \gamma_2), & y = (0, 2, 1) \\ -\exp(-4\gamma_1 + 2\ell \gamma_2), & y = (0, 2, 2) \\ 1, & y = (2, 0, \cdot) \end{cases}.$$

For $y_0 = 1$ and $y_{-1} = \ell$ (for $\ell = 0, 1, 2$), we have the moment functions

$$m_{\ell 1}^a(y; \gamma_1, \gamma_2) = \begin{cases} -\exp(-\gamma_1 - (\ell - 1) \gamma_2) & y = (1, 0, 0) \\ -\exp(-(\ell - 1) \gamma_2) & y = (1, 0, 1) \\ -\exp(\gamma_1 - (\ell - 1) \gamma_2) & y = (1, 0, 2) \\ 1 & y = (0, 1, \cdot) \end{cases}$$

and

$$m_{\ell 1}^b(y; \gamma_1, \gamma_2) = \begin{cases} -\exp(\gamma_1 + (\ell - 1)\gamma_2) & y = (1, 2, 0) \\ -\exp((\ell - 1)\gamma_2) & y = (1, 2, 1) \\ -\exp(-\gamma_1 + (\ell - 1)\gamma_2) & y = (1, 2, 2) \\ 1 & y = (2, 1, \cdot) \end{cases}.$$

For $y_0 = 2$ and $y_{-1} = \ell$ (for $\ell = 0, 1, 2$), we have the moment functions

$$m_{\ell 2}^a(y; \gamma_1, \gamma_2) = \begin{cases} -\exp(-4\gamma_1 - 2(\ell - 2)\gamma_2) & y = (2, 0, 0) \\ -\exp(-2\gamma_1 - 2(\ell - 2)\gamma_2) & y = (2, 0, 1) \\ -\exp(-2(\ell - 2)\gamma_2) & y = (2, 0, 2) \\ 1 & y = (0, 2, \cdot) \end{cases}$$

and

$$m_{\ell 2}^b(y; \gamma_1, \gamma_2) = \begin{cases} -\exp(-2\gamma_1 - (\ell - 2)\gamma_2) & y = (2, 1, 0) \\ -\exp(-\gamma_1 - (\ell - 2)\gamma_2) & y = (2, 1, 1) \\ -\exp(-(\ell - 2)\gamma_2) & y = (2, 1, 2) \\ 1 & y = (1, 2, \cdot) \end{cases}.$$

When $y_0 = y_{-1} = \ell$ for $\ell = 0, 2$ the moment functions above are monotone in γ_1 and do not depend on γ_2 . The corresponding moment conditions therefore identify the γ_1 even though that parameter would not be identified from a conditional likelihood approach that uses one additional time period. Moreover, the moment functions above that do depend on γ_2 are monotone in that parameter, so they identify γ_2 (once γ_1 has been identified).

This is similar to what happens in the binary case, but an interesting difference is that the binomial case also gives moment conditions that are not monotone in γ_1 .

3.2 Indicator AR Model

In this section we show that it is possible to find moment conditions for alternative specifications for the autoregressive binomial logit model. Consider, for example, a model with $M = 2$ and two lags in which the terms $Y_{t-1}\gamma_1$ and $Y_{t-2}\gamma_2$ are replaced with $1\{Y_{t-1} = 1\}\gamma_{11} + 1\{Y_{t-1} = 2\}\gamma_{12}$ and $1\{Y_{t-2} = 1\}\gamma_{21} + 1\{Y_{t-2} = 2\}\gamma_{22}$, respectively. For $M = 2$, $J = 2$, and a total of five time periods, there are again moment functions for each set of initial conditions, and a subset of these moment conditions identify the parameters of the model. Specifically, the moment function include

$$m_{00}(y; \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}) = \begin{cases} -1 & y = (0, 1, 0) \\ -\exp(-\gamma_{11}) & y = (0, 1, 1) \\ -\exp(-2\gamma_{11}) & y = (0, 1, 2) \\ 1 & y = (1, 0, \cdot) \end{cases},$$

$$m_{02}(y; \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}) = \begin{cases} -\exp(-2\gamma_{12} + 2\gamma_{22}) & y = (2, 0, 0) \\ -\exp(-\gamma_{12} + 2\gamma_{22}) & y = (2, 0, 1) \\ -\exp(2\gamma_{22}) & y = (2, 0, 2) \\ 1 & y = (0, 2, \cdot) \end{cases},$$

$$m_{12}(y; \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}) = \begin{cases} -\exp(-2\gamma_{12} - 2\gamma_{21} + 2\gamma_{22}) & y = (2, 0, 0) \\ -\exp(-\gamma_{12} - 2\gamma_{21} + 2\gamma_{22}) & y = (2, 0, 1) \\ -\exp(-2\gamma_{21} + 2\gamma_{22}) & y = (2, 0, 2) \\ 1 & y = (0, 2, \cdot) \end{cases},$$

and

$$m_{22}(y; \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}) = \begin{cases} -\exp(-2\gamma_{12}) & y = (2, 0, 0) \\ -\exp(-\gamma_{12}) & y = (2, 0, 1) \\ -1 & y = (2, 0, 2) \\ 1 & y = (0, 2, \cdot) \end{cases}.$$

The first of these identifies γ_{11} , while the last identifies γ_{12} . Having identified γ_{12} , the second identifies γ_{22} and γ_{21} is then finally identified from the moment condition based on the third moment function.

3.3 Poisson-Binomial Model

In the previous sections the outcome in any given time period consisted of a sum of Bernoulli random variables that are independent and identically distributed, conditional on a set of explanatory and fixed effects, so the outcome followed a Binomial distribution. In this section, we consider the implications of relaxing the assumption on identical distribution, so the outcomes consist of (conditionally) independent but non-identically distributed Bernoulli random variables. The resulting distribution of the outcome in any

given time period is therefore a Poisson-Binomial.⁵ This further bridges the gap between the literature on logit panel data and the literature on simple entry games. In simple entry games, as discussed in Section 3.4, the Bernoulli random variables may represent the entry decision of a firm. As firms differ in their primitives, such as the number of employees, resulting in different X s, or in their brand loyalty, resulting in different A s, the Bernoulli random variables will have non-identical distributions. Unlike the models considered in this section, static entry games are characterized by also allowing for “strategic interaction,” resulting in the Bernoulli random variables no longer being independent.

We consider the outcomes $Y_t = \sum_{m=1}^M B_{mt}$ where B_{mt} are conditionally independent within a given time period but non-identically distributed outcomes from a binary logit model. That is,

$$B_{mt} = 1 \{X'_{mt}\beta + Y_{t-1}\gamma + A_m + \varepsilon_{mt} \geq 0\}$$

where ε_{mt} are *i.i.d.* and logistically distributed. Note that the fixed effects, A_m , are allowed to differ across the M binary outcomes. This makes the model different from the model in section 3.1 even when $\beta = 0$.

Let $Y = (Y_1, \dots, Y_T)$, $X = (X_{1,1}, \dots, X_{M,1}, X_{1,2}, \dots, X_{M,2}, X_{1,T}, \dots, X_{M,T})$ and $A = (A_1, \dots, A_M)$. The outcome in a given period, Y_t , follows a Poisson-Binomial distribution with probability of success parameters conditional on the past values of the outcomes, the X 's, and the A 's,

$$Y_t \mid (\{Y_s\}_{s < t}, X, A) \sim PB(p_1, \dots, p_M), \quad p_m = \frac{\exp(X'_{mt}\beta + Y_{t-1}\gamma + A_m)}{1 + \exp(X'_{mt}\beta + Y_{t-1}\gamma + A_m)}. \quad (3.5)$$

⁵The Poisson-Binomial distribution is sometimes called “Poisson’s Binomial distribution,” (see, for example, Wang, 1993).

