Machine Learning in International Trade Research – Evaluating the Impact of Trade Agreements*

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

4 May 2022

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 problems 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 propose novel data-driven methods for selecting the most important provisions and quantifying their impact on trade flows. The proposed methods have the advantage of not requiring ad hoc assumptions on how to aggregate individual provisions and offer improved selection accuracy over the standard lasso. We find that provisions related to technical barriers to trade, antidumping, trade facilitation, subsidies, and competition policy are associated with enhancing the trade-increasing effect of trade agreements.

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

*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, Maia Linask, Yoto Yotov, and seminar participants at the World Bank Economics of Deep Trade Agreements Seminar Series for useful comments. Alvaro Espitia, Diego Ferreras-Garrucho, Jiayi Ni, and Nicolas Apfel 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.

†University of Surrey, CEP and CEPR. Email: h.breinlich@surrey.ac.uk
‡University of Surrey. Email: v.corradi@surrey.ac.uk
§World Bank. Email: nrocha@worldbank.org.
¶World Bank. Email: mruta@worldbank.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 difficulties facing multilateral trade negotiations through the World Trade Organization (WTO), in the last two decades, countries have increasingly turned their focus to preferential trade agreements (PTAs) involving only one or a small number of partners. At the same time, attention has shifted from the reduction of import tariffs to the role of non-tariff barriers and behind-the-border policies, such as differences in regulations, technical standards, or intellectual property rights protections. Accordingly, modern trade agreements contain a host of provisions besides tariff reductions, in areas as diverse as services trade, competition policy, trade-related investment measures, or public procurement (Hofmann, Osnago, and Ruta, 2017).

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 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. Dhingra, Freeman, and Mavroeidi (2018) overcome multicollinearity problems by grouping services, investment, and competition provisions and examining the effect of these "provision bundles" on trade flows.

In this paper, we build upon recent developments in the machine learning and variable selection literature to propose novel data-driven methods to select the most important provisions and quantify their impact on trade flows. These methods address difficulties arising from the high degree of correlation between individual PTA provisions, without requiring ad hoc assumptions on how to aggregate individual provisions. Though, to be clear, they do not completely answer the question of "which provisions matter for trade?", our proposed methods do lead to substantial improvements in our ability to find the more relevant provisions while narrowing down the large number of potential explanatory variables.

We start by proposing an extension of the well-known lasso (Least Absolute Shrinkage and Selection Operator) method for variable selection (see, e.g., Hastie, Tibshirani, and Friedman, 2009) to the case of nonlinear models with high-dimensional fixed effects, which have become standard in the analysis of trade flows (see, e.g., Yotov, Piermartini, Monteiro, and Larch, 2016). Specifically, we use a Poisson pseudo-maximum likelihood

(PPML) version of the lasso and show how to choose the tuning parameter of this estimator using either cross-validation or the "plug-in" (or "theory-driven") approach of Belloni, Chernozhukov, Hansen, and Kozbur (2016), which accounts for heteroskedasticity and clustered errors.[1]

We apply our PPML-lasso estimators to a comprehensive data set on PTA provisions 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, due to template effects and possible synergies between groups of provisions, the 305 provision variables in our data can be highly correlated with one another. 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. One of the methods, that we call "iceberg lasso", involves regressing each of the provisions selected by the plug-in lasso on all other provisions, with the purpose of identifying relevant variables that were initially missed due to their collinearity with the provisions selected in the initial step. The other method, termed "bootstrap lasso", augments the set of variables selected by the plug-in lasso with the variables selected when the plug-in lasso is bootstrapped. As we show using simulations, these new methods present a favorable balance between the parsimony of the plug-in lasso and the lenience of cross-validation methods in small-to-moderate data sets where the true causal variables may be highly correlated with an unknown number of other variables.

To provide some headline results, the PPML-lasso based on cross-validation selects 133 provisions as being relevant, whereas using the plug-in penalty we find that only 8 provisions are associated with enhancing the trade-increasing effect of trade agreements. In turn, the iceberg lasso procedure identifies a set of 42 provisions and, depending on the cutoff used, the bootstrap lasso identifies between 30 and 74 provisions that may be impacting trade. Therefore, our iceberg lasso and bootstrap lasso methods select sets of provisions that are small enough to be interpretable and large enough to give us some confidence that they include the more relevant provisions, something that is confirmed by the simulation evidence we provide. Reassuringly, both the iceberg lasso and bootstrap lasso select similar sets of provisions, mainly related to technical barriers to trade, antidumping, trade facilitation, subsidies, and competition policy. Having identified the set of provisions that are more likely to have an impact on trade, we also discuss how our

[1] An R package (penppml) implementing penalized PPML regressions with high-dimensional fixed 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).

findings can be used to estimate the effects of different PTAs and to predict the impact of future ones, as well as the risks associated with such exercises.

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. The new methods we propose allow 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 the machine learning literature interested in variable selection and prediction. 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. As noted above, this requires an extension of their original method to the estimation of nonlinear models with high-dimensional fixed effects using PPML. The iceberg lasso and bootstrap 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. Therefore, these new methods can potentially be useful in other contexts, especially when the available sample is relatively small and contains a large number of highly-correlated potential explanatory variables.

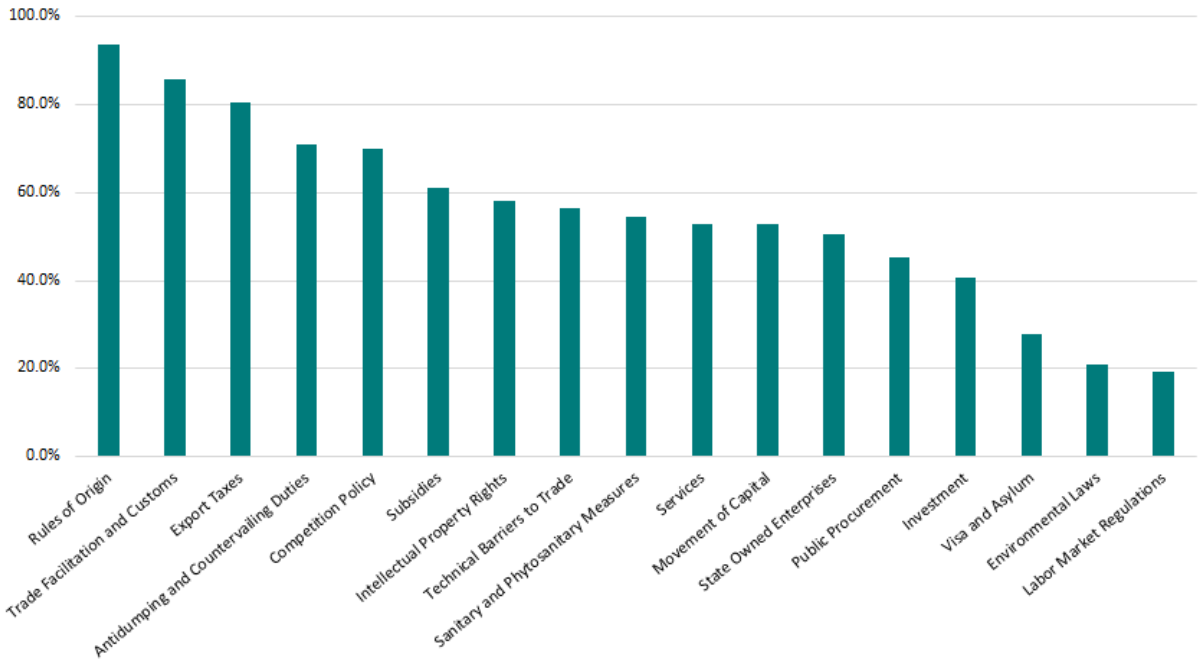
Finally, we contribute to a small existing literature that has used machine learning and other related methods to study the effects of trade agreements in a gravity context. For example, Regmi and Baier (2020) use an unsupervised learning method to group PTAs by textual similarity, so as to provide a 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.

