Determining which trade agreement provisions matter for trade*

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

27 May 2026

Abstract

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

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

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

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

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

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

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

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

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2 Data

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

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

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

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

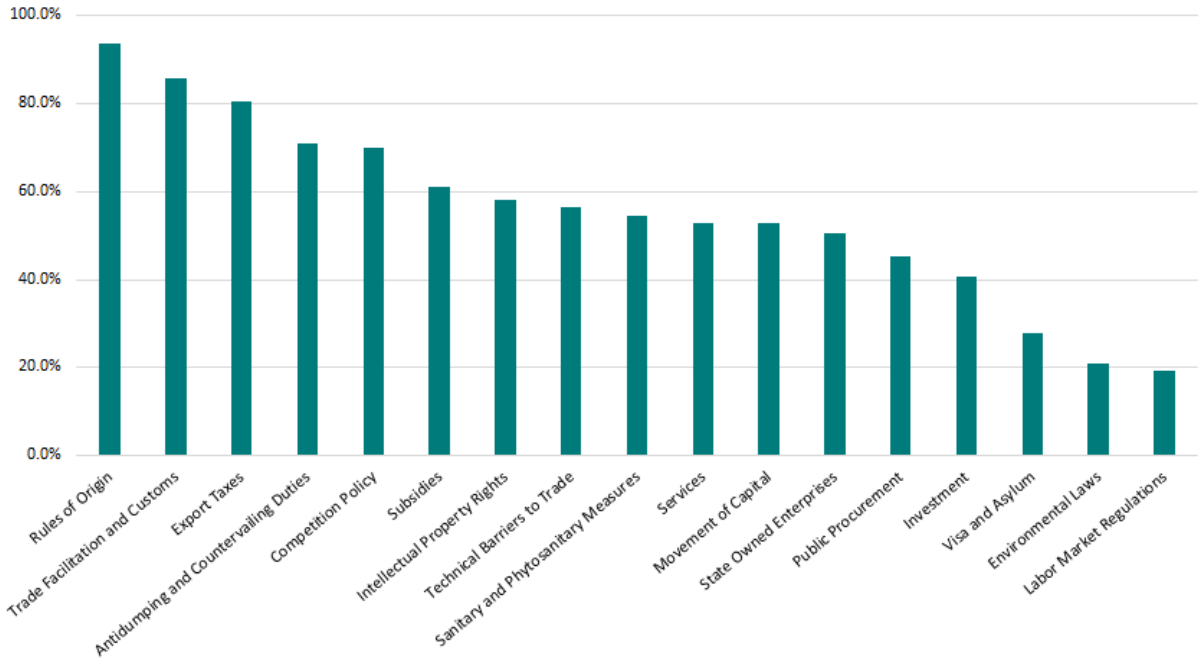
[4] In this data set, information coming from secondary law (the body of law that derives from the principles and objectives of the treaties) has not been coded. This is of particular importance for agreements such as the EU, since most policy areas covered have used secondary law such as regulations,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

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

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

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

Figure 1: Share of PTAs that cover selected policy areas



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

Table 1: Distribution of essential provisions by policy area

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

3 Determining Which Provisions Matter for Trade

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

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

3.1 The Gravity Model

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

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

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

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

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

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

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

3.2 Variable Selection and Gravity

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

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

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

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

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

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

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

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

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

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

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

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

3.3 Implementing the Lasso

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

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

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

Plug-in Lasso

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

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

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

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

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

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

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

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

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

Cross-Validation

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

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

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

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

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

Extensions of the plug-in lasso

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

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

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

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

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

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

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

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

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

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

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

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

3.4 Discussion and caveats

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

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

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

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

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

4 Simulation Evidence

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5 Empirical Results

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

5.1 Plug-in Lasso Results

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2 Bootstrap Lasso Results

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.3 Iceberg Lasso Results

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

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

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

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

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

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

Table 6: Iceberg lasso results

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

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

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

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

6 Conclusions

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

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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