Using the approach from Section 3.1.3, we find that when $\gamma = 0$ and $M = T = 2$, there is one moment condition that does not depend on (A_1, A_2) , which is given by

$$m(Y, X; \beta) = \begin{cases} -\exp((X_{11} - X_{12} + X_{21} - X_{22})' \beta), & Y = (0, 2) \\ 1, & Y = (2, 0) \\ 0, & \text{otherwise} \end{cases}$$

The moment condition above corresponds to the first order condition for maximizing the conditional likelihood conditional on $\{Y = (0, 2)\} \cup \{Y = (2, 0)\}$.

The finding that this kind of conditioning can be used to estimate β generalizes to more general values of M and T . To see this, let $y = (y_1, \dots, y_T)$ and note that

$$P(Y = y \mid X, A) = \prod_{t=1}^T \sum_{\left\{b \in \{0,1\}^M : \sum_m b_m = y_t\right\}} \prod_{m=1}^M \frac{\exp(b_m (X'_{mt} \beta + A_m))}{1 + \exp(X'_{mt} \beta + A_m)}. \quad (3.6)$$

Notice that the denominator is independent of the outcome sequence and if $y_t \in \{0, M\}$ for all t , the summation sign disappears. Furthermore, whenever $y_t = M$ the numerator factorizes to become $\exp\left(\sum_{m=1}^M X'_{mt} \beta\right) \exp\left(\sum_{m=1}^M A_m\right)$, which implies that for two sequences y and $\tilde{y}$, the ratio

$$\frac{P(Y = y \mid X, A)}{P(Y = \tilde{y} \mid X, A)}$$

does not depend on any of the A_m if $y_t, \tilde{y}_t \in \{0, M\}$ and $\sum_{t=1}^T y_t = \sum_{t=1}^T \tilde{y}_t$, so the corresponding statistic is sufficient for A_m (see Remark 3.1 in Appendix 3.B).

The conditioning argument above superficially suggests that the same sufficient statistic as in Proposition 3.1 applied to the Poisson-Binomial case. To see that this is not

true, suppose that $T = 2$ and $M = 2$ and consider outcome sequences $Y = (1, 2)$ and $Y = (2, 1)$, we have

$$P(Y = (1, 2) | X, A) = \frac{\exp(X'_{11}\beta + A_1) + \exp(X'_{21}\beta + A_2)}{\prod_{m=1}^2 (1 + \exp(X'_{m1}\beta + A_m))} \frac{\exp(X'_{12}\beta + A_1) \exp(X'_{22}\beta + A_2)}{\prod_{m=1}^2 (1 + \exp(X'_{m2}\beta + A_m))},$$

and

$$P(Y = (2, 1) | X, A) = \frac{\exp(X'_{11}\beta + A_1) \exp(X'_{21}\beta + A_2)}{\prod_{m=1}^2 (1 + \exp(X'_{m1}\beta + A_m))} \frac{\exp(X'_{12}\beta + A_1) + \exp(X'_{22}\beta + A_2)}{\prod_{m=1}^2 (1 + \exp(X'_{m2}\beta + A_m))},$$

so

$$\frac{P(Y = (1, 2) | X, A)}{P(Y = (2, 1) | X, A)} = \frac{\exp(X'_{12}\beta) \exp(X'_{22}\beta)}{\exp(X'_{11}\beta) \exp(X'_{21}\beta)} \frac{\exp(X'_{11}\beta) + \exp(X'_{21}\beta + A_2 - A_1)}{\exp(X'_{12}\beta) + \exp(X'_{22}\beta + A_2 - A_1)}.$$

This implies that $P(Y = (1, 2) | X, A, \{Y = (1, 2)\} \cup \{Y = (2, 1)\})$ will generally depend on $A_2 - A_1$. In other words, conditioning on the sufficient statistic in Proposition 3.1 will not generally eliminate the fixed effects.

In contrast, for the case of a pure AR(1) Poisson-Binomial model Proposition 3.2 does carry over and a sufficient statistic exists.

Proposition 3.5. *Consider the special case of (3.5) in which $\beta = 0$. The statistic*

$$\left(Y_0, \sum_{t=1}^T 1 \{Y_t = 0\}, \sum_{t=1}^T 1 \{Y_t = 1\}, \sum_{t=1}^T 1 \{Y_t = 2\}, \dots, \sum_{t=1}^T 1 \{Y_t = M\}, Y_T \right)$$

is minimally sufficient for A .

Numerical calculations suggest that no moment conditions exist for the general model in (3.5) when $M = 2$ and $T \leq 4$. It does however seem to be possible, when $T \geq 5$, but in that case there are 243 possible outcome sequences, and our numerical calculations suggest that the candidate moments depend on a great number of outcome sequences, rendering symbolic calculations too computationally costly.

Conjecture 3.1. *Consider the unrestricted version of (3.5). We conjecture that*

- *With $M = 2$ and $T \leq 4$, there are no conditional moment functions that do not depend on (A_1, A_2) .*
- *With $M = 2$ and $T = 5$, there are 96 potential moments that do not depend on (A_1, A_2) . They all depend on β and γ .*

3.4 Models With Strategic Interactions

In this section, we relate the Poisson-Binomial model from the previous section to the literature on simple entry games.

Consider a static game in the spirit of Bresnahan and Reiss (1991) and Tamer (2003). There are two potential entrants with

$$\begin{aligned} B_{1t} &= 1 \{X'_{1t}\beta - \delta_1 B_{2t} + A_1 + \varepsilon_{1t} > 0\} \\ B_{2t} &= 1 \{X'_{2t}\beta - \delta_2 B_{1t} + A_2 + \varepsilon_{2t} > 0\}, \end{aligned} \tag{3.7}$$

Here δ_1 and δ_2 are positive so that if one unit enters, then that decreases the probability that the other firm will enter. This model is an incomplete econometric model in the sense that for given X 's and A 's there is no mapping from the ε 's to unique outcomes (B_{1t}, B_{2t}) . This prevents one from deriving the distribution of (B_{1t}, B_{2t}) conditional on the X 's and the A 's. However, the sum of the Y 's is typically uniquely determined, and in those cases one can derive the distribution of the sum. Assume, for example, that the ε 's are i.i.d. logistic, then conventional calculations (see for example de Paula (2013)) deliver:

$$\begin{aligned} P(Y_t = 2 \mid X_1, X_2, A_1, A_2) &= \frac{\exp(X'_{1t}\beta - \delta_1 + A_1)}{1 + \exp(X'_{1t}\beta - \delta_1 + A_1)} \frac{\exp(X'_{2t}\beta - \delta_2 + A_2)}{1 + \exp(X'_{2t}\beta - \delta_2 + A_2)}, \\ P(Y_t = 0 \mid X_1, X_2, A_1, A_2) &= \frac{1}{1 + \exp(X'_{1t}\beta + A_1)} \frac{1}{1 + \exp(X'_{2t}\beta + A_2)}, \end{aligned}$$

and

$$\begin{aligned}
& P(Y_t = 1 \mid X_1, X_2, A_1, A_2) \\
& = 1 - P(Y_t = 2 \mid \{X_{1s}, X_{2s}\}_{s=1}^2, A_1, A_2) - P(Y_t = 0 \mid \{X_{1s}, X_{2s}\}_{s=1}^2, A_1, A_2),
\end{aligned}$$

where $Y_t = B_{1t} + B_{2t}$.

When $\delta_1 = \delta_2 = 0$, the model is equivalent to the Poisson-Binomial model in equation (3.5) with $\gamma = 0$ and $M = 2$. If we further constrain $X_{1t} = X_{2t}$, $A_1 = A_2$, this becomes the simple binomial model (3.3).