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 individual provisions are the strongest predictors of 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 new database on the content of PTAs that has been collected by Mattoo, Rocha and Ruta (2020). On trade, we use merchandise trade exports between 1964 and 2016 from 220 exporters to 270 importers. 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 trade agreements – these are areas that are close or above the 20 percent share of the trade agreements that have been 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).

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).

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.[3]

[3]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.

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%

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 a 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".[4] 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 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).

The coverage of essential provisions also varies widely across trade agreements and disciplines, indicating that not all PTAs cover the same set of essential provisions. As shown in Table 2, more than $^3/_4$ of agreements cover 25 percent or less of essential provisions included in policy areas such as environmental laws, anti-dumping, sanitary and phytosanitary measures, and technical barriers to trade. Conversely, for policy areas such as visa and asylum, rules of origin, and trade facilitation and customs, more than 70 percent of the mapped agreements cover between 25 and 75 percent of essential

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

provisions. With the exception of services and investment, coverage of more than 75 percent of essential provisions is rare and happens in less than 15 percent of the mapped agreements.

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 in effect in present day, 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). The EU is treated differently in our analysis for this reason, as we discuss further in Section 4. To identify agreements no longer in effect, we consult the NSF-Kellogg database created by Jeff Bergstrand and Scott Baier cross-checked 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.

Table 2: Coverage of essential provisions by policy area

	Share of agreements covering:		
Policy Area	0 to 25%	25% to 75%	over 75%
Anti-dumping and Countervailing Duties	99%	1%	0%
Competition Policy	48%	47%	5%
Environmental Laws	88%	12%	0%
Export Taxes	41%	59%	0%
Intellectual Property Rights	76%	23%	1%
Investment	6%	64%	30%
Labor Market Regulations	68%	17%	15%
Movement of Capital	44%	42%	13%
Public Procurement	53%	40%	7%
Rules of Origin	7%	93%	0%
Sanitary and Phytosanitary Measures	87%	13%	0%
Services	6%	62%	33%
State-Owned Enterprises	45%	54%	1%
Subsidies	59%	41%	0%
Technical Barriers to Trade	93%	7%	0%
Trade Facilitation and Customs	21%	78%	0%
Visa and Asylum	27%	70%	3%

Note: Coverage ratio refers to the share of essential provisions for a policy area contained in a given agreement relative to the maximum number of essential provisions in that policy area. Source: Mattoo, Rocha and Ruta (2020)

3 Determining Which Provisions Matter for Trade

We now outline the methodology we use to identify which PTA provisions have the largest 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), 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 iceberg lasso and bootstrap 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:

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

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. The inclusion of 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 recommended 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.[5] 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) recommend it as the workhorse method for estimating the effects of trade policies. It is frequently applied to 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 introducing interactions 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). However, a major estimation challenge that arises in our setting is that we must treat the number of provisions as being very large. As we will show, 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 innovations is to suppose that only a handful of our provision variables have a non-negligible effect on trade flows. To be more precise, we have $p = 305$ essential provision variables, coded as dummies, of which a subset $s < p$ are assumed to have non-zero effects, where s is typically small with respect to the sample size.[6] We do not know s beforehand, nor do we know the identities of any of 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

[5] 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).

[6] 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.

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,[7] 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 consider $\lambda \to \infty$, 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 to iteratively refine 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. This is mainly because there is no reason to believe that most of the fixed effects parameters are actually zero. In addition, it turns out they do not pose special issues for computation. This is because they do not depend on the penalty. As such, for any given β, the fixed effects 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 the Appendix, 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.[8]

[7] 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.

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

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, the two existing approaches we will first examine are the "plug-in" lasso of Belloni, Chernozhukov, Hansen, and Kozbur (2016) and the traditional cross-validation approach, 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 "iceberg lasso" and "bootstrap lasso", that are intended to address one of the plug-in lasso's key shortcomings in this context.

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 the Appendix, 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,

iii. Values for λ and $\widehat{\phi}_k$ set high enough so that the absolute value of the score for β_k must be large relative to its standard error 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 adjust to reflect the standard error of the score. 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.

A principal 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, the "post-lasso" estimates obtained using unpenalized PPML on the covariates selected by the plug-in lasso have a "near-oracle" property that ensures they will capture the correct model if the sample is sufficiently large relatively to the number of potential regressors (see Belloni, Chen, Chernozhukov, and Hansen, 2012).[9]

However, the plug-in lasso's parsimony can also be a weakness it 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 are 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 (or perhaps both) 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 fit 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, 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, so that we can always form the necessary predictions for the omitted trade flows associated with the PTA that have been dropped.[10]

[9] 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.

[10] 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 $k = 25$.

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.[11] 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.[12]

3.3.1 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. 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 other candidate regressors (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. 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 important complication, we now introduce two methods that add variables to the set of regressors selected by the plug-in lasso, and in the next section we evaluate their performance in a simulation experiment; we call these methods the "iceberg lasso" and the "bootstrap lasso".

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

[11] 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.

[12] Alternatively, we could consider the adaptive lasso (Zou, 2006), which adds a second tuning parameter and is known to deliver consistent variable selection. 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.

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 may help to identify these bundles. 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.[13]

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., 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 more 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, averaging estimates and predictions across bootstrap samples may reduce overfitting due to the sampling error in the original data; in the machine learning literature, this approach is known as "bootstrap aggregating", or "bagging" for short (see, e.g., Hastie, Tibshirani, and Friedman, 2009).

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 proportion of relevant variables selected 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%.[14]

[13] The iceberg lasso 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.

[14] In the simulations we use $B = 20$ and use only the 5% cutoff.

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 the iceberg lasso and the bootstrap lasso can be said to have causal effects. Whether or not these methods are more informative than other methods that are already known to over-select regressors is an empirical matter and the answer will depend on the application. Second, in general, we need to be very humble about potential causal interpretation of our results. We view our approach as a statistical method to select a group of variables that is likely to include the ones most relevant to the fit of the three-way gravity model. This of course requires taking the model to be an appropriate representation of the determinants of trade. 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. However, the initial PPML-lasso with the tuning parameter selected by the plug-in method is likely to omit relevant variables, and that obviously complicates interpretation of those estimates. The additional steps in the iceberg lasso and in the bootstrap lasso are explicitly designed to address this latter issue and should at least partially alleviate this problem, at the cost of possibly selecting some variables that effectively have little or no 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, and illustrate how these performances change when we progressively move towards 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.

