Discussion Papers in Economics

DETERMINING WHICH TRADE AGREEMENT PROVISIONS MATTER FOR TRADE

By

Holger Breinlich
(University of Surrey),

Valentina Corradi
(NYU Abu Dhabi),

Nadia Rocha
(Inter American Development Bank),

Michele Ruta
(International Monetary Fund),

J.M.C. Santos Silva
(University of Surrey),

Tom Zylkin
(University of Richmond).

DP 04/26

School of Economics
University of Surrey
Guildford
Surrey GU2 7XH, UK
Telephone +44 (0)1483 689380
Facsimile +44 (0)1483 689548
Web https://www.surrey.ac.uk/school-economics
ISSN: 1749-5075

Determining which trade agreement provisions matter for trade*

Holger Breinlich† Valentina Corradi‡ Nadia Rocha§ Michele Ruta¶
J.M.C. Santos Silva‖ Tom Zylkin**

27 May 2026

Abstract

Modern trade agreements contain a large number of provisions besides tariff reductions, in areas as diverse as services trade, competition policy, trade-related investment measures, or public procurement. Existing research has struggled with overfitting and severe multicollinearity when trying to estimate the effects of these provisions on trade flows. In this paper, we build on recent developments in the machine learning and variable selection literature to develop data-driven methods for selecting the most important provisions and quantifying their impact on trade flows and apply them to a recent database with highly detailed information on the provisions included in trade agreements. We find that provisions related to technical barriers to trade, antidumping, competition policy, subsidies, trade facilitation, sanitary and phytosanitary measures, and export taxes are associated with enhancing the trade-increasing effects of trade agreements. Interestingly, we find that the majority of the 305 provision variables in our data have no measurable impact on goods trade, including virtually all provisions that fall within commonly included policy areas such as services, labor markets, and public procurement.

KEY WORDS: Deep Trade Agreements, Lasso, Machine Learning, Preferential Trade Agreements.
JEL CLASSIFICATION: F14, F15, F17, C52, C55.

*An earlier version of this paper was circulated with the title "Machine Learning in International Trade Research – Evaluating the Impact of Trade Agreements".

†University of Surrey, CEP, CEPR and CESifo. Email: h.breinlich@surrey.ac.uk

‡NYU Abu Dhabi. Email: vc2718@nyu.edu.

§Inter American Development Bank. Email: nadiaro@iadb.org.

¶International Monetary Fund. Email: mruta@imf.org.

‖University of Surrey. Email: jmcss@surrey.ac.uk.

**University of Richmond. Email: tzylkin@richmond.edu.

1 Introduction

International trade is of vital importance for modern economies, and governments around the world try to shape their countries' export and import patterns through numerous interventions. Given the increasing difficulties facing the World Trade Organization (WTO), the usual forum for multilateral trade negotiations, over the last several decades, countries have become increasingly reliant on preferential trade agreements (PTAs) involving only one or a small number of partners. At the same time, modern PTAs have increasingly shifted their focus from the reduction of tariffs to instead targeting non-tariff barriers and behind-the-border policies, and they now often explicitly include a number of provisions addressing issues such as differences in technical standards, competition regulations, and intellectual property rights.

Against this background, researchers and policy makers interested in the effects of trade agreements face difficult challenges. In particular, recent research has tried to move beyond estimating the overall impact of PTAs and to establish the relative importance of individual trade agreement provisions in determining an agreement's overall impact (e.g., Kohl, Brakman, and Garretsen, 2016, Mulabdic, Osnago, and Ruta, 2017, Dhingra, Freeman, and Mavroeidi, 2018, Regmi and Baier, 2020, and Falvey and Foster-McGregor, 2022). However, such attempts face the difficulty that the large number of provisions, and the fact that similar provisions appear in different trade agreements, create severe multicollinearity problems, which make it very difficult to identify the effects of individual PTA provisions. Traditional methods such as gravity regressions of trade flows on dummies for individual provisions are not able to deal with such multicollinearity. Instead, researchers have grouped or aggregated provisions in different ways. For example, Mattoo, Mulabdic, and Ruta (2017) use the count of provisions in an agreement as a measure of its depth, hence implicitly giving equal weight to each measure, whereas Dhingra, Freeman, and Mavroeidi (2018) overcome multicollinearity problems by grouping provisions into a small number of bundles.

In this paper, we provide more detailed evidence on which types of trade agreement provisions are most important for trade. We do so by using data-driven methods that build upon recent developments in the machine learning and variable selection literature. Specifically, we incorporate variable selection methods based on the lasso (Least Absolute Shrinkage and Selection Operator; see, e.g., Hastie, Tibshirani, and Friedman, 2009) into a panel data gravity model, considered standard in the analysis of how PTAs affect trade flows; see, e.g., Yotov, Piermartini, Monteiro, and Larch (2016) and Larch, Shikher, and Yotov (2025). By estimating gravity models with

high-dimensional fixed effects using a Poisson pseudo-maximum likelihood (PPML) estimator with an added lasso penalty on the provision-variable coefficients, we show that the trade-enhancing effects of PTAs are concentrated in only a small number of policy domains. In particular, we find that provisions related to technical barriers to trade, anti-dumping procedures, competition policy, subsidy policies, trade facilitation and customs procedures, sanitary and phytosanitary measures, and export taxes collectively have the largest impacts on trade, whereas provisions falling within domains less directly related to goods trade—such as services, labor markets, and public procurement—are much less likely to be selected by our methods and have a much smaller aggregate trade impact. Broadly, our findings therefore suggest that what matters for trade is not simply how many provisions a PTA contains, but rather which policy areas those provisions address.

To carry out our analysis, we use a high-resolution data set on the provision content of PTAs recently made available by the World Bank (Mattoo, Rocha and Ruta, 2020). Importantly, this database is very detailed, to the point where the number of provision variables we consider is larger than the number of PTAs we observe in our data. In addition, in many cases groups of provisions, even seemingly unrelated provisions, appear together with high frequency due to common templates and other similarities across PTAs. These features of the data are crucial for understanding how to frame our results: even with lasso-based methods, it is not possible for us to identify the impact of each specific provision in the presence of such strong multicollinearity. Nonetheless, our methods are still able to provide some clear answers to the question of which provisions matter for trade that are new to the literature. In particular, we find that the majority of provision types (more than two-thirds of the 305 provision variables in our data) have no measurable impact on trade. Moreover, we do not find any evidence that some provisions reduce trade while others increase it; a small number collectively increase it while most have little to no measurable effect.

The starting point for our empirical approach is a standard gravity specification with exporter-time, importer-time, and country-pair fixed effects, following established best practices in the empirical trade literature for addressing multilateral resistance and pair-specific heterogeneity. To address the high dimensionality of our provision variables, we augment the PPML estimator normally used in this setting with a lasso penalty that selects the subset of provision variables that are most strongly associated with trade flows. Since implementing the lasso requires careful consideration of how to choose the parameters that govern the penalty term, we consider two main approaches: cross-validation, which involves maximizing out-of-sample fit, and the plug-in (or theory-driven) approach of Belloni, Chernozhukov, Hansen, and Kozbur (2016), which has the important advantage

that it accounts for heteroskedasticity and clustered errors.[1] Because of the high multicollinearity in our data, we find that the number of provisions selected when using the PPML-lasso with the tuning parameter chosen by cross-validation is too large for the model to have a meaningful interpretation and that, in contrast, the number of provisions identified when using the plug-in penalty is too small to allow us to be confident that it includes the majority of relevant provisions.[2]

To address these issues, we introduce two additional methods that seek to identify potentially important variables that may have been missed in an initial lasso step based on the plug-in penalty. Our preferred approach, which we call the bootstrap lasso, augments the set of variables selected by the plug-in lasso with the variables selected with sufficiently high frequency when the plug-in lasso is applied to many bootstrap samples. In the spirit of both Meinshausen and Bühlmann (2010) and Wang et al (2011), who each apply a similar principle to the standard lasso, this type of resampling procedure exploits the instability of variable selection in finite samples and thereby allows us to identify additional provisions that can act as close substitutes for the ones selected using the original sample. As a complementary approach, we also propose what we call the iceberg lasso, which involves regressing each of the provisions selected by the plug-in lasso on all other provisions. The advantage of the latter method is that it explicitly aims to detect relevant variables that were missed in the initial step due to collinearity. However, unlike the bootstrap lasso, it does not offer information about the stability of selection or about the combined effects of groups of selected regressors. The bootstrap lasso can, for example, recover the aggregate effects of the selected provisions that fall within each of the provision classifications used by Mattoo, Rocha and Ruta (2020), which serve as the categories emphasized in our main results.

Our work contributes to several different literatures. Most directly, we contribute to the large and growing literature on the effects of PTAs on trade flows. As previously discussed, recently this literature has tried to decompose the overall PTA effect by disentangling the effects of individual trade agreement provisions, but problems related to multicollinearity and high dimensionality have led researchers to group or aggregate provisions before estimation. By contrast, our approach

[1] An R package (penppml) implementing penalized PPML regressions with high-dimensional xed effects is available from CRAN and can be installed with install.packages("penppml"). For more details see https://github.com/tomzylkin/penppml.

[2] Our simulation results in Section 4 suggest that the lasso with a penalty parameter chosen by the plug-in method often fails to select the relevant regressors. A similar result, in a different context, is reported by Wüthrich and Zhu (2021).

allows us to select the most important provisions and to quantify their impact on trade flows while avoiding the need to make essentially arbitrary assumptions about how to aggregate individual provisions (see Mattoo, Mulabdic, and Ruta, 2017; Dhingra, Freeman, and Mavroeidi, 2018).

In addition, we contribute to a small existing literature that has used machine learning and other related methods to study the effects of trade agreements in a gravity context. For example, Regmi and Baier (2020) use an unsupervised learning method to group PTAs by textual similarity, so as to provide a more nuanced notion of PTA depth. Following from a similar motivation, Hofmann, Osnago, and Ruta (2017) propose an earlier depth measure for PTAs based on principal components analysis applied to their provisions data. In contrast, Baier, Yotov, and Zylkin (2019) use a two-step methodology where pair-specific PTA effects are estimated in a first stage and then predicted out of sample using country- and pair-specific variables.

Finally, we make a contribution to the methodological literature interested in variable selection. In particular, we extend and adapt existing work by Belloni, Chernozhukov, Hansen, and Kozbur (2016) on the use of the lasso in the presence of heteroskedasticity and clustered errors to make it applicable to the context of international trade flows and trade agreements. This requires an extension of their original method to the estimation of nonlinear models with high-dimensional fixed effects using PPML. The bootstrap lasso and iceberg lasso that we propose build on the results obtained using the plug-in penalty and identify additional sets of provisions that may have a causal effect on trade. Both methods add to the information provided by the standard lasso approaches and, as illustrated in our simulations, are better able to identify the provisions that have a causal effect.

The rest of this paper is structured as follows. Section 2 presents the data on PTA provisions and provides a descriptive analysis of these data, highlighting a number of stylized facts about the provisions present in recent trade agreements. Section 3 introduces the variable selection problem in the three-way gravity model context and explains how we implement PPML-lasso estimation with high-dimensional fixed effects. Section 4 presents the results of a simulation study comparing the relative performance of different lasso methods in a simplified setting with high correlation between regressors. Section 5 applies our methods to our database on PTA provisions and shows which provisions have the strongest impact on trade flows. Section 6 concludes and technical details are gathered in an Appendix.

2 Data

Our analysis combines data on international trade flows from Comtrade with the database on the content of PTAs that has been collected by Mattoo, Rocha and Ruta (2020). On trade, we use four-yearly merchandise exports between 1964 and 2016 from 220 exporters to 270 importers. Flows for country pairs without export information are considered as zeros. The database on the content of trade agreements includes information on 282 PTAs that have been signed and notified to the WTO between 1958 and 2017. The data focus on the sub-sample of 17 policy areas that are most frequently covered in the trade agreements mapped in Hofmann, Osnago, and Ruta (2017). These policy areas range from environmental laws and labor market regulations, that are covered in roughly 20 percent of the PTAs, to areas such as rules of origin and trade facilitation that are present in over 80 percent of the agreements (see Figure 1).[3]

For each agreement and policy area, the database provides granular information on the specific provisions covering stated objectives and substantive commitments, as well as aspects relating to transparency, procedures and enforcement. The coding exercise focuses on the legal text of the agreements and therefore excludes information on the actual implementation of the commitments included in the agreements.[4]

To alleviate the problems caused by the high dimensionality of the data and the high level of correlation across the provisions included in the agreements, the analysis presented in this paper focuses on the sub-set of *essential* provisions. This includes the set of substantive provisions (those that require specific integration/liberalization commitments and obligations) plus the disciplines among procedures, transparency, enforcement or objectives, which are viewed as indispensable and complementary to achieving the substantive commitments. Non-essential provisions are referred to as *corollary*.[5] The share of essential provisions in the total number of provisions included in an agreement ranges from less than 10 percent for public procurement, movement of capital and visa and asylum, to more than 50 percent for policy areas such as environmental laws and labor market

[3] All data is publicly available and can also be obtained from the authors upon request.

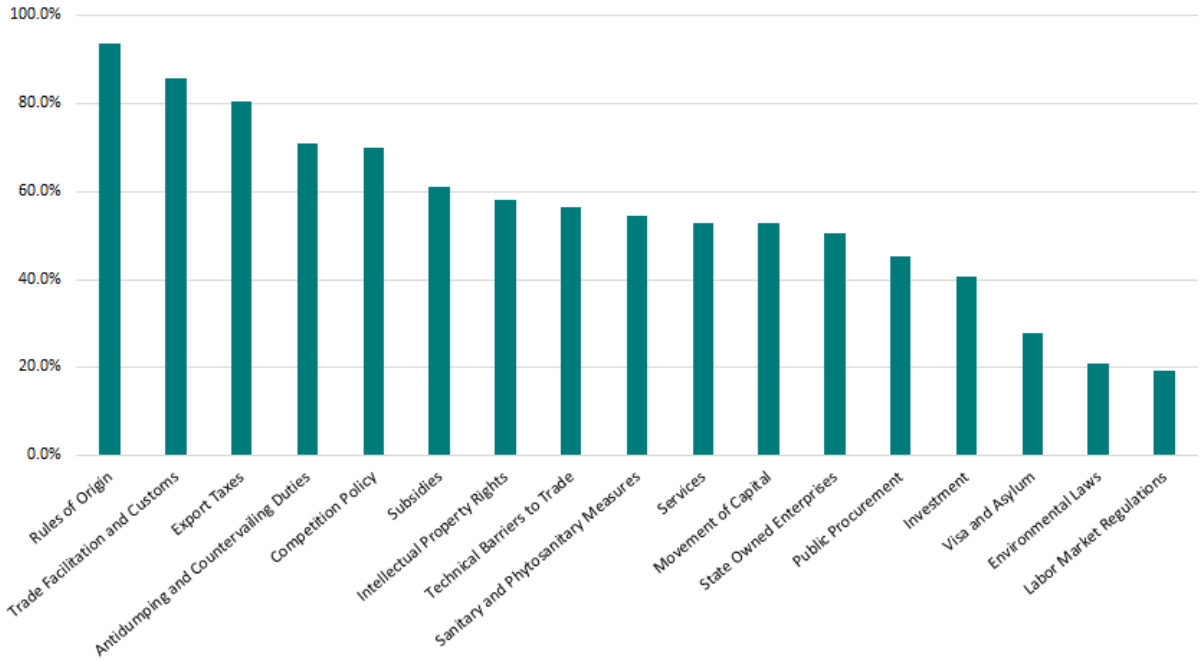
[4] In this data set, information coming from secondary law (the body of law that derives from the principles and objectives of the treaties) has not been coded. This is of particular importance for agreements such as the EU, since most policy areas covered have used secondary law such as regulations, directives, and other legal instruments to pursue integration.

[5] The classification into essential and corollary in the database is based on experts' knowledge and, hence, has an element of subjectivity.

regulations. Overall, the sub-set of essential provisions represents almost one-third (305/937) of the total number of provisions coded in this exercise (see Table 1).

One important caveat regarding this data set is that it does not cover all of the trade agreements that have been in force during the period under study. Specifically, our information on provisions is limited to agreements that are currently in effect, i.e., excluding any agreements that are no longer in effect. For this reason, we drop observations associated with agreements no longer in effect. This means that the effects of newer agreements are identified by changes in trade relative to when that pair did not have any agreement rather than relative to pre-existing agreements. The majority of the observations that are dropped are due to pre-accession agreements that new European Union (EU) members sign before joining the EU. Thus, to use one of these cases as an example, Italy-Croatia is included in our data for years 1992-2000 (after Croatian independence and before the initial EU-Croatia PTA in 2001) and for year 2016 (after Croatia joins the EU in 2013). For this reason, in our analysis we experiment with treating the EU differently, as discussed further in Section 5. To identify agreements no longer in effect, we consult the NSF-Kellogg database created by Jeff Bergstrand and Scott Baier crosschecked with data from the WTO. The EU and the earlier European Community are treated as the same agreement for these purposes, though it is allowed to evolve as new provisions are added.

Figure 1: Share of PTAs that cover selected policy areas



Note: Figure shows the share of PTAs that cover a policy area. Source: Mattoo, Rocha and Ruta (2020).

Table 1: Distribution of essential provisions by policy area

Policy Area	Number of provisions	Number of Essential provisions	Share
Anti-dumping and Countervailing Duties	53	11	28.8%
Competition Policy	35	14	40.0%
Environmental Laws	48	27	56.3%
Export Taxes	46	23	50.0%
Intellectual Property Rights	120	67	55.8%
Investment	57	15	26.3%
Labor Market Regulations	18	12	66.7%
Movement of Capital	94	8	8.5%
Public Procurement	100	5	5.0%
Rules of Origin	38	19	50.0%
Sanitary and Phytosanitary Measures	59	24	40.7%
Services	64	21	32.8%
State-Owned Enterprises	53	13	24.5%
Subsidies	36	13	36.1%
Technical Barriers to Trade	34	19	55.9%
Trade Facilitation and Customs	52	11	21.2%
Visa and Asylum	30	3	10.0%
Total	937	305	32.6%

3 Determining Which Provisions Matter for Trade

We now outline the methodology we use to identify which PTA provisions are most likely to have an impact on bilateral trade. To preview our approach, we will first specify a typical panel data gravity model for trade flows. Following the latest recommendations from the methodological literature (Yotov, Piermartini, Monteiro, and Larch, 2016, Weidner and Zylkin, 2021, Larch, Shikher, and Yotov, 2025), we will use a multiplicative model where expected trade flows are given by an exponential function of our covariates of interest plus three sets of fixed effects. Drawing on this standard framework, we will then consider the estimation challenges that arise when the number of covariates (here, provision variables) is allowed to be very large. As we will discuss, it will be convenient to reformulate the usual estimation problem as a "variable selection" problem, where we suppose that many of the provisions have zero or approximately zero effect.

Bringing together these elements will require that we extend recent computational advances in high-dimensional fixed effects estimation to incorporate lasso and lasso-type penalties. It will also require that we introduce our own innovations, the bootstrap lasso and iceberg lasso methods, which we will motivate as providing a balance between cross-validation approaches that tend to select too many variables and more parsimonious plug-in methods that may select too few.

3.1 The Gravity Model

Our starting point for estimation is the following multiplicative gravity model:

$$\mu_{ijt} := \mathrm{E}(y_{ijt}|x_{ijt}, \alpha_{it}, \gamma_{jt}, \eta_{ij}) = \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij}). \tag{1}$$

Here, i, j, and t respectively index exporter, importer, and time. Bilateral trade flows from exporter i to importer j at time t are therefore given by y_{ijt}, x_{ijt} are our covariates of interest, and α_{it}, γ_{jt}, and η_{ij} are, respectively, exporter-time, importer-time, and exporter-importer ("pair") fixed effects.

Because of the three fixed effects, the model in (1) is often called the "three-way gravity model". Intuitively, the exporter-time and importer-time fixed effects α_{it} and γ_{jt} may be thought of as controlling for changes over time in the "gravitational pull" that the exporter and importer each exert on world trade flows. More formally, these fixed effects can be shown to depend on the market sizes of the two countries as well as on what Anderson and van Wincoop (2003) call "multilateral resistance", a theoretical measure of each country's connectedness to the overall trade network. A variety of trade models can be shown to give rise to these fixed effects; see Head and Mayer (2014) for a comprehensive review. The inclusion of the pair fixed effect η_{ij} was suggested by Baier and Bergstrand (2007), who convincingly argue that estimates of the effect of trade agreements and other similar variables would otherwise be biased due to omitted cross-sectional heterogeneity. In terms of a trade model, this omitted heterogeneity is often motivated as coming from unobserved trade costs.

An important point about (1) is that it motivates estimating the model in its original nonlinear form using PPML; see Gourieroux, Monfort and Trognon (1984). In principle, one could instead use a linear model after taking logs, but Santos Silva and Tenreyro (2006) have pointed out that this estimator is generally inconsistent and recommend that (1) should instead be estimated by PPML. Though the resulting model is nonlinear with three sets of high-dimensional fixed effects, estimation is feasible due to recent computation innovations by Correia, Guimarães, and Zylkin

(2020) and others.[6] Weidner and Zylkin (2021) have recently established the consistency and asymptotic distribution of the three-way PPML estimator, and Yotov, Piermartini, Monteiro, and Larch (2016) and Larch, Shikher, and Yotov (2025) recommend it as the workhorse method for estimating the effects of trade policies. It is frequently applied in the context of trade agreements in particular.

Having established these details, our focus is on the set of covariates, x_{ijt}. In most applications in the trade agreements literature, x_{ijt} is often either a single variable—i.e., a dummy for the presence of a trade agreement—or minor variants thereof, such as interactions of a trade agreement dummy with either the depth of the agreement or the bilateral characteristics of the two countries (Baier, Bergstrand, and Feng, 2014; Baier, Bergstrand, and Clance, 2018). Instead of using a single PTA dummy, we include 305 dummies, one for each of the provisions contained in our data. Therefore, a major estimation challenge that arises in our setting is that we must treat the number of provisions as being very large. As we show in Appendices D and E, in our data set this high dimensionality, combined with the relatively small number of PTAs, creates strong multicollinearity that results in implausibly large and uninterpretable estimates when a standard estimator is used. Furthermore, the estimated model has poor predictive performance due to overfitting. We therefore must discuss how the standard gravity estimation approach must be modified in order to deal with this additional source of high dimensionality.

3.2 Variable Selection and Gravity

The starting point for our methodological innovation is the fact that it is very likely that some provision variables will have a causal effect on trade flows, but the number of such provisions is small. To be more precise, we have $p = 305$ essential provision variables, coded as dummies, of which a subset $0 < s < p$ are assumed to have non-zero effects, where s is typically small with respect to the sample size.[7] We do not know s beforehand, nor do we know the identities of any of

[6] Correia, Guimarães, and Zylkin (2020) and Stammann (2018) have each proposed algorithms for estimating nonlinear fixed effects models based on iteratively re-weighted least squares (IRLS). Heuristically, this type of algorithm exploits the linearity of the weighted least squares step in the IRLS algorithm to wipe out the fixed effects in each iteration, then uses an application of the Frisch-Waugh-Lovell theorem to update the weights, repeating until convergence. For a different approach, see Larch, Wanner, Yotov, and Zylkin (2019).

[7] Note that of the 305 provisions in our data, 8 are always equal to zero. Therefore, the effective number of provisions we consider is 297.

the s provisions that substantively affect trade. Our goal then is to use statistical methods along with the model described in (1) in order to identify these provisions.