Honoré and de Paula (2021) discusses conditions under which β in (3.7) is identified if the distribution of $(\varepsilon_{1t}, \varepsilon_{2t})$ is left unspecified. Specifically, they find that if the X 's are the same for the two players, (so $X_{1t} = X_{2t}$), then one can use an argument similar to that in Manski (1987) to identify β using two time-periods.⁶ On the other hand, Honoré and de Paula (2021) is unable to provide conditions for point identification of the interaction parameters, δ_1 and δ_2 , if the distribution of $(\varepsilon_{1t}, \varepsilon_{2t})$ is left unspecified. Despite this, the similarity between simplified versions of (3.7) and the binomial and Poisson-Binomial model above gives hope that one can obtain moment conditions for (3.7) in the case where ε_{1t} and ε_{2t} are assumed to be independent and logistic. Numerical calculations suggest that this is indeed the case. While some of the potential moments only depend on a limited number of outcome sequences, solving for these and simplifying to a readable output currently makes the computer stall.

⁶Honoré and de Paula (2021) also have results based on observations for which the X 's change over time for one individual, but not for the other.

Conjecture 3.2. *Suppose that $X_{1t} = X_{2t}$. When $T = 2$, there are no moment conditions that do not depend on (A_1, A_2) .*

For $T = 3$,

- *If δ_1 is restricted to equal δ_2 , then there are eight moment conditions that do not depend on (A_1, A_2) . They all depend on β , but two do not depend on the common δ .*
- *If δ_1 is not restricted to equal δ_2 , then there are two moment conditions that do not depend on (A_1, A_2) . They both depend on β , but they do not depend on the δ 's.*

For $T = 4$,

- *If δ_1 is restricted to equal δ_2 , then there are 48 moment conditions that do not depend on (A_1, A_2) . They all depend on β , but 24 do not depend on the common δ .*
- *If δ_1 is not restricted to equal δ_2 , then there are 31 moment conditions that do not depend on (A_1, A_2) . They all depend on β , but 18 do not depend on the δ 's.*

Conjecture 3.3. *Suppose that X_{1t} is allowed to differ from X_{2t} . When $T \leq 3$, there are no moment conditions that do not depend on (A_1, A_2) . For $T = 4$, there are 25 moment conditions that do not depend on (A_1, A_2) . They all depend on β , but four do not depend on (δ_1, δ_2) . Imposing that $\delta_1 = \delta_2$ does not change the number of moment conditions.*

3.5 Conclusion

In this paper, we considered panel data models where the outcome in a given period is the sum of logit Bernoulli random variables that are independent and identically distributed conditional on a set of common explanatory variables and fixed effects. We first showed that many of the results known from the literature on binary outcome logit panel data models readily extends to this framework, including the existence of (some) sufficient statistics for the fixed effects, and which parameters are estimable using a conditional likelihood approach. Applying the methodology from Honoré and Weidner (2024), we found moment conditions in the cases where conditional likelihood estimation was not possible. We subsequently relaxed the assumption of identical distribution for the Bernoulli random variables and show that under certain restrictions of the model estimation of the common parameters are still possible, either with a conditional likelihood approach or by the existence of moments. Finally, we discussed the relationship between the models we have considered so far and the simple entry game models, where the Bernoulli random variables are no longer independent. For this framework, we conjecture, using numerical procedures, the existence of moment conditions, although the form of these are seemingly too complicated to solve algebraically using currently available tools.

References

- Bresnahan, T., & Reiss, P. (1991). Empirical models of discrete games. *Journal of Econometrics*, 48.
- Chamberlain, G. (1978). *On the use of panel data* [Department of Economics, Harvard University].
- Chamberlain, G. (1993). *Feedback in panel data models* [Unpublished Manuscript, Department of Economics, Harvard University].
- Chamberlain, G. (1984). Panel data. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (pp. 1248–1318). North-Holland.
- Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In J. J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 3–38). Cambridge University Press.
- Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, 10(1), 20–26. <https://www.tandfonline.com/doi/abs/10.1080/07350015.1992.10509881>
- Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica*, 78(1), 159–168.
- Charlier, E., Melenberg, B., & van Soest, A. H. O. (1995). A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation. *Statistica Neerlandica*, 49(3), 324–342.
- Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society*, 20, 215–242.

- d'Addio, A. C., & Honoré, B. E. (2010). Duration dependence and timevarying variables in discrete time duration models. *Brazilian Review of Econometrics*, 30(2), 487–527.
- Dano, K. (2023). Transition probabilities and moment restrictions in dynamic fixed effects logit models. *arXiv preprint arXiv:2303.00083*.
- de Paula, Á. (2013). Econometric analysis of games with multiple equilibria. *Annual Review of Economics*, 5, 107–131.
- Dobronyi, C., Gu, J., il Kim, K., & Russell, T. M. (2024). Identification of dynamic panel logit models with fixed effects.
- Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica*, 52(4), 909–38.
- Heckman, J. J. (1978). Simple statistical models for discrete panel data developed and applied to tests of the hypothesis of true state dependence against the hypothesis of spurious state dependence. *Annales de l'INSEE*, 227–269.
- Heckman, J. J. (1981a). Heterogeneity and state dependence. *Studies in Labor Markets*, S. Rosen (ed).
- Heckman, J. J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. *Structural Analysis of Discrete Panel Data with Econometric Applications*, C. F. Manski and D. McFadden (eds), 179–195.
- Heckman, J. J. (1981c). Statistical models for discrete panel data. *Structural Analysis of Discrete Panel Data with Econometric Applications*, C. F. Manski and D. McFadden (eds), 114–178.

- Honoré, B. E. (1992). Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica*, 60, 533–565.
- Honoré, B. E. (1993). Orthogonality conditions for tobit models with fixed effects and lagged dependent variables. *Journal of Econometrics*, 59, 35–61.
- Honoré, B. E., & de Paula, Á. (2021). Identification in simple binary outcome panel data models. *The Econometrics Journal*, 24(2), C78–C93.
- Honoré, B. E., & Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68, 839–874.
- Honoré, B. E., & Weidner, M. (2024). Moment conditions for dynamic panel logit models with fixed effects. *The Review of Economic Studies*, rdae097.
- Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60, 505–531.
- Hu, L. (2002). Estimation of a censored dynamic panel data model. *Econometrica*, 70(6), pp. 2499–2517. <http://www.jstor.org/stable/3081995>
- Khan, S., Ponomareva, M., & Tamer, E. (2023). Identification of dynamic binary response models. *Journal of Econometrics*, 237(1), 105515.
- Kitazawa, Y. (2022). Transformations and moment conditions for dynamic fixed effects logit models. *Journal of Econometrics*, 229(2), 350–362.
- Kruiniger, H. (2023). Further results on the estimation of dynamic panel logit models with fixed effects. <https://arxiv.org/abs/2010.03382>
- Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65, 1335–1364.

- Kyriazidou, E. (2001). Estimation of dynamic panel data sample selection models. *Review of Economic Studies*, 68, 543–572.
- Magnac, T. (2000). State dependence and unobserved heterogeneity in youth employment histories. *Economic Journal*, 110, 805–837.
- Manski, C. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55, 357–62.
- Ouyang, F., & Yang, T. T. (2024). Semiparametric estimation of dynamic binary choice panel data models. *Econometric Theory*, 1–40.
- Qu, Q., Sun, J., & Wright, J. (2014). Finding a sparse vector in a subspace: Linear sparsity using alternating directions. *CoRR*, abs/1412.4659. <http://arxiv.org/abs/1412.4659>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Denmark: Pædagogiske Institut.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, 4, 321–333.
- Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies*, 70.
- Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, 3(2), 295–312.
- Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, 13(5), 667–678.