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 $\kappa \in \{5, 10, 20\}$, $\rho \in \{0.75, 0.90, 0.99\}$, $n \in \{250, 1000, 4000\}$, 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.[15]

In these simulations we considered each of the four methods presented before: cross-validation lasso, plug-in lasso, the iceberg lasso, and the bootstrap 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 lasso of Zou (2006), with the penalty parameter chosen by cross-validation in both steps.[16] 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.[17] We repeat the simulations $1,000$ times and study both the ability of each method to correctly select x_1 as a regressor and their predictive performance.

4.1 Variable selection

For each of the cases considered, Table 3 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 3 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 adaptive-lasso (AL) leads to better results, but its performance is also very poor when $\rho = 0.99$. Lasso with the penalty chosen by cross-validation (CV) provides a substantial improvement, but it also struggles for larger values of ρ. The bootstrap lasso (BL) is at

[15] 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.

[16] We also performed simulations using Zou and Hastie's (2005) elastic-net. However, those results are not particularly interesting and are not reported to conserve space and to simplify the exposition.

[17] Note, however, that the plug-in lasso has a related near-oracle property.

least as successful as CV, but clearly dominates it for the higher values of ρ. Finally, the iceberg lasso (IL) is marginally outperformed by CV and BL when $\rho = 0.75$, outperforms CV but is again marginally outperformed by BL when $\rho = 0.9$, but has a substantial advantage over all other methods for $\rho = 0.99$.[18]

Table 3: Percentage of times x_1 is selected & average number of variables selected

		$\rho = 0.75$			$\rho = 0.90$			$\rho = 0.99$		
n		$\kappa=5$	$\kappa=10$	$\kappa=20$	$\kappa=5$	$\kappa=10$	$\kappa=20$	$\kappa=5$	$\kappa=10$	$\kappa=20$
250	CV	100.0 (8.65)	99.7 (8.55)	99.3 (8.74)	96.6 (8.87)	91.8 (8.66)	85.5 (8.64)	55.2 (8.52)	37.7 (8.22)	23.4 (7.93)
	AL	99.7 (7.22)	99.4 (7.21)	97.9 (7.05)	93.9 (7.34)	87.4 (7.21)	80.4 (7.05)	45.3 (6.99)	29.4 (6.72)	17.7 (6.26)
	PI	91.6 (1.26)	89.9 (1.52)	88.1 (1.89)	80.6 (1.45)	72.1 (1.73)	63.7 (2.06)	41.1 (1.23)	26.8 (1.33)	16.9 (1.41)
	BL	100.0 (11.11)	100.0 (12.81)	99.8 (15.27)	96.6 (11.31)	98.4 (13.25)	96.7 (15.66)	90.4 (11.27)	79.2 (12.77)	64.2 (14.03)
	IL	95.7 (4.80)	95.9 (9.14)	95.2 (15.97)	95.9 (4.81)	95.8 (9.43)	93.0 (17.00)	95.3 (4.78)	93.4 (9.32)	80.1 (15.65)
1000	CV	100.0 (9.43)	100.0 (9.59)	100.0 (10.05)	100.0 (9.76)	100.0 (10.10)	99.9 (10.69)	81.0 (9.92)	69.8 (10.11)	56.4 (10.51)
	AL	100.0 (3.93)	100.0 (4.19)	100.0 (4.49)	100.0 (4.71)	99.7 (5.22)	99.7 (5.85)	68.3 (5.37)	54.8 (5.97)	40.8 (6.22)
	PI	99.8 (1.31)	99.8 (1.54)	99.7 (1.88)	99.2 (1.63)	98.4 (2.02)	97.5 (2.57)	71.4 (1.75)	55.9 (2.02)	41.4 (2.34)
	BL	100.0 (8.88)	100.0 (10.89)	100.0 (13.91)	100.0 (9.26)	100.0 (11.67)	100.0 (15.23)	98.0 (9.36)	93.70 (11.85)	87.1 (14.81)
	IL	100.0 (5.01)	100.0 (10.00)	100.0 (19.22)	100.0 (5.00)	100.0 (10.01)	100.0 (19.69)	100.0 (5.01)	100.0 (10.01)	98.8 (19.72)
4000	CV	100.0 (10.46)	100.0 (10.85)	100.0 (11.24)	100.0 (10.78)	100.0 (11.28)	100.0 (11.88)	99.0 (11.18)	97.8 (12.06)	94.9 (12.63)
	AL	100.0 (1.00)	100.0 (1.00)	100.0 (1.00)	100.0 (1.03)	100.0 (1.00)	100.0 (1.03)	91.9 (1.18)	86.0 (1.30)	79.1 (1.70)
	PI	100.0 (1.23)	100.0 (1.43)	100.0 (1.68)	100.0 (1.53)	100.0 (1.96)	100.0 (2.42)	98.0 (2.00)	93.9 (2.60)	88.1 (3.18)
	BL	100.0 (7.86)	100.0 (9.91)	100.0 (13.03)	100.0 (8.44)	100.0 (11.04)	100.0 (14.94)	100.0 (8.93)	99.9 (11.94)	99.8 (16.27)
	IL	100.0 (5.00)	100.0 (10.00)	100.0 (19.99)	100.0 (5.00)	100.0 (10.00)	100.0 (20.00)	100.0 (5.01)	100.0 (10.00)	100.0 (20.00)

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, the two extensions of the PI we consider, the BL and IL, lead to greatly improved ability of identifying the relevant regressor, and there is generally little to choose between them, except in the extreme cases with $n = 250$ and $\rho = 0.99$ where the IL has a very clear advantage.

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

[18] 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.

extensions of the PI method we consider, 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 method performs as intended. 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 is the preferred method; this is justified both by our simulation results and 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, if selecting the relevant regressor is important, the bootstrap lasso and the iceberg lasso are safe bets, with the iceberg lasso being clearly preferable only for smaller samples where there is extremely high collinearity between the relevant variable and other potential controls.

4.2 Prediction

We now consider 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 these observations. In this context, we can consider both lasso predictions, using the penalized lasso estimates, and post-lasso predictions, using unpenalized estimates.[19] 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 PL, IL, and BL.

Table 4 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 includes all p regressors and with the PPML estimates of the "oracle" model that just includes x_1.

The results in Table 4 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

[19] 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.

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 4 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.