Because of the high dimensionality of x_{ijt}, experimenting with different subsets of provisions to see which has the best performance is unlikely to be fruitful. Instead, we adopt a penalized regression (or "regularization") approach that involves appending a penalty term to the Poisson pseudo-likelihood one would use to estimate the unpenalized gravity model. The idea is that the penalty term "shrinks" all estimated coefficients towards zero and forces some of them to be exactly equal to zero. The higher the penalty, the fewer the variables that are found to have non-zero coefficients and are therefore "selected". By design, the variables that are selected should be those that exert the strongest influence on the fit of the model; coefficients for variables that are not as relevant should end up getting shrunken to zero completely.

Because of its computational feasibility, the most frequently used approach to this type of variable selection problem is the lasso, introduced by Tibshirani (1996). In our setting, the penalized objective function that defines the three-way PPML-lasso is

$$\mathcal{PL}(\beta,\alpha,\gamma,\eta) = \underbrace{\frac{1}{n}\left(\sum_{i,j,t}\left(\mu_{ijt} - y_{ijt}\ln\mu_{ijt}\right)\right)}_{-1\times\text{PPML pseudo likelihood}} + \underbrace{\frac{1}{n}\sum_{k=1}^{p}\widehat{\phi}_k\lambda|\beta_k|}_{\text{Lasso penalty}}, \tag{2}$$

where n is the number of observations,[8] as in (1) above, $\mu_{ijt} = e^{\alpha_{it}+\gamma_{jt}+\eta_{ij}+x'_{ijt}\beta}$ is the conditional mean, and $\lambda \geq 0$ and $\widehat{\phi}_k \geq 0$ are tuning parameters that determine the penalty. As indicated in (2), the first term in this expression is the standard PPML objective function one would minimize in order to estimate the three-way gravity model. Thus, the PPML-lasso nests PPML as a special case when λ is set to zero.

The second term in (2) is a modified lasso penalty that allows for regressor-specific penalty weights as opposed to having λ as the only tuning parameter as in the standard lasso. Intuitively, larger penalties increasingly shrink the estimated β-coefficients towards zero. The coefficients for any variables that do not sufficiently increase the likelihood are set to exactly zero, thereby giving us a way of identifying which variables to include in the final model. For some illustration, if we choose a large enough λ, the only way to minimize $\mathcal{PL}$ is to set all $\widehat{\beta}_k$s equal to zero, meaning that no variables are selected. As in Belloni, Chernozhukov, Hansen, and Kozbur (2016), we will use the regressor-specific $\widehat{\phi}_k$ penalty terms, which are explained in more detail below, to iteratively refine

[8] Naturally, the number of observations will depend on the number of countries for which we have data and on the number of years we observe them. For simplicity, we do not make that relation explicit.

the model while also reflecting any heteroskedasticity and within-cluster correlation featured in the data.

Importantly, the fixed effects parameters α, γ, and η are not penalized because there is no reason to believe that most of these parameters are actually zero. In addition, it turns out they do not pose special issues for computation. This is because the estimation of the fixed effects does not depend directly on the penalty. As such, for any given β, the fixed effects estimates can be obtained by solving their usual PPML first-order conditions from the standard unpenalized regression approach. In practice, this means that the fixed effects can actually be dealt with in the exact same manner as in Correia, Guimarães, and Zylkin (2020). More details on the computational methods are provided in Appendix B, but, basically, we use the original HDFE-IRLS algorithm of Correia, Guimarães, and Zylkin (2020) to take care of the fixed effects but replace the weighted linear regression step from that algorithm with a weighted lasso regression.[9]

3.3 Implementing the Lasso

The next question of course is how to determine the tuning parameters λ and $\widehat{\phi}_k$. As a starting point, we will examine the traditional cross-validation approach and the plug-in lasso of Belloni, Chernozhukov, Hansen, and Kozbur (2016), both of which we have modified to fit the demands of the three-way PPML setting. As we will discuss, each of these methods has its strengths and weaknesses. Therefore, we will then turn to describing two extensions of the plug-in lasso, which we call the "bootstrap lasso" and "iceberg lasso", that are intended to address one of the plug-in lasso's key shortcomings in this context. The performance of the proposed methods is evaluated in the next section using a simulation experiment. Appendices B and C present additional details on their implementation and properties, including the proof that our proposed plug-in estimator has a near-oracle property in spite of the fixed effects.[10]

[9] For the lasso regression step, we use the coordinate descent algorithm of Friedman, Hastie, and Tibshirani (2010).

[10] The oracle property of estimators such as the adaptive lasso of Zou (2006) refers to their ability to correctly recover which parameters are zero and non-zero in a setting where the number of potential regressors is fixed and the number of observations is large. The near-oracle property of the plug-in lasso is similar, but its rate of convergence is slower and depends on the number of potential regressors because in the setting considered by Belloni, Chen, Chernozhukov, and Hansen (2012) the number of potential regressors is allowed to grow with the sample size.

Plug-in Lasso

The plug-in lasso is so-named because it specifies appropriate functional forms for the penalty parameters based on statistical theory and then uses plug-in estimates for these parameters. It is therefore a "theory-driven" approach to the variable selection problem, whereas cross-validation, discussed next, is a more traditional machine learning method that relies on out-of-sample prediction. The plug-in lasso was first proposed by Belloni, Chen, Chernozhukov, and Hansen (2012), though the specific implementation we build on is the "cluster lasso" method of Belloni, Chernozhukov, Hansen, and Kozbur (2016), which allows for correlated errors within clusters.

Without delving too much into technical details, which we defer to Appendices B and C, variable selection using the plug-in lasso can be thought of as involving the following three ingredients:

i. The absolute value of the score for each β_k when evaluated at 0,

ii. The standard error of the score for each β_k, i.e., $\widehat{\phi}_k$,

iii. A value for λ such that the absolute value of the score for β_k must be large relative to its standard error (as captured by $\widehat{\phi}_k$) in order for regressor $x_{ijt,k}$ to be selected.

Intuitively, the value of the score reflects the impact that a small change in β_k has on the fit of the model. When evaluated at 0, it tells us how much the fit of the model improves when we make β_k non-zero. The standard logic of the lasso is that this improvement in fit must be large relative to the penalty in order for $\widehat{\beta}_k$ to be non-zero. One of the main innovations of the plug-in lasso is to allow the regressor-specific penalty $\widehat{\phi}_k$ to reflect the standard error of the score.[11] This way, we counteract the possibility that regressors could be mistakenly selected due to estimation noise rather than because of their true impact on the model. These regressor-specific penalties play an important role in the presence of heteroskedasticity, which of course is an important feature of trade data. Because the provision sets in x_{ijt} vary by agreement, and because we expect errors to be serially correlated over time, we use the cluster lasso approach to constructing these weights as in Belloni, Chernozhukov, Hansen, and Kozbur (2016). Specifically, we cluster all observations belonging to pairs that form agreements by the agreement they eventually belong to, including before the agreement begins. Other observations are clustered by pair.

[11] The standard error of the score for each β_k is defined as the square root of its sample variance. It is closely related to the heteroskedasticity correction that would normally be used for computing a cluster-robust sandwich estimator for the standard errors of the coefficient estimates. See Appendix B for more details.

An important advantage of the plug-in lasso is that it is very parsimonious in terms of the number of variables it selects. As shown by Drukker and Liu (2019), the plug-in method offers superior performance versus cross-validation approaches in finite samples, in large part because these other methods tend to select too many variables. Furthermore, Theorem 1 in Appendix C proves that the "post-lasso" estimates obtained using unpenalized PPML on the covariates selected by our implementation of the plug-in lasso have a near-oracle property, in the sense that the L_1 norm of estimation error, for the selected variables, goes to zero at a sufficiently fast rate (see also Belloni, Chen, Chernozhukov, and Hansen, 2012).[12]

However, the plug-in lasso's parsimony can also be a weakness in that it may select too few variables. In general, it attempts to select a small number of variables that are most useful for predicting the outcome. However, in data settings where there is a substantial number of regressors that are highly correlated, as is the case with our provisions data, it is possible that the plug-in lasso will wrongly select a regressor that does not affect the outcome but is strongly correlated with another regressor that does, since either can have similar predictive value for fitting the model. We discuss this issue in more detail when we introduce our extensions of the plug-in lasso.

Cross-Validation

As an alternative to the plug-in method, we also consider a more traditional approach based on cross-validation. Under cross-validation, one repeatedly holds out some of the data and chooses λ in order to maximize the predictive performance of the model when evaluated on the held-out data. The regressor-specific $\widehat{\phi}_k$ do not play a role and are set equal to 1.

Because of the size of the data and the nature of our model, implementing this approach presents some interesting challenges. A standard implementation would be a "k-fold" approach that randomly partitions the sample into k folds and then uses $k-1$ subsets to estimate the parameters and the excluded subset to evaluate the predictive ability of the model. To adapt this idea to our setting, we validate our model by repeatedly dropping the observations corresponding to randomly selected groups of agreements in our data when they are in effect, and then use their provisions to predict trade for the dropped observations, similar to the approach taken by Baier, Yotov, and Zylkin (2019). In this case, all fixed effects are always present in each practice sample,

[12] Specifically, we show that, despite the nonlinearity of the model, the bias of the three-way PPML estimator documented in Weidner and Zylkin (2021) does not alter the usual near-oracle convergence rate of the plug-in lasso method.

so that we can always form the necessary predictions for the omitted trade flows associated with the PTA that have been dropped.[13]

The main advantage of cross-validation is that it is explicitly designed to optimize predictive performance. Thus, it may offer a conceptual advantage where forecasting tasks are concerned. However, a known weakness of the standard lasso with cross-validation is that it often errs on the side of selecting too many variables that are not relevant.[14] Furthermore, it does not take into account heteroskedasticity when performing the selection, and it generally does not have either an oracle or near-oracle property in large samples. For these reasons, cross-validation is not our preferred method for answering the question of which provisions matter for trade; we consider it mainly to illustrate the basic mechanics of the lasso and as a check on our plug-in results.[15]

Extensions of the plug-in lasso

One important feature of the lasso is that it selects variables that are good predictors of the outcome, but these are not necessarily variables that have a causal impact on the outcome and may just be correlated with regressors or with unobserved factors that have a causal impact. Indeed, Zhao and Yu (2006) show that only when the so-called "irrepresentability condition" is valid can we expect the variables selected by lasso to have a causal interpretation; the condition essentially imposes limits on the degree of collinearity between the variables with a causal effect on the outcome and the candidate regressors with no causal effect (see also Wainwright, 2009).

As we have noted, in the case of our data set, there is a very high degree of collinearity between some of the variables, and therefore we cannot expect the irrepresentability condition to hold.

[13] It may, however, happen that some provisions are not included in the agreements used in the estimation sample. This is less likely to happen if k is large, and therefore we use $k = 25$.

[14] In linear models, tuning λ using cross-validation is analogous to selection based on the Akaike information criterion, which ensures that the probability of selecting too few variables goes to zero but does not eliminate the possibility of selecting too many. Relatedly, Drukker and Liu (2019) find that selecting λ using cross-validation also leads to the inclusion of too many regressors in Poisson regressions. In our own application, we too find that the cross-validation method selects many more provisions than the plug-in method.

[15] Alternatively, we could consider the adaptive lasso (Zou, 2006), which adds a second tuning parameter and is known to deliver consistent variable selection, i.e, it has the oracle property. However, in our application we have found that the adaptive lasso is similar to the standard cross-validation lasso in that it is much too lenient and it keeps too many regressors that are not relevant. The simulations reported in the next section suggest that this is likely to be the case in relatively small samples.

Furthermore, for the plug-in lasso especially, which tends to select a very parsimonious model, we should be worried whether the selected provisions mask the effects of a potentially more complex set of other provisions that are often included in the same agreements as the provisions that are selected.

To address this complication, we consider two extensions of the plug-in lasso: the bootstrap lasso and the iceberg lasso. Both approaches expand the initial set of plug-in-selected provisions by adding variables that may help predict trade but are omitted by the more parsimonious plug-in method. The motivation in each case is that, although the reverse is not true, variables with a significant causal impact should be good predictors of the outcome. Therefore, expanding the selected set in this way may help uncover additional relevant provisions whose effects are otherwise masked by their collinearity with the variables selected in the first step.

The Bootstrap Lasso It is well documented that in small to moderate samples the set of variables selected by the lasso can be somewhat unstable, in the sense that it is very sensitive to perturbations of the sample (see, e.g., Meinshausen and Bühlmann, 2010; Mullainathan and Spiess, 2017). We use this feature of the lasso to try to alleviate the tendency of the plug-in lasso to select too few variables. In what we call the bootstrap lasso, we apply the plug-in lasso to an additional set of $B-1$ samples obtained by bootstrap, and define the set of variables selected by this method as the variables that are most frequently selected in the B samples considered. Doing so has several conceptual benefits.

First, because this method is likely to uncover variables that substitute for the originally selected variables in approximating the patterns found in the data in different versions of the sample, the augmented set of variables it selects is likely to contain more of the relevant variables than the initial set selected by the plug-in lasso. Second, the frequency with which each variable is selected provides useful information about the stability of its selection and thus the degree of confidence we should have in its importance to the model. Third, when multiple provisions from a natural grouping of provisions, such as those from a common policy area, are selected with low to moderate frequency, the post-lasso estimates from each bootstrap sample can be aggregated across samples to obtain an estimate of the trade effect associated with the provision grouping as a whole. This feature

provides a feasible way of ranking, for example, how provisions on trade facilitation collectively affect trade as compared to provisions relating to investment or services.[16]

Naturally, the performance of the bootstrap lasso will depend on B and on the frequency cutoff used to select the variables, with lower cutoffs increasing the probability of selecting relevant variables but also the number of irrelevant variables included in the model. In our application, we use $B = 250$ and restrict our attention to variables that are selected with a frequency exceeding 5% or 1%.[17]

The Iceberg Lasso Simply put, the iceberg lasso involves performing a subsequent set of plug-in lasso regressions in which each of the provisions selected by the plug-in lasso estimator is regressed on all of the provisions that were excluded; the set of variables selected by the iceberg lasso is the union of the set selected in the first step with the sets selected in each of the regressions of the second step. The purpose of the second-step regressions is to identify bundles of provisions that are highly correlated with the ones selected in the first step, and therefore may be representable by them, in the sense of Zhao and Yu (2006). That is, each of the variables selected by the PPML-lasso with the plug-in tuning parameter may be just "the tip of the iceberg" of a bundle of variables that have a causal impact on trade, and the lasso regressions in the second step is designed to identify these bundles.[18]

3.4 Discussion and caveats

Having described the ideas behind our methods, several further caveats are in order. First, by construction, not all of the provisions selected by lasso-based methods can be said to have causal

[16] This averaging is related to the idea of bootstrap aggregation, or "bagging," in the machine learning literature; see, e.g., Hastie, Tibshirani, and Friedman (2009). As shown in the Appendix, averaging estimates and predictions across bootstrap samples may also reduce overfitting by averaging out some of the sampling error in the original data.

[17] In the simulations reported in Section 4 we use $B = 20$ and use only the 5% cutoff.

[18] As such, the iceberg lasso may be interpreted as a data-driven alternative to the method used in Dhingra, Freeman, and Mavroeidi (2018) to construct provision bundles. The iceberg lasso also complements the approach adopted by Regmi and Baier (2020), who use machine learning tools to construct groups of provisions and then use these clusters in a gravity equation. The main difference between the two approaches is that Regmi and Baier (2020) use what is called an unsupervised machine learning method, which uses only information on the provisions to form the clusters. In contrast, the iceberg lasso selects the provisions using a supervised method that considers the impact of the provisions on trade, and then adds another step which can be interpreted as unsupervised learning.

effects. Whether or not the plug-in lasso and the extensions we introduce are more informative than other methods that are already known to over-select regressors (such as the cross-validation approach described earlier) is an empirical matter and the answer will depend on the application. Second, the causal interpretation of our results depends on the maintained assumption that the three-way gravity model isolates the relevant variation in trade flows. Under this assumption, and provided that some provisions do in fact affect trade, our approach is designed to select a group of variables that is likely to contain the provisions with causal effects. The three-way gravity model has the considerable advantage that it isolates a particular variation in the data that is empirically relevant for the study of trade agreements, namely the within-pair variation that is time-varying and independent of country-specific changes in trade. At the same time, the strong collinearity among the provisions in our data means that the selected variables should not be interpreted as a definitive list of causal provisions. The initial PPML-lasso with the tuning parameter selected by the plug-in method is likely to omit relevant variables, while the bootstrap lasso and iceberg lasso are designed to recover additional provisions whose effects may be masked by their collinearity with those selected in the first step. The broader sets of variables selected by the latter two methods are therefore more likely to include the relevant provisions, but at the cost of also selecting some provisions that may have little or no direct causal impact on trade.

4 Simulation Evidence

In this section we report the results of a simulation exercise investigating the finite-sample properties of the variable-selection methods discussed before. The simulation design we use covers a range of scenarios that, to different degrees, combine two important features of our application: a relatively small sample and a high degree of collinearity between several potential explanatory variables. The results we obtain, therefore, provide information on the performance of the different methods in conditions similar to those we face. They also illustrate how these performances change when we progressively move towards either more or less challenging environments.

In all experiments, the n observations of the dependent variable are generated as

$$y = \exp\left(1 + \beta x_1 + z + \sigma\varepsilon\right),$$

where β and σ are parameters and x_1, z, and ε are independent random draws from the standard normal distribution. In the estimation, performed by PPML-lasso, ε is not included as a regressor (it is the error term), z is always included as a regressor whose coefficient is not penalized, and

we use different methods to select other regressors from a set of p potential explanatory variables $x_1, \ldots, x_p$. Therefore, in this design, x_1 plays the role of the presumably small number of provisions that effectively affect trade, $x_2, \ldots, x_p$ represent the provisions that have no impact on trade, and z mimics the role of the fixed effects that explain a significant share of the variation of trade and are included without penalty.[19]

The parameters β and σ determine the relevance of x_1 and the signal-to-noise ratio: because gravity equations typically have an excellent fit, we set $\beta = 0.2$ and $\sigma = 0.3$, which ensures that model has a reasonably high R^2 and that the effect of x_1 is neither too small (which makes its role very difficult to detect) nor too large (in which case all approaches have an excellent performance).

The p potential explanatory variables are obtained as random draws from the normal distribution; the first κ variables $x_1, \ldots, x_\kappa$ are equicorrelated with correlation coefficient ρ, and the remaining ones are independent of all other variables. All regressors have zero mean and variance 1 and we perform simulations with $\kappa \in \{5, 10, 20\}$, $\rho \in \{0.90, 0.95, 0.99\}$, $n \in \{250, 1000, 4000\}$, and set p to $5\lceil\sqrt{n}\rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function; that is, depending on the value of n, p is either 80, 160, or 320. Note that, although only x_1 actually influences y, we include all p potential explanatory variables in our regressions and check whether our methods can find the one variable that matters.[20]

In these simulations we considered each of the four methods presented before: cross-validation lasso, plug-in lasso, the bootstrap lasso and the iceberg lasso. The bootstrap lasso is performed with $B = 20$ and we include in the set of selected variables any variable that is selected in at least one sample (that is, we use a cutoff of 5%). Additionally, we also considered the adaptive

[19] Abstracting from the full specification of fixed effects in this way greatly facilitates computation and enables us to conduct a more focused demonstration of how the different variable selection approaches we consider are affected by collinearity. Our Appendix C provides some formal discussion of how the estimation of the full set of fixed effects affects the selection properties of the PPML plug-in lasso, which provides the basis for several of these methods.

[20] A noticeable difference between the simulation design we use and our application is that in the simulations the potential explanatory variables have a continuous distribution whereas in the application they are dummies. We preformed some experiments where the potential explanatory variables are dummies generated using the method described by Lunn and Davies (1998) and found broadly comparable results. However, we prefer to report the results obtained using the normally distributed variables because when dummies are used we frequently encounter numerical issues and cases of perfect collinearity that make it more difficult to keep track of the variables selected.

lasso of Zou (2006), with the penalty parameter chosen by cross-validation in both steps.[21] Unlike the other methods we consider, the adaptive lasso has the so-called oracle property, implying that asymptotically it will choose the right set of regressors, and therefore it provides an interesting benchmark against which the performance of the other methods can be judged.[22] We repeat the simulations $1,000$ times to study the ability of each method to correctly select x_1 as a regressor. In Appendix E, we use the same simulations to also assess the predictive performance of each method.

For each of the cases considered, Table 2 presents the percentage of times the regressor x_1 is selected and, in parentheses, the average number of regressors selected by each method. The results in Table 2 reveal that the various methods can have very different performances.

Starting with the ability of each method to correctly select x_1 as a regressor, we find that for $n = 250$ the lasso with penalty chosen by the plug-in method (PI) is the method with the worst results, and its performance deteriorates quickly as κ and ρ increase. The standard lasso with the penalty chosen by cross-validation (CV) and the adaptive lasso (AL) lead to better results, but their performances also degrade as ρ increases, becoming very poor in the extreme case where $\rho = 0.99$. Both the bootstrap lasso (BL) and the iceberg lasso (IL) are at least competitive across all settings. BL is the most accurate overall method for the cases with $\rho = 0.90$ or $\rho = 0.95$ but is outdone by IL for $\rho = 0.99$.[23]

The performance of all methods improves for the larger sample sizes, but the IL maintains its advantage in the more challenging cases with $\rho = 0.99$, with BL having a very similar performance. Overall, these results confirm that both the BL and IL lead to greatly improved ability of identifying the relevant regressor versus the other variable selection approaches we consider. Other than the extreme cases where $n = 250$ and $\rho = 0.99$, where the IL has a clear advantage, there is generally little to choose between them.

The results for the average number of variables selected are also interesting.[24] In all cases considered, CV tends to lead to a high average number of selected regressors. On the other extreme,

[21] We also performed simulations using Zou and Hastie's (2005) elastic-net. However, that method does not lead to particularly good results and we do not reported them to conserve space and to simplify the exposition.

[22] This property improves on the related near-oracle property we have noted for the plug-in method by guaranteeing a faster rate of convergence.

[23] Part of the reason why in some cases IL does not perform well is that sometimes PI selects no regressors at all, and in those cases IL cannot improve on it but BL can.

[24] Recall that there is only one regressor, x_1, with a causal impact on the outcome.

PI is generally the most parsimonious except for when n is large, in which case the oracle property of AL starts to become salient. Turning to the performance of the BL and IL, we observe that the average number of regressors selected by BL is always reasonably high and that, for the values of ρ we consider, the average number of variables selected by the IL increases with κ, suggesting that the latter method performs as intended: it identifies the set of variables highly correlated with the relevant ones. Naturally, this behavior will be less pronounced for lower values of ρ, and we have confirmed that in unreported simulations.