Appendix

3.A Moment Conditions

Binomial Model With Fixed Effects

For notational economy we define $\Gamma = \exp(\gamma)$ and $\check{x}_t = \exp(x'_t \beta)$ allowing us to (more) compactly write the following moment conditions that have all been derived from the case where $y_0 = 1$.

The first moment condition is

$$m_1(y, x; \beta, \gamma) = \begin{cases} \frac{4(\Gamma - 1)\check{x}_1\check{x}_2(\check{x}_1^2 + \check{x}_2\check{x}_3)}{\Gamma(\check{x}_1^2 - \check{x}_2\check{x}_3)^2}, & y = (0, 0, 1) \\ \frac{4\check{x}_2\check{x}_3^2(\Gamma\check{x}_1^4 + 2(\Gamma - 2)\check{x}_2\check{x}_3\check{x}_1^2 + \Gamma\check{x}_2^2\check{x}_3^2)}{\Gamma\check{x}_1^3(\check{x}_1^2 - \check{x}_2\check{x}_3)^2}, & y = (0, 0, 2) \\ -\frac{4(\Gamma - 1)\check{x}_1\check{x}_2\check{x}_3^2(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)}{\Gamma(\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (0, 1, 0) \\ \frac{\Gamma^2\check{x}_1^7 + (\Gamma - 4)\Gamma\check{x}_2\check{x}_3\check{x}_1^5 + 3\check{x}_2^2\check{x}_3^2\check{x}_1^3 - \check{x}_2^3\check{x}_3^3\check{x}_1}{\Gamma\check{x}_3(\check{x}_2\check{x}_3 - \check{x}_1^2)^3}, & y = (0, 1, 1) \\ \frac{4(\Gamma - 1)\check{x}_1^2\check{x}_2\check{x}_3(\Gamma\check{x}_2\check{x}_3 - \check{x}_1^2)}{\Gamma^2(\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (1, 0, 0) \\ \frac{\check{x}_2\check{x}_3(\Gamma^2\check{x}_2^3\check{x}_3^3 + (\Gamma - 4)\Gamma\check{x}_2^2\check{x}_3^2\check{x}_1^2 - \check{x}_1^6 + 3\check{x}_2\check{x}_3\check{x}_1^4)}{\Gamma\check{x}_1^2(\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (1, 0, 1) \\ 1, & y = (1, 1, 0) \\ 0, & \text{otherwise} \end{cases}.$$

The second moment condition is

$$m_2(y, x; \beta, \gamma) = \begin{cases} \frac{4\Gamma\check{x}_1(\check{x}_1^2 - \check{x}_2\check{x}_3)^2}{(\Gamma - 1)\check{x}_3^3(\check{x}_1^2 + \check{x}_2\check{x}_3)}, & y = (0, 0, 2) \\ -\frac{\Gamma^2\check{x}_1^6 + (\Gamma - 4)\Gamma\check{x}_2\check{x}_3\check{x}_1^4 + 3\check{x}_2^2\check{x}_3^2\check{x}_1^2 - \check{x}_2^3\check{x}_3^3}{(\Gamma - 1)\check{x}_1\check{x}_2(\check{x}_1^4 - \check{x}_2^2\check{x}_3^2)}, & y = (0, 1, 1) \\ \frac{4\check{x}_1\check{x}_2\check{x}_3(\check{x}_2 - \Gamma\check{x}_3)}{\Gamma(\check{x}_1^4 - \check{x}_2^2\check{x}_3^2)}, & y = (0, 1, 2) \\ -\frac{\Gamma^2\check{x}_2^3\check{x}_3^3 - (\Gamma - 4)\Gamma\check{x}_2^2\check{x}_3^2\check{x}_1^2 + \check{x}_1^6 - 3\check{x}_2\check{x}_3\check{x}_1^4}{(\Gamma - 1)\check{x}_3^2(\check{x}_2^2\check{x}_3^2 - \check{x}_1^4)}, & y = (1, 0, 1) \\ -\frac{4\check{x}_2\check{x}_3(\check{x}_1^2 - \Gamma\check{x}_2\check{x}_3)}{\check{x}_1^4 - \check{x}_2^2\check{x}_3^2}, & y = (1, 0, 2) \\ \frac{\Gamma + \check{x}_1^2\left(\frac{\Gamma}{\check{x}_2\check{x}_3} - \frac{4}{\check{x}_1^2 + \check{x}_2\check{x}_3}\right)}{\Gamma - 1}, & y = (1, 1, 0) \\ 1, & y = (1, 1, 1) \\ 0, & \text{otherwise} \end{cases}.$$

The third moment condition is

$$m_3(y, x; \beta, \gamma) = \begin{cases} \frac{(\Gamma - 1)\check{x}_1\check{x}_2(\Gamma\check{x}_1^2 + \check{x}_2\check{x}_3)}{\Gamma(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)^2}, & y = (0, 1, 1) \\ \frac{\check{x}_1\check{x}_2(\Gamma^2\check{x}_3^2 + 2(\Gamma - 2)\check{x}_3\check{x}_2 + \check{x}_2^2)}{\Gamma(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)^2}, & y = (0, 1, 2) \\ -\frac{4(\Gamma - 1)\check{x}_1\check{x}_2\check{x}_3^2(\Gamma^2\check{x}_1^2 - \check{x}_2\check{x}_3)}{(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (0, 2, 0) \\ -\frac{\Gamma^5\check{x}_1^7 + (\Gamma - 4)\Gamma^3\check{x}_2\check{x}_3\check{x}_1^5 + 3\Gamma\check{x}_2^2\check{x}_3^2\check{x}_1^3 - \check{x}_2^3\check{x}_3^3\check{x}_1}{\Gamma\check{x}_3(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (0, 2, 1) \\ \frac{(\Gamma - 1)\check{x}_1^2\check{x}_2\check{x}_3(\check{x}_2\check{x}_3 - \check{x}_1^2)}{(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (1, 1, 0) \\ -\frac{\check{x}_2\check{x}_3(\check{x}_1^2 - \check{x}_2\check{x}_3)(\Gamma\check{x}_1^4 + (\Gamma - 3)\check{x}_2\check{x}_3\check{x}_1^2 + \check{x}_2^2\check{x}_3^2)}{4\check{x}_1^2(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (1, 1, 1) \\ 1, & y = (1, 2, 0) \\ 0, & \text{otherwise} \end{cases}.$$

The fourth moment conditions is

$$m_4(y, x; \beta, \gamma) = \begin{cases} \frac{\check{x}_1 (\Gamma \check{x}_1^2 - \check{x}_2 \check{x}_3)^2}{(\Gamma - 1) \check{x}_3^3 (\Gamma \check{x}_1^2 + \check{x}_2 \check{x}_3)}, & y = (0, 1, 2) \\ -\frac{\Gamma^5 \check{x}_1^6 + (\Gamma - 4) \Gamma^3 \check{x}_2 \check{x}_3 \check{x}_1^4 + 3 \Gamma \check{x}_2^2 \check{x}_3^2 \check{x}_1^2 - \check{x}_2^3 \check{x}_3^3}{(\Gamma - 1) \check{x}_1 \check{x}_2 (\Gamma^2 \check{x}_1^4 - \check{x}_2^2 \check{x}_3^2)}, & y = (0, 2, 1) \\ \frac{4 \check{x}_1 \check{x}_2 \check{x}_3 (\check{x}_2 - \Gamma^2 \check{x}_3)}{\Gamma^3 \check{x}_1^4 - \Gamma \check{x}_2^2 \check{x}_3^2}, & y = (0, 2, 2) \\ \frac{\Gamma (-\Gamma \check{x}_1^6 + (\Gamma - 4) \check{x}_2^2 \check{x}_3^2 \check{x}_1^2 + 3 \check{x}_2 \check{x}_3 \check{x}_1^4 + \check{x}_2^3 \check{x}_3^3)}{4(\Gamma - 1) \check{x}_3^2 (\Gamma^2 \check{x}_1^4 - \check{x}_2^2 \check{x}_3^2)}, & y = (1, 1, 1) \\ \frac{\Gamma \check{x}_2 \check{x}_3 (\check{x}_2 \check{x}_3 - \check{x}_1^2)}{\Gamma^2 \check{x}_1^4 - \check{x}_2^2 \check{x}_3^2}, & y = (1, 1, 2) \\ \frac{\Gamma \left(\check{x}_1^2 \left(\frac{\Gamma}{\check{x}_2 \check{x}_3} - \frac{4}{\Gamma \check{x}_1^2 + \check{x}_2 \check{x}_3} \right) + 1 \right)}{\Gamma - 1}, & y = (1, 2, 0) \\ 1, & y = (1, 2, 1) \\ 0, & \text{otherwise} \end{cases}.$$