Table 4: MSE for prediction errors

	$\rho = 0.75$			$\rho = 0.90$			$\rho = 0.99$		
	κ=5	κ=10	κ=20	κ=5	κ=10	κ=20	κ=5	κ=10	κ=20
				$n = 250$					
CV	6.85	6.83	6.86	6.87	6.88	6.88	6.83	6.83	6.80
AL	7.27	7.23	7.22	7.29	7.26	7.24	7.17	7.18	7.08
PI	6.57	6.53	6.66	6.59	6.63	6.71	6.53	6.52	6.52
BL	6.63	6.60	6.66	6.64	6.62	6.66	6.57	6.53	6.53
IL	6.71	6.83	7.21	6.71	6.84	7.25	6.72	6.85	7.23
All regressors	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
				$n = 1000$					
CV	6.34	6.35	6.35	6.34	6.34	6.35	6.33	6.32	6.34
AL	6.34	6.31	6.30	6.35	6.39	6.40	6.39	6.41	6.47
PI	6.19	6.19	6.22	6.18	6.19	6.22	6.16	6.17	6.20
BL	6.19	6.18	6.21	6.18	6.18	6.21	6.16	6.16	6.18
IL	6.22	6.31	6.48	6.22	6.31	6.47	6.22	6.31	6.48
All regressors	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
				$n = 4000$					
CV	6.37	6.37	6.37	6.36	6.37	6.38	6.37	6.38	6.38
AL	6.34	6.34	6.34	6.33	6.33	6.34	6.34	6.34	6.34
PI	6.34	6.35	6.36	6.34	6.35	6.35	6.33	6.34	6.35
BL	6.35	6.36	6.37	6.36	6.37	6.37	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 regressors	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.

4.3 Summary of the findings

The simulation results presented above, which confirm and extend the findings of Drukker and Liu (2019), have important implications for our work.

Given that in our application we only have data on 282 trade agreements,[20] we cannot expect any of the methods considered to be able to precisely identify the set of provisions that matter for trade. The task of identifying the correct set of explanatory variables is particularly challenging in our application because many of the provisions have very strong correlations with others, and there are even cases of perfect collinearity. In this challenging context, the iceberg lasso and bootstrap lasso emerge as providing a good compromise between parsimony and the ability to identify the relevant variables. The iceberg lasso has the practical advantage of being easier to implement and not requiring the choice of additional parameters, such as the number of bootstrap samples. Consequently, the iceberg lasso is our preferred approach to select the relevant variables, but the bootstrap lasso is a credible alternative that can be used, at least, as a robustness check. Additionally, the bootstrap lasso is the only approach we have considered that can provide information about model uncertainty, but exploring that is beyond the scope of this paper.

If the objective of the researcher is to accurately predict the trade impact of a given PTA, 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.[21] However, the bootstrap lasso is also a credible alternative in this context and it can serve as a useful robustness check.

5 Empirical Results

In this section, we present the lasso results obtained using the methods described and studied in the previous sections. We first present results for the plug-in method before briefly discussing the results obtained using cross-validation. We then turn to the iceberg and bootstrap lasso results, which each build in their own way on the selection done by the plug-in lasso. We also include a brief discussion of using these methods for prediction.

[20] Note that the information on the effect of the different provisions is limited by the relatively small number of PTAs that are observed. Therefore, despite having a large number of observations, we effectively only have a small sample to identify the effect of the different provisions.

[21] 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.

5.1 Plug-in Lasso Results

Table 5 presents results for the plug-in lasso and post-lasso regressions discussed before.[22] 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% ($\exp(0.131) - 1 = 0.14$).

Column (2) then shows the results of the plug-in PPML-lasso regression, showing only the coefficients that are found to be non-zero. Using this approach, the lasso selects 8 provisions related to anti-dumping, competition policy, technical barriers to trade (TBT), and trade facilitation. Broadly speaking, these variables all 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.[23] 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 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 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% ($\exp(0.349) - 1 = 0.42$). Interestingly, not all of the provisions selected by the lasso step are found to be statistically significant in the post-lasso step. This apparent contradiction 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 different from zero. Second, because the lasso shrinks all coefficients towards zero simultaneously, 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 re-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

[22] Both the PPML standard errors and the plug-in lasso estimates 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.

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

Table 5: 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 by PPML using 316,317 observations. Columns labelled "Post-lasso" report PPML coefficients for all variables selected by a plug-in lasso method in a prior step. All other columns report further experiments using PPML. 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.

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).

Next, column (5) returns to our original simple model from column (1) but adds a dummy variable for the EU.[24] Our reasons for treating the EU separately from other agreements are three-fold. First, we suspect that not all of the EU's efforts to promote trade are captured in how their provisions variables are coded in our data. There could also be unobserved effects that are channeled through the EU's secondary law process, in which the EU's governing institutions are empowered to pass new regulations and directives on an ongoing basis. Second, our provisions data set does not include agreements that are no longer in effect. For the most part, the agreements that cannot be included are EU pre-accession agreements, which obviously are subsumed by the EU agreement once each new member joins the EU. As discussed in Section 2, we deal with this data issue in practice by dropping all observations associated with obsolete agreements. Nonetheless, this could lead to biased estimates of the EU agreement and the provisions associated with it. Third, the latest EU agreement has in place six of the eight provisions selected in column 2 (all except AD14 and TBT7); thus, we want to make sure we are not simply picking up an "EU effect" in the provisions that are selected.

As the PPML results in column (5) show, the estimated EU effect is large, several times that of non-EU PTAs in fact. However, when we treat the EU as a possible predictor in the lasso, we find that is not selected and consequently the set of provision variables selected is identical to that in column (2), which is our preferred set to work with in the subsequent iceberg lasso and bootstrap lasso analyses.

5.2 Cross-Validation Lasso Results

As discussed before, 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 use the cross-validation approach.[25]

Figure 2 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 3 and 4 show the corresponding regularization paths for selected provisions.[26] That is, the figures show how the value of the estimated (post-lasso) coefficient on the selected provisions changes as we vary λ. As expected, fewer provisions are selected as we increase λ and, for values of $\lambda/\sum_{ijt} y_{ijt}$ around 0.01, which

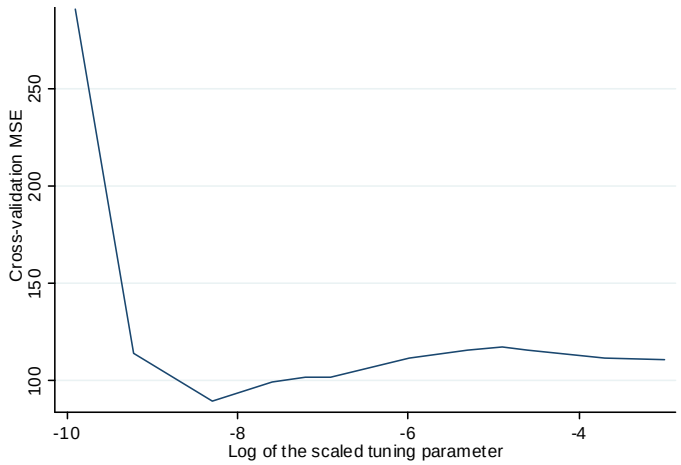
[24] We use EU as shorthand for the EU and EC agreements.

[25] As explained before and in the Appendix, the cross validation is performed clustering by agreement.

[26] In each panel of the figures, the fourth set of estimates from the right corresponds to the variables selected by the cross-validation method.

is forty times larger than the optimal value, we generally see a close correspondence between the results in Figures 3 and 4 and those that we found earlier using the plug-in method.

Figure 2: Cross-validation MSE *vs.* tuning parameter



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 3 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.

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 3 and 4 that many of the estimated effects for the provisions selected by cross-validation are not plausible when interpreted on their own. These observations reflect the known shortcomings of the cross-validation approach that we stated earlier and found support for in our simulations.