In summary, for very large samples, the adaptive lasso with penalty parameter selected by cross-validation would be the preferred method; this is justified both by our simulation results and

Table 2: Percentage of times x_1 is selected & average number of variables selected

		$\rho = 0.90$			$\rho = 0.95$			$\rho = 0.99$		
n		$\kappa=5$	$\kappa=10$	$\kappa=20$	$\kappa=5$	$\kappa=10$	$\kappa=20$	$\kappa=5$	$\kappa=10$	$\kappa=20$
250	CV	96.6 (8.87)	91.8 (8.66)	85.5 (8.64)	86.9 (8.78)	77.9 (8.65)	67.0 (8.45)	55.2 (8.52)	37.7 (8.22)	23.4 (7.93)
	AL	93.9 (7.34)	87.4 (7.21)	80.4 (7.05)	78.7 (7.22)	67.2 (7.07)	56.0 (6.77)	45.3 (6.99)	29.4 (6.72)	17.7 (6.26)
	PI	80.6 (1.45)	72.1 (1.73)	63.7 (2.06)	67.2 (1.43)	54.3 (1.65)	44.1 (1.90)	41.1 (1.23)	26.8 (1.33)	16.9 (1.41)
	BL	96.6 (11.31)	98.4 (13.25)	96.7 (15.66)	98.5 (11.42)	95.7 (13.26)	88.8 (15.30)	90.4 (11.27)	79.2 (12.77)	64.2 (14.03)
	IL	95.9 (4.81)	95.8 (9.43)	93.0 (17.00)	95.5 (4.81)	95.1 (9.40)	88.8 (16.81)	95.3 (4.78)	93.4 (9.32)	80.1 (15.65)
1000	CV	100.0 (9.76)	100.0 (10.10)	99.9 (10.69)	99.8 (10.01)	99.0 (10.32)	97.6 (10.90)	81.0 (9.92)	69.8 (10.11)	56.4 (10.51)
	AL	100.0 (4.71)	99.7 (5.22)	99.7 (5.85)	98.5 (5.20)	96.2 (5.91)	93.6 (6.51)	68.3 (5.37)	54.8 (5.97)	40.8 (6.22)
	PI	99.2 (1.63)	98.4 (2.02)	97.5 (2.57)	96.8 (1.78)	93.2 (2.26)	88.9 (2.78)	71.4 (1.75)	55.9 (2.02)	41.4 (2.34)
	BL	100.0 (9.26)	100.0 (11.67)	100.0 (15.23)	100.0 (9.35)	100.0 (11.96)	99.6 (15.65)	98.0 (9.36)	93.70 (11.85)	87.1 (14.81)
	IL	100.0 (5.00)	100.0 (10.01)	100.0 (19.69)	100.0 (5.01)	100.0 (10.01)	99.9 (19.79)	100.0 (5.01)	100.0 (10.01)	98.8 (19.72)
4000	CV	100.0 (10.78)	100.0 (11.28)	100.0 (11.88)	100.0 (10.94)	100.0 (11.59)	100.0 (12.29)	99.0 (11.18)	97.8 (12.06)	94.9 (12.63)
	AL	100.0 (1.03)	100.0 (1.00)	100.0 (1.03)	100.0 (1.03)	100.0 (1.10)	100.0 (1.17)	91.9 (1.18)	86.0 (1.30)	79.1 (1.70)
	PI	100.0 (1.53)	100.0 (1.96)	100.0 (2.42)	100.0 (1.73)	99.9 (2.27)	99.8 (2.83)	98.0 (2.00)	93.9 (2.60)	88.1 (3.18)
	BL	100.0 (8.44)	100.0 (11.04)	100.0 (14.94)	100.0 (8.67)	100.0 (11.53)	100.0 (15.75)	100.0 (8.93)	99.9 (11.94)	99.8 (16.27)
	IL	100.0 (5.00)	100.0 (10.00)	100.0 (20.00)	100.0 (5.00)	100.0 (10.01)	100.0 (20.00)	100.0 (5.01)	100.0 (10.00)	100.0 (20.00)

by its oracle property. However, for small to medium samples, and especially with high correlation between potential explanatory variables, the adaptive lasso is outperformed by other methods. In these cases, the choice of method depends on whether we favor selecting the relevant regressors or having a parsimonious model. If parsimony is paramount, the lasso with penalty parameter selected by the plug-in method is difficult to beat. However, since the goal in our application is to identify the relevant regressors, even at the cost of selecting some irrelevant ones, the bootstrap lasso and iceberg lasso emerge as the preferred methods. Which of the two approaches has the best performance will depend on the nature of the data, and therefore we see the two methods as complements.

5 Empirical Results

We now present the results obtained when applying the methods studied in the previous sections to our real world dataset of trade flows and trade provision dummies, as described in Section 2. We first present results for the plug-in method and then turn to the bootstrap and iceberg lasso results, which each build in their own way on the selection done by the plug-in lasso. To conserve space, the results obtained using cross-validation are reported in Appendix D, and Appendix E includes a brief discussion of using these methods for prediction.

5.1 Plug-in Lasso Results

Table 3 presents results for the plug-in lasso and post-lasso regressions. As discussed before, we should expect these results to capture provisions that are highly predictive of larger PTA effects and that are at least correlated with the causal ones. They will also serve as a precursor to the bootstrap lasso and iceberg lasso, which should be better suited for identifying all of the provisions that are likely to have a causal impact. Both the plug-in lasso estimates and the PPML standard errors account for clustering, which is done at the agreement level for observations that correspond to agreements, and at the pair level for the remaining observations.

In column (1), we start by presenting the results of a traditional PPML gravity estimation with a dummy for the presence of a PTA between the trading partners. This shows that we can replicate the usual finding that PTAs lead to a significant increase in trade flows. Specifically, we find that the PTAs in our data increase trade by 14% ($\exp(0.131) - 1 = 0.14$).

Table 3: PPML, PPML-lasso, and post-lasso PPML results for plug-in approach

Dependent variable: Bilateral Trade Flows (1964-2016, every 4 years)					
	PPML	Lasso	Post-lasso	PPML	PPML
	(1)	(2)	(3)	(4)	(5)
PTA	0.131***			−0.008	0.087**
	(0.044)			(0.062)	(0.041)
EU					0.658***
					(0.087)
AD14. Anti-dumping – Material Injury		0.329	0.349***	0.347***	
			(0.117)	(0.119)	
CP23. Competition Policy – Transparency / Coordination		0.002	0.118	0.118	
			(0.077)	(0.078)	
TBT2 / TBT29. Mutual Recognition†		0.142	0.184	0.182	
			(0.142)	(0.144)	
TBT7. Technical Reg's: use International Standards		0.016	0.032	0.034	
			(0.078)	(0.080)	
TBT8. Conformity Assessment: Mutual Recognition		0.028	0.123	0.124	
			(0.099)	(0.099)	
TBT33. Standards: use Regional Standards		0.109	0.113*	0.116*	
			(0.061)	(0.064)	
TF45. Issuance of Proof of Origin		0.000	0.089***	0.095*	
			(0.032)	(0.053)	

Gravity equations with exporter-time, importer-time, and exporter-importer FE, estimated using 316,317 observations. The column labelled "Lasso" reports all non-zero coefficient estimates obtained using the plug-in lasso. The column labelled "Post-lasso" reports PPML estimates for a model including only the variables selected by the plug-in lasso. All other columns report PPML estimates. Cluster-robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < .05$, *** $p < .01$. † TBT2 is perfectly collinear with TBT29: TBT2 refers to mutual recognition of technical regulations; TBT29 refers to mutual recognition of standards.

Column (2) then shows the results of the plug-in PPML-lasso regression, showing only the coefficients that are found to be non-zero.[25] Using this approach, the lasso selects 8 provisions related to anti-dumping, competition policy, technical barriers to trade (TBT), and trade facilitation.[26] As explained above, the selection of these 8 provisions does not indicate that these are the provisions that causally impact trade, or that there are only 8 such provisions, but rather that the inclusion of these provisions is especially predictive of larger PTA effects. Nonetheless, broadly speaking, all these variables can be rationalized as having intuitive effects on trade. The selected anti-dumping and competition policy provisions create more certainty as to how disciplinary investigations and proceedings will be carried out in these policy areas.[27] This increased certainty may increase entry by foreign exporting firms. The inclusion of provisions related to technical barriers to trade and trade facilitation is likewise intuitive, but the selection of TF45, which facilitates obtaining certificates of origin, seems of particular note in that it highlights the costs of complying with rules of origin. It is also worth noting that the plug-in PPML-lasso selects TBT2 and TBT29, two provisions that are perfectly collinear in our data set. This illustrates both the ability of the method to select variables that are perfectly collinear as well as the challenges faced when trying to interpret the results in this setting.

We next estimate a "post-lasso" PPML regression—a standard PPML regression using only the provisions that were selected in the previous step. These post-lasso PPML results, presented in column (3), show that some of the selected provisions have large predictive effects when estimated in the conventional way. For example, the inclusion of anti-dumping provision AD14, which requires that anti-dumping proceedings establish "material injury" to domestic producers, is associated with an increase in trade flows of about 42% ($\exp(0.349) - 1 = 0.42$).[28] Interestingly, not all of the provisions selected by the lasso step are found to be statistically significant in the post-lasso step. This arises for two reasons. First, the lasso focuses on the contribution of each variable to the pseudo-likelihood function, which is not the same as testing whether its coefficient is statistically

[25] More detailed descriptions of all variables highlighted in our analysis can be found in Table A1, in Appendix A.

[26] Note that only 7 coefficients are reported because TBT2 and TBT29 are perfectly collinear with each other, and therefore are counted as two selected provisions.

[27] For more on the effect of anti-dumping provisions, see Prusa, Teh, and Zhu (2022).

[28] It should be reiterated that we do not interpret this as the causal effect of AD14 on trade; our interpretation is that the presence of AD14 is predictive of a larger PTA effect because it is, at the very least, strongly correlated with provisions that have causal effects. We explore this further when discussing the bootstrap and iceberg lasso results.

different from zero. Second, because the lasso shrinks the coefficients on all variables towards zero, it reduces the influence of the collinearity between them and can allow individual provisions that are not significant in the conventional regressions to speak more loudly.

In column (4), we re-estimate the model using the same covariates as column (3) but now adding our original PTA dummy from column 1. In this case, the coefficient on PTA captures any effect on trade flows that is not already captured by the provision variables that were selected by the lasso. With this in mind, we take the insignificant and near-zero coefficient on PTA in column (4) as an encouraging indication that the selected provisions completely explain the average PTA effect reported in column (1).

Finally, we check whether the selected provisions are simply capturing the trade effects that are actually specific to the EU.[29] This is a natural concern because the EU contains six of the eight provisions selected by the plug-in lasso (all except AD14 and TBT7) and because its institutional features and secondary law process may generate trade effects not fully captured by the provision coding in our data. To investigate this possibility, column (5) adds an EU dummy to the original simple model from column (1). As the unpenalized PPML results in column (5) show, the estimated EU effect is large, several times that of non-EU PTAs in fact. However, when we include the EU dummy as a possible predictor in the lasso, we find that it is not selected. Consequently, the set of selected provision variables when the EU is included is identical to that in column (2), which is our preferred set to work with in the subsequent bootstrap lasso and iceberg lasso analyses.

5.2 Bootstrap Lasso Results

Tables 4 and 5 summarize the results obtained from the bootstrap lasso, using 250 bootstrap samples. To be consistent with the sampling assumptions used in constructing the plug-in lasso penalty, the resampling process for each bootstrap sample treats pairs belonging to the same agreement as belonging to the same resampling block and treats each remaining pair as its own resampling block. In each replication, we perform selection using plug-in lasso and record which variables are selected and their post-lasso PPML coefficient estimates.

Table 4 presents the average coefficients (on the left) as well as the selection frequencies (on the right) for the 30 provisions selected at least 5% of the times by the bootstrap lasso. Expanded versions of these results showing all 75 provisions selected at least 1% of the time are provided

[29] We use EU as shorthand for the EU and European Community agreements.

in Appendix D. Before turning to the selected provisions in more detail, the first main takeaway is that, regardless of the cutoff used, the bootstrap lasso finds that more than two-thirds of the provisions in our data are not relevant for trade. That is, they are not selected by the plug-in lasso as being predictive of changes in trade flows, nor are they found by the bootstrap lasso to be close substitutes for the ones initially selected by the plug-in lasso. This does not literally mean that the remaining 230 provisions in our data have no effect on trade. It does, however, indicate that these provisions are likely to have small effects, and thus a highly selective trade agreement that focuses mainly on the most trade-promoting provisions can in principle produce most of the same trade effect as a more comprehensive agreement.

Another immediate takeaway that stands out from Table 4 is that even the provisions that are selected most frequently in relative terms are selected less that half of the time. For example, AD14 is the most commonly selected provision, and it has the largest coefficient estimate of the variables selected by the plug-in lasso (see Table 3), but it is only selected in 37% of replications. This

Table 4: Bootstrap Lasso results: largest average coefficients and selection frequencies

Provisions with largest average coefficients				Provisions selected most frequently			
AD14	0.079	SUB10	0.020	AD14	0.372	ET09	0.100
CP23	0.065	MOC27	0.019	CP23	0.320	MOC27	0.084
CP22	0.063	ET43	0.013	TBT07	0.308	SUB13	0.080
AD05	0.055	TF45	0.012	SPS06	0.228	TF42	0.076
TBT07	0.054	SUB13	0.011	TBT08	0.208	SUB10	0.072
TBT02/29	0.048	ENV33	0.011	SUB12	0.184	ET43	0.072
TBT08	0.037	TBT15	0.010	TBT02/29	0.168	MIG14	0.068
SUB12	0.030	MIG14	0.010	TBT33	0.160	STE32	0.068
TBT34	0.028	STE32	0.008	CP22	0.156	TBT15	0.064
SPS06	0.028	ET09	0.007	TBT34	0.152	ROR04	0.060
TF42	0.028	SUB11	0.005	TBT06	0.148	SUB11	0.056
AD07	0.027	ROR04	0.005	AD05	0.140	SUB09	0.056
TBT33	0.023	SUB09	0.004	CP21	0.124	STE30	0.056
TBT06	0.021	STE30	0.003	TF45	0.116	AD07	0.052
CP21	0.020			ENV33	0.116		

Notes: Bootstrap plug-in lasso performed using cluster-bootstrap resampling with 250 replications. The numbers shown are (left) average coefficient estimates for the provisions selected at least 5% of the time across all replications and (right) selection frequencies for the same set of provisions. Both lists are ordered from greatest to least. A description of the provisions in this table can be found in Appendix A.

illustrates that, as discussed before, we should only have limited confidence that AD14 is the provision the delivers the effect indicated by the original plug-in estimates for AD14. At the same time, since these frequencies can be interpreted as measures of selection stability, AD14 is found to be more likely to matter than other provisions.

Overall, the results in Table 4 expand on the results we found earlier using the plug-in lasso in several ways. Though the bootstrap lasso generally confirms that many of the same provisions shown in the plug-in lasso results in Table 3 are among those most likely to be relevant—indeed, AD14, CP23, and TBT7 are the three most frequently selected variables—there are a wide variety of other variables not selected by the plug-in lasso that nonetheless emerge as being selected a comparable amount of the time in the bootstrap lasso analysis. Notably, the latter set of variables not only includes some provisions from the same policy areas (e.g., AD05, CP21, CP22, TBT34), but also draws from some areas not represented in Table 3, such as sanitary and phytosanitary standards (SPS06), subsidy policies (SUB12 and SUB13), environmental protection (ENV33), and export taxes (ET09 and ET43).

Table 5 further summarizes the bootstrap lasso results by documenting the broad provision categories in which provisions were most likely to be selected as well as the sum of the average coefficients within each category.[30] These results, therefore, show which provision categories, when taken as a whole, are likely to have the biggest impact on trade. These results can therefore inform policymakers about which general areas in trade agreements are most worth their efforts if their goal is to increase trade. Interestingly, these results are not quite what one would expect based on our earlier plug-in lasso results in Table 3, which would tend to suggest that anti-dumping is the category with the most trade-increasing potential. As shown in the last column of Table 5, the category with the biggest total trade impact turns out to be TBTs, which arguably makes sense since provisions in this category explicitly target so-called non-tariff trade barriers. The next two largest impacts are for anti-dumping and competition policy, followed by subsidies, trade facilitation, sanitary and phytosanitary measures, and export taxes. Notably, the differences between categories seem to comport with intuition (very small impacts for services and labor markets, for example).

Finally, one last interesting finding from our bootstrap lasso results is that, even for the expanded set of selected provisions shown in Table A2, all provisions that are selected as having non-zero coefficients are estimated to have positive effects. Using the results from Table 5 for

[30] For example, the value 0.171 reported for the anti-dumping category in Table 5 is the sum of the average post-lasso effect of the provisions within the AD category that are selected in any of the bootstrap samples.

Table 5: Bootstrap lasso results: Summarizing results by provision category

	Number of provisions selected more than 5% of the time	Number of provisions selected more than 1% of the time	Sum of average post-lasso effects across categories
Anti-dumping	3	5	0.171
Competition Policy	3	5	0.151
Environment	1	5	0.017
Export Taxes	2	5	0.049
Investment	0	2	0.020
IPR	0	5	0.019
Labor Markets	0	0	0.000
Migration	1	1	0.012
Movement of Capital	1	2	0.023
Public Procurement	0	1	0.013
Rules of Origin	1	4	0.021
Services	0	1	0.004
SPS	1	10	0.062
State aid	2	2	0.011
Subsidies	5	7	0.076
TBTs	8	13	0.237
Trade Facilitation	2	5	0.064
Total	30	75	0.951

Note: The table documents the categories in which provisions were most likely to be selected and the total of the average coefficients of each provision within each category.

illustration, it is notable that we do not find, for example, that environmental or labor market provisions reduce trade. Again, this results should not be construed as definitively saying that none of the trade agreement provisions in our data reduce trade. It does, however, suggest that any such negative effects are likely to be small.

5.3 Iceberg Lasso Results

As a complementary exercise, Table 6 reports the results from the iceberg lasso. This method takes the eight provisions selected by the initial plug-in lasso as starting points and then, for each of them, uses a second-step plug-in lasso regression to identify other provisions that help predict the initially

selected provision.[31] The first row of Table 6 lists the provisions selected in the first step. The remaining rows list the additional provisions selected in the second step, with the raw correlation between each provision and the corresponding first-step provision reported in parentheses. The final row reports the R^2 from each second-step regression.

The main value of the iceberg lasso is that it provides a transparent way to examine the collinearity structure behind the initial plug-in lasso results. In total, the procedure identifies 42 distinct provisions: the eight selected by the initial plug-in lasso and 34 additional provisions that help predict them. This number is again far smaller than the 133 provisions selected by cross-validation, but larger than the initial plug-in set. Thus, as in the bootstrap lasso results, the iceberg lasso points to a relatively concentrated set of provisions that may be difficult for the initial plug-in lasso to distinguish from one another.

The results also underscore why we do not treat the iceberg lasso as our preferred empirical method. The additional provisions selected in the second step should not be interpreted as provisions with independently estimated positive effects on trade. Rather, they are provisions that help explain the cross-agreement variation in the provisions selected by the plug-in lasso. In some cases this interpretation is straightforward. For example, AD14 is strongly associated with other anti-dumping provisions, suggesting that the initial anti-dumping result should be interpreted as evidence for the importance of this broader policy area rather than as evidence that AD14 itself is uniquely responsible for the estimated trade effect. Similarly, several of the TBT provisions selected in the first step are associated with other TBT-related provisions, as well as with other types of provisions, consistent with the idea that the lasso is identifying broader clusters of provisions aimed at reducing technical barriers to trade.

In other cases, however, the selected provisions are more difficult to interpret. Some provisions selected in the second step belong to policy areas that are not obviously related to the first-step provision, and several have small or even negative raw correlations with the first-step provision they help predict. These cases are best understood as reflecting the complex template structure of the data rather than as evidence that these provisions have negative trade effects or that they are close substitutes in a simple bivariate sense. Because the second-step regressions are multivariate prediction exercises, a provision can be selected because it helps distinguish between overlapping

[31] These linear plug-in lasso regressions are performed using only the $34,370$ observations for which PTAs are in force. This is because the provisions are identically zero for the remaining observations, which therefore are not informative about the relations of interest. As a consequence, the clustering is by agreement.

Table 6: Iceberg lasso results

(1)	(2)	(3)	(4)	(5)	(6)	(7)
AD14	CP23	TBT02/29	TBT07	TBT08	TBT33	TF45
AD06 (0.98)	AD06 (0.40)	AD06 (-0.07)	AD06 (0.51)	SUB10 (0.84)	AD11 (-0.05)	AD06 (0.16)
AD08 (0.98)	AD08 (0.40)	AD08 (-0.07)	AD08 (0.51)	TF42 (0.93)	ENV44 (-0.02)	AD08 (0.16)
ENV42 (0.98)	CP22 (0.80)	CP14 (0.61)	ENV42 (0.51)		MOC26 (-0.10)	AD11 (0.08)
	CP24 (0.89)	CP21 (0.77)	ENV44 (0.08)		PP08 (-0.01)	CP15 (0.71)
	ENV42 (0.40)	CP22 (0.80)	SPS21 (0.16)		SUB07 (0.08)	ENV19 (0.40)
	PP08 (0.05)	ENV22 (-0.01)	SUB07 (0.10)		TBT05 (0.69)	ENV27 (0.50)
	SPS24 (-0.05)	ENV42 (-0.07)	TBT15 (0.68)		TBT06 (0.98)	ENV42 (0.16)
	STE31 (0.54)	ENV44 (-0.01)	TBT34 (0.93)		TBT14 (0.89)	MOC26 (0.16)
	TBT10 (-0.01)	SPS11 (-0.00)			TBT15 (0.58)	STE37 (0.06)
	TF42 (0.65)	STE32 (0.66)			TBT32 (0.69)	SUB07 (0.03)
	TF43 (-0.04)	SUB09 (0.78)			TBT34 (0.42)	SUB10 (0.28)
	TF44 (0.38)	SUB10 (0.90)			TF42 (0.64)	TF44 (0.98)
		TF42 (0.98)				
0.95	0.82	0.97	0.86	0.86	0.97	0.96

Notes: Table shows PTA provisions identified by the plug-in lasso as being associated with increases in bilateral trade flows (row 1), together with other provisions that predict the provision in row 1 (rows 2-14; numbers in brackets are raw correlations with the provision from row 1). The last row displays the R^2 of the regression of each provision selected by the plug-in lasso on the corresponding correlated provisions. A description of the provisions in this table can be found in Appendix A.

agreement templates even when its raw correlation with the first-step provision is weak or negative. This feature is useful for diagnosing multicollinearity, but it also limits the substantive interpretation of the iceberg lasso results.

Reassuringly, the main qualitative conclusions from the iceberg lasso overlap with those from the bootstrap lasso. Both methods point to TBTs, anti-dumping, competition policy, trade facilitation, and subsidies as areas where the initial plug-in lasso appears to be capturing broader clusters of related provisions, and the iceberg lasso provides little reason to revise the main conclusion from the bootstrap lasso that the trade effects of PTAs are concentrated in a relatively small number of policy domains. Moreover, though the correspondence between the provisions selected by the iceberg lasso and those selected by the bootstrap lasso is not exact, in many cases the additional provisions shown in Table 6 are also found in our selection results for the bootstrap lasso (especially in the expanded set of results shown in Tables A2 and A3 in Appendix D). At the same time, the iceberg lasso surfaces some additional provisions not identified by the bootstrap lasso as being potentially important, such as CP14, CP21, STE31, and TBT32. Again, the selection of these additional variables does not change any of our main conclusions.

6 Conclusions

In this paper, we have proposed new methods for assessing the impact of individual trade agreement provisions on trade flows. While other work in this area has relied on summary measures of agreement depth or on specific provision bundles of interest, our approach is instead to study the rich provision content of PTAs as a variable selection problem. By combining the three-way PPML estimator that is popular in the study of PTAs with lasso methods for variable selection, we are able to identify a relatively parsimonious set of provisions that are most likely to impact trade. Specifically, using our bootstrap lasso we identify between 30 and 75 (out of 305) provisions that are likely to have a trade-enhancing effect, while with the iceberg lasso the number of provisions selected is 42. These numbers are in sharp contrast with the results obtained by the two benchmark lasso methods that we consider, which vary between 8 when using the plug-in PPML-lasso, and 133 when cross-validation is used to select the penalty. While the provisions we identify with the bootstrap and iceberg lasso approaches span a range of policy areas, the results using our preferred method, the bootstrap lasso, support the conclusion that a select number of provisions related to technical barriers to trade, antidumping, competition policy, subsidies, trade facilitation, sanitary

and phytosanitary measures, and export taxes are likely to be the most effective at promoting trade as compared to other types of provisions that appear in PTAs.