The fifth moment condition is

$$m_5(y, x; \beta, \gamma) = \begin{cases} \frac{(\Gamma - 1)\check{x}_1\check{x}_2(\Gamma\check{x}_2\check{x}_3 + \check{x}_1^2)}{(\check{x}_1^2 - \Gamma\check{x}_2\check{x}_3)^2}, & y = (1, 0, 1) \\ \frac{\Gamma\check{x}_2\check{x}_3^2(\Gamma^2\check{x}_2^2\check{x}_3^2 + 2(\Gamma - 2)\check{x}_2\check{x}_3\check{x}_1^2 + \check{x}_1^4)}{\check{x}_1^3(\check{x}_1^2 - \Gamma\check{x}_2\check{x}_3)^2}, & y = (1, 0, 2) \\ -\frac{(\Gamma - 1)\Gamma\check{x}_1\check{x}_2\check{x}_3^2(\check{x}_1^2 - \check{x}_2\check{x}_3)}{(\check{x}_1^2 - \Gamma\check{x}_2\check{x}_3)^3}, & y = (1, 1, 0) \\ \frac{\Gamma\check{x}_1(\check{x}_1^2 - \check{x}_2\check{x}_3)((\Gamma - 3)\check{x}_2\check{x}_3\check{x}_1^2 + \Gamma\check{x}_2^2\check{x}_3^2 + \check{x}_1^4)}{4\check{x}_3(\Gamma\check{x}_2\check{x}_3 - \check{x}_1^2)^3}, & y = (1, 1, 1) \\ \frac{4(\Gamma - 1)\check{x}_1^2\check{x}_2\check{x}_3(\check{x}_1^2 - \Gamma^2\check{x}_2\check{x}_3)}{\Gamma(\Gamma\check{x}_2\check{x}_3 - \check{x}_1^2)^3}, & y = (2, 0, 0) \\ \frac{\check{x}_2\check{x}_3(\Gamma^5\check{x}_2^3\check{x}_3^3 + (\Gamma - 4)\Gamma^3\check{x}_2^2\check{x}_3^2\check{x}_1^2 + 3\Gamma\check{x}_2\check{x}_3\check{x}_1^4 - \check{x}_1^6)}{\check{x}_1^2(\check{x}_1^2 - \Gamma\check{x}_2\check{x}_3)^3}, & y = (2, 0, 1) \\ 1, & y = (2, 1, 0) \\ 0, & \text{otherwise} \end{cases}.$$

The sixth moment condition is

$$m_6(y, x; \beta, \gamma) = \begin{cases} \frac{\Gamma \check{x}_1 (\check{x}_1^2 - \Gamma \check{x}_2 \check{x}_3)^2}{(\Gamma - 1) \check{x}_3^3 (\Gamma \check{x}_2 \check{x}_3 + \check{x}_1^2)}, & y = (1, 0, 2) \\ \frac{\Gamma ((\Gamma - 4) \check{x}_2 \check{x}_3 \check{x}_1^4 - \Gamma \check{x}_2^3 \check{x}_3^3 + \check{x}_1^6 + 3 \check{x}_2^2 \check{x}_3^2 \check{x}_1^2)}{4(\Gamma - 1) \check{x}_1 \check{x}_2 (\Gamma^2 \check{x}_2^2 \check{x}_3^2 - \check{x}_1^4)}, & y = (1, 1, 1) \\ \frac{\Gamma \check{x}_1 \check{x}_2 \check{x}_3 (\check{x}_3 - \check{x}_2)}{\Gamma^2 \check{x}_2^2 \check{x}_3^2 - \check{x}_1^4}, & y = (1, 1, 2) \\ \frac{-\Gamma^5 \check{x}_2^3 \check{x}_3^3 - (\Gamma - 4) \Gamma^3 \check{x}_2^2 \check{x}_3^2 \check{x}_1^2 - 3 \Gamma \check{x}_2 \check{x}_3 \check{x}_1^4 + \check{x}_1^6}{(\Gamma - 1) \check{x}_3^2 (\Gamma^2 \check{x}_2^2 \check{x}_3^2 - \check{x}_1^4)}, & y = (2, 0, 1) \\ \frac{4 \Gamma \check{x}_2 \check{x}_3 (\check{x}_1^2 - \Gamma^2 \check{x}_2 \check{x}_3)}{\Gamma^2 \check{x}_2^2 \check{x}_3^2 - \check{x}_1^4}, & y = (2, 0, 2) \\ \frac{\Gamma + \check{x}_1^2 \left(\frac{1}{\check{x}_2 \check{x}_3} - \frac{4}{\Gamma \check{x}_2 \check{x}_3 + \check{x}_1^2} \right)}{\Gamma - 1}, & y = (2, 1, 0) \\ 1, & y = (2, 1, 1) \\ 0, & \text{otherwise} \end{cases}.$$

The seventh moment condition is

$$m_7(y, x; \beta, \gamma) = \begin{cases} \frac{(\Gamma - 1)\check{x}_1\check{x}_2(\check{x}_1^2 + \check{x}_2\check{x}_3)}{4\Gamma(\check{x}_1^2 - \check{x}_2\check{x}_3)^2}, & y = (1, 1, 1) \\ \frac{\check{x}_1\check{x}_2(\Gamma\check{x}_2^2 + 2(\Gamma - 2)\check{x}_3\check{x}_2 + \Gamma\check{x}_3^2)}{4\Gamma(\check{x}_1^2 - \check{x}_2\check{x}_3)^2}, & y = (1, 1, 2) \\ -\frac{(\Gamma - 1)\check{x}_1\check{x}_2\check{x}_3^2(\Gamma\check{x}_1^2 - \check{x}_2\check{x}_3)}{\Gamma(\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (1, 2, 0) \\ \frac{\Gamma^2\check{x}_1^7 + (\Gamma - 4)\Gamma\check{x}_2\check{x}_3\check{x}_1^5 + 3\check{x}_2^2\check{x}_3^2\check{x}_1^3 - \check{x}_2^3\check{x}_3^3\check{x}_1}{4\Gamma\check{x}_3(\check{x}_2\check{x}_3 - \check{x}_1^2)^3}, & y = (1, 2, 1) \\ \frac{(\Gamma - 1)\check{x}_1^2\check{x}_2\check{x}_3(\Gamma\check{x}_2\check{x}_3 - \check{x}_1^2)}{\Gamma^2(\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (2, 1, 0) \\ \frac{\check{x}_2\check{x}_3(\Gamma^2\check{x}_2^3\check{x}_3^3 + (\Gamma - 4)\Gamma\check{x}_2^2\check{x}_3^2\check{x}_1^2 - \check{x}_1^6 + 3\check{x}_2\check{x}_3\check{x}_1^4)}{4\Gamma\check{x}_1^2(\check{x}_1^2 - \check{x}_2\check{x}_3)^3}, & y = (2, 1, 1) \\ 1, & y = (2, 2, 0) \\ 0, & \text{otherwise} \end{cases}.$$