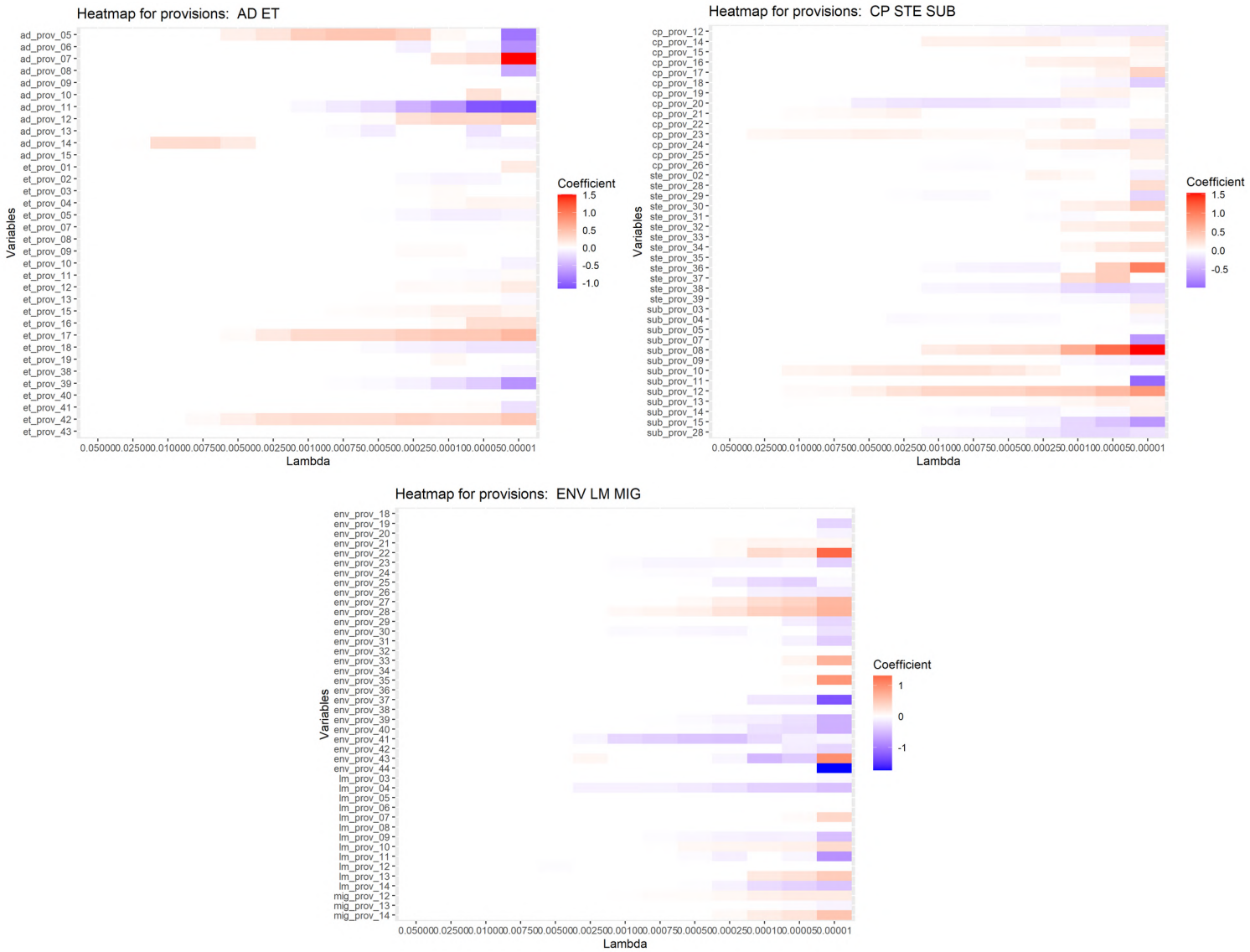


Figure 3: Regularization path for selected provisions (AD, ET, CP, STE, SUB, ENV, LM, and MIG)

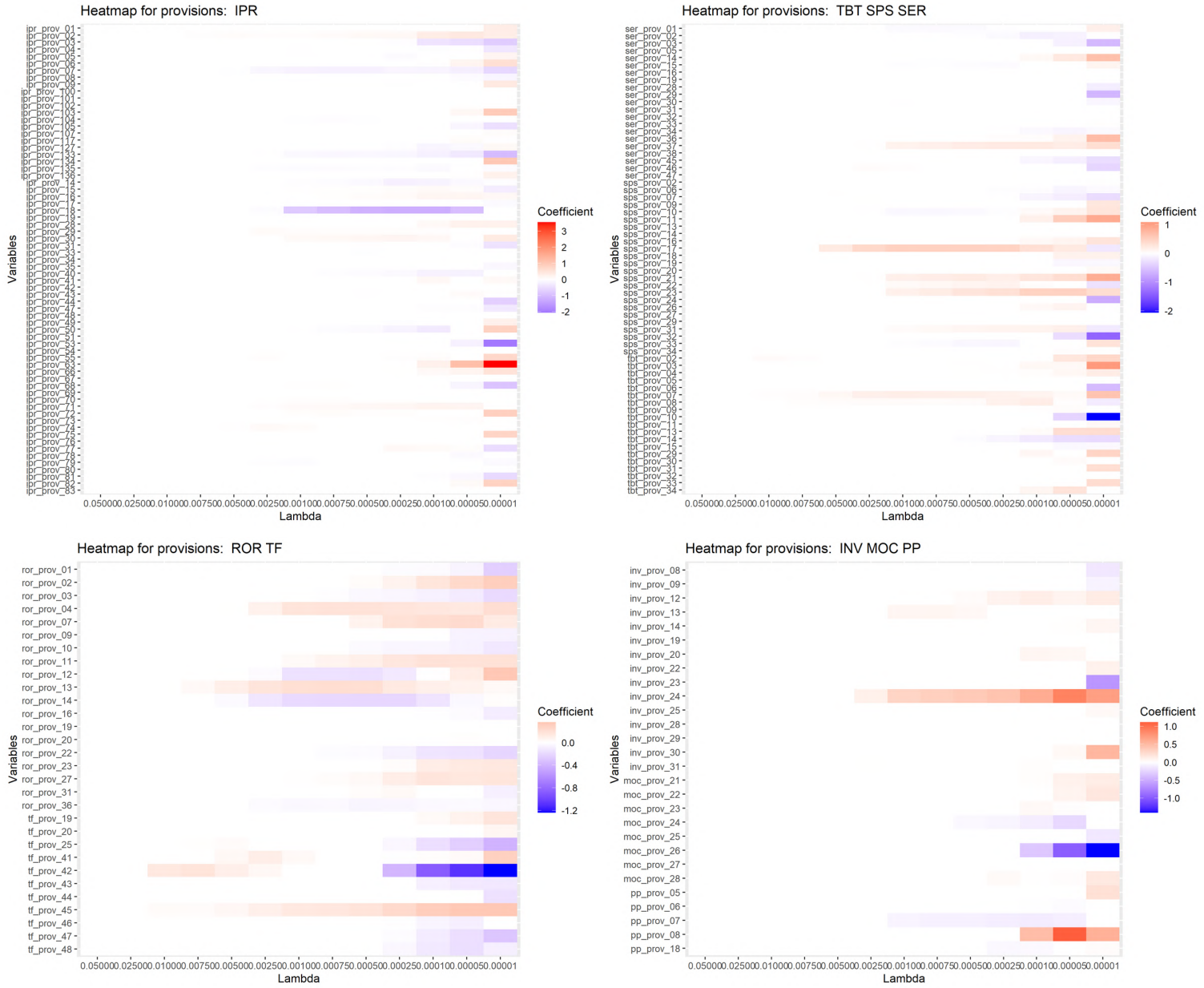


Figure 4: Regularization path for selected provisions (IPR, TBT, SPS, SER, ROR, TF, INV, MOC, and PP)