As we have emphasized, the high collinearity among provision variables means that the provisions selected by our methods should not be interpreted as a definitive list of individual causal provisions, but rather as a data-driven set of provisions likely to include those most relevant for the trade effects of PTAs. The results obtained with our methods therefore clearly suggest that the trade effects of modern PTAs are driven not by the sheer number of provisions they contain, but by a relatively small set of provisions in policy areas closely connected to goods trade.

Acknowledgements

Research for this paper has been in part supported by the World Bank's Multidonor Trust Fund for Trade and Development. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. We gratefully acknowledge financial support through ESRC grant EST013567/1, and thank Scott Baier, Federico Bandi, Tibor Besedes, Christian Brownless, Maia Linask, Dennis Novy, Roberto Reno, Yoto Yotov, and seminar participants at the World Bank Economics of Deep Trade Agreements Seminar Series, the Empirical Investigations in Trade 2022 Annual Conference, the Midwest International Trade Conference 2022, the ICEEE 2023, Cagliari, the Barcelona Workshop in Financial Econometrics, and at the Rensselaer Polytechnic Institute, for useful comments. Nicolas Apfel, João Cruz, Alvaro Espitia, Diego Ferreras-Garrucho, and Jiayi Ni provided excellent research assistance. The usual disclaimer applies. An R package (penppml) implementing penalized PPML regressions with high-dimensional fixed effects is available from CRAN.

References

Anderson, J. and E. Van Wincoop (2003). "Gravity with gravitas: A solution to the border puzzle," *American Economic Review,* 93, 170-192.

Baier, S.L. and J.H. Bergstrand (2007). "Do free trade agreements actually increase members' international trade?," *Journal of International Economics*, 71, 72-95.

Baier, S.L., J.H. Bergstrand, and M.W. Clance (2018). "Heterogeneous effects of economic integration agreements," *Journal of Development Economics*, 135, 587-608.

Baier, S.L., J.H. Bergstrand, and M. Feng (2014). "Economic integration agreements and the margins of international trade," *Journal of International Economics*, 93, 339-350.

Baier, S.L, Y.V. Yotov, and T. Zylkin (2019). "On the widely differing effects of free trade agreements: Lessons from twenty years of trade integration," *Journal of International Economics,* 116, 206-228.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369-2429.

Belloni, A., V. Chernozhukov, C. Hansen, D. Kozbur (2016). "Inference in high dimensional panel models with an application to gun control," *Journal of Business & Economic Statistics,* 34, 590-605.

Chatterjee, A. and S.N. Lahiri (2010). "Asymptotic properties of residual bootstrap for lasso estimators," *Proceedings of the American Mathematical Society,* 138, 4497-4509.

Correia, S., P. Guimarães and T. Zylkin (2020). "Fast Poisson estimation with high dimensional fixed effects," *STATA Journal,* 20, 90-115.

Dhingra, S., R. Freeman, and E. Mavroeidi (2018). "Beyond tariff reductions: What extra boost to trade from agreement provisions?," LSE Centre for Economic Performance Discussion Paper 1532.

Drukker, D.M and D. Liu (2019). "A plug-in for Poisson lasso and a comparison of partialing-out Poisson estimators that use different methods for selecting the lasso tuning parameters," mimeo.

Falvey, R., N. Foster-McGregor (2022). "The breadth of preferential trade agreements and the margins of exports," *Review of World Economics*, 158, 181-251.

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33, 1-22.

Garrucho, D.F., T. Zylkin, J. Cruz, and N. Apfel (2021). `penppml`: Penalized Poisson Pseudo Maximum Likelihood Regression, `https://tinyurl.com/penppml`.

Gourieroux, C., A. Monfort, A. Trognon (1984). "Pseudo maximum likelihood methods: Applications to Poisson models," *Econometrica,* 52, 701-720.

Hastie, T., R. Tibshirani, and J.H. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction.* New York (NY): Springer.

Head, K. and T. Mayer (2014). "Gravity Equations: Workhorse, Toolkit, and Cookbook," in Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, eds., *Handbook of International Economics*, Volume 4, Chapter 3, Amsterdam: Elsevier, 131-195.

Hofmann, C., A. Osnago, M. Ruta (2017). "Horizontal depth. A new database on the content of preferential trade agreements," World Bank Policy Research Working Paper 7981.

Kohl, T., S. Brakman, H. Garretsen (2016). "Do trade agreements stimulate international trade differently? Evidence from 296 trade agreements," *The World Economy*, 39, 97-131.

Larch, M., S. Shikher, and Y.V. Yotov (2025). "Estimating gravity equations: Theory implications, econometric developments, and practical recommendations," *Review of International Economics*, 33, 1066-1092

Larch, M., J. Wanner, Y.V. Yotov, T. Zylkin (2019). "Currency unions and trade: a PPML reassessment with high dimensional fixed effects," *Oxford Bulletin of Economics and Statistics,* 81, 487-510.

Lunn, A.D. and S.J. Davies (1998). "A note on generating correlated binary variables," *Biometrika*, 85, 487-490.

Mattoo, A., A. Mulabdic, and M. Ruta (2017). *Trade creation and trade diversion in deep agreements.* Policy Research Working Paper Series 8206, The World Bank.

Mattoo, A., N. Rocha, M. Ruta (2020). "Handbook of deep trade agreements." Washington, DC: World Bank.

Meinshausen, N., and P. Bühlmann (2010). "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417-473.

Mulabdic, A., A. Osnago, and M. Ruta (2017). "Deep integration and UK-EU trade relations," World Bank Policy Research Working Paper Series 7947.

Mullainathan, S. and J. Spiess, (2017). "Machine learning: An applied econometric approach," *Journal of Economic Perspectives*, 31, 87-106.

Neri-Lainé, M., G. Orefice, and M. Ruta (2023). "Deep Trade Agreements and Heterogeneous Firms Exports," CESifo Working Paper No. 10436.

Prusa, T., R. Teh, and M. Zhu (2022). "PTAs and the incidence of antidumping disputes," `https://tinyurl.com/PTA-PTZ-2022`.

Regmi, N. and S. Baier (2020). "Using machine learning methods to capture heterogeneity in free trade agreements," mimeograph.

Santos Silva, J.M.C. and S. Tenreyro (2006). "The log of gravity," *Review of Economics and Statistics,* 88, 641-658.

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

Tibshirani, R. (1996). "Regression shrinkage and selection via lasso," *Journal of the Royal Statistical Society, Ser B.* 59, 267-288.

Wainwright, M.J. (2009). "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso)," *IEEE Transactions on Information Theory*, 55, 2183-2202.

Wang, S., B. Nan, S. Rosset, J. Zhu, (2011) "Random lasso," Annals of Applied Statistics, 5, 468-485.

Weidner, M., T. Zylkin (2021). "Bias and consistency in three-way gravity models," *Journal of International Economics*, 132, 103513.

Wüthrich, K. and Y. Zhu, (2021). "Omitted variable bias of Lasso-based inference methods: A finite sample analysis," *Review of Economics and Statistics*, forthcoming.

Yotov, Y.V., R. Piermartini, J.-A. Monteiro, M. Larch (2016). *An advanced guide to trade policy analysis: The structural gravity model.* Geneva: World Trade Organization.

Zhao, P. and B. Yu (2006). "On model selection consistency of lasso," *Journal of Machine Learning Research*, 7, 2541-2563.

Zou, H. (2006). "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418-1429.

Zou, H. and T. Hastie, (2005). "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301-320.

Determining which trade agreement provisions matter for trade

Online Appendix – Not for publication

Holger Breinlich* Valentina Corradi† Nadia Rocha‡ Michele Ruta§
J.M.C. Santos Silva¶ Tom Zylkin‖

24 May 2026

This Appendix is divided into 5 parts:

- Appendix A provides more information on each of the provision variables that appear in the paper.
- Appendix B describes the computational algorithms we use to implement our PPML-lasso methods.
- Appendix C provides more formal results supporting the validity of our methods.
- Appendix D presents additional bootstrap lasso results and the results obtained with the cross-validation approach.
- Appendix E discusses the use the lasso methods that we consider for the prediction of trade flows.

Appendix A: More details on provision variables

In Table A1, we show more detailed descriptions for each of the provision variables that appear in the results generated by our preferred methods. For space reasons, we do not provide descriptions for all the variables in the data. However, a complete dictionary of the World Bank Deep Trade Agreements data is available at https://jmcss.som.surrey.ac.uk/Provision_Names.xlsx.

*University of Surrey, CEP, CEPR and CESifo. Email: h.breinlich@surrey.ac.uk
†NYU Abu Dhabi. Email: vc2718@nyu.edu.
‡Inter American Development Bank. Email: nadiaro@iadb.org.
§International Monetary Fund. Email: mruta@imf.org.
¶University of Surrey. Email: jmcss@surrey.ac.uk.
‖University of Richmond. Email: tzylkin@richmond.edu.

Table A1: Main Provisions

Anti-dumping	
AD05	Determination of dumping - Export price less than comparable price when destined for consumption in the exporting country
AD06	Determination of dumping - If there are no sales in the normal course of trade in the domestic market of the exporting country
AD07	Determination of dumping - A comparable price of the like product when exported to an appropriate third country
AD08	Determination of dumping - Cost of production in the country of origin plus a reasonable amount
AD10	Determination of injury - Volume of dumped imports
AD11	Determination of injury - Price effects of dumped imports
AD14	Determination of injury - Material injury
Competition Policy	
CP14	Does the agreement require the establishment or existence of competition policy (either economy wide or sector specific)?
CP15	Does the agreement prohibit/regulate cartels/concerted practices?
CP16	Does the agreement prohibits/regulates abuse of market dominance?
CP21	Does the agreement regulate mergers and acquisitions?
CP22	Does the agreement contain provisions that promote predictability?
CP23	Does the agreement contain provisions that promote transparency?
CP24	Does the agreement contain provisions that promote the right of defense?
Environmental Laws	
ENV19	Does the agreement regulate pollution by ships?
ENV22	Does the agreement regulate fishing subsidies?
ENV27	Does the agreement promote renewable energy and improving energy efficiency?
ENV28	Does the agreement require states to implement water management?
ENV30	Does the agreement specify supremacy of MEA obligations over PTA obligations?
ENV33	Does the agreement require states to comply with the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal?
ENV42	Does the agreement require states to comply with the UN Conference on Environment and Development?
ENV44	Does the agreement require states to comply with the International Energy Program?
Export Taxes	
ET09	Prohibits all export taxes between the Parties, but with reference to certain exceptions mentioned in the provision that are WTO-plus
ET15	Prohibits an increase in the rate of any existing export tax
ET17	Exempts re-exports from import or export duties
ET42	Requires flexibility in the port/point of departure used for export for certain goods
ET43	Prohibits different treatment based on port/point of departure for exports
Investment	
INV12	Does the investment chapter prohibit or limit the use of performance requirements?
INV24	Does the FET clause prohibit arbitrary, unreasonable or discriminatory measures?

Table A1 (continued): Main Provisions

Intellectual Property Rights	
IPR02	Paris Convention
IPR03	Berne Convention
IPR51	Requires provision of legal means to prevent commercial use of country name that misleads consumers
IPR68	Provides minimum term of protection for new clinical info for a new indication/formulation/administration method of a previously-approved pharmaceutical product
IPR127	Protocol Amending the TRIPS Agreement (2005)
Visa and Asylum	
MIG12	Does the agreement provide a mutual recognition scheme (on qualifications, training, work experience)?
MIG14	Does the agreement provide a quota on number of visas to be issued to natural persons of parties?
Movement of Capital	
MOC26	Does the transfer provision explicitly exclude "good faith and non-discriminatory application of its laws" related to prevention of deceptive and fraudulent practices?
MOC27	Does the transfer provision explicitly exclude "good faith and non-discriminatory application of its laws" governing capital account regulations?
MOC28	Does the agreement contain Country annexes with specific transfers reservations by individual parties?
Public Procurement	
PP08	Does the agreement contain explicit provisions on MFN treatment of third parties?
Rules of Origin	
ROR02	Does the certificate have to be issued by competent authorities of the exporting party, including customs administrations, other government authorities, and designated private ones?
ROR04	Is there a certificate exemption?
ROR13	Does the agreement allow for diagonal cumulation?
ROR16	Does the agreement allow for full cumulation?
Services	
SER46	Does the agreement contain a ratchet provision- implying all unilateral liberalization is legally bound?
Sanitary and Phytosanitary Measures	
SPS06	Are there specified existing standards to which countries shall harmonize?
SPS07	Is equivalence recognised?
SPS09	Is mutual recognition recognised?
SPS11	Is the creation of concerted/regional standards referenced?
SPS16	Do parties reference international standards?
SPS18	Is there reference to international standards/procedures?
SPS21	Are there specified existing standards to which countries shall harmonize?
SPS22	Does the importing party have the right to audit the exporting party's competent authorities, inspection systems, or production procedure?
SPS23	Is mutual recognition in force?
SPS24	Is the burden of justifying non-equivalence on the importing country?
SPS33	Is there a provision on control and inspection?
SPS34	Do parties have to notify each other prior to the entry into force of a new standard or regulation?

Table A1 (continued): Main Provisions

State-Owned Enterprises	
STE30	Does the agreement regulate subsidization to state enterprises?
STE31	Does the agreement prohibit anti-competitive behavior of state enterprises?
STE32	Does the agreement require state enterprises not to distort trade?
STE37	Does the agreement indicate the geographical market where the objectionable conduct or the effect takes place?
Subsidies	
SUB03	Does the agreement prohibit or regulate export subsidies?
SUB07	Does the agreement introduce any ceiling to permitted subsidies?
SUB09	Does the agreement include any specific regulation of agricultural subsidies?
SUB10	Does the agreement include any specific regulation of fisheries subsidies?
SUB11	Does the agreement include any specific discipline for public services?
SUB12	Does the agreement include any other specific discipline for certain sectors or objectives?
SUB13	Does the agreement include any national treatment obligation (goods) for subsidies?
SUB14	Does the agreement include any national treatment obligation (services or establishment) for subsidies?
Technical Barriers to Trade	
TBT02	Technical Regulations - Is mutual recognition in force?
TBT04	Technical Regulations - Is the burden of justifying non-equivalence on the importing country?
TBT05	Technical Regulations - Are there specified existing standards to which countries shall harmonize?
TBT06	Technical Regulations - Is the use or creation of regional standards promoted?
TBT07	Technical Regulations - Is the use of international standards promoted?
TBT08	Conformity Assessment - Is mutual recognition in force?
TBT10	Conformity Assessment - Do parties participate in international or regional accreditation agencies?
TBT11	Conformity Assessment - Is the burden of justifying non-equivalence on the importing country?
TBT14	Conformity Assessment - Is the use or creation of regional standards promoted?
TBT15	Conformity Assessment - Is the use of international standards promoted?
TBT29	Standards - Is mutual recognition in force?
TBT31	Standards - Is the burden of justifying non-equivalence on the importing country?
TBT32	Standards - Are there specified existing standards to which countries shall harmonize?
TBT33	Standards - Is the use or creation of regional standards promoted?
TBT34	Standards - Is the use of international standards promoted?
Trade Facilitation and Customs	
TF25	Freedom of transit for goods
TF41	Does the agreement regulate harmonization and common legal framework
TF42	Does the agreement regulate customs and other duties collection?
TF43	Does the agreement require the sharing of customs revenues?
TF44	Do trade facilitation provisions simplify requirements for proof of origin?
TF45	Does trade facilitation provisions simplify procedures to issue proof of origin?

Appendix B: More details on implementation

This Appendix describes how we implement our PPML-Lasso methods from a computational and conceptual perspective. Our procedures and code are available to use via the R package penppml (Garrucho, Zylkin, Cruz and Apfel, 2021). Additionally, we provide some heuristic discussion of the bootstrap lasso and iceberg lasso, and present a simple example that illustrates the selection performance advantages of these methods over the plug-in lasso when we have high correlation between regressors.

Details on HDFE-PPML-lasso estimation

The minimization problem that defines the three-way PPML-lasso is

$$(\widehat{\alpha},\widehat{\gamma},\widehat{\eta},\widehat{\beta}) := \arg\min_{\alpha,\gamma,\eta,\beta} \left[\frac{1}{n}\sum_{i,j,t} \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij}) \right.$$
$$\left. -\frac{1}{n}\sum_{i,j,t} y_{ijt}\left(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij}\right) + \frac{1}{n}\sum_{k=1}^{p}\widehat{\phi}_k\lambda|\beta_k| \right], \tag{1}$$

where $x'_{ijt} = (x_{1,ijt},\ldots,x_{p,ijt})$ and $\widehat{\phi}_k$, to be precisely defined below, is identically equal to 1 except when the plug-in method is used.

The first-order conditions (FOCs) for this problem are

$$\widehat{\alpha}_{it} : \frac{1}{n}\sum_{j} y_{ijt} - \widehat{\mu}_{ijt} = 0, \qquad \forall i,t,$$
$$\widehat{\gamma}_{jt} : \frac{1}{n}\sum_{i} y_{ijt} - \widehat{\mu}_{ijt} = 0, \qquad \forall j,t,$$
$$\widehat{\eta}_{ij} : \frac{1}{n}\sum_{t} y_{ijt} - \widehat{\mu}_{ijt} = 0, \qquad \forall i,j,$$
$$\widehat{\beta}_k : \begin{cases} \frac{1}{n}\sum_{i,j,t}\left(y_{ijt} - \widehat{\mu}_{ijt}\right)x_{k,ijt} - \frac{1}{n}\widehat{\phi}_k\lambda\text{sign}\left(\widehat{\beta}_k\right) = 0 & \text{if} \quad \widehat{\beta}_k \neq 0, \\ \left|\frac{1}{n}\sum_{i,j,t}\left(y_{ijt} - \widehat{\mu}_{ijt}\right)x_{k,ijt}\right| - \frac{1}{n}\widehat{\phi}_k\lambda \leq 0 & \text{if} \quad \widehat{\beta}_k = 0. \end{cases} \qquad k = 1\ldots p,$$

where $\widehat{\mu}_{ijt}$ denotes $\mu_{ijt} := e^{x_{ijt}\beta+\alpha_{it}+\gamma_{jt}+\eta_{ij}}$ evaluated at $\widehat{\alpha}$, $\widehat{\gamma}$, $\widehat{\eta}$, $\widehat{\beta}$. The condition in the FOC for $\widehat{\beta}_k$ captures the possibility that some variables may not improve the fit of the model enough to justify their impact on the penalty. Notice that the penalty only affects the FOCs for the main covariates of interest. The FOCs for the fixed effects are exactly the same as they would be in unpenalized PPML. That said, further simplification is still needed because it is generally not practical to estimate all of the parameters directly, with or without the penalty. Instead, we first need to "concentrate out" the fixed effect parameters. That is, instead of minimizing (1) over all of the parameters, we treat $\widehat{\alpha}_{it}$, $\widehat{\gamma}_{it}$, and $\widehat{\eta}_{it}$ as functions of $\widehat{\beta}$ that are implicitly defined by their FOCs. The resulting "concentrated" minimization problem is

$$\widehat{\beta} := \arg\min_{\beta} \left[\frac{1}{n}\sum_{i,j,t} \exp\left(x'_{ijt}\beta + \widehat{\alpha}_{it}(\beta) + \widehat{\gamma}_{jt}(\beta) + \widehat{\eta}_{ij}(\beta)\right) \right.$$
$$\left. -\frac{1}{n}\sum_{i,j,t} y_{ijt}\left(x'_{ijt}\beta + \widehat{\alpha}_{it}(\beta) + \widehat{\gamma}_{jt}(\beta) + \widehat{\eta}_{ij}(\beta)\right) + \frac{1}{n}\sum_{k=1}^{p}\widehat{\phi}_k\lambda|\beta_k| \right], \tag{2}$$

such that β is now the only argument we need to solve for. The FOC for each non-zero $\widehat{\beta}_k$ associated with this modified problem is:

$$\widehat{\beta}_k : \frac{1}{n}\sum_{i,j,t}\left(y_{ijt} - \exp\left(x'_{ijt}\widehat{\beta} + \widehat{\alpha}_{it}\left(\widehat{\beta}\right) + \widehat{\gamma}_{jt}\left(\widehat{\beta}\right) + \widehat{\eta}_{ij}\left(\widehat{\beta}\right)\right)\right)\widetilde{x}_{k,ijt} - \frac{1}{n}\widehat{\phi}_k\lambda\text{sign}\left(\widehat{\beta}_k\right) = 0,$$

where

$$\widetilde{x}_{k,ijt} := x_{k,ijt} + \frac{d\widehat{\alpha}_{it,k}}{d\beta} + \frac{d\widehat{\gamma}_{it,k}}{d\beta} + \frac{d\widehat{\eta}_{ij,k}}{d\beta} \tag{3}$$

captures both the direct and indirect effects of a change in β on the conditional mean of y_{ijt}. Similarly, the associated FOC in instances where $\widehat{\beta}_k = 0$ is

$$\left| \frac{1}{n} \sum_{i,j,t} \left(y_{ijt} - \exp\left(x'_{ijt}\widehat{\beta} + \widehat{\alpha}_{it}\left(\widehat{\beta}\right) + \widehat{\gamma}_{jt}\left(\widehat{\beta}\right) + \widehat{\eta}_{ij}\left(\widehat{\beta}\right) \right) \right) \widetilde{x}_{k,ijt} \right| - \frac{1}{n}\widehat{\phi}_k \lambda \leq 0.$$

To explain how we deal with the fixed effects, assume for the moment that we know the true values of μ_{ijt} that we will eventually estimate. If that is the case, then the penalized PPML solution $(\beta, \alpha, \gamma, \eta)$ is also the solution to the following weighted least squares problem

$$\min_{\beta} \left[\frac{1}{2n} \sum_{i,j,t} \mu_{ijt} \left(z_{ijt} - \alpha_{it} - \gamma_{jt} - \eta_{ij} - x'_{ijt}\beta \right)^2 + \frac{1}{n} \sum_{k=1}^{p} \widehat{\phi}_k \lambda \left| \beta_k \right| \right],$$

where

$$z_{ijt} = \frac{y_{ijt} - \mu_{ijt}}{\mu_{ijt}} + \ln \mu_{ijt}$$

is the transformed dependent variable that is used to motivate estimation via iteratively re-weighted least squares (IRLS). The convenient thing about this representation of the problem is that we can rewrite it as

$$\min_{\beta} \left[\frac{1}{2} \sum_{i,j,t} \mu_{ijt} \left(\widetilde{z}_{ijt} - \widetilde{x}'_{ijt}\beta \right)^2 + \sum_{k=1}^{p} \lambda \widehat{\phi}_k \left| \beta_k \right| \right], \tag{4}$$

where $\widetilde{z}_{ijt}$ and $\widetilde{x}_{ijt}$ are respectively defined as the "partialled-out" versions of x_{ijt} and z_{ijt}, which are obtained by within-transforming x_{ijt} and z_{ijt} with respect to it, jt, and ij and weighting by μ_{ijt}. The within-transformation steps involved in computing $\widetilde{z}_{ijt}$ and $\widetilde{x}_{ijt}$ are the same as in Correia, Guimarães, and Zylkin (2020) and can be computed quickly using the methods of Gaure (2013). Furthermore, one can show that the $\widetilde{x}_{ijt}$ that appears in (4) is consistent with the definition given for $\widetilde{x}_{k,ijt}$ in (3).

The nice thing about expressing the problem as in (4) is that it now resembles a simple penalized regression problem. It can thus be quickly solved using the coordinate descent algorithm of Friedman, Hastie, and Tibshirani (2010). Furthermore, though we do not know the correct estimation weights (the μ_{ijt}s) beforehand, we can follow the approach of Correia, Guimarães, and Zylkin (2020) by repeatedly updating them until convergence after each new estimate of β, as in IRLS estimation. Altogether, our algorithm closely follows Correia, Guimarães, and Zylkin (2020) and otherwise only involves swapping out their weighted least squares step for a penalized weighted least squares step, as shown in (4). In principle, this algorithm can be easily adapted to other settings that feature multi-way fixed effects in order to simplify estimation.