The eighth moment condition is

$$m_8(y, x; \beta, \gamma) = \begin{cases} \frac{\Gamma \check{x}_1 (\check{x}_1^2 - \check{x}_2 \check{x}_3)^2}{4(\Gamma - 1) \check{x}_3^3 (\check{x}_1^2 + \check{x}_2 \check{x}_3)}, & y = (1, 1, 2) \\ -\frac{\Gamma^2 \check{x}_1^6 + (\Gamma - 4) \Gamma \check{x}_2 \check{x}_3 \check{x}_1^4 + 3 \check{x}_2^2 \check{x}_3^2 \check{x}_1^2 - \check{x}_2^3 \check{x}_3^3}{4(\Gamma - 1) \check{x}_1 \check{x}_2 (\check{x}_1^4 - \check{x}_2^2 \check{x}_3^2)}, & y = (1, 2, 1) \\ \frac{\check{x}_1 \check{x}_2 \check{x}_3 (\check{x}_2 - \Gamma \check{x}_3)}{\Gamma (\check{x}_1^4 - \check{x}_2^2 \check{x}_3^2)}, & y = (1, 2, 2) \\ -\frac{\Gamma^2 \check{x}_2^3 \check{x}_3^3 + (\Gamma - 4) \Gamma \check{x}_2^2 \check{x}_3^2 \check{x}_1^2 - \check{x}_1^6 + 3 \check{x}_2 \check{x}_3 \check{x}_1^4}{4(\Gamma - 1) \check{x}_3^2 (\check{x}_2^2 \check{x}_3^2 - \check{x}_1^4)}, & y = (2, 1, 1) \\ \frac{\check{x}_2 \check{x}_3 (\check{x}_1^2 - \Gamma \check{x}_2 \check{x}_3)}{\check{x}_2^2 \check{x}_3^2 - \check{x}_1^4}, & y = (2, 1, 2) \\ \frac{\Gamma + \check{x}_1^2 \left(\frac{\Gamma}{\check{x}_2 \check{x}_3} - \frac{4}{\check{x}_1^2 + \check{x}_2 \check{x}_3} \right)}{\Gamma - 1}, & y = (2, 2, 0) \\ 1, & y = (2, 2, 1) \\ 0, & \text{otherwise} \end{cases}.$$

3.B Proofs

We start by making the following remark that we return to in the proofs

Remark 3.1. *Let $y = (y_1, \dots, y_T)$ be an outcome sequence. If $S(Y)$ is a sufficient statistic for A , then (and only then) will the distribution for Y factor as $P(Y = y \mid X, A, \theta) = f(y \mid X, \theta) g(S(y) \mid X, A, \theta)$, by the Fisher-Neyman factorization theorem. If $S(y) = S(\tilde{y})$ a.s., then*

$$\frac{P(Y = y \mid X, A, \theta)}{P(Y = \tilde{y} \mid X, A, \theta)} = \frac{f(y \mid X, \theta)}{f(\tilde{y} \mid X, \theta)}.$$

So to check whether $S(Y)$ is sufficient for A , we can check whether the ratio $P(Y = y \mid X, A, \theta) / P(Y = \tilde{y} \mid X, A, \theta)$ is independent of A .

If the reverse implication is also true, such that for every pair of outcome sequences, y and $\tilde{y}$, where the ratio is independent of A satisfy $S(y) = S(\tilde{y})$ a.s., then S is minimally sufficient.

Additionally, we provide the following Lemma

Lemma 3.1. *Let $z = (z_1, z_2, \dots, z_T)$ and $\tilde{z} = (\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_T)$ denote two sequences of real numbers. Then the fraction*

$$\prod_{t=1}^T \frac{1 + \tilde{z}_t a}{1 + z_t a}$$

is constant in $a \in \mathbb{R}$ if and only if z is a permutation of $\tilde{z}$.

For the proof we make the following definition

Definition 3.1. *The elementary symmetric polynomials in N variables, $x_1, x_2, \dots, x_N$ are defined as*

$$e_0(x_1, \dots, x_N) = 1,$$

and

$$e_k(x_1, \dots, x_N) = \sum_{1 \leq m_1 < \dots < m_k \leq N} x_{m_1} x_{m_2} \cdots x_{m_k}$$

for $k \in \{0, 1, \dots, N\}$, and

$$e_k(x_1, \dots, x_N) = 0$$

for $k > N$.

and we need the following helpful lemma

Lemma 3.2. *Let $x = (x_1, x_2, \dots, x_N)$ and $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_N)$ denote two sequences of real numbers.*

Then the elementary symmetric polynomials for the two sequences are equal, i.e. $e_k(x_1, \dots, x_N) = e_k(\tilde{x}_1, \dots, \tilde{x}_N)$ for all k , if and only if the sequence $\tilde{x}$ is a permutation of the sequence x .

Proof of Lemma 3.2. Let $x = (x_1, x_2, \dots, x_N)$ and $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_N)$ denote two sequences of real numbers.

Let the sequence $\tilde{x}$ be a permutation of the sequence x , then $e_k(x_1, \dots, x_N) = e_k(\tilde{x}_1, \dots, \tilde{x}_N)$ follows from the symmetry of the functions for all k .

Assume instead that $e_k(x_1, \dots, x_N) = e_k(\tilde{x}_1, \dots, \tilde{x}_N)$ for all k . We want to show that this implies that $\tilde{x}$ is a permutation of x . To that end, consider the monic polynomials

$$\begin{aligned} P(\lambda) &= \prod_{n=1}^N (\lambda - x_n) \\ &= \lambda^N - e_1(x_1, \dots, x_N) \lambda^{N-1} + e_2(x_1, \dots, x_N) \lambda^{N-2} - \dots + (-1)^N e_N(x_1, \dots, x_N) \end{aligned}$$

and analogously $\tilde{P}(\lambda) = \prod_{n=1}^N (\lambda - \tilde{x}_n)$. By the equality of the elementary symmetric polynomials, the two polynomials have identical coefficients, and are therefore identical $P = \tilde{P}$. From the representation $P(\lambda) = \prod_{n=1}^N (\lambda - x_n)$, it is clear that the roots for P is the set of numbers $\{x_1, \dots, x_N\}$, and the roots for $\tilde{P}$ is the set of numbers $\{\tilde{x}_1, \dots, \tilde{x}_N\}$. As identical polynomials have identical roots, and these sets must therefore contain the same numbers and the sequence $\tilde{x}$ must be a permutation of the sequence x . $\square$

Proof of Lemma 3.1. Let $z = (z_1, z_2, \dots, z_T)$ and $\tilde{z} = (\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_T)$ denote two sequences of real numbers.

Let the sequence $\tilde{z}$ be a permutation of the sequence z , then the factors cancel and the fraction is constant and equal to 1.

