Discussion Papers in Economics

Machine Learning in International Trade Research - Evaluating the Impact of Trade Agreements

By

Holger Breinlich

(University of Surrey, CEP and CEPR),

Valentina Corradi

(University of Surrey),

Nadia Rocha

(World Bank),

Michele Ruta

(World Bank),

J.M.C. Santos Silva

(University of Surrey),

&

Tom Zylkin

(University of Richmond).

DP 05/21

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

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[**]

24 March 2021

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 develop a new method to estimate the impact of individual provisions on trade flows that does not require ad hoc assumptions on how to aggregate individual provisions. Building on recent developments in the machine learning and variable selection literature, we propose data-driven methods for selecting the most important provisions and quantifying their impact on trade flows. We find that provisions related to antidumping, competition policy, technical barriers to trade, and trade facilitation 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 also 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 and Jiayi Ni provided excellent research assistance. The usual disclaimer applies.

[†] 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 protection. 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 policymakers 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, and Regmi and Baier, 2020). 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 propose a new method to estimate the impact of individual provisions on trade flows that does not require ad hoc assumptions to aggregate individual provisions. Instead, we propose a data-driven method based on recent developments in the machine learning and variable selection literature to select the most important provisions and quantify their impact on trade flows.

In doing so, we build on recent advances in variable selection methods that address difficulties arising from a key feature exhibited by trade data, namely the high degree of correlation between individual PTA provisions. We propose an extension of the Belloni, Chernozhukov, Hansen, and Kozbur (2016) approach to the case of nonlinear models with high-dimensional fixed effects, which have become standard in the analysis of trade flows in recent years (see, for example, Head and Mayer, 2014, Yotov, Piermartini, Monteiro, and Larch, 2016). In particular, we use a Poisson pseudo-maximum likelihood (PPML) version of the well-known lasso (Least Absolute Shrinkage and Selection Operator) method for variable selection (see, for example, Hastie, Tibshirani, and Friedman, 2009) and show how to choose the tuning parameter of this estimator using either a plug-in method based on Belloni, Chernozhukov, Hansen, and Kozbur (2016) or cross-validation. Notably, this requires overcoming a number of practical problems inherent in the nature of trade data, such as the nonlinearity of the underlying gravity model and the need to control for multilateral resistance and unobserved trade barriers.

We apply our method to a comprehensive dataset on PTA provisions recently made available by the World Bank (Mattoo, Rocha and Ruta, 2020). Importantly, this database is very rich, 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, these provision variables can be highly correlated with one another. For these reasons, we complement our plug-in lasso results with a novel methodology that seeks to identify potentially important variables that may have been missed in the initial lasso step. As we show using simulation evidence, this new method, termed the “iceberg lasso”, presents a favorable balance between the rigor of the plug-in lasso and the lenience of cross-validation methods in small-to-moderate samples where the true causal variables may be highly correlated with an unknown number of other variables in the dataset. To be clear, this two-step approach does not completely answer the question of “which provisions matter most for trade?”, but it does lead to substantial improvements in our ability to find the correct provision variables and narrow down the number of potential candidates in the presence of such rich data.

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. This literature has been predominantly interested in estimating the overall effects of trade agreements rather than individual provisions (see, for example, Baier and Bergstrand, 2007). More recently, attention has shifted to trying to decompose the overall PTA effect and to disentangle the effects of individual trade agreement provisions. As previously discussed, this literature often needs to make strong assumptions about the importance of individual provisions or needs to aggregate them in essentially arbitrary ways (see Mattoo, Mulabdic, and Ruta, 2017; Dhingra, Freeman, and Mavroeidi, 2018). We propose instead a novel set of methods to select the most important provisions and to quantify their impact on trade flows. To provide some headline results, our plug-in lasso results find that 6 provisions related to antidumping, competition policy, technical barriers to trade, and trade facilitation are associated with enhancing the trade-increasing effect of trade agreements. When we then use our iceberg lasso procedure to look beyond the “tip” of the proverbial iceberg, we subsequently identify a set of 43 provisions out of 305 provision variables in our data that may be impacting trade. For some comparison, a more conventional approach based on cross-validation selects 124 provisions as being relevant and, based on our simulations, is actually less likely to include all of the “correct” provisions.

In addition, we contribute to the subset of the machine learning literature interested in variable selection. In particular, we extend and adapt existing work by Belloni, Chernozhukov, Hansen, and Kozbur (2016) to make it applicable to the context of international trade flows and trade agreements. This requires an extension to the estimation of nonlinear models with high-dimensional fixed effects using PPML. The international trade context also throws up some interesting challenges when trying to select the tuning parameter that governs the extent to which our PPML-lasso estimator penalizes coefficients on included variables and hence selects included variables. In particular, standard cross-validation methods such as k-fold or leave-one-out approaches are not feasible in practice, requiring us to propose a novel approach based on out-of-sample predictions of the effects of PTAs. We find that the number of provisions selected when the tuning parameter is chosen by cross-validation is too large 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. The two-step method that we propose builds on the results obtained using the plug-in penalty and identifies an additional set of provisions that may have a causal effect on trade.