Details on cross-validation

Cross-validation (CV) is the traditional method to choose the penalty parameter. The idea behind CV is to repeatedly hold out a subset of the sample during estimation and then use it to validate the resulting estimates. In our setup, rather than holding out observations in an unstructured way, we keep together all observations for which a given agreement is in effect, and hold out subsets of agreements. Doing so allows us to obtain estimates for the all the fixed effects in the model.

To describe the implementation of CV, suppose that the observations associated with trade agreements are partitioned into G subsets indexed by $g = 1, \ldots, G$. Each resulting hold-out sample g will have n_g observations, where n_g is the number of observations associated with agreements that are held out in partition g. Because our variables of interest are all dummies, a problem that may occur is that over some subsamples some regressors may not be present, but that is less likely to happen when G is large.

The CV approach sets all regressor-specific penalty weights $\widehat{\phi}_k$ equal to 1. Let $\widehat{\beta}_g(\lambda)$ be the lasso estimator obtained via the minimization of (2) when holding out the n_g observations contained in partition g. Then, the CV bandwidth is defined as

$$\lambda_{CV} = \arg\min_{\lambda \in \Lambda} \left[\frac{1}{G} \sum_{g=1}^{G} \frac{1}{n_g} \sum_{(i,j,t) \in g} \Big(y_{ijt} - \exp\left(x'_{ijt} \widehat{\beta}_g(\lambda) + \alpha_{it}\left(\widehat{\beta}_g(\lambda)\right) + \gamma_{jt}\left(\widehat{\beta}_g(\lambda)\right) + \eta_{ij}\left(\widehat{\beta}_g(\lambda)\right)\right)\Big)^2 \right].$$

Since λ_{CV} is based on the minimization of the average out-of-sample mean square error over different subsamples, we expect it to deliver a much more lenient variable selection. There is some disagreement in the literature over whether dummy variables, such as the ones used in our application, should be standardized before applying the CV lasso. This consideration is in contrast to the plug-in lasso, since standardization of the covariates simply causes the $\widehat{\phi}_k$ terms to be re-scaled without otherwise affecting estimation in that case. We have computed CV lasso results with and without first standardizing and found that the results with standardization are noticeably more similar to the plug-in lasso results. Thus, our preference is to work with standardized dummy covariates.

Details on plug-in lasso

Rather than relying on out-of-sample performance, the Belloni, Chernozhukov, Hansen, and Kozbur (2016) "plug-in" lasso method chooses the penalty parameters λ and $\widehat{\phi}_k$ using statistical arguments. Their specific framework is a simple linear panel data model, but their reasoning involves modifying the standard lasso penalty to reflect the variance of the score. These concepts are quite general; thus, we can modify their approach to take into account the more complex case of a nonlinear model with multiple fixed effects.

The key condition in choosing these penalty parameters is that, for all k and for some $c > 1$, they should satisfy the following inequality with probability tending to one

$$\frac{\lambda \widehat{\phi}_k}{n} \geq c \left| \frac{1}{n} \sum_{i,j,t} (y_{ijt} - \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij})) \widetilde{x}_{k,ijt} \right| \quad \forall k. \tag{5}$$

Intuitively,

$$\left| \frac{1}{n} \sum_{i,j,t} (y_{ijt} - \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij})) \widetilde{x}_{k,ijt} \right|$$

is the absolute value of the score for β_k. When evaluated at $\beta_k = 0$, it tells us to what degree moving each β_k away from zero will affect the fit of the model. If it does not produce a sufficient improvement in fit as compared to the penalty $\lambda \widehat{\phi}_k$, then regressor $x_{k,ijt}$ will not be selected.

Next, suppose again that the observations associated with trade agreements are partitioned into G clusters, and let $o = (i, j, t)$ serve as the unique index for each observation. Set

$$\widehat{\phi}_k^2 = \frac{1}{n} \sum_{g} \left(\sum_{o \in g} \widetilde{x}_{k,o} \widehat{\epsilon}_o \right)^2 = \frac{1}{n} \sum_{g} \sum_{o \in g} \sum_{o' \in g} \widetilde{x}_{k,o} \widetilde{x}_{k,o'} \widehat{\epsilon}_o \widehat{\epsilon}_{o'},$$

where $\widehat{\epsilon}_o = \widehat{\epsilon}_{ijt} = y_{ijt} - \exp(x'_{ijt}\widehat{\beta} + \widehat{\alpha}_{it} + \widehat{\gamma}_{jt} + \widehat{\eta}_{ij})$, but can also be obtained as $\widehat{\epsilon}_o = \widehat{\epsilon}_{ijt} = \widehat{\mu}_{ijt}(\widetilde{z}_{ijt} - \widetilde{x}'_{ijt}\widehat{\beta})$. By inspection, this expression provides an estimate of the variance of the score for β_k allowing the errors to be correlated within their respective clusters. Under suitable regularity conditions, $\widehat{\phi}_k^2 - \left(\phi_k^0\right)^2 = o_p(1)$ uniformly in k, where ϕ_k^0 is the analogue of $\widehat{\phi}_k^2$ evaluated at the true values of ϵ_{ijt}. By choosing $\widehat{\phi}_k$ in this way we ensure that the score for β_k when evaluated at zero must be large as compared to its standard error in order for regressor k to be selected.

The choice of λ then involves selecting a value that is sufficiently large to ensure that the probability of an irrelevant regressor being selected is small. By the maximal inequality for self-normalized sums (see Jing, Shao, and Wang, 2003), it follows that

$$\frac{\Pr\left(\widehat{\phi}_k^{-1}\frac{1}{\sqrt{n}}\sum_{i,j,t}\widetilde{x}_{k,ijt}\epsilon_{ijt}\geq m\right)}{\Pr\left(N(0,1)\geq m\right)}=o(1),$$

for $|m|=o(n^{1/6})$, thus establishing a bound for the tails of the normalized sum. This suggests that by choosing a λ that is sufficiently large to dominate a p-dimensional standard normal, the inequality in (5) is satisfied. Hence, following Belloni, Chernozhukov, Hansen, and Kozbur (2016), we set

$$\lambda=\lambda_{plug}=2c\sqrt{n}\Phi^{-1}\left(1-\gamma/2p\right), \tag{6}$$

where $c=1.1$ and $\gamma=0.1/\ln(n)$.

As discussed in the main text, after the lasso step, we then perform an unpenalized PPML estimation using the selected covariates, a so-called "post-lasso" regression. Let $\widehat{\beta}_{PL}$ be the estimator of the parameters associated with the s selected covariates. Such an estimator is said to have the "oracle property" if the asymptotic distribution of $\widehat{\beta}_{PL}$ coincides with that of the estimator we would obtain if we knew exactly which coefficients were equal to zero, i.e., for large enough samples we would have $\widehat{\beta}_{PL,k}=0$ if and only if $\beta_k^0=0$ for $k=1,\ldots,p$, where we use the superscript 0 to signify the true values. Hence, for estimators with the oracle property, asymptotically the post-lasso model is indeed the right model. In general, the lasso does not satisfy the oracle property. Nevertheless, under some additional conditions, the use of the plug-in lasso method just described ensures the following "near-oracle" property for $\widehat{\beta}_{PL}$,

$$\left\|\widehat{\beta}_{PL}-\beta^0\right\|_1=O_p\left(\sqrt{\frac{s^2\ln\left(n\right)}{n}}\right),$$

and hence the post-lasso estimates are consistent at a rate that differs from the oracle rate only up to the log factor $\ln(n)$. In Appendix C below, we establish that this property holds in our setting.

In practice, the plug-in lasso method mainly requires adding one additional step to the procedure used for the estimation of the PPML-lasso with high-dimensional fixed effects described before. Though the $\widehat{\phi}_k$ penalty terms are not known beforehand, they, too, can be iterated on in the same fashion as μ_{ijt}. Simply use the most recent values of $\widehat{\epsilon}_{ijt}$ (obtained using post-lasso PPML) in each iteration to construct new values for $\widehat{\phi}_k$. It also requires an initial value for $\widehat{\mu}_{ijt}$. For this, we first estimate a three-way gravity model with a single dummy for PTA using PPML.

Bootstrap lasso and iceberg lasso

When two or more regressors are highly collinear, the plug-in lasso may select only one variable from a set of close substitutes. This reflects a well-known property of the lasso: in the presence of a group of variables with very high pairwise correlation, it tends to select only one variable from the group rather than selecting the entire group (see, e.g., Zou and Hastie, 2005).[1] As we have noted, this issue is highly relevant for our setting because the high collinearity we observe between many of the provision variables in our provisions data.

Our primary approach to addressing this issue is the bootstrap lasso, which uses bootstrap resampling to exploit the instability of variable selection in finite samples to recover additional relevant variables. This idea is closely related to the stability-based selection method of Meinshausen and Bühlmann (2010) and to the random lasso approach of Wang et al. (2011). We share with Wang et al. (2011) the idea of using resampling as a way of identifying additional variables in the presence of multicollinearity.

The key intuition behind the bootstrap lasso is that when two or more regressors are highly correlated, small changes in the sample, such as those induced by resampling, can lead the lasso to select different

[1] To put in formal terms, consider our Assumption A.4 in the preceding theoretical discussion that minimum (sparse) eigenvalue of the design matrix $M_n=\frac{1}{n}\sum_{i,j,t}\mu_{ijt}\widetilde{x}_{ijt}\widetilde{x}'_{ijt}$ to be bounded above zero. If two or more regressors are very highly correlated, the smallest eigenvalue can be close to zero.

members of the same group. This can be framed as a finite sample issue because, when a relevant variable is highly correlated with an irrelevant one (or with a group of irrelevant ones), the variable that gets selected is likely to depend narrowly on the sampling error in the data we observe. More practically speaking, it is likely to depend on a small number of influential observations that determine the sample correlation between any irrelevant variables that are highly correlated with relevant ones and the noise in the data. By reshuffling the data, we may uncover more relevant variables previously hidden by their correlations with other variables, though at the expense of selecting additional irrelevant ones.

For simplicity, we ignore cross sectional correlation in the following, though our empirical implementation does allow for correlation across certain pairs. In this case, at each replication, we can make $N(N-1)$ independent draws of the full time series for each pair, thereby allowing for arbitrary time dependence within pairs.[2] Let $\widehat{\mathcal{S}}_0$ denote the set of variables originally selected by plug-in lasso, and let $\widehat{\mathcal{S}}_b$ denote the set of variables selected by the plug-in lasso using bootstrap sample b, with $b = 1, \ldots, B-1$. The final set of variables selected by the bootstrap lasso, denoted $\widehat{\mathcal{S}}^B$, is the set of variables selected in at least a fraction α of the B samples considered, i.e.,

$$\widehat{\mathcal{S}}^B = \left\{ k : \left(\frac{1}{B} \sum_{b=0}^{B-1} 1\left[k \in \widehat{\mathcal{S}}_b \right] \right) \geq \alpha \right\}.$$

The use of α as a threshold ensures that the procedure does not mechanically select all variables as $B \to \infty$. We still prefer a relatively low value for α to ensure that we can recover the variables that are selected with low frequency across all samples. In our empirical implementation, we have used $\alpha = 0.01$ and 0.05 as thresholds.[3]

As a complementary approach, we also introduce what we call the iceberg lasso, which directly searches for variables that are highly predictive of those selected by the plug-in lasso. Again, let $\widehat{\mathcal{S}}_0$ denote the set of variables selected by plug-in lasso and let $\widehat{s}$ be the cardinality of $\widehat{\mathcal{S}}_0$. For $k = 1, \ldots, \widehat{s}$, we estimate an auxiliary plug-in lasso regression of the k-th element of $\widehat{\mathcal{S}}_0$ on the set of non-selected variables. That is, for each $k = 1, \ldots, \widehat{s}$, we estimate

$$\widehat{\delta}_n^{(k)} = \arg\min_{\delta^{(k)}} \frac{1}{n} \sum_{i,j,t} \left(x_{k,ijt} - \sum_{k' \notin \widehat{\mathcal{S}}_0} \delta_{k'}^{(k)} x_{k',ijt} \right)^2 + \frac{\lambda}{n} \sum_{k' \notin \widehat{\mathcal{S}}_0} \widehat{\psi}_{k'}^{(k)} \left| \delta_{k'}^{(k)} \right|,$$

where $\widehat{\psi}_{k'}^{(k)} = \sum_{i,j} \left(\sum_t x_{k',ijt} \widehat{e}_{ijt}^{(k)} \right)^2$, with $\widehat{e}_{ijt}^{(k)}$ being the residual from the k-th regression.[4]

Let $\widehat{\mathcal{S}}_k$ denote the set of variables selected in the k-th regression. Then, the final selected set of variables is $\widehat{\mathcal{S}}^{IL} = \cup_{k=0}^{\widehat{s}} \widehat{\mathcal{S}}_k$; that is, the union of the set of variables in $\widehat{\mathcal{S}}_0$ with the sets of those selected in each of the $\widehat{s}$ regressions in the second step. The variables selected in these regressions are good predictors of at least one of the variables in $\widehat{\mathcal{S}}_0$, and this may have led to their exclusion from $\widehat{\mathcal{S}}_0$ even if they have a causal impact on the outcome y_{ijt}. At the same time, similar to how we interpret the results from the bootstrap lasso, the iceberg lasso is agnostic about which of the variables in $\widehat{\mathcal{S}}_k^{IL}$ are relevant for y_{ijt}; by construction, it will include irrelevant variables that are closely associated with the relevant ones.

A simple example

To illustrate how the bootstrap lasso and iceberg lasso can help recover important variables not selected by the plug-in lasso, suppose that

$$\mathrm{E}(y_{ijt} | x_{1,ijt}, \ldots, x_{p,ijt}, \alpha_{it}^0, \gamma_{jt}^0, \eta_{ij}^0) = \exp\left(\alpha_{it}^0 + \gamma_{jt}^0 + \eta_{ij}^0 + x_{1,ijt} \right), \tag{7}$$

[2] More precisely, a typical bootstrap sample is $(y_{ij}^*, x_{ij}^*, \alpha_i^*, \gamma_j^*, \eta_{ij}^*)$, with $y_{ij}^* = \left(y_{ij,1}^*, \ldots y_{ij,T}^* \right)$, and $y_{ij,t}^* = y_{I_{ij},t}$ with I_{ij} being independent discrete uniform on $[1, \ldots, N(N-1)]$; $x_{ij,t}^*$, $\alpha_{i,t}^*$, $\gamma_{j,t}^*$, η_{ij}^* are defined analogously. We treat the fixed effects no differently than any other regressor when re-sampling.

[3] Though the bootstrap does not provide a valid approximation to the limiting distribution of post-selection estimators in this setting (see, e.g., Chatterjee and Lahiri, 2010), our goal here is not to conduct inference but rather to use selection frequencies as a diagnostic for variable relevance under multicollinearity.

[4] Again, for simplicity, we ignore cross-sectional correlation in our presentation here, but we do allow for this type of correlation in our implementation.

where $x_{1,ijt}, \ldots, x_{p,ijt}$ are potential explanatory variables with zero mean and unit variance. Moreover, assume that $x_{2,ijt}$ is the only potential regressor correlated with $x_{1,ijt}$, which is the only variable with a non-zero coefficient in (7).

Using the plug-in penalty, the decision of whether to include a certain variable is based on the value of its score, evaluated after taking into account the contributions of the regressors already included in the model. When $x_{1,ijt}$ and $x_{ijt,2}$ are highly collinear, the scores of $x_{1,ijt}$ and $x_{ijt,2}$ are very close, and thus which of these variables gets selected is likely to be driven by the small number of observations where the differences between $x_{1,ijt}$ and $x_{ijt,2}$ are larger.

This explains why the bootstrap lasso can be an effective remedy in this setting. In cases where the difference between the scores of $x_{1,ijt}$ and $x_{ijt,2}$ is disproportionately determined by only a small number of observations, bootstrap resampling changes the weight placed on the observations that contribute more to this difference. Therefore, which of the variables is selected will change from one sample to another. Consequently, the bootstrap lasso is likely to select both $x_{1,ijt}$ and $x_{2,ijt}$ if the correlation between them is sufficiently high, which is desired outcome in this setting. Additionally, some of the remaining, irrelevant, regressors $x_{3,ijt}, \ldots, x_{p,ijt,}$ may also be selected by chance in some samples, but our use of a threshold rule should help to screen out variables whose selection is driven by idiosyncratic noise as opposed to the systematic instability caused by multicollinearity.

The iceberg lasso also has straightforward implications in this simple setting. To illustrate the different possibilities, suppose first that the plug-in lasso selects only $x_{1,ijt}$ or $x_{2,ijt}$. Then, we proceed to the second step of the iceberg lasso and we are very likely to select the variable correlated with the one selected in the first step. It is also possible that the plug-in lasso selects both $x_{1,ijt}$ or $x_{2,ijt}$, in which case the second step is unlikely to select further variables. Therefore, as with the bootstrap lasso, in these situations the iceberg lasso is likely to select both $x_{1,ijt}$ and $x_{1,ijt}$, which is the desired outcome. It is, of course, possible that some of the remaining variables are also selected, but this is less likely when they have low correlation with $x_{1,ijt}$. A third possibility is that the plug-in lasso does not select neither $x_{1,ijt}$ nor $x_{2,ijt}$. If this happens, the second step is unlikely to help and we may end up not selecting any of the relevant variables. However, this is not the only case where the iceberg lasso fails.

For example, consider a case where there are three or more highly correlated variables in the original set, and only one of them affects the outcome. In this case, if the plug-in lasso selects one of the irrelevant variables, then the second step may choose another of the irrelevant variables, and then the variable with causal effect is not selected by the two-step procedure. In these situations, the bootstrap lasso is more likely to be able to select the entire group but, as noted above, the usefulness of the bootstrap lasso will depend on making a wise choice of the threshold parameter α so that only the relevant variables and those highly correlated with it are selected, while the other regressors are not.

Appendix C: Validity of plug-in lasso for the three-way gravity model

In this appendix, we prove that, in spite of the incidental parameter bias documented in Weidner and Zylkin (2021), our PPML lasso estimator has the same "near-oracle" property as in Belloni, Chernozhukov, Hansen, and Kozbur (2016) for the setting of a three-way gravity model.

To add some needed precision, we will henceforth maintain that a 0 superscript denotes a true parameter value or, analogously, a function evaluated at the true parameter values. The gravity model with origin-time, destination-time, and pair fixed effects may therefore be written as

$$y_{ijt} = \exp\left(x'_{ijt}\beta^0 + \alpha^0_{it} + \gamma^0_{jt} + \eta^0_{ij}\right) u^0_{ijt} = \mu^0_{ijt} u_{ijt} = \mu^0_{ijt} + \epsilon^0_{ijt}, \tag{8}$$

where $x'_{ijt} = (x_{1,ijt}, \ldots, x_{p,ijt})$ and it is assumed that $\mathrm{E}\left(u^0_{ijt} | x_{ijt}, \alpha^0_{it}, \gamma^0_{jt}, \eta^0_{ij}\right) = 1$. Equivalently, $E\left(\epsilon^0_{ijt} | x_{ijt}, \alpha^0_{it}, \gamma^0_{jt}, \eta^0_{ij}\right) = 0$.

As in the definition for z_{ijt} in Appendix B, let

$$z^0_{ijt} = \frac{y_{ijt} - \mu^0_{ijt}}{\mu^0_{ijt}} + \ln \mu^0_{ijt}.$$

Similarly, let $\widetilde{z}^0_{ijt}$ and $\widetilde{x}^0_{ijt}$ be the corresponding partialled-out versions of z^0_{ijt} and x_{ijt}, i.e., after partialling them out with respect to the fixed effects.[5] The data is assumed to be a balanced panel with $i = 1, \ldots N$, $j = 1, \ldots N$, $t = 1, \ldots T$, and we define $n = N(N-1)T$. To simplify proofs, we allow only for time dependence within pairs, whereas in our application we allow for cross-sectional dependence across certain pairs.

One key difference with Belloni, Chernozhukov, Hansen, and Kozbur (2016) is that, in our nonlinear setting, we need to estimate the fixed effects rather than simply difference them out. Thus, we will proceed by first establishing an infeasible estimator that will allow us to build a bridge between their setting, which assumes a linear model, and our setting of an exponential gravity model with three-way fixed effects estimated via PPML. Accordingly, let $\widetilde{\beta}_n$ be the infeasible estimator that treats α^0_{it}, γ^0_{jt}, η^0_{ij}, and μ^0_{ijt} as known rather than as quantities to be estimated. Using a weighted least squares (WLS) representation of the PPML-lasso estimator, we can express $\widetilde{\beta}_n$ as

$$\widetilde{\beta}_n = \arg\min_{\beta} \left(\frac{1}{n} \sum_{i,j,t} \mu^0_{ijt} \left(\widetilde{z}^0_{ijt} - \widetilde{x}^{0\prime}_{ijt}\beta \right)^2 + \frac{\lambda}{n} \sum_{k=1}^{p} \widetilde{\phi}_k |\beta_k| \right), \tag{9}$$

where

$$\lambda = 2c\sqrt{n}\Phi^{-1}\left(1 - \gamma/2p\right), \tag{10}$$

with $c > 1$ and $\gamma = 0.1/\ln(n)$. Furthermore, let

$$\widetilde{\phi}^2_k = \frac{1}{n} \sum_{i,j} \left(\sum_t \widetilde{x}^0_{k,ijt} \widetilde{\epsilon}_{ijt} \right)^2, \tag{11}$$

where $\widetilde{\phi}_k$ is constructed following the Appendix in Belloni, Chernozhukov, Hansen, and Kozbur (2016), with $\widetilde{\epsilon}_{ijt} = \mu^0_{ijt}\left(\widetilde{z}^0_{ijt} - \widetilde{x}^{0\prime}_{ijt}\overline{\beta}_n\right)$, and $\overline{\beta}_n$ is some preliminary estimator; for example, $\overline{\beta}_n$ may be obtained in a previous step by setting $\widetilde{\phi}_k = 1$. As explained before, the construction of $\widetilde{\phi}_k$ has been simplified relative to the earlier derivation used in Appendix B by allowing for time dependence only.

An example of a feasible estimator is computed using estimated weights, i.e.,

$$\widehat{\beta}^{(2)}_n = \arg\min_{\beta} \left(\frac{1}{n} \sum_{i,j,t} \widehat{\mu}^{(1)}_{ijt} \left(\widehat{\widetilde{z}}^{(1)}_{ijt} - \widehat{\widetilde{x}}^{(1)\prime}_{ijt}\beta \right)^2 + \frac{\lambda}{n} \sum_{k=1}^{p} \widehat{\phi}^{(1)}_k |\beta_k| \right). \tag{12}$$

[5] Recall that we weight by μ_{ijt} when partialling out both z_{ijt} and x_{ijt}. Thus, it is necessary to specify that, e.g., $\widetilde{x}^0_{ijt}$ is the partialled-out version of x_{ijt} that uses μ^0_{ijt} as weights.