Assume instead that the fraction is constant in $a \in \mathbb{R}$, then evidently it must equal to 1, as it is equal to 1 when $a = 0$. Therefore

$$\prod_{t=1}^T (1 + \tilde{z}_t a) = \prod_{t=1}^T (1 + z_t a),$$

for any value of a . which can be rewritten as

$$\sum_{k=1}^T a^k e_k(\tilde{z}_1, \dots, \tilde{z}_T) = \sum_{k=1}^T a^k e_k(z_1, \dots, z_T).$$

Since both sides of the equation are polynomials in a , they are equal only if the respective coefficients are equal, implying that we must have $e_k(z_1, \dots, z_T) = e_k(\tilde{z}_1, \dots, \tilde{z}_T)$ for all k . As e_k are the elementary symmetric polynomials these are equal if and only if $(\tilde{z}_1, \dots, \tilde{z}_T)$ is a permutation of $(z_1, \dots, z_T)$ by Lemma 3.2 as desired. $\square$

Proof of Proposition 3.1. Consider the static binomial model. Let $y = (y_1, \dots, y_T)$ and $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_T)$ denote two outcome sequences. The ratio

$$\begin{aligned} \frac{P(Y = y \mid X, A)}{P(Y = \tilde{y} \mid X, A)} &= \prod_t \frac{\binom{M}{y_t} \exp(y_t X'_t \beta) \exp(y_t A) (1 + \exp(X'_t \beta + A))^M}{\binom{M}{\tilde{y}_t} \exp(\tilde{y}_t X'_t \beta) \exp(\tilde{y}_t A) (1 + \exp(X'_t \beta + A))^M} \\ &= \exp \left(A \left(\sum_{t=1}^T y_t - \sum_{t=1}^T \tilde{y}_t \right) \right) \prod_t \frac{\binom{M}{y_t} \exp(y_t X'_t \beta)}{\binom{M}{\tilde{y}_t} \exp(\tilde{y}_t X'_t \beta)} \end{aligned}$$

does not depend on A if and only if $\sum_{t=1}^T y_t = \sum_{t=1}^T \tilde{y}_t$, hence $\sum_{t=1}^T y_t$ is a minimally sufficient statistic for A . $\square$

Proof of Proposition 3.2. Consider the AR(1) binomial model and let $y = (y_1, \dots, y_T)$ and $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_T)$ be two outcome sequences, and consider the ratio

$$\begin{aligned} &\frac{P(Y = y \mid Y_0, A)}{P(Y = \tilde{y} \mid Y_0, A)} \\ &= \frac{\prod_{t=1}^T \binom{M}{y_t} \exp(y_t y_{t-1} \gamma)}{\prod_{t=1}^T \binom{M}{\tilde{y}_t} \exp(\tilde{y}_t \tilde{y}_{t-1} \gamma)} \exp \left(A \left(\sum_{t=1}^T y_t - \sum_{t=1}^T \tilde{y}_t \right) \right) \left(\frac{\prod_{t=1}^T (1 + \exp(\tilde{y}_{t-1} \gamma + A))}{\prod_{t=1}^T (1 + \exp(y_{t-1} \gamma + A))} \right)^M. \end{aligned}$$

The first term does not depend on A . It is clear that for the second term to not depend on A , we require $\sum_{t=1}^T y_t = \sum_{t=1}^T \tilde{y}_t$. Note then for the last term, that

$$\prod_{t=1}^T (1 + \exp(y_{t-1} \gamma) \exp(A)) = \prod_{m=1}^M (1 + \exp(m \gamma) \exp(A))^{\sum_{t=1}^T \mathbb{1}_{\{y_{t-1}=m\}}}.$$

Therefore the terms cancel out if $\sum_{t=1}^T \mathbb{1}\{y_{t-1} = m\} = \sum_{t=1}^T \mathbb{1}\{\tilde{y}_{t-1} = m\}$ for all m .

This would imply that the ratio is constant in A , and therefore that

$$\left(Y_0, \sum_{t=1}^T \mathbb{1}\{Y_{t-1} = 0\}, \sum_{t=1}^T \mathbb{1}\{Y_{t-1} = 1\} \dots, \sum_{t=1}^T \mathbb{1}\{Y_{t-1} = M\}, \sum_{t=1}^T Y_t \right)$$

is a sufficient statistic for A (see Remark 3.1). Finally, note that this is equivalent to conditioning on

$$\left(Y_0, \sum_{t=1}^T \mathbb{1}\{Y_t = 0\}, \sum_{t=1}^T \mathbb{1}\{Y_t = 1\}, \sum_{t=1}^T \mathbb{1}\{Y_t = 2\}, \dots, \sum_{t=1}^T \mathbb{1}\{Y_t = M\}, Y_T \right)$$

yielding the more familiar expression in the proposition. $\square$

Proof of Proposition 3.3. Consider the simple binomial model with $\beta \neq 0$ and $\gamma \neq 0$. Let

$y = (y_1, \dots, y_T)$ and $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_T)$ denote two outcome sequences, then

$$\begin{aligned} & \frac{P(Y = y \mid Y_0, X, A)}{P(Y = \tilde{y} \mid Y_0, X, A)} \\ &= \exp \left(A \left(\sum_{t=1}^T y_t - \sum_{t=1}^T \tilde{y}_t \right) \right) \prod_{t=1}^T \frac{(1 + \exp(X'_t \beta + \tilde{y}_{t-1} \gamma + A))^M}{(1 + \exp(X'_t \beta + y_{t-1} \gamma + A))^M} C_t, \end{aligned}$$

where

$$C_t = \frac{\binom{M}{y_t}}{\binom{M}{\tilde{y}_t}}.$$

It is therefore clear that for a statistic to be sufficient for A , it must satisfy $\sum_{t=1}^T y_t = \sum_{t=1}^T \tilde{y}_t$, to make the first term drop out.

Note then that for the second term we can rewrite the fraction, ignoring the superscript M , as

$$\prod_{t=1}^T \frac{1 + \exp(X'_t \beta + \tilde{y}_{t-1} \gamma + A)}{1 + \exp(X'_t \beta + y_{t-1} \gamma + A)} = \prod_{t=1}^T \frac{1 + \exp(X'_t \beta + \tilde{y}_{t-1} \gamma) \exp(A)}{1 + \exp(X'_t \beta + y_{t-1} \gamma) \exp(A)}.$$

These terms will only cancel out if the sequence $\{\exp(X'_t \beta + \tilde{y}_{t-1} \gamma)\}_{t=1}^T$ is a permutation of the sequence $\{\exp(X'_t \beta + y_{t-1} \gamma)\}_{t=1}^T$, or equivalently if $\{X'_t \beta + \tilde{y}_{t-1} \gamma\}_{t=1}^T$ is a permutation of $\{X'_t \beta + y_{t-1} \gamma\}_{t=1}^T$ (Lemma 3.1). If this has to be true for all $(X'_0 \beta, X'_1 \beta, \dots, X'_T \beta)$ in some open ball, the only permissible permutation is the trivial (identity) permutation. $\square$

Proof of Proposition 3.4. Let $y = (y_{1-j}, y_{2-j}, \dots, y_T)$ and $\tilde{y} = (\tilde{y}_{1-j}, \tilde{y}_{2-j}, \dots, \tilde{y}_T)$ denote two outcome sequences, and consider the ratio

$$\begin{aligned} & \frac{P(Y = y \mid A, Y_0, Y_{-1}, \dots, Y_{J-1})}{P(Y = \tilde{y} \mid A, Y_0, Y_{-1}, \dots, Y_{J-1})} \\ &= \prod_{t=1}^T \frac{\binom{M}{y_t} \exp(y'_{t-} \gamma + A)^{y_t} (1 + \exp(\tilde{y}'_{t-} \gamma + A))^M}{\binom{M}{\tilde{y}_t} \exp(\tilde{y}'_{t-} \gamma + A)^{\tilde{y}_t} (1 + \exp(y'_{t-} \gamma + A))^M} \\ &= \exp \left(A \left(\sum_{t=1}^T y_t - \sum_{t=1}^T \tilde{y}_t \right) \right) \prod_{t=1}^T \frac{(1 + \exp(\tilde{y}'_{t-} \gamma + A))^M}{(1 + \exp(y'_{t-} \gamma + A))^M} C_t \end{aligned}$$

where

$$C_t = \frac{\binom{M}{y_t} \exp(y_t y'_{t-} \gamma)}{\binom{M}{\tilde{y}_t} \exp(\tilde{y}_t \tilde{y}'_{t-} \gamma)}$$