5.3 Iceberg Lasso Results

As previously mentioned, we cannot be certain whether the variables selected by the lasso have a causal effect on trade or are simply highly correlated with the variables that have a causal effect. In this section, we investigate this issue further by carrying out the iceberg lasso analysis we proposed earlier. That is, for each of the provisions from our preferred set of estimates (those from the third column of Table 5), we run an additional plug-in lasso regression where we regress each selected provision on all of the provisions excluded by our first-stage lasso.[27] As discussed, the purpose of these auxiliary regressions is to construct bundles of provisions that, at least when combined together, are likely to have a causal impact on trade flows when included in trade agreements. As we have noted, the reader should be cautioned that we will not be able to say with high certainty whether a given provision is important for promoting trade, but, as we will see, this method gives us significantly increased parsimony versus relying on cross-validation. Furthermore, as we have seen from our simulations, it should also give us more confidence in the results.

Table 6 presents the results of our iceberg lasso analysis. The first two rows of Table 6 list each of the eight provisions selected by the first-stage plug-in lasso, as well as their estimated impact on trade flows from column (3) of Table 5. The subsequent rows of Table 6 report all provisions that were not selected by the lasso in the first step but are identified in the second step of the iceberg lasso; we also report the correlation of each of these provisions with the selected provision in the first row. Finally, the last row reports the R^2 of the regression of each selected provision on the corresponding correlated provisions. For example, column (1) shows that anti-dumping provision AD14 is highly correlated with two further anti-dumping provisions (AD06 and AD08), as well as with one provision on environmental protection (ENV42); the R^2 of the regression of AD14 on these three provisions is 0.95.

The results in Table 6 show that the iceberg-lasso identifies a total of 42 ($= 8 + 34$) distinct provisions that are likely to be associated with increased trade. This finding contrasts with the 133 provisions identified by the cross-validation lasso and the 8 provisions selected by the plug-in lasso. Therefore, as in the simulations in the preceding section, the iceberg lasso appears to provide a good compromise between the cross-validation lasso, which selects so many provisions that makes it difficult to interpret its results, and the plug-in lasso, which is likely to miss important provisions.

Looking in more detail at the results in Table 6, we find that provision AD14 is correlated with other anti-dumping provisions; this correlation is not surprising because all these provisions fulfill a similar purpose, which is to increase transparency in the use of anti-dumping duties. In that sense, one conclusion to be drawn from this exercise is that anti-dumping provisions are likely to increase trade flows, although we cannot say which of them has the biggest effect. Table 6 shows that, more surprisingly, AD14 is also strongly correlated with ENV42. This correlation seems to be due to what might

[27] 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 now is only by agreement.

Table 6: Iceberg lasso results

(1)	(2)	(3)	(4)	(5)	(6)	(7)
AD14	CP23	TBT02/29	TBT07	TBT08	TBT33	TF45
(+41.7%)	(+12.5%)	(+20.2%)	(+3.2%)	(+13.1%)	(+12.0%)	(+9.3%)
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 associated with increases in bilateral trade flows (row 1), together with the estimated increase in trade flows (row 2), as well as other provisions that predict the provision in row 1 (rows 3-15; numbers in brackets are raw correlations with the provision from line 1). The last row displays the R^2 of the regression of each selected provision on the corresponding correlated provisions.

be called a template effect, that is, the tendency of important trading blocs such as the EU and the US to use similar provisions in all their agreements. For example, most agreements signed by the EU include provisions on anti-dumping and the environment, hence leading to a high correlation between the corresponding provisions in our data.[28]

The same provisions that were found to be correlated with AD14 also have a reasonably high correlation with CP23, which serves to promote transparency in competition policy. That said, the variables with the strongest correlations with CP23 are other competition policy provisions, namely CP22 and CP24. Thus, it seems likely that the presence of provisions on competition policy is behind the observed trade increasing effect of CP23, although we are again unable to say which provision exactly is driving this effect.

We find that TBT07 also has a substantial correlation with the above-mentiond AD6, AD8, and ENV42 provisions but, not surprising, the strongest correlations are with other TBT provisions (TBT15, TBT34) that also relate to the use of international standards. Thus, it seems that provisions encouraging the use of international standards in the area of technical barriers to trade are likely to be behind the trade increases associated with provision TBT07, although we cannot say which of the individual TBT provisions is driving the observed effect.

As for the other TBT provisions selected in the first step, TBT02/29, TBT08, and TBT33, they are all strongly related to TF42, a trade facilitation provision, with TBT02/29 being also correlated with provisions related to competition policy (CP14. CP21, and CP22), state-owned enterprises (STE32), and subsidies (SUB09 and SUB10), and TBT33 to other TBT provisions such as TBT06 and TBT14. This set of results makes clear that provisions related to TBT are likely to have a significant trade facilitation effect, but we are not able to identify precisely which ones are relevant.

The plug-in PPML-lasso also selects a provision related to the simplification of procedures to issue proof of origin (TF45), and this provision is highly correlated with TF44, which relates to the simplification of requirements for proof of origin. As noted above, Table 6 also indicates that other trade facilitation provisions are correlated with some of the provisions selected in the first stage; this is true for CP23, TBT33, and especially for TBT02/29 and TBT8. Thus, our results suggest that trade facilitation procedures, particularly those related to rules of origin, are likely to play an significant role in increasing trade flows.

Finally, the iceberg lasso also identifies provisions from other areas that help predict the provisions identified in the first step. For example, provisions in policy areas such as movement of capital and public procurement are related to TBT33, but these types of provisions are associated with smaller raw correlations. By the logic of the lasso, it is likely that these provisions are informative for predicting the presence of TBT33 in a relatively small number of agreements where other provisions with higher raw correlations are not found.

In summary, although it is not possible to identify with certainty which provisions are most important for increasing trade, our results allow us to find a relatively small

[28] In our data, ENV42 is perfectly colinear with AD06 and AD08.

bundle of provisions that are likely to have the desired effect. In particular, provisions related to TBTs, anti-dumping, trade facilitation, subsidies, and competition policy are likely to enhance the trade-increasing effect of trade agreements.

5.4 Bootstrap Lasso Results

As an alternative to the iceberg lasso, we now present the results obtained with the bootstrap lasso. Tables 7 and 8 summarize the results obtained from 250 bootstrap samples. The resampling process treats pairs belonging to the same agreement as belonging to the same cluster, treating pairs as clusters otherwise. In each replication, we performed selection using plug-in lasso and record which variables are selected and their their post-lasso PPML coefficient estimates.