Here, $\widehat{\mu}_{ijt}^{(1)} = \exp\left(\widehat{\alpha}_{it}\left(\widehat{\beta}_n^{(1)}\right) + \widehat{\gamma}_{jt}\left(\widehat{\beta}_n^{(1)}\right) + \widehat{\eta}_{ij}\left(\widehat{\beta}_n^{(1)}\right) + x_{ijt}'\widehat{\beta}_n^{(1)}\right)$ where we make it explicit that the fixed effects estimates used to construct $\widehat{\mu}_{ijt}^{(1)}$ are obtained using the chosen preliminary estimate $\beta = \widehat{\beta}_n^{(1)}$. The term $\widehat{\widetilde{z}}_{ijt}^{(1)}$ represents the partialled-out version of

$$\widehat{z}_{ijt}^{(1)} = \frac{y_{ijt} - \widehat{\mu}_{ijt}^{(1)}}{\widehat{\mu}_{ijt}^{(1)}} + \ln \widehat{\mu}_{ijt}^{(1)},$$

$\widehat{\widetilde{x}}_{ijt}^{(1)}$ is correspondingly the partialled-out version of x_{ijt}, and $\widehat{\phi}_k^{(1)}$ is constructed in the same manner as $\widetilde{\phi}_k$ in (11). Both $\widehat{\widetilde{z}}_{ijt}^{(1)}$ and $\widehat{\widetilde{x}}_{ijt}^{(1)}$ use $\widehat{\mu}_{ijt}^{(1)}$ as weights when partialling out.

This feasible estimator differs from the full PPML-lasso estimator because we have only specified two steps, a preliminary estimate $\widehat{\beta}_n^{(1)}$ and a single update $\widehat{\beta}_n^{(2)}$. The full iterative algorithm, which we allow for in our proof, uses the following IRLS-like updating step for each iteration $r \geq 2$:

$$\widehat{\beta}_n^{(r)} = \arg\min_{\beta}\left(\frac{1}{n}\sum_{i,j,t}\widehat{\mu}_{ijt}^{(r-1)}\left(\widehat{\widetilde{z}}_{ijt}^{(r-1)} - \widehat{\widetilde{x}}_{ijt}^{(r-1)\prime}\beta\right)^2 + \frac{\lambda}{n}\sum_{k=1}^{p}\widehat{\phi}_k^{(r-1)}|\beta_k|\right), \tag{13}$$

which is the same as the feasible estimator we gave previously in (12) using $\widehat{\beta}^{(1)} = \widehat{\beta}^{(r-1)}$, $\widehat{\mu}^{(1)} = \widehat{\mu}^{(r-1)}$, and so on, where the $(r-1)$s denote calculations made using the estimates from the previous iteration. As $r \to \infty$, the iterative procedure based on the penalized WLS estimator in (13) converges to the same estimates as the PPML-lasso estimator of (8), like in Correia, Guimarães, and Zylkin (2020); that is,

$$\widehat{\beta}_n = \lim_{r\to\infty}\widehat{\beta}_n^{(r)}, \tag{14}$$

with $\widehat{\phi} = \lim_{r\to\infty}\widehat{\phi}^{(r-1)}$ as its associated vector of penalty loading terms.[6] Furthermore, the feasible WLS-based estimator in (12) also produces the PPML-lasso estimate if we define the preliminary estimate $\widehat{\beta}^{(1)}$ as the next-to-last update from the full algorithm before numerical convergence is achieved.

To proceed, we need to state some assumptions on the data generating process. Noting that it follows from Correia, Guimarães, and Zylkin (2020) that $\epsilon_{ijt}^0 = \mu_{ijt}^0\left(\widetilde{z}_{ijt}^0 - \widetilde{x}_{ijt}^{0\prime}\beta^0\right)$, hereafter let $\phi_k^0 = \frac{1}{n}\left(\sum_{i,j}\left(\sum_t \widetilde{x}_{k,ijt}^0\epsilon_{ijt}^0\right)^2\right)^{1/2}$ and

$$\varpi_k^0 = \left(\mathrm{E}\left[\left|\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widetilde{x}_{k,ijt}^0\epsilon_{ijt}^0\right)\right|^3\right]\right)^{1/3}.$$

Also, let $\widehat{M}_n = \frac{1}{n}\sum_{i,j,t}\widehat{\mu}_{ijt}\widehat{\widetilde{x}}_{ijt}\widehat{\widetilde{x}}_{ijt}'$, $M_n = \frac{1}{n}\sum_{i,j,t}\mu_{ijt}^0\widetilde{x}_{ijt}^0\widetilde{x}_{ijt}^{0\prime}$, and note that $M_n = \widehat{M}_n + o_p(1)$, with M_n being a $p \times p$ matrix. It is sufficient to require that blocks of M_n of size depending on the degree of sparsity have eigenvalues bounded below and above. Following Belloni, Chernozhukov, Hansen, and Kozbur (2016), we define the minimal and maximal m-sparse eigenvalue of M_n as

$$\varphi_{\min}(m)(M_n) = \min_{\delta\in\Delta(m)}\delta' M_n\delta \text{ and } \varphi_{\max}(m)(M_n) = \max_{\delta\in\Delta(m)}\delta' M_n\delta,$$

where

$$\Delta(m) = \left\{\delta \in R^p : 1\left\{\sum_{k=1}^{p}\delta_j \neq 0\right\} \leq m \text{ and } \|\delta\|_2 = 1\right\}.$$

Assumption A4 below controls the sparse eigenvalues of subsets of M_n.

To state all of our assumptions together, our results are obtained under the following conditions.

[6] Because $\widehat{\phi}$ is obtained by iterating on $\widehat{\phi}^{(r-1)}$, it again is constructed in the same manner as $\widetilde{\phi}$ in (11). Each $\widehat{\phi}_k$ therefore has the same form as the earlier generic formula given for $\widehat{\phi}_k$ in Appendix B but here is specialized to the case of cross-sectional independence across pairs (while still allowing for time dependence within pairs).

A1: (i) $\mathrm{E}\left(y_{ijt}|x_{ijt},\alpha_{it},\gamma_{jt},\eta_{ij}\right)=\exp\left(x_{ijt}'\beta^0+\alpha_{it}^0+\gamma_{jt}^0+\eta_{ij}^0\right)$,
(ii) $y_{ijt}|\alpha_{it},\gamma_{jt},\eta_{ij},x_{ijt}$ is independently distributed across i,j but not across t,[7]
(iii) $\mathrm{E}\left(y_{ijt}^8\right)<\infty$, and x_{ijt} has bounded support,
(iv) for all i, j, and t, $\left(\beta^0,\alpha_{it}^0,\gamma_{jt}^0,\eta_{ij}^0\right)$ belongs to a compact set,
(v) for all k, $\mathrm{E}(\phi_k^0)$ is bounded below and above, where as in Appendix B, ϕ_k^0 is the analogue of $\widehat{\phi}_k$ evaluated at ϵ_{ijt}^0, the true values of ϵ_{ijt}.

A2: $\sum_{k=1}^{p}1\left\{\beta_k\neq 0\right\}=s$, with $p=o(n)$, $s\ln(n)=o\left(\sqrt{n}\right)$, where $n=N(N-1)T$.

A3: (i) For $k=1,\ldots,p$, $\left(\frac{1}{T}\sum_{t=1}\mathrm{E}\left[\left(\widetilde{x}_{k,ijt}^0\epsilon_{ijt}^0\right)^2\right]\right)+\left(\frac{1}{T}\sum_{t=1}\mathrm{E}\left[\left(\widetilde{x}_{k,ijt}^0\epsilon_{ijt}^0\right)^2\right]\right)^{-1}=O\left(1\right)$,
(ii) $1\leq\max_{k=1,\ldots,p}\phi_k^0/\min_{k=1,\ldots,p}\phi_k^0=O\left(1\right)$,
(iii) $1\leq\max_{k=1,\ldots,p}\varpi_k\Big/\sqrt{\mathrm{E}\left(\phi_k^0\right)^2}=O\left(1\right)$,
(iv) $\max_{k=1,\ldots,p}\left|\phi_k-\sqrt{\mathrm{E}\left[\left(\phi_k^0\right)^2\right]}\right|\Big/\sqrt{\mathrm{E}\left[\left(\phi_k^0\right)^2\right]}=o(1)$,
(v) $\max_{k=1,\ldots,p}\left|\frac{1}{N^2-N}\sum_{i,j}\frac{1}{T}\left(\sum_t\mu_{ijt}^0\widetilde{x}_{k,ijt}^0\widetilde{z}_{ijt}^0\right)^2-\mathrm{E}\left[\frac{1}{T}\left(\sum_t\mu_{ijt}^0\widetilde{x}_{k,ijt}^0\widetilde{z}_{ijt}^0\right)^2\right]\right|\Big/\mathrm{E}\left[\left(\phi_k^0\right)^2\right]=o_p(1)$,
(vi) $\max_{k=1,\ldots,p}\left|\frac{1}{N^2-N}\sum_{i,j}\frac{1}{T}\left(\sum_t\widetilde{x}_{k,ijt}^0\epsilon_{ijt}^0\right)^2\right|\Big/\mathrm{E}\left[\left(\phi_k^0\right)^2\right]=O_p(1)$,
(vii) $\left(\max_{k=1,\ldots,p}\mu_{ijt}^0\left(\widetilde{x}_{k,ijt}^0\right)^2\Big/\mathrm{E}\left[\left(\phi_k^0\right)^2\right]\right)\frac{s\ln(n)}{N(N-1)\iota_T}=o_p(1)$ with $\iota_T=T$ if there no time dependence or weak dependence within each ij, while $\iota_T=1$ if there is strong or perfect dependence.

A4: For any constant C there exists κ_L and κ_U, possibly dependent on C but independent of n, such that $\kappa_L\leq\varphi_{\min}(Cs)\left(M_n\right)\leq\varphi_{\max}(Cs)\left(M_n\right)\leq\kappa_U$.

Note that A4 is violated when there is perfect collinearity among the variables with predictive power, since in that case $\varphi_{\min}(Cs)\left(M_n\right)$ would be zero. Moreover, in order to properly define blocks of size Cs, we need to rule out strong dependence between relevant and irrelevant variables. Note that we do not make any assumption on the type of dependence within ij pairs, i.e., strong versus weak. Nor do we require $T\to\infty$, though we do require $N\to\infty$ for consistency. Finally, note that ϕ_k^0 is not a nonrandom population parameter but rather is the infeasible sample analogue of $\widehat{\phi}_k$ that uses the true conditional mean and true error term. Thus, expectations involving ϕ_k^0 are taken over the sampling variation in the true error terms.

The following Lemma establishes the near-oracle property of Belloni, Chen, Chernozhukov and Hansen (2012) for the infeasible estimator in (9).

Lemma 1 *Let $\widetilde{\beta}_n$ be defined as in (9) and λ be defined as in (10), and let Assumptions A1-A4 hold. Then as $N\to\infty$,*

$$\left\|\widetilde{\beta}_n-\beta^0\right\|_1=O_p\left(\sqrt{\frac{s^2\ln\left(n\right)}{N(N-1)\iota_T}}\right). \tag{15}$$

The same statement applies to the post-lasso counterpart of $\widetilde{\beta}_n$. Given A2, the infeasible estimator in (9) is consistent provided $N\to\infty$ and, as usual, the convergence rate is faster in the case of weak (or no) time dependence in which case $\iota_T=T$.

We then have the following result for $\widehat{\beta}_n$, as defined in (14).

Theorem 1 *Let λ be defined as in (10) and let Assumptions A1-A4 hold.*
(a) If both $N,T\to\infty$ and $\iota_T=T$, then

$$\left\|\widehat{\beta}_n-\beta^0\right\|_1=O_p\left(\sqrt{\frac{s^2\ln\left(n\right)}{N(N-1)T}}\right)+O_p\left(\frac{s}{NT}\right), \tag{16}$$

[7] Recall again that, in the our empirics, we instead allow for some cross sectional correlation across certain pairs. We use A1(ii) to simplify the proof of Theorem 1 below.

(b) Let $\mathcal{S} = \left\{k : \beta_k^0 \neq 0\right\}$ *and* $\mathcal{S}^c$ *be the complement of* $\mathcal{S}$ *: for all* $k \in \mathcal{S}$, $\left|\beta_k^0\right| \geq N^{-(1-\delta)}$ *and for any* $k \in \mathcal{S}^c$, $\left|\beta_k^0\right| \leq N^{-(1+\delta)}$ *with* $0 < \delta < 1/2$. *Then, if either* $\iota_T = T$ *and* T *is fixed or* $\iota_T = 1$, *as* $N \to \infty$,

$$\left\|\widehat{\beta}_n - \beta^0\right\|_1 = O_p\left(\sqrt{\frac{s^2 \ln(n)}{N(N-1)}}\right) + O_p\left(\frac{s}{N}\right). \tag{17}$$

Note that in the case of either T fixed or strong time dependence, as in part (b) of the theorem, we need a beta-min type condition, requiring that the coefficients of the relevant variables and irrelevant variables are sufficiently apart. This additional condition is due to the incidental parameter bias that arises in these cases. It ensures that this bias does not affect variable selection, i.e., no variable can be selected or discarded just because of the bias due to incidental parameter bias. On the other hand, when we have weak time dependence and $T \to \infty$, as in part (a) of the theorem, such a condition is not required since in this case the incidental parameter bias becomes negligible.

The following comments apply to statements (a) and (b) in Theorem 1. For the plug-in lasso we cannot establish the variable selection oracle property, according to which $\left\{k : \beta_k^0 = 0\right\} = \left\{k : \widehat{\beta}_{k,n} = 0\right\}$; see, e.g., Zou (2006). However, note that the plug-in lasso cannot fail to select a variable with a sufficiently large coefficient. Suppose that $s \leq \overline{s} < \infty$. If for some constant c and for some $\psi > 0$, $\left|\beta_1^0\right| > c\left(\frac{\ln(n)}{N}\right)^{\frac{1}{2}-\psi}$ then $\widehat{\beta}_{1,n}$ cannot be set equal to zero, since otherwise the statement in (16) or in (17) would have been violated because $\left(\frac{\ln(n)}{N}\right)^{\frac{1}{2}-\psi} > \left(\sqrt{\frac{\ln(n)}{N(N-1)\iota_T}}\right) + O_p\left(\frac{1}{N\iota_T}\right)$. In fact,

$$N^{\psi}\iota_T^{1/2} \Pr\left(\widehat{\beta}_{1,n} = 0 \,\middle|\, \left|\beta_1^0\right| > c\left(\frac{\ln(n)}{N}\right)^{\frac{1}{2}-\psi}\right) \to 0.$$

On the other hand, if $\beta_1^0 = 0$, then also the probability of selecting $\widehat{\beta}_{1,n} > c\sqrt{\frac{\ln(n)}{N(N-1)\iota_T}}$ will go to zero. Finally, the second term on the right-hand side of equations (16) and (17), which does not appear in Lemma 1, reflects the bias due to the incidental parameter problem described in Weidner and Zylkin (2021). In both cases, it should never dominate the first term.

Proof of Lemma 1: For μ_{ijt}^0 known, $\widetilde{\beta}_n$ is a weighted version of the OLS estimators in equation (2.1) in Belloni, Chernozhukov, Hansen, and Kozbur (2016). Hence, the statement in the Proposition follows once we show that their Assumptions are satisfied. Assumptions A2 and A4 correspond to their assumptions ASM and SE and to (iv) of their Condition R. Assumptions A3(i)-(iii) correspond to (i)-(iii) of their Condition R and A3(iv) corresponds to (v) of their Condition R. Finally, Assumptions A3(v)-(vii) correspond to their Condition R', which is needed to ensure that, for all k, $l\widehat{\phi}_k \leq \phi_k^0 \leq u\widehat{\phi}_k$ for $l, u \to 1$ as the sample increases. Since $p < n$, the law of large numbers would apply and ensure that $\widehat{\phi}_k - \phi_k^0 = o_p(1)$. ■

Proof of Theorem 1: Let

$$\beta_n^{*(r)} = \arg\min_{\beta}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)} - \widetilde{x}_{ijt}^{\prime(r-1)}\beta\right)^2 + \frac{\lambda}{n}\sum_{k=1}^{p}\phi_k^{*(r-1)}\left|\beta_k\right|\right), \tag{18}$$

where $\widetilde{z}_{ijt}^{(r)}$ and $\widetilde{x}_{ijt}^{(r)}$ are the partialled-out analogs of

$$z_{ijt}^{(r)} = \frac{y_{ijt} - \mu_{ijt}^{(r)}}{\mu_{ijt}^{(r)}} + \ln\mu_{ijt}^{(r)},$$

and x_{ijt}, $\phi_k^{*(r-1)}$ has been computed using $\widetilde{z}_{ijt}^{(r-1)} - \widetilde{x}_{ijt}^{\prime(r-1)}\beta_n^{*(r-1)}$, and

$$\mu_{ijt}^{(r)} = \exp\left(x_{ijt}'\beta + \alpha_{it}\left(\beta_n^{*(r)}\right) + \gamma_{jt}\left(\beta_n^{*(r)}\right) + \eta_{ij}\left(\beta_n^{*(r)}\right)\right),$$

with $\alpha_{it}\left(\beta_n^{*(r)}\right)$ being the probability limit as $N,T\to\infty$ of $\widehat{\alpha}_{it}\left(\beta\right)$ when evaluated at $\beta=\beta_n^{*(r)}$, the same applies to $\gamma_{jt}\left(\beta_n^{*(r)}\right)$ and $\eta_{ij}\left(\beta_n^{*(r)}\right)$.

Also, note that the difference between $\widehat{\beta}_n^{(r)}$, as defined in (13), and $\beta_n^{*(r)}$ is that the latter is based on pseudo-true fixed effects and thus is not influenced by the bias caused by the estimation noise in the fixed effects estimates. Now,

$$\left\|\widehat{\beta}_n^{(r)}-\beta^0\right\|_1=\left\|\left(\widehat{\beta}_n^{(r)}-\beta_n^{*(r)}\right)+\left(\beta_n^{*(r)}-\widetilde{\beta}_n\right)+\left(\widetilde{\beta}_n-\beta^0\right)\right\|_1,$$

so that

$$\lim_{r\to\infty}\left\|\widehat{\beta}_n^{(r)}-\beta^0\right\|_1=\left\|\widehat{\beta}_n-\beta^0\right\|_1\leq\lim_{r\to\infty}\left\|\widehat{\beta}_n^{(r)}-\beta_n^{*(r)}\right\|_1+\lim_{r\to\infty}\left\|\beta_n^{*(r)}-\widetilde{\beta}_n\right\|_1+\left\|\widetilde{\beta}_n-\beta_0\right\|_1, \tag{19}$$

and by Lemma 1,

$$\left\|\widetilde{\beta}_n-\beta_0\right\|_1=O_p\left(\sqrt{\frac{s^2\ln\left(n\right)}{N(N-1)\iota_T}}\right).$$

The remainder of the proof is thus dedicated to characterizing the other two terms on the RHS of (19). We first prove that

$$\lim_{r\to\infty}\left\|\beta_n^{*(r)}-\widetilde{\beta}_n\right\|_1=o_p\left(\sqrt{\frac{s^2\ln\left(n\right)}{N(N-1)\iota_T}}\right), \tag{20}$$

with $\widetilde{\beta}_n$ defined as in (9). The proof of (20) is the same for $\iota_T=T$ or $\iota_T=1$, and T fixed or $T\to\infty$. Second, we prove that

$$\lim_{r\to\infty}\left\|\beta_n^{*(r)}-\widehat{\beta}_n^{(r)}\right\|_1=O_p\left(\frac{s}{N\iota_T}\right). \tag{21}$$

The proof of (21) requires the additional beta-min condition in part (b), for the case of either $\iota_T=1$ or T fixed. Given (20) and (21), the statements in the Theorem then follow from Lemma 1.

We begin by proving (20). In the sequel, we shall also need the following result

$$\lim_{r\to\infty}\left(\beta_n^{*(r)}-\beta_0\right)=o_p(1). \tag{22}$$

Thus, we first need to show (22). Let

$$\beta^{(r)}=\arg\min_{\beta\in\mathcal{B}}\mathrm{E}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{\prime(r-1)}\beta\right)^2+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_k^{*(r-1)}\left|\beta_k\right|\right),$$

where by A1(iv) $\mathcal{B}$ is a compact set. Also, note that given A1-A4, and recalling that $p<n$, we have that by the uniform law of large numbers,

$$\sup_{\beta\in\mathcal{B}}\left|\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{\prime(r-1)}\beta\right)^2+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_k^{*(r-1)}\left|\beta_k\right|\right.$$
$$\left.-\mathrm{E}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{\prime(r-1)}\beta\right)^2+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_k^{*(r-1)}\left|\beta_k\right|\right)\right|=o_p(1).$$

Hence, since the arg min is a continuous operation,

$$\arg\min_{\beta\in\mathcal{B}}\left(\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{\prime(r-1)}\beta\right)^2+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_k^{*(r-1)}\left|\beta_k\right|\right)\right.$$
$$\left.-\arg\min_{\beta\in\mathcal{B}}\mathrm{E}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{\prime(r-1)}\beta\right)^2+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_k^{*(r-1)}\left|\beta_k\right|\right)\right)=o_p(1).$$

Since the objective function is convex and so there is a unique minimum, it follows that $\lim_{r\to\infty}\left(\beta_n^{*(r)}-\beta^{(r)}\right)=o_p(1)$. From Nelder and Wedderburn (1972), it follows that

$$\lim_{r\to\infty}\left(\beta^{(r)}-\beta^{(r-1)}\right)=0.$$

To proceed, let $\mu_{ijt}^{\beta}=\exp\left(\alpha_{it}\left(\beta\right)+\gamma_{jt}\left(\beta\right)+\eta_{ij}\left(\beta\right)+x_{ijt}'\beta\right)$, where β denotes a generic vector of parameters, define $z_{ijt}^{\beta}=\frac{y_{ijt}-\mu_{ijt}^{\beta}}{\mu_{ijt}^{\beta}}+\ln\left(\mu_{ijt}^{\beta}\right)$, and let $\widetilde{z}_{ijt}^{\beta}$ and $\widetilde{x}_{ijt}^{\beta}$ be defined accordingly. To show (22), we first show that

$$\lim_{r\to\infty}\left(\arg\min_{\beta\in\mathcal{B}}\mathrm{E}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{\prime(r-1)}\beta\right)^{2}+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_{k}^{*(r-1)}\left|\beta_{k}\right|\right)\right.$$
$$\left.-\arg\min_{\beta\in\mathcal{B}}\mathrm{E}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{\beta}\left(\widetilde{z}_{ijt}^{\beta}-\widetilde{x}_{ijt}^{\prime\beta}\beta\right)^{2}+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_{k}^{0}\left|\beta_{k}\right|\right)\right)=0.$$

This occurs since we first fix $\beta^{(r)}$ and minimize over $\beta\in\mathcal{B}$ to obtain $\beta^{(r+1)}$, which we then use to maximize over $\beta\in\mathcal{B}$, and get $\beta^{(r+2)}$ and so on. Thus, by iterating, since $\beta^{(r)}-\beta^{(r-1)}$ shrinks to zero, it is as if we solved (1) directly. Finally, note that

$$\arg\min_{\beta\in\mathcal{B}}\mathrm{E}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{\beta}\left(\widetilde{z}_{ijt}^{\beta}-\widetilde{x}_{ijt}^{\prime\beta}\beta\right)^{2}+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_{k}^{0}\left|\beta_{k}\right|\right)$$
$$-\arg\min_{\beta\in\mathcal{B}}\mathrm{E}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\left(\widetilde{z}_{ijt}^{0}-\widetilde{x}_{ijt}^{\prime0}\beta\right)^{2}+\frac{\lambda}{n}\sum_{k=1}^{p}\phi_{k}^{0}\left|\beta_{k}\right|\right)=0,$$

which establishes (22).