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 a 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. It also includes simulation evidence comparing relative performance of different lasso methods in a simplified setting with high correlation between regressors. Section 4 applies our methods to our database on PTA provisions and shows which individual provisions are the strongest predictors of trade flows. Section 5 concludes.

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 283 PTAs that have been signed and notified to the WTO between 1958 and 2017. The data focus on the sub-sample of 18 policy areas that are most frequently covered in trade agreements – defined as areas that are present in 20 percent or more of the 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 export taxes 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

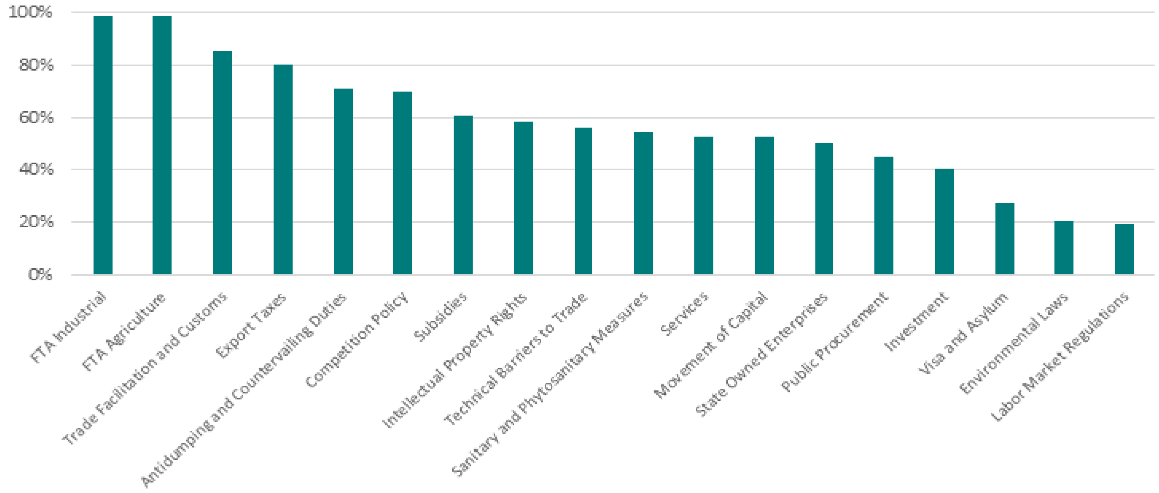


Figure shows the share of PTAs that cover a policy area. Source: Mattoo, Rocha and Ruta (2020).

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.[1]

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".[2] 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).

[1] In this dataset, 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.

[2] The classification into essential and corollary in the database is based on experts' knowledge and, hence, is subjective.

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	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%

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, antidumping, 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 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 dataset 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 an agreement 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 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.

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 Yotov, Piermartini, Monteiro, and Larch, 2016, Weidner and Zylkin, 2020), 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 innovation, the iceberg lasso method, which we will motivate as providing a balance between "cross-validation" approaches that tend to select too many variables and more rigorous, "plug-in" methods that may select too few. We also include simulation evidence comparing the performance of these various methods.

3.1 The Gravity Model

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

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

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". The use of the term "gravity" is most closely associated with the exporter-time and importer-time fixed effects α_{it} and γ_{jt}. Intuitively, these two fixed effects 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 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 two main pitfalls with this approach. First, if the correct model for trade flows is given by (1), OLS estimation is consistent only if the distribution of the error term satisfies very strong conditions. Second, it cannot deal with zero trade flows. Given the exponential mean form, there are good reasons to instead estimate using 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.[3] Weidner and Zylkin (2020) have recently

[3] 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 theorm to update the weights, repeating until convergence. For a different approach, see Larch, Wanner, Yotov, and Zylkin (2019).

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. With our dataset, this high dimensionality, combined with the relatively small number of PTAs, leads to implausibly large and uninterpretable estimates due to multicollinearity. Furthermore, the estimated model will have 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. 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 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,[4] 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 x_{ijt} 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 constructing them appropriately to reflect any heteroskedasticity and within-pair 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 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 our 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.[5]

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 an extension of the plug-in lasso, termed the "iceberg lasso", that is intended to address one of the plug-in lasso's key shortcomings in this context.

[4] 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.

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

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 relatively "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 panel data lasso method of Belloni, Chernozhukov, Hansen, and Kozbur (2016), which allows for correlated errors within cross-sectional units.

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 statistically 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. Since the gravity context often assumes that errors are correlated over time within pair, we take this correlation into account as well in constructing these penalty weights.

A principal advantage of the plug-in lasso is that it is very rigorous 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).[6]

[6] 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 regresors is fixed and the number of observations is large. The "near-oracle" property of the plug-in

However, the plug-in method's rigor can also be a weakness. 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 the iceberg lasso method.