Table 7: Bootstrap lasso results

Provisions with largest average coefficients		Provisions selected most frequently	
AD14	0.079	AD14	0.372
CP23	0.065	CP23	0.320
CP22	0.063	TBT07	0.308
AD05	0.055	SPS06	0.228
TBT07	0.054	TBT08	0.208
TBT02/29	0.048	SUB12	0.184
TBT08	0.038	TBT02/29	0.168
SUB12	0.030	TBT33	0.160
TBT34	0.029	CP22	0.156
SPS06	0.028	TBT34	0.152
TF42	0.027	TBT06	0.148
TBT33	0.023	AD05	0.140
TF41	0.023	CP21	0.124
TBT06	0.021	TF45	0.116
CP21	0.020	ENV33	0.116

Notes: Bootstrap plug-in lasso performed using cluster-bootstrap resampling with 250 replications. The numbers shown are (left) the 15 largest average post-lasso coefficient estimates across all replications, and (right) selection frequencies for the 15 most frequently selected provisions.

Table 7 presents the average coefficients for the provisions with the 15 largest average coefficient across all replications (on the left) as well as the selection frequencies for the 15 most frequently selected provisions (on the right). It is worth noting 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 5), but it is only selected in 37% of replications. This 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, if we take the method literally, AD14 is found quantitatively to be more likely to matter than other provisions.

Overall, the results in Table 7 are reassuring in that they broadly confirm our earlier findings using the iceberg lasso. Indeed, most of the provisions in Table 7 were previously identified as potentially relevant by the iceberg lasso. Moreover, there are multiple provisions related to anti-dumping, competition policy, trade facilitation, and TBTs that tend to be selected with relatively high frequency and have relatively high average coefficients (when averaged across all the bootstrap replications), reminiscent of the provision groupings that were indicated with the iceberg method.

Table 8: 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	74	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.

Table 8 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. These results, therefore, show which provision categories, when taken as a whole, are likely to have the biggest impact on trade. The category with the biggest total impact turns out to be TBTs, followed by anti-dumping and competition policy. Next after that are subsidies, sanitary and phytosanitary measures, trade facilitation, and export taxes. Overall, the differences between categories seem to comport with intuition (very small impacts for services and

labor markets, for example). They also are, again, broadly in line with the findings obtained with iceberg lasso.

5.5 Predicting the effect of trade agreements

Having identified sets of provisions that are more likely to positively affect trade flows, it is natural to think of ways to use this information to evaluate the effects of different PTAs, and even to predict the impact of new ones. In the reminder of this section we discuss ways to perform these prediction exercises and the associated caveats.[29]

The simulation results presented in Section 4 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 5. 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 5, which is 86% ($\exp(0.618) - 1 = 0.86$).[30]

In results that are summarized in the third column of Table 9, we repeat this exercise for each of the PTAs in our data.[31] 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.[32] 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 deviation is thousands of times that of the estimates produced using plug-in lasso. 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 in Section 4, 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 two extensions of the plug-in lasso we have proposed, the iceberg lasso and the bootstrap lasso. The iceberg lasso has the advantage

[29] As in Section 4, in this section we compute penalized predictions when using cross-validation, and post-lasso unpenalized predictions for the plug-in, iceberg, and bootstrap lasso. For the bootstrap lasso, the predictions are obtained by averaging the post-lasso predictions in each of the bootstrap samples.

[30] 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.

[31] 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 7.

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

that it is likely to select more of the provisions with a causal impact than the plug-in lasso. Moreover, it performed reasonably well as a predictive method in our simulations. However, as is apparent from column 4 of Table 9, in this application, predictions based on iceberg lasso lead to some unrealistic estimates. Intuitively, the provisions selected by the iceberg lasso will, by design, include multiple regressors that are highly collinear with one another. Therefore, although it may be possible to estimate the joint effect of these variables with reasonable precision, the same is unlikely to be the case for each individual effect. This implies that the iceberg lasso is likely to be a good predictor of the effect of PTAs that include all these variables, but it may lead to unreliable results for PTAs that only include a subset of the highly collinear provisions.

Table 9: Summarizing Estimates of Heterogeneous PTA Effects

	(1)	(2)	(3)	(4)	(5)
	All variables	CV	Plug-in	Iceberg	Bootstrap
Descriptive statistics					
Min	−81.2%	−50.4%	0.0%	−62.8%	0.0%
Max	> 1e6%	387.0%	144.4%	284.9%	101.0%
Mean	328774.6%	32.1%	13.8%	17.2%	12.5%
Median	26.4%	14.4%	9.3%	6.7%	7.2%
Stdev.	300514.7pp	63.0pp	20.7pp	42.4pp	15.3pp
Correlations					
PPML	1	0.146	−0.054	0.233	0.041
CV	0.146	1	0.391	0.550	0.513
Plug-in	−0.054	0.391	1	0.507	0.925
Iceberg	0.233	0.550	0.507	1	0.679
Bootstrap	0.041	0.513	0.925	0.679	1
Estimated partial effects for selected PTAs					
EU	104.9%	105.4%	87.1%	101.6%	64.2%
EEA	80.4%	90.5%	9.3%	94.4%	18.3%
Eurasian Econ. Union	−21.8%	71.8%	144.4%	38.5%	101.0%
NAFTA	77.9%	77.5%	79.9%	81.5%	52.9%
MERCOSUR	145.5%	115.9%	42.1%	76.2%	39.6%
ECOWAS	469.6%	379.2%	9.3%	23.3%	19.4%
ASEAN	1.8%	−9.4%	0.0%	0.0%	3.3%

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

The predictions based on the bootstrap lasso performed well in our simulations, and this approach also shows promise here. As shown in column 5 of Table 9, 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 9 reveal that the estimated effects obtained with the plug-in lasso and bootstrap lasso can differ substantially for individual PTAs.

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.

In summary, the plug-in lasso is our preferred method to estimate the effect of individual PTAs, but the bootstrap lasso may be a worthwhile check at the very least. 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 in Section 4 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 measured to have zero impact despite having many of the true causal provisions.

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. While these provisions span a range of policy areas, our results generally support the conclusion that a select number of provisions related to technical barriers to trade, anti-dumping, trade facilitation, subsidies, and competition policy are most effective at promoting trade as compared to other types of provisions that appear in PTAs.

In spite of the obvious appeal that lasso methods have in this context, we need to be clear that interpreting their results requires some important caveats. In particular, we know that it is possible that even our preferred lasso methods may fail to discover important trade-promoting provisions, and that they are almost certain to lead to the inclusion of provisions that are not relevant. The iceberg lasso and bootstrap lasso

methods do, however, improve upon both the standard cross-validation lasso and the plug-in lasso as variable selection methods.

In terms of broader applications, our methods are not limited to just PTAs or even just to trade. There are many other contexts in which the iceberg lasso and bootstrap lasso methods we have introduced could be helpful tools for researchers wishing to determine which of a large number of variables are worth focusing on as most relevant for the outcome. Furthermore, by integrating the lasso into a nonlinear model with high-dimensional fixed effects, we show how machine learning methods for variable selection and related tasks can be utilized in much more generalized settings than what had been possible previously.