Note that, (22) implies that

$$\lim_{r\to\infty}\left(\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{(r)}\widetilde{x}_{ijt}^{(r)}\widetilde{x}_{ijt}^{(r)\prime}\right)^{-1}-\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}\right)=o_p(1),\tag{23}$$

and

$$\mathrm{plim}_{r\to\infty}s^{(r)}=s.\tag{24}$$

We now can prove (20). For all r, the probability order of $\beta_n^{*(r)}-\widetilde{\beta}_n$ is given by the probability order of the difference of the respective scores evaluated at β^0, which, ignoring for the time being the penalization term, are

$$n^{-1}\sum_{ijt}\mu_{ijt}^{(r-1)}\widetilde{x}_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{(r-1)\prime}\beta^{0}\right)$$

and

$$n^{-1}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\left(\widetilde{z}_{ijt}^{0}-\widetilde{x}_{ijt}^{0\prime}\beta^{0}\right),$$

respectively. Let $\epsilon_{ijt}^{(r-1)}=\mu_{ijt}^{(r-1)}\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{(r-1)\prime}\beta^{0}\right)$ and, as before, $\epsilon_{ijt}^{0}=\mu_{ijt}^{0}\left(\widetilde{z}_{ijt}^{0}-\widetilde{x}_{ijt}^{0}\beta^{0}\right)$, then

$$n^{-1}\sum_{ijt}\widetilde{x}_{ijt}^{(r-1)}\epsilon_{ijt}^{(r-1)}-n^{-1}\sum_{ijt}\widetilde{x}_{ijt}^{0}\epsilon_{ijt}^{0}$$
$$\approx n^{-1}\sum_{ijt}\left(\widetilde{x}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{0}\right)\epsilon_{ijt}^{0}+n^{-1}\sum_{ijt}\widetilde{x}_{ijt}^{0}\left(\epsilon_{ijt}^{(r-1)}-\epsilon_{ijt}^{0}\right)$$
$$+\text{ smaller order term},$$

where $\approx$ means of the same probability order and where the smaller order term captures the product $\left(\widetilde{x}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{0}\right)\left(\epsilon_{ijt}^{(r-1)}-\epsilon_{ijt}^{0}\right)$. Given (22), (23) and (24), for the $s^{(r)}$ selected variables we have

$$\begin{aligned}
&\beta_n^{*(r)}-\widetilde{\beta}_n \\
&\quad\approx\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}\left(n^{-1}\sum_{ijt}\left(\widetilde{x}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{0}\right)\epsilon_{ijt}^{0}+n^{-1}\sum_{ijt}\widetilde{x}_{ijt}^{0}\left(\epsilon_{ijt}^{(r-1)}-\epsilon_{ijt}^{0}\right)\right. \\
&\qquad\left.+\frac{\lambda}{n}\sum_{k=1}^{s^{(r)}}\left(\phi_k^{(r-1)}-\phi_k^{0}\right)\operatorname{sign}\left(\beta_k^{0}\right)\right)+\text{ smaller order terms,}
\end{aligned}\tag{25}$$

where $\phi_k^{(r-1)}$ has been computed using $\epsilon_{ijt}^{(r-1)}$, as in (18), and ϕ_k^0 using ϵ_{ijt}^0. Now, recalling the definition of $\widetilde{x}_{k,ijt}$ given by (3), we can write the first term in (25) as

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}n^{-1}\sum_{ijt}\left(\widetilde{x}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{0}\right)\epsilon_{ijt}^{0} \\
&\quad\approx\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}n^{-1}\sum_{ijt}\epsilon_{ijt}^{0}\left(\left(\frac{\partial\alpha_{it}(\beta)}{\partial\beta}\Big|_{\beta=\beta_n^{*(r-1)}}-\frac{\partial\alpha_{it}(\beta)}{\partial\beta}\Big|_{\beta=\beta^0}\right)\right. \\
&\qquad+\left(\frac{\partial\gamma_{jt}(\beta)}{\partial\beta}\Big|_{\beta=\beta_n^{*(r-1)}}-\frac{\partial\gamma_{jt}(\beta)}{\partial\beta}\Big|_{\beta=\beta^0}\right) \\
&\qquad\left.+\left(\frac{\partial\eta_{ij}(\beta)}{\partial\beta}\Big|_{\beta=\beta_n^{*(r-1)}}-\frac{\partial\eta_{ij}(\beta)}{\partial\beta}\Big|_{\beta=\beta^0}\right)\right).
\end{aligned}\tag{26}$$

By a Taylor expansion of the RHS of (26) around β^0, and using $\partial_{\beta\beta}$ to denote second derivatives, we have that

$$\begin{aligned}
&\lim_{r\to\infty}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}n^{-1}\sum_{ijt}\left(\widetilde{x}_{ijt}^{(r-1)}-\widetilde{x}_{ijt}^{0}\right)\epsilon_{ijt}^{0} \\
&\approx\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}n^{-1}\sum_{ijt}\epsilon_{ijt}^{0}\,\partial_{\beta\beta}\alpha_{it}(\beta)|_{\beta=\beta^0}\lim_{r\to\infty}\left(\beta_n^{*(r-1)}-\beta^0\right) \\
&\quad+\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}n^{-1}\sum_{ijt}\epsilon_{ijt}^{0}\,\partial_{\beta\beta}\gamma_{jt}(\beta)\big|_{\beta=\beta^0}\lim_{r\to\infty}\left(\beta_n^{*(r-1)}-\beta^0\right) \\
&\quad+\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^{0}\widetilde{x}_{ijt}^{0}\widetilde{x}_{ijt}^{0\prime}\right)^{-1}n^{-1}\sum_{ijt}\epsilon_{ijt}^{0}\,\partial_{\beta\beta}\eta_{ij}(\beta)\big|_{\beta=\beta^0}\lim_{r\to\infty}\left(\beta_n^{*(r-1)}-\beta^0\right) \\
&\quad+\text{ smaller order terms} \\
&=O_p\left(\sqrt{\frac{s^2\ln(n)}{N(N-1)\iota_T}}\right)o_p(1)+\text{ smaller order terms} \\
&=o_p\left(\sqrt{\frac{s^2\ln(n)}{N(N-1)\iota_T}}\right)+\text{ smaller order terms,}
\end{aligned}$$

where the $o_p(1)$ term comes from (22) and the order of the O_p term comes by the same argument as in the proof of Lemma 1. Also, note that for the second term in (25) we have

$$\begin{aligned}
&\lim_{r\to\infty}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^0\widetilde{x}_{ijt}^0\widetilde{x}_{ijt}^{0\prime}\right)^{-1} n^{-1}\sum_{ijt}\widetilde{x}_{ijt}^0\left(\epsilon_{ijt}^{(r-1)}-\epsilon_{ijt}^0\right)\\
&\quad=\lim_{r\to\infty}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^0\widetilde{x}_{ijt}^0\widetilde{x}_{ijt}^{0\prime}\right)^{-1} n^{-1}\sum_{ijt}\widetilde{x}_{ijt}^0\left(\mu_{ijt}^0\left(\left(\widetilde{z}_{ijt}^{(r-1)}-\widetilde{z}_{ijt}^0\right)-\left(\widetilde{x}_{ijt}^{\prime(r-1)}-\widetilde{x}_{ijt}^{\prime 0}\right)\beta^0\right)\right.\\
&\qquad\qquad\left.+\left(\mu_{ijt}^0-\mu_{ijt}^{(r-1)}\right)\left(\widetilde{z}_{ijt}^0-\widetilde{x}_{ijt}^{\prime 0}\beta^0\right)\right) \;+\; \text{smaller order terms}\\
&\quad=o_p\left(\sqrt{\frac{s^2\ln(n)}{N(N-1)\iota_T}}\right)+\;\text{smaller order terms},
\end{aligned}$$

by the same argument used above. Finally, for the third term in (25), given (24), Assumption A2, and recalling that λ is of order $O(\sqrt{n})$ up to a log term and $s=o(\sqrt{n})$,

$$\begin{aligned}
&\lim_{r\to\infty}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^0\widetilde{x}_{ijt}^0\widetilde{x}_{ijt}^{0\prime}\right)^{-1}\frac{\lambda}{n}\sum_{k=1}^{p}\left(\phi_k^{*(r-1)}-\phi_k^0\right)\text{sign}\left(\beta_k^0\right)\\
&\quad=\lim_{r\to\infty}\left(\frac{1}{n}\sum_{ijt}\mu_{ijt}^0\widetilde{x}_{ijt}^0\widetilde{x}_{ijt}^{0\prime}\right)^{-1}\frac{\lambda}{n}\sum_{k=1}^{s}\left(\phi_k^{*(r-1)}-\phi_k^0\right)\text{sign}\left(\beta_k^0\right)\;+\;\text{smaller order terms}\\
&\quad=O_p(1)o_p\left(\sqrt{\frac{s^2\ln(n)}{N(N-1)\iota_T}}\right)+\;\text{smaller order terms}.
\end{aligned}$$

This completes the proof of (20).

It remains to show (21). Let $\widehat{\beta}_n^{(r)}$ be defined as in (13). The difference between $\widehat{\beta}_n^{(r)}$ and $\beta_n^{*(r)}$ is that the former is computed using estimated fixed effects, and thus the difference between the two is due to the bias due to the estimation of the incidental parameters. To characterize this difference as $r\to\infty$, we utilize a penalized "concentrated likelihood" representation for the PPML-lasso estimator that concentrates out the fixed effects like in our earlier equation (2):

$$Q_n(\beta):=L_n\left(\beta,\widehat{\alpha}(\beta),\widehat{\gamma}(\beta),\widehat{\eta}(\beta)\right)+\frac{\lambda}{n}\sum_k\widehat{\phi}_k\left|\beta_k\right|, \tag{27}$$

Note that L_n is the (concentrated) unpenalized PPML pseudo-likelihood times minus one. We also need to use an analogue of this penalized likelihood that uses the probability limits as $N,T\to\infty$ of the fixed effects in place of their estimated values:

$$Q_n^*(\beta):=L_n\left(\beta,\alpha(\beta),\gamma(\beta),\eta(\beta)\right)+\frac{\lambda}{n}\sum_k\phi_k^*\left|\beta_k\right|, \tag{28}$$

where $\phi_k^*=\lim_{r\to\infty}\phi_k^{*(r-1)}$. Note that minimizing $Q_n^*(\beta)$ with respect to β results in $\beta_n^*=\lim_{r\to\infty}\beta_n^{*(r)}$.

Recall the definition of $\mathcal{S}$ given in Theorem 1 and let $\beta_{\mathcal{S}}^0$, $\widehat{\beta}_{n,\mathcal{S}}$, and $\beta_{n,\mathcal{S}}^*$ denote, respectively, the non-zero elements of β^0 and the corresponding elements of $\widehat{\beta}_n$ and β_n^*. Asymptotically, the coefficient vectors $\widehat{\beta}_{n,\mathcal{S}}$ and $\beta_{n,\mathcal{S}}^*$ are defined, respectively, by the following FOCs

$$\begin{aligned}
\partial_{\beta_{\mathcal{S}}}Q_n &= 0,\\
\partial_{\beta_{\mathcal{S}}}Q_n^* &= 0.
\end{aligned}$$

These FOCs hold with equality asymptotically because the coefficients in $\beta_{\mathcal{S}}^0$ are asymptotically bounded away from zero, so that the lasso penalty is locally differentiable around $\beta_{\mathcal{S}}^0$. For $k\in\mathcal{S}^c$, meanwhile, the

lasso FOCs for each $\widehat{\beta}_{k,n}$ and $\beta^*_{k,n}$ are inequalities rather than equalities because of how the lasso penalty is not differentiable at zero.

Taking Taylor expansions on the support $\mathcal{S}$ for both of these systems around β_0 yields

$$0 = \partial_{\beta_{\mathcal{S}}} Q_n\left(\beta^0\right) + \partial_{\beta_{\mathcal{S}}\beta_{\mathcal{S}}} Q_n\left(\beta^0\right)\left(\widehat{\beta}_{n,\mathcal{S}} - \beta^0_{\mathcal{S}}\right) + \text{smaller order terms},$$
$$0 = \partial_{\beta_{\mathcal{S}}} Q^*_n\left(\beta^0\right) + \partial_{\beta_{\mathcal{S}}\beta_{\mathcal{S}}} Q^*_n\left(\beta^0\right)\left(\beta^*_{n,\mathcal{S}} - \beta^0_{\mathcal{S}}\right) + \text{smaller order terms}.$$

These expressions can then be combined as

$$\partial_{\beta_{\mathcal{S}}\beta_{\mathcal{S}}} Q^*_n\left(\beta^0\right)\left(\widehat{\beta}_{n,\mathcal{S}} - \beta^*_{n,\mathcal{S}}\right) = \partial_{\beta_{\mathcal{S}}} Q_n\left(\beta^0\right) - \partial_{\beta_{\mathcal{S}}} Q^*_n\left(\beta^0\right) + \text{smaller order terms}.$$

Now, let $L_n(\beta)$ henceforth denote the first term on the RHS of (27) (i.e., the unpenalized likelihood term) and let $L^*_n(\beta)$ denote the corresponding term in (28). Then, the difference between $\widehat{\beta}_{n,\mathcal{S}}$ and $\beta^*_{n,\mathcal{S}}$ (i.e., the difference between the two estimators for the coefficients whose true values are non-zero) can be expressed as

$$\begin{aligned}\widehat{\beta}_{n,\mathcal{S}} - \beta^*_{n,\mathcal{S}} = &\left[\partial_{\beta_{\mathcal{S}}\beta_{\mathcal{S}}} Q^*_n(\beta^0)\right]^{-1}\left(\partial_{\beta_{\mathcal{S}}} L_n(\beta^0) - \partial_{\beta_{\mathcal{S}}} L^*_n(\beta^0)\right) \\ &+ \left[\partial_{\beta_{\mathcal{S}}\beta_{\mathcal{S}}} Q^*_n(\beta^0)\right]^{-1}\left[\frac{\lambda}{n}\left(\widehat{\phi}_{\mathcal{S}} - \phi^*_{\mathcal{S}}\right) \odot \text{sign}(\beta^0_{\mathcal{S}})\right] \\ &+ \text{smaller order terms}. \end{aligned} \tag{29}$$

Provided the Hessian term $\partial_{\beta_{\mathcal{S}}\beta_{\mathcal{S}}} Q^*_n(\beta^0)$ is non-singular, which our assumptions ensure, the first term on the RHS of (29) is the same order as the bias of the unpenalized PPML estimator characterized in Weidner and Zylkin (2021).[8] For the second term, note that $\widehat{\phi}_{\mathcal{S}} \to \phi^*_{\mathcal{S}} \iff \widehat{\beta}_{n,\mathcal{S}} \to \beta^*_{n,\mathcal{S}}$; differentiability of $\widehat{\phi}$ and ϕ^* with respect to β then ensures that $\widehat{\phi}_{\mathcal{S}} - \phi^*_{\mathcal{S}} = O_p(\widehat{\beta}_{n,\mathcal{S}} - \beta^*_{n,\mathcal{S}}) \implies \frac{\lambda}{n}(\widehat{\phi}_{\mathcal{S}} - \phi^*_{\mathcal{S}}) = o_p(\widehat{\beta}_{n,\mathcal{S}} - \beta^*_{n,\mathcal{S}})$, such that the first term dominates.

Therefore, by a similar argument as in Proposition 3 and Remark 2 in Weidner and Zylkin (2021), it follows that for each $k \in \mathcal{S}$, we have

$$\lim_{r\to\infty}\left(\widehat{\beta}^{(r)}_{k,n} - \beta^{*(r)}_{k,n}\right) = \widehat{\beta}_{k,n} - \beta^*_{k,n} = O_p\left(\frac{1}{N\iota_T}\right). \tag{30}$$

Summing over $k \in \mathcal{S}$ yields

$$\left\|\widehat{\beta}_{n,\mathcal{S}} - \beta^*_{n,\mathcal{S}}\right\|_1 = O_p\left(\frac{s}{N\iota_T}\right).$$

To extend this to the full coefficient vector, note that

$$\left\|\widehat{\beta}_n - \beta^*_n\right\|_1 = \left\|\widehat{\beta}_{n,\mathcal{S}} - \beta^*_{n,\mathcal{S}}\right\|_1 + \left\|\widehat{\beta}_{n,\mathcal{S}^c} - \beta^*_{n,\mathcal{S}^c}\right\|_1. \tag{31}$$

For the difference on $\mathcal{S}^c$ note that the incidental parameter-induced bias in the profile score is of order $1/N\iota_T$, while the lasso penalty threshold is of order $1/\sqrt{N(N-1)\iota_T}$. Hence, for each $\beta^*_{k,n}$ of order larger than $1/N\iota_T$, the bias does not affect the selection of the k-th variable; on the other hand, if $\beta^*_{k,n}$ is of order equal or smaller than $1/N\iota_T$, variable k will never be selected. Then, whenever $\iota_T = T$ and $T \to \infty$, the penalty threshold is of a larger order than the bias and thus the estimation of each $\beta_k \in \beta_{\mathcal{S}^c}$ does not contribute to the incidental parameter bias of $\widehat{\beta}_n$, so that the second term in (31) is negligible asymptotically. On the other hand, when $\beta^*_{k,n}$ is exactly of order $1/N$, $\iota_T = 1$, and/or T is fixed, the contribution of variable k to the profile likelihood and the contribution to the penalty are of the same order. For this reason, we require that all relevant variables, i.e., $k \in \mathcal{S}$, are such that $\left|\beta^0_k\right| \geq N^{-(1-\delta)}$ and the irrelevant $k \in \mathcal{S}^c$, are such that $\left|\beta^0_k\right| \leq N^{-(1+\delta)}$, with $1/2 > \delta > 0$. The beta-min condition rules out the possibility of $\beta^*_{k,n}$ being of order $1/N$, then ensuring that the bias due to incidental parameters does not affect variable selection. ■

[8] Note that the full expansion in (29) would also include terms depending on $\left(\widehat{\beta}_{n,\mathcal{S}^c} - \beta^*_{n,\mathcal{S}^c}\right)$, such as $\partial_{\beta_{\mathcal{S}^c}\beta_{\mathcal{S}}} Q_n\left(\beta^0\right)\left(\widehat{\beta}_{n,\mathcal{S}^c} - \beta^*_{n,\mathcal{S}^c}\right)$. These terms are controlled by an analogous argument to the one used below to bound the $\mathcal{S}^c$ component in (31).

Appendix D: Further empirical results

Additional boostrap-lasso results

Table A2: Provisions selected by bootstrap lasso, sorted by average coefficient

AD14	0.079	INV24	0.016	SUB11	0.005	CP15	0.002
CP23	0.065	ET17	0.014	ROR04	0.005	SPS22	0.002
CP22	0.063	ET43	0.013	TBT05	0.004	IPR51	0.002
AD05	0.055	PP08	0.013	TBT14	0.004	IPR03	0.002
TBT07	0.054	TF45	0.012	MOC28	0.004	TF44	0.002
TBT02/29	0.048	SUB13	0.011	TBT11	0.004	SPS34	0.002
TBT08	0.037	ENV33	0.011	ET15	0.004	ENV19	0.002
SUB12	0.030	TBT15	0.010	SUB09	0.004	TBT31	0.002
TBT34	0.028	MIG14	0.010	SPS07	0.003	INV12	0.001
SPS06	0.028	ET42	0.009	SPS23	0.003	ROR16	0.001
TF42	0.028	AD06/08/ENV42	0.008	SPS09	0.003	ENV30	0.001
AD07	0.027	STE32	0.008	STE30	0.003	AD10	0.001
TBT33	0.023	ROR13	0.008	SUB03	0.003	IPR127	0.001
TF41	0.023	ET09	0.007	IPR68	0.003	SER46	0.001
TBT06	0.021	IPR02	0.007	MIG12	0.003	TBT04	0.001
CP21	0.020	SPS21	0.007	SPS33	0.002	SUB14	0.001
SUB10	0.020	SPS18	0.006	SPS16	0.002	CP16	0.001
MOC27	0.019	ROR02	0.006	ENV28	0.002	TF25	0.001

Notes: Bootstrap plug-in lasso performed using cluster-bootstrap resampling with 250 replications. The numbers shown are average coefficient estimates for the provisions selected at least 1% of the time across all replications, ordered from greatest to least. A description of the provisions in this table can be found in Appendix A.

Table A3: Provisions selected by bootstrap lasso, sorted by selection frequency

AD14	0.372	TF42	0.076	MIG12	0.036	SPS21	0.016
CP23	0.320	SUB10	0.072	SPS18	0.032	TBT05	0.016
TBT07	0.308	ET43	0.072	SPS07	0.032	TBT14	0.016
SPS06	0.228	MIG14	0.068	CP15	0.032	SPS09	0.016
TBT08	0.208	STE32	0.068	ET17	0.028	IPR68	0.016
SUB12	0.184	TBT15	0.064	ROR02	0.028	SPS16	0.016
TBT02/29	0.168	ROR04	0.060	SPS33	0.028	TBT31	0.016
TBT33	0.160	SUB11	0.056	ENV28	0.028	AD10	0.016
CP22	0.156	SUB09	0.056	ENV19	0.028	TBT04	0.016
TBT34	0.152	STE30	0.056	PP08	0.024	SUB14	0.016
TBT06	0.148	AD07	0.052	TBT11	0.024	TF25	0.016
AD05	0.140	INV24	0.048	SUB03	0.024	IPR51	0.012
CP21	0.124	ET42	0.048	SPS34	0.024	INV12	0.012
TF45	0.116	ET15	0.048	MOC28	0.020	ROR16	0.012
ENV33	0.116	IPR02	0.044	SPS23	0.020	ENV30	0.012
ET09	0.010	SPS22	0.044	IPR03	0.002	IPR127	0.012
MOC27	0.084	ROR13	0.040	TF44	0.020	SER46	0.012
SUB13	0.080	TF41	0.036	AD06/08/ENV42	0.016	CP16	0.012

Notes: Bootstrap plug-in lasso performed using cluster-bootstrap resampling with 250 replications. The numbers shown are selection frequencies for the provisions selected at least 1% of the time across all replications, ordered from greatest to least. A description of the provisions in this table can be found in Appendix A.

Cross-validation lasso results

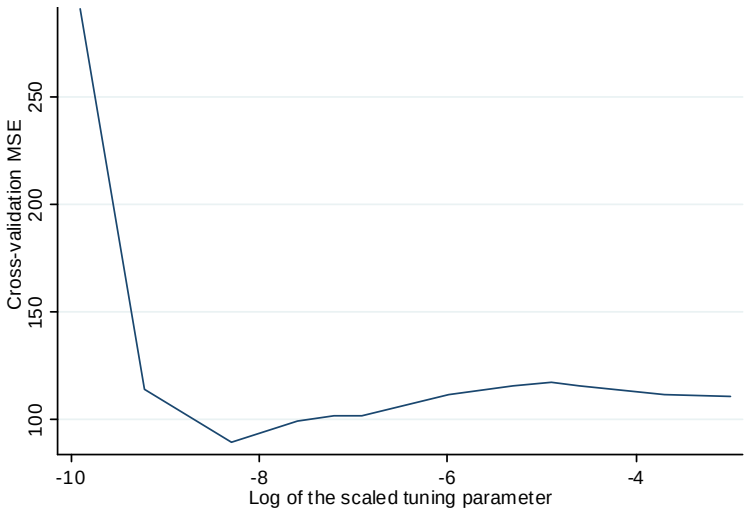
As discussed in the paper, the plug-in approach to choosing penalty parameter tends to choose a relatively small set of regressors and may fail to pick the "correct" regressors. For comparison, we now discuss the choice of regressors when we apply the cross-validation approach to our data.[9]

Figure A1 shows how the out-of-sample mean square error (MSE) varies with the log of the tuning parameter, which is scaled by $\sum_{ijt} y_{ijt}$ so that the results do not depend on the scale of the data. At the optimal value of the tuning parameter, $\lambda/\sum_{ijt} y_{ijt} = 0.00025$, the cross-validation approach selects 128 provisions to have non-zero effects. Additionally, some of the selected provisions are perfectly collinear with variables that are not selected; if we take this into account, the effective number of provisions selected is 133, which is many more than what we found using the plug-in approach.