The first term cancels if $\sum_{t=1}^T y_t = \sum_{t=1}^T \tilde{y}_t$. Note then that for the second term we can rewrite the fraction, ignoring the superscript M , as

$$\prod_{t=1}^T \frac{1 + \exp(\tilde{y}'_{t-} \gamma + A)}{1 + \exp(y'_{t-} \gamma + A)} = \prod_{t=1}^T \frac{1 + \exp(\tilde{y}'_{t-} \gamma) \exp(A)}{1 + \exp(y'_{t-} \gamma) \exp(A)}$$

and these terms cancel out if the sequence $\{\exp(\tilde{y}'_{t-} \gamma)\}_{t=1}^T$ is a permutation of the sequence $\{\exp(y'_{t-} \gamma)\}_{t=1}^T$ or, equivalently, if $\{\tilde{y}'_{t-} \gamma\}_{t=1}^T$ is a permutation of $\{y'_{t-} \gamma\}_{t=1}^T$ (see Lemma 3.1). The statistic in Proposition 3.4 is equal for two outcome sequences when this is the case. $\square$

Proof of Proposition 3.5. Consider the pure AR(1) Poisson-Binomial model. The probability mass function for an outcome sequence $y = (y_1, \dots, y_T)$ conditional on y_0, A is given by

$$P(Y = y \mid y_0, A) = \prod_{t=1}^T \sum_{\left\{b \in \{0,1\}^M : \sum_m b_m = y_t\right\}} \prod_{m=1}^M \frac{\exp(b_m(y_{t-1}\gamma + A_m))}{1 + \exp(y_{t-1}\gamma + A_m)}$$

note first that the denominator can be rewritten as

$$\prod_{m=1}^M \prod_{k=1}^M (1 + \exp(k\gamma) \exp(A_m))^{\sum_{t=1}^T \mathbb{1}\{y_{t-1}=k\}}$$

making it clear that for the ratio of probabilities of two outcome sequences

$$\frac{P(Y = y \mid y_0, A)}{P(Y = \tilde{y} \mid \tilde{y}_0, A)}$$

to be constant in A , we need $\sum_{t=1}^T \mathbb{1}\{y_{t-1} = k\} = \sum_{t=1}^T \mathbb{1}\{\tilde{y}_{t-1} = k\}$ for all k .

Since $\sum_{m=1}^M b_m = y_t$, we can rewrite the numerator as

$$\prod_{t=1}^T \exp(y_t y_{t-1} \gamma) \prod_{t=1}^T \sum_{\left\{b \in \{0,1\}^M : \sum_m b_m = y_t\right\}} \exp(b_m A_m)$$

Let $\mathcal{M}_k$ denote the set of all subsets of size k generated from the set of indices $\{1, 2, \dots, M\}$, such that for e.g. $k = 2$

$$\mathcal{M}_2 = \{\{1, 2\}, \{1, 3\}, \dots, \{M-2, M-1\}, \{M-2, M\}, \{M-1, M\}\}$$

then we can rewrite

$$\prod_{t=1}^T \sum_{\left\{b \in \{0,1\}^M : \sum_m b_m = y_t\right\}} \exp(b_m A_m) = \prod_{t=1}^T \sum_{\mu \in \mathcal{M}_{y_t}} \exp\left(\sum_{m \in \mu} A_m\right)$$

as these get multiplied, the ordering of the sequence of y_t is unimportant, what matters is

$$\left\{ \sum_{t=1}^T \mathbb{1}\{Y_t = 0\}, \sum_{t=1}^T \mathbb{1}\{Y_t = 1\}, \dots, \sum_{t=1}^T \mathbb{1}\{Y_t = M\} \right\}.$$

combining this with the previously established conditioning requirements for the denominator, we get the sufficient statistic described in the proposition (see Remark 3.1). $\square$

3.C Moment Finding Procedure

This section outlines the procedure used to numerically evaluate the existence of moment functions for a subset of the outcome sequences. In most cases it is computationally intractable to symbolically solve for moment functions when searching over all outcome sequences, rendering it necessary to numerically search for sub-matrices with a non-zero nullity.

Consider the simple binomial model in (3.1). For a given T and we can construct a vector of probability for each outcome sequence, $p(X, Y_0, A; \beta, \gamma)$, so, e.g., the first two entries of p may correspond to outcome sequences $y = (0, 0, 0)$ and $y = (0, 0, 1)$.

Say that we are interested in investigating whether moment functions exist that depends only on a subset of outcomes of length k . We will therefore construct matrices where each of the k columns correspond to an outcome sequence and each of the $k + 1$ rows is differentiated only by having different values of the fixed effects. This ensures that vectors in the nullspace of the matrix are invariant to the choice of fixed effects.

We fix Y_0 , and proceed by drawing T random values to construct the X vector, as well as $k + 1$ random values for A and fix β, γ to take on some values. Each row in the probability matrix would correspond to the same k entries of the probability vector, differing only in their values for A . If the rank of the resulting matrix is then $r < k$, by the rank-nullity theorem the nullspace is non-trivial with dimension $k - r$ and vectors in the nullspace are potential moments.

If the numerical procedure finds a rank deficient matrix, we would then find the nullspace of this matrix symbolically in Mathematica afterwards. This is detailed below.

Numerical Procedure

We do the numerical calculations with higher precision floating point numbers to minimize false positives as symbolically solving for moments is computationally expensive.

There are $(M + 1)^T$ different outcome sequences. The first sequence, consisting of all zeros, and last sequence, consisting of all M s, are not informative of the parameters as these can be achieved by fixed effects diverging to either plus or negative infinity. We therefore drop the first and last column of the probability matrix

For $M = 2, T = 3$, solving symbolically for the nullspace of the probability matrix with even just $3^3 - 2 = 25$ columns and 26 rows is computationally too expensive. We therefore want to numerically check whether there are potential moments that rely only on a subset of the columns. This, in turn, becomes computationally expensive by the sheer number of subsets that one can create.

However many subsets of a given length share many of the same columns. For example, if considering subsets of length 4, the subsets with indices

$$[2, 3, 4, 5]$$

$$[2, 3, 4, 6]$$

$$[2, 3, 4, 7]$$

$$[2, 3, 4, 8]$$

have the columns with indices $[2, 3, 4]$ in common. It is therefore possible to construct the annihilator matrix pertaining to these three columns and in turn multiply with the fifth, sixth, seventh and eighth column respectively to check whether the product yields zero, in

which case that column is in the span of the sub-matrix and it is therefore rank deficient in which case we have found a potential moment.

In case we find a potential moment from a subsequence of length k , then there is no need to search over sequences that constitute a super-sequence of the potential moment, unless, of course, the goal is to enumerate *all* moment conditions. Specifically, if e.g. the subsequence $[2, 3, 4, 5]$ gives rise to a moment, then it is redundant to look at $[2, 3, 4, 5, 6]$, $[2, 3, 4, 5, 7]$, and $[2, 3, 4, 5, 8]$ as they are guaranteed to be rank-deficient.

Our values for fixed effects and explanatory variables are constructed as the ratio of two (random) primes with the sign chosen at random, in order to avoid unintended cancellations.

Analytical Procedure

Once we have determined a potential candidate moment numerically, we proceed to using a computer algebraic system such as Mathematica to finding the an analytical expression for the moment function. We do so by leaving the values for the parameters and the explanatory variables as symbols, but still drawing numerical values for the fixed effects, as the moments should not depend on these anyway, and solving with those in a numerical format is more rapid. Furthermore, it is advantageous to apply a log transformation of the coefficients to bring them out of the exponential function prior to solving.