Appendix

Provisions list

Table A1: Provisions selected by the iceberg lasso

Anti-dumping	
AD06	If there are no sales in the normal course of trade in the domestic market of the exporting country
AD08	Cost of production in the country of origin plus a reasonable amount
AD11	Price effects of dumped imports
AD14	Requirement to establish material injury to domestic producers
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?
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?
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?
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?
Public Procurement	
PP08	Does the agreement contain explicit provisions on MFN treatment of third parties?
Sanitary and Phytosanitary Measures	
SPS11	Does the agreement promote the creation of concerted/regional standards?
SPS21	Risk Assessment: Is there reference to international standards/procedures?
SPS24	Is the burden of justifying non-equivalence on the importing country?

Table A1 (cont'd): Provisions selected by the iceberg lasso

State-Owned 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	
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?
Technical Barriers to Trade	
TBT02	Technical Regulations - Is mutual recognition in force?
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?
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?
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	
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?

More Details on HDFE-PPML-Lasso Estimation

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

$$(\widehat{\alpha},\widehat{\gamma},\widehat{\eta},\widehat{\beta}) := \arg\min_{\alpha,\gamma,\eta,\beta}\left[\frac{1}{n}\sum_{i,j,t}\exp(x'_{ijt}\beta+\alpha_{it}+\gamma_{jt}+\eta_{ij})\right.$$
$$\left.-\frac{1}{n}\sum_{i,j,t}y_{ijt}\left(x'_{ijt}\beta+\alpha_{it}+\gamma_{jt}+\eta_{ij}\right)+\frac{1}{n}\sum_{k=1}^{p}\widehat{\phi}_k\lambda|\beta_k|\right], \tag{3}$$

where $\widehat{\phi}_k$, to be precisely defined below, is identical to 1 except when the plug-in method is used.

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

$$\begin{aligned}
\widehat{\alpha}_{it}:&\frac{1}{n}\sum_{j}y_{ijt}-\widehat{\mu}_{ijt}=0, && \forall i,t,\\
\widehat{\gamma}_{jt}:&\frac{1}{n}\sum_{i}y_{ijt}-\widehat{\mu}_{ijt}=0, && \forall j,t,\\
\widehat{\eta}_{ij}:&\frac{1}{n}\sum_{t}y_{ijt}-\widehat{\mu}_{ijt}=0, && \forall i,j,\\
\widehat{\beta}_{k}:&\frac{1}{n}\sum_{i,j,t}\left(y_{ijt}-\widehat{\mu}_{ijt}\right)x_{ijt,k}-\frac{1}{n}\widehat{\phi}_k\lambda\,\mathrm{sign}(\hat{\beta}_k)=0, && k=1...p,
\end{aligned}$$

where $\widehat{\mu}_{ijt}$ denotes μ_{ijt} evaluated at $\widehat{\alpha}$, $\widehat{\gamma}$, $\widehat{\eta}$, $\widehat{\beta}$. 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 possible 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 (3) 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

$$\widehat{\beta} := \arg\min_{\beta}\left[\frac{1}{n}\sum_{i,j,t}\exp\left(x'_{ijt}\beta+\widehat{\alpha}_{it}(\beta)+\widehat{\gamma}_{jt}(\beta)+\widehat{\eta}_{ij}(\beta)\right)\right.$$
$$\left.-\frac{1}{n}\sum_{i,j,t}y_{ijt}\left(x'_{ijt}\beta+\widehat{\alpha}_{it}(\beta)+\widehat{\gamma}_{jt}(\beta)+\widehat{\eta}_{ij}(\beta)\right)+\frac{1}{n}\sum_{k=1}^{p}\widehat{\phi}_k\lambda|\beta_k|\right], \tag{4}$$

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

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

where

$$\widetilde{x}_{ijt,k} := x_{ijt,k} + \frac{d\widehat{\alpha}_{it,k}}{d\beta} + \frac{d\widehat{\gamma}_{it,k}}{d\beta} + \frac{d\widehat{\eta}_{ij,k}}{d\beta} \tag{5}$$

captures both the direct and indirect effects of a change in β on the conditional mean of y_{ijt}.

To explain how we deal with the fixed effects, assume for the moment that we know the true values of $\mu_{ijt} := e^{x_{ijt}\beta+\alpha_{it}+\gamma_{jt}+\eta_{ij}}$ 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

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

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

$$\min_{\beta} \left[\frac{1}{2} \sum_{i,j,t} \mu_{ijt} \left(\widetilde{z}_{ijt} - \widetilde{x}'_{ijt}\beta \right)^2 + \sum_{k=1}^{p} \lambda \widehat{\phi}_k \left| \beta_k \right| \right], \tag{6}$$

where $\widetilde{z}_{ijt}$ and $\widetilde{x}_{ijt}$ are respectively defined as the "partialed-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 (6) is consistent with the definition given for $\widetilde{x}_{ijt,k}$ in (5).

The nice thing about expressing the problem as in (6) 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 (6). In principle, this algorithm can be easily modified to other settings that feature multi-way fixed effects in order to simplify estimation.

More 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 they should satisfy the following inequality for all k:

$$\frac{\lambda\widehat{\phi}_k}{n} \geq c\left|\frac{1}{n}\sum_{i,j,t}(y_{ijt} - \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij}))\widetilde{x}_{ijt,k}\right| \quad \forall k, \tag{7}$$

for some $c > 1$. Intuitively,

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

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 $x_{ijt,k}$ will not be selected.

Next, suppose that the observations associated with trade agreements are partitioned into G clusters indexed by $g = 1, \ldots, G$, and let $o = (i, j, t)$ serve as the unique index for each observation. Set

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

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 under the assumption that errors are correlated within their respective clusters. Under suitable regularity conditions, $\widehat{\phi}_k^2 - \phi_k^2 = o_p(1)$ uniformly in k, where ϕ_k^2 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 deviation in order for regressor k to be selected.

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

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

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 (7) is satisfied. Hence, following Belloni, Chernozhukov, Hansen, and Kozbur (2016), we set

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

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

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 $\beta_k = 0$ for $k = 1, ..., p$. 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 regularization conditions, the use of the plug-in lasso method just described ensures the following "near-oracle" property for $\widehat{\beta}_{PL}$,

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

and hence the post-lasso estimates are consistent at a rate that differs from the oracle rate only up to the log factor $\max(\log n, \log p)$.

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.

More Details on Cross-Validation

As discussed in the main text, the idea behind cross-validation (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. 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 $\widehat{\beta}_{L,g}(\lambda)$ be the lasso estimator obtained via the minimization of (4) when holding out the n_g

observations contained in partition g. Define the CV bandwidth as

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

Since λ_{CV} is based on the minimization of the average MSE over different subsamples, we expect it to deliver a much more lenient variable selection. There is some disagreement 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.

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.

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.

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

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.

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.

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.

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., J. Wanner, Y.V. Yotov, T. Zylkin (2019). "Currency unions and trade: a PPML re-assessment 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.

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.

Prusa, T., R. Teh, and M. Zhu (2022). "PTAs and the incidence of antidumping disputes," available at `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.

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.