For more illustration, Figures A2 and A3 show the corresponding regularization paths for selected provisions.[10] That is, the figures show how the value of the estimated (post-lasso) coefficient on the selected provisions changes as we vary λ. For the smallest value of the tuning parameter most provisions are retained, so the results are close to those obtained with the unpenalised PPML estimator. In this case we clearly see that the combination of high dimensionality and high collinearity generates many coefficient estimates that are implausibly large and difficult to interpret. For example, a single IPR provision has an estimated coefficient greater than 3, suggesting that single provision increases trade by more than 1000%, and numerous provisions have negative coefficients less than -1, implying a reduction in trade exceeding 60%. As expected, fewer provisions are selected as we increase λ and, for values of $\lambda/\sum_{ijt} y_{ijt}$ around 0.01, which is forty times larger than the optimal value, we generally see a close correspondence between the results in Figures A2 and A3 and those that we found earlier using the plug-in method.

Note, however, that it is not necessarily the case that the set of provisions selected at lower levels of λ includes the set of provisions selected at higher levels. For example, Figure A2 shows that provision AD14, which was one of the provisions selected by the plug-in approach, is selected with a negative coefficient for the smallest value of λ we consider, drops out when we increase the penalty, and is selected with a positive coefficient for higher values for λ. Intuitively, for small values of λ, the procedure selects many provisions, and the high collinearity between the variables selected makes it difficult to precisely identify their effect. As we increase λ, some provisions are dropped; because many provisions are correlated with AD14, it can be dropped without significant deterioration of the out-of-sample forecasts during cross-validation, and hence it is no longer selected. It is only when the provisions correlated with AD14 are purged from the model, as λ increases even more, that AD14 on its own gains predictive power and is again included.

Figure A1: Cross-validation MSE *vs.* tuning parameter



[9] As explained before, the cross validation is performed clustering by agreement.

[10] In each panel of the figures, the fourth set of estimates from the right is the one for the optimal value of lambda and the associated provisions are thus the ones selected by the CV method.

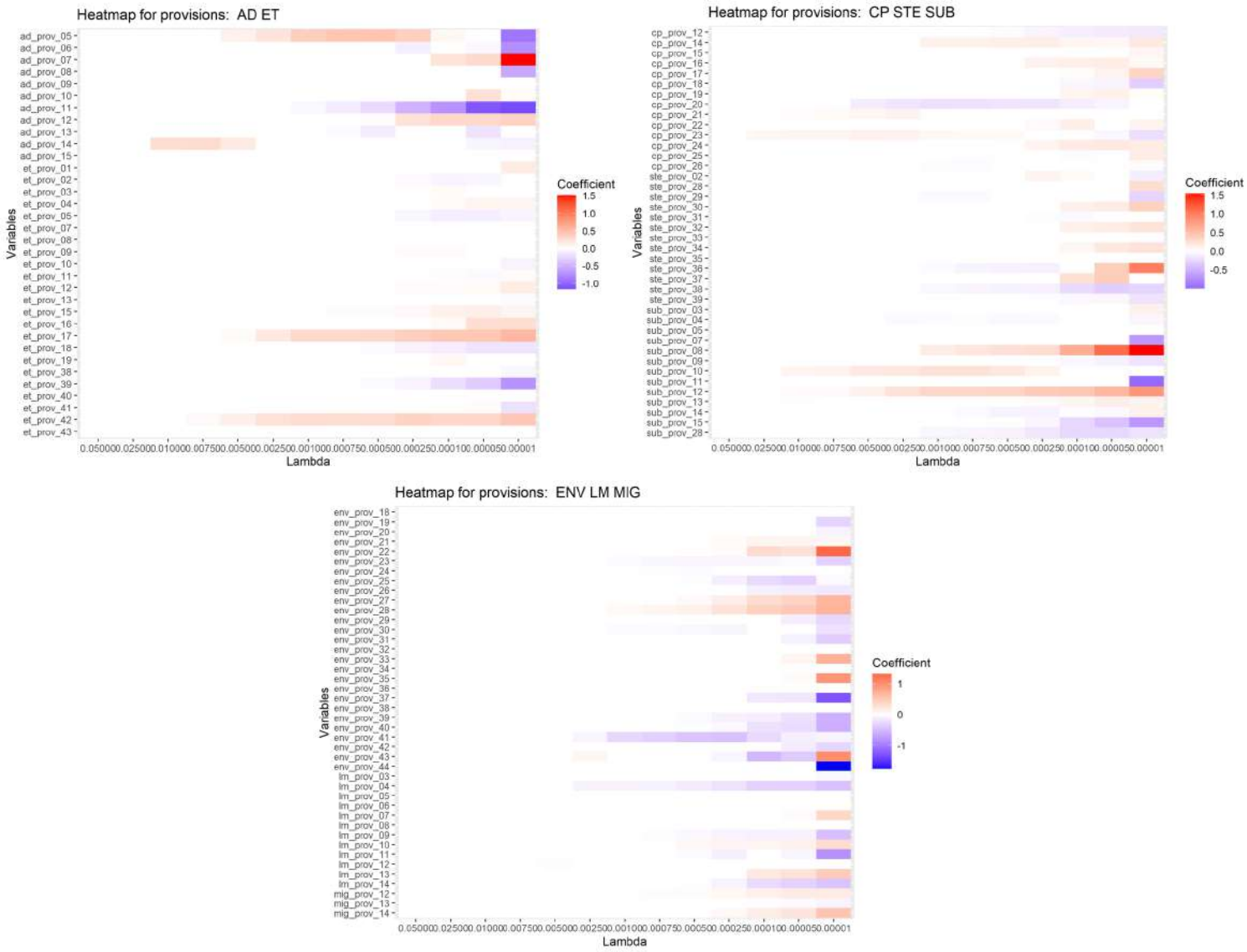


Figure A2: Regularization path for selected provisions (AD, ET, CP, STE, SUB, ENV, LM, and MIG)

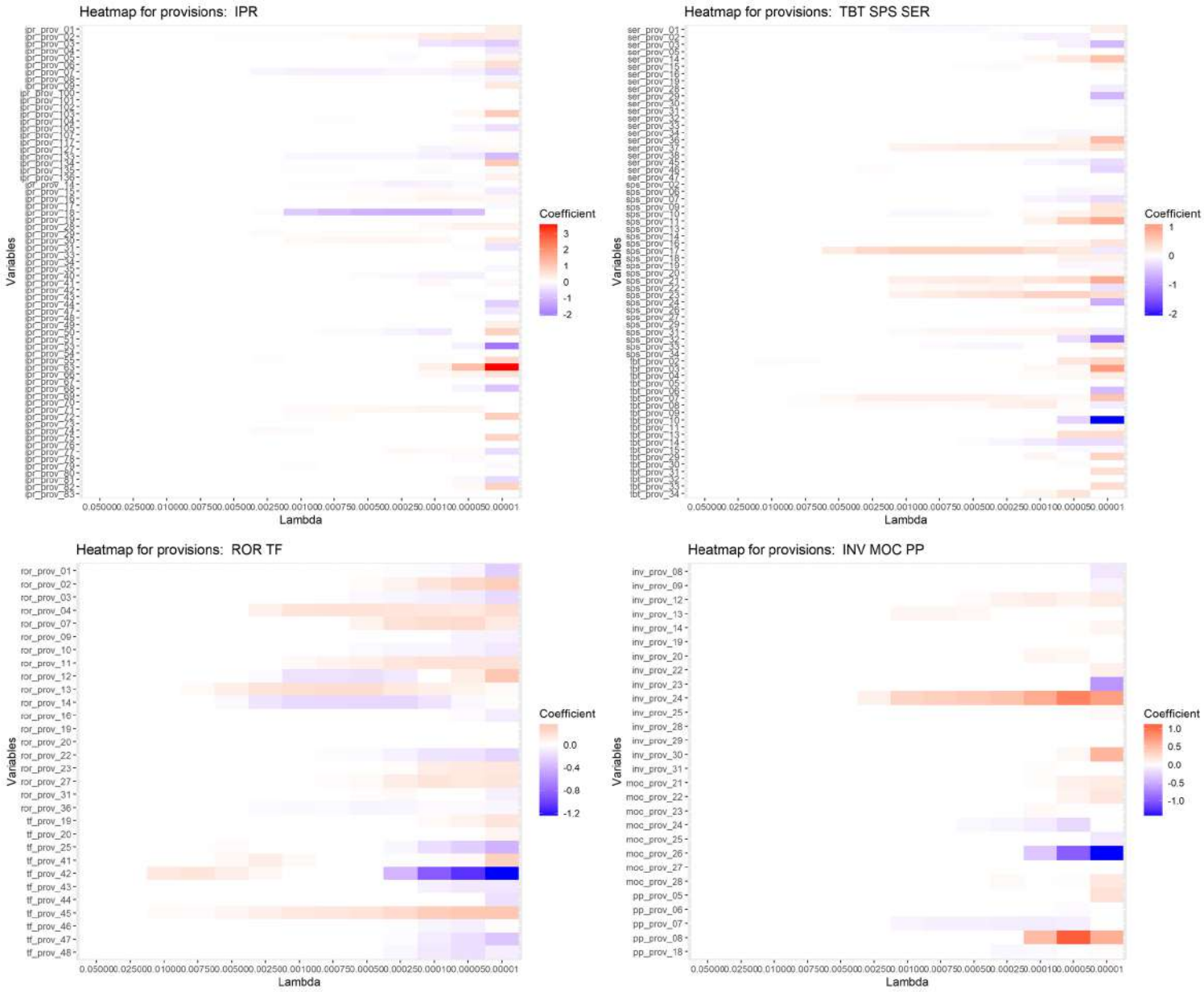


Figure A3: Regularization path for selected provisions (IPR, TBT, SPS, SER, ROR, TF, INV, MOC, and PP)

Overall, the plug-in and cross-validation approaches lead to the selection of very different sets of trade agreement provisions. While some provision, such as TBT07 or TF45 are selected by both approaches, others, such as AD14, are only selected by the plug-in method, and many provisions are only selected using cross-validation, such as anti-dumping provisions AD05 and AD06. Furthermore, we also see in Figures A2 and A3 that many of the estimated effects for the provisions selected by cross-validation are not plausible when interpreted on their own (e.g., TF42 has a negative coefficient). These observations reflect the known shortcomings of the cross-validation approach that we stated earlier and found support for in our simulations.

Appendix E: Prediction

In this appendix we consider the usefulness of the different lasso-based methods to predict trade flows. We start by presenting simulation results comparing the different methods, and then present empirical results using the data discussed before.

Simulation results

We now extend the simulation results presented in Section 4 of the paper to evaluate the predictive ability of the models obtained with the different variable-selection methods. To that end, for each replica of the simulations we generated 100 additional observations and used the different models to predict them. In this context, we can consider both lasso predictions, using the penalized lasso estimates, and post-lasso predictions, using unpenalized estimates.[11] We computed penalized and unpenalized predictions for all approaches and found that for CV and AL penalized predictions tend to dominate unpenalized ones, and the reverse holds for PI, IL, and BL.

Table A4 summarizes these results and reports the mean square error (MSE) of the prediction error for each of the models considered. To conserve space, we only report the results obtained with the penalized predictors for the CV and AL, and unpenalized predictors for PI, IL, and BL. For comparison, the table also presents the MSE of the predictions obtained with the unpenalized PPML estimates of the models that include all p regressors and with the PPML estimates of the "oracle" model that just includes x_1.

The results in Table A4 show that the predictions obtained with the unpenalized estimator of the full model are clearly outperformed by all lasso-based predictions, with the difference being particularly stark in the smaller sample. The results also suggest that the predictive performance of the different methods depends little on the values of κ and ρ, but generally improves with n. The exception to this is the IL, for which we see a small but systematic drop in performance as κ increases. This is not surprising because the method is designed to select all the regressors that are sufficiently correlated with the ones identified by the PI, and therefore for large κ the IL selects many irrelevant predictors.

Perhaps the most striking feature of the results in Table A4 is, however, the excellent performance of PI, which can be comparable to that of the oracle model even in cases where PI often fails to identify x_1 as a predictor. It is also noteworthy that the performance of the BL is also very good and better than that of the IL, especially for the larger values of κ. For the larger sample, however, there is little to choose between the different lasso methods, but AL has the best performance.

In summary, if the goal of the researcher is to accurately predict the trade impact of a given PTA, as opposed to selecting the correct set of provisions that impact trade, the preferred approach is to compute the predictions using the post-lasso estimates obtained with the plug-in penalty. Indeed, this approach performs extremely well in all cases, and is only marginally outperformed by the adaptive lasso in the larger sample we considered.[12] However, the bootstrap lasso is also a credible alternative in this context and it can serve as a useful robustness check.

[11] The unpenalized predictions for the IL are computed from the PPML estimates of a model including the full set of variables selected by IL; for BL they are computed as the average of the predictions corresponding to the post-lasso PPML estimates in each sample. The penalized predictor for the IL is obtained from a plug-in lasso based on the full set of variables selected by IL; for BL, the penalized predictor is the average of the predictions obtained with the penalized estimates in each of the bootstrap samples. For the PI, the penalized predictor is obtained directly from the penalized estimates, whereas the unpenalized predictions are computed from the PPML estimates of a model that includes only the variables selected by PI.

[12] One may wonder why the PPML-lasso with the tuning parameter chosen by the plug-in method predicts so well, even if it often fails to select the right regressor. The answer, of course, is that when the purpose is simply to predict the outcome, the results change little if the regressor with a causal impact is replaced by another that is highly correlated with it.

Table A4: MSE for prediction errors

		$\rho = 0.90$			$\rho = 0.95$			$\rho = 0.99$		
n		$\kappa=5$	$\kappa=10$	$\kappa=20$	$\kappa=5$	$\kappa=10$	$\kappa=20$	$\kappa=5$	$\kappa=10$	$\kappa=20$
250	CV	6.87	6.88	6.88	6.86	6.87	6.85	6.83	6.83	6.80
	AL	7.29	7.26	7.24	7.26	7.25	7.16	7.17	7.18	7.08
	PI	6.59	6.63	6.71	6.62	6.61	6.67	6.53	6.52	6.52
	BL	6.64	6.62	6.66	6.64	6.59	6.61	6.57	6.53	6.53
	IL	6.71	6.84	7.25	6.73	6.85	7.23	6.72	6.85	7.23
	All	10.98	10.98	10.98	10.98	10.98	10.98	10.98	10.98	10.98
	Oracle	6.39	6.39	6.39	6.39	6.39	6.39	6.39	6.39	6.39
1000	CV	6.34	6.34	6.35	6.34	6.33	6.36	6.33	6.32	6.34
	AL	6.35	6.39	6.40	6.46	6.51	6.47	6.39	6.41	6.47
	PI	6.18	6.19	6.22	6.18	6.18	6.23	6.16	6.17	6.20
	BL	6.18	6.18	6.21	6.18	6.18	6.20	6.16	6.16	6.18
	IL	6.22	6.31	6.47	6.22	6.31	6.48	6.22	6.31	6.48
	All	8.44	8.44	8.44	8.44	8.44	8.44	8.44	8.44	8.44
	Oracle	6.19	6.19	6.19	6.19	6.19	6.19	6.19	6.19	6.19
4000	CV	6.36	6.37	6.38	6.36	6.37	6.37	6.37	6.38	6.38
	AL	6.33	6.33	6.34	6.33	6.33	6.33	6.34	6.34	6.34
	PI	6.34	6.35	6.35	6.33	6.34	6.36	6.33	6.34	6.35
	BL	6.36	6.37	6.37	6.36	6.37	6.38	6.36	6.37	6.38
	IL	6.34	6.35	6.43	6.34	6.35	6.43	6.34	6.35	6.43
	All	7.39	7.39	7.39	7.39	7.39	7.39	7.39	7.39	7.39
	Oracle	6.34	6.34	6.34	6.34	6.34	6.34	6.34	6.34	6.34

Note: The table reports the mean square error of the prediction error obtained using penalized predictors for the CV and AL, and unpenalized predictors for PI, IL, and BL. For comparison, the table also presents the mean square error of the predictions obtained with the model with all regressors and with the "oracle" model that just includes the relevant regressor.

Empirical results

We now use the different methods and our data to predict the effects of different PTAs and discuss the associated caveats.[13]

The simulation results presented earlier suggest that, in small to moderate samples, the most reliable predictions are the ones based on the (post-lasso) PPML estimates of a model whose regressors are the provisions selected by the plug-in lasso. This kind of prediction can easily be obtained using the results in column (3) of Table 3 of the paper. For example, we have noted that the latest EU agreement includes all the provisions selected by the plug-in lasso, with the exception of AD14 and TBT7. Therefore, the effect of the latest EU agreement is estimated to be 87% ($\exp(0.118 + 0.184 + 0.123 + 0.113 + 0.089) - 1 = 0.87$). This result is comparable to the effect estimated when the EU dummy is included in the model as in column (5) of Table 3 of the paper, which is 86% ($\exp(0.618) - 1 = 0.86$).[14]

In results that are summarized in the third column of Table A5, we repeat this exercise for each of the PTAs in our data.[15] As in Baier, Yotov, and Zylkin (2019), we find a wide variety of effects, ranging from very large impacts in agreements such as the Eurasian Economic Union, which includes all of the selected provisions, to no effect at all in agreements that do not include any, such as ASEAN.[16] In comparison with column 1 of this Table, which describes results for PPML with the full set of provision variables, we see an immediate advantage of using the plug-in method to model PTA heterogeneity: it greatly cuts down on overfitting. The range spanned by the estimates obtained with the full set of provision reaches implausibly large positive and negative values at the extremes, and their standard error is thousands of times that of

[13] As before, we compute penalized predictions when using cross-validation, and post-lasso unpenalized predictions for the plug-in, and bootstrap lasso. For the bootstrap lasso, the predictions are obtained by averaging the post-lasso predictions in each of the bootstrap samples.

[14] Of course, using the delta method it is possible to obtain confidence intervals for these effects. However, such confidence intervals do not take into account model uncertainty, which is likely to be the main source of uncertainty in this context. We consider this issue below.

[15] Note that the average estimated effect is 13.8%, which is very close to the estimated PTA effect of 14.0% corresponding to result in column 1 of Table 3 of the paper.

[16] In contrast to Baier, Yotov, and Zylkin (2019), we are able to identify heterogeneity across different PTAs but not within PTAs.

the estimates produced using plug-in lasso. This is a consequence of the fact that the high dimensionality of the set of regressors creates strong multicollinearity, which leads to erratic estimates that are generally uninterpretable. As shown in column 2, overfitting may also be a problem for the predictions generated by cross-validation lasso, which also lead to some implausible estimates. These results resonate with what we found in the simulations reported earlier, where both the model with all regressors and the model with regressors selected by cross-validation performed poorly.

We next consider the performance of the bootstrap lasso. The predictions based on the bootstrap lasso performed well in our simulations, and this approach also shows promise here. As shown in column 4 of Table A5, the PTA estimates produced by bootstrap lasso lead to less extreme predictions and have the lowest dispersion of any the methods we consider, consistent with what would be expected for a method based on bootstrap aggregation. Though they are highly correlated with the estimates produced by plug-in lasso, the selected PTA estimates shown in the bottom panel of Table A5 reveal that the estimated effects obtained with the plug-in lasso and bootstrap lasso can differ substantially for individual PTAs.

Table A5: Summarizing Estimates of Heterogeneous PTA Effects

	(1)	(2)	(3)	(4)
	All	CV	Plug-in	Bootstrap
Descriptive statistics				
Min	−81.2%	−50.4%	0.0%	0.0%
Max	> 1e6%	387.0%	144.4%	101.0%
Mean	328774.6%	32.1%	13.8%	12.5%
Median	26.4%	14.4%	9.3%	7.2%
Stdev.	300514.7pp	63.0pp	20.7pp	15.3pp
Correlations				
All	1	0.146	−0.054	0.041
CV	0.146	1	0.391	0.513
Plug-in	−0.054	0.391	1	0.925
Bootstrap	0.041	0.513	0.925	1
Estimated partial effects for selected PTAs				
EU	104.9%	105.4%	87.1%	64.2%
EEA	80.4%	90.5%	9.3%	18.3%
Eurasian Econ. Union	−21.8%	71.8%	144.4%	101.0%
NAFTA	77.9%	77.5%	79.9%	52.9%
MERCOSUR	145.5%	115.9%	42.1%	39.6%
ECOWAS	469.6%	379.2%	9.3%	19.4%
ASEAN	1.8%	−9.4%	0.0%	3.3%

Notes: This table summarizes estimated partial effects for individual PTAs produced by the different methods we consider. Results labelled "All" refer to an unpenalized PPML regression with all 305 provision variables. The other columns refer to variants of the lasso discussed in Section 3.

It should be noted that the bootstrap lasso is the only approach we have considered that can provide information about model uncertainty. Indeed, as a by-product of the bootstrap sampling procedure, it can provide confidence intervals showing how sensitive predictions of individual PTA effects are to the particular sample that is used in the estimation. We have not rigorously evaluated the validity of such confidence intervals for bounding prediction uncertainty, but it is certainly an avenue worth exploring.

The results of this exercise, however, need to be treated with some caution. As we have repeatedly noted, the results of the plug-in lasso do not have a causal interpretation. Therefore, their accuracy for predicting effects of individual PTAs will depend, at least to some extent, on whether the selected provisions themselves have a causal impact on trade or serve as a signal of the presence of provisions that have a causal effect. When this condition holds, the predictions based on this method are likely to be reasonably accurate and, indeed, the simulation results reported before show that this approach can work well even in situations where the variables having a causal impact on the outcome are not selected by the plug-in lasso. That said, it is possible to envision scenarios where predictions based on the plug-in lasso fail dramatically. For example, it could be the case that a PTA is incorrectly predicted to have zero impact despite having many

of the true causal provisions. Though we do not consider inferences for these predicted values, obtaining valid confidence intervals for them is complicated by the biases examined in Rakshit and Guo (2024).

References

Baier, S.L, Y.V. Yotov, and T. Zylkin (2019). "On the widely differing effects of free trade agreements: Lessons from twenty years of trade integration," *Journal of International Economics,* 116, 206-228.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369-2429.

Belloni, A., V. Chernozhukov, C. Hansen, D. Kozbur (2016). "Inference in high dimensional panel models with an application to gun control," *Journal of Business & Economic Statistics,* 34, 590-605.

Chatterjee, A. and S.N. Lahiri (2010). "Asymptotic properties of residual bootstrap for lasso estimators," *Proceedings of the American Mathematical Society,* 138, 4497-4509.

Correia, S., P. Guimarães and T. Zylkin (2020). "Fast Poisson estimation with high dimensional fixed effects," *STATA Journal,* 20, 90-115.

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33, 1-22.

Garrucho, D.F., T. Zylkin, J. Cruz, and N. Apfel (2021). `penppml`: Penalized Poisson Pseudo Maximum Likelihood Regression, `https://tinyurl.com/penppml`.

Gaure, S (2013). "OLS with multiple high dimensional category variables," *Computational Statistics & Data Analysis* 66, 8-18.

Jing, B.Y., Q.M. Shao, and Q. Wang (2003). "Self-normalized Cramér-type large deviations for independent random variables," *The Annals of Probability*, 31, 2167-2215.

Meinshausen, N., and P. Bühlmann (2010). "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417-473.

Nelder, J.A. and R.W.M. Wedderburn (1972). "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A (General)*, 135, 370-384.

Rakshit, P. and Guo, Z. (2024). "Statistical Inference in High-dimensional Poisson Regression with Applications to Mediation Analysis," arXiv:2410.20671.

Wang, S., B. Nan, S. Rosset, J. Zhu, (2011) "Random lasso," Annals of Applied Statistics, 5, 468-485.

Weidner, M., T. Zylkin (2021). "Bias and consistency in three-way gravity models," *Journal of International Economics*, 132, 103513.

Zou, H. (2006). "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418-1429.

Zou, H. and T. Hastie, (2005). "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301-320.