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 one to evaluate the predictive ability of the model. To adapt this idea to our setting, we validate our model by repeatedly dropping random groups of agreements in our data, and then predicting their effects on trade out of sample, 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 has been dropped.[7]

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.[8] 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

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.

[7] 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 $k = 25$.

[8] 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.

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.[9]

The Iceberg 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.

As we have noted, in the case of our dataset, 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 introduce what we call the "iceberg lasso". Simply put, it involves performing a subsequent set of plug-in lasso regressions in which each of the provisions selected by the plug-in PPML-lasso estimator is regressed on all of the provisions that were excluded. The purpose of these regressions is to identify bundles of provisions that are highly correlated with the selected ones 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 these additional lasso regressions 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.[10]

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 can be said to have causal effects. Whether or not this is 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

[9] Alternatively, we could consider the adaptive lasso (Zou, 2006), which adds a second tuning parameter and is known to deliver consistent variable selection. However, we have still found that the adaptive lasso is similar to the standard lasso in that it is much too lenient and it keeps too many regressors that are not relevant.

[10] Our approach complements the one 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, we select the provisions using a supervised method that also considers the impact of the provisions on trade, and then add another step which can be interpreted as unsupervised learning.

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 step in the iceberg lasso is 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.

3.4 Simulation evidence

In this section we report the results of a small simulation exercise investigating the finite-sample properties of the three methods we will use to identify the set of PTA provisions that are likely to have more impact on trade flows. 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 the experiments we use n observations on a set of p potential explanatory variables; we consider cases with sample size $n \in \{250, 1000, 4000\}$, 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. The p potential explanatory variables are obtained as random draws from the normal distribution; the first κ variables are correlated with each other, and the remaining ones are independent of all other variables. The covariance matrix of the first κ regressors is given by $U'U$, where U is a $\kappa \times \kappa$ matrix where each entry is a draw from the uniform distribution on the interval $(u, 1)$. All regressors have zero mean and variance 1 and we perform simulations with $\kappa \in \{5, 10, 20\}$ and $u \in \{0.0, 0.3, 0.6\}$.[11]

For all combinations of n, u and κ, the dependent variable is generated as

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

where x_1 is the first of the p potential explanatory variables described above, β and σ are parameters, and z and ε are independent random draws from the standard normal distribution. 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

[11] These values of u imply average correlations between the first κ variables of around 0.75, 0.91, and 0.98, respectively.

case all approaches have an excellent performance). When performing the selection of the relevant elements of the p potential explanatory variables, z is always included as a regressor whose coefficient is not penalized. Therefore, in this design, x_1 plays the role of the presumably small number of provisions that effectively affect trade and are correlated with others that do not, and z mimics the role of the fixed effects that explain a significant share of the variation of trade and are always included without penalty.

The selection of the relevant explanatory variables is performed using each of the three methods presented before: plug-in lasso, cross-validation lasso, and the proposed iceberg lasso, which uses the plug-in penalty in both steps. Additionally, we also perform the variable selection using the adaptive lasso of Zou (2006), with penalty chosen by cross validation.[12] 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.[13]

We repeat the simulations 1000 times, recording the number of times the variable x_1 is correctly selected as a regressor, and the total number of variables selected by each method. For each of the cases considered, Tables 3 and 4 present the percentage of times the regressor x_1 is selected and the average and standard error of the number of regressors selected by each method.

The results in Table 3 reveal a number of interesting patterns. For $n = 250$, lasso with the penalty chosen by the plug-in method (PI) is the method that most often fails to identify x_1 as a relevant regressor, and its performance deteriorates quickly as u increases. The adaptive-lasso (AL) performs better, but its performance is also very poor when u is high. Lasso with the penalty chosen by cross-validation (CV) provides a substantial improvement, but it also struggles for larger values of u. The iceberg lasso (IL) is marginally outperformed by CV when $u = 0.0$, but in the more challenging cases it can have a substantial advantage over all other methods.[14] The performance of all methods improves for the larger sample sizes, but the iceberg lasso maintains its advantage in the more challenging cases.

The results in Table 4 are equally interesting. In all cases considered, CV tends to lead to a high average number of selected regressors; this method also generally leads to high variability in the number of selected regressors. Remarkably, the average number of regressors picked by CV increases with n, and therefore with p, but is almost insensitive to κ. The average number of regressors selected by PI is always very small, and we do not see a clear pattern as n and κ vary. In contrast, the average number of variables selected by AL drops with the sample size and for $n = 4000$ it is always very close to 1, as we would expect from its oracle property. Finally, not surprisingly, the average number of variables selected by the IL increases with κ, and this is the feature that allows it to more frequently identifying x_1 as a relevant regressor.

[12] The adaptive lasso requires a set of initial estimates; we used those obtained by the cross-validation lasso.

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

[14] 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.

Table 3: Percentage of times x_1 is selected

		$u = 0.0$			$u = 0.3$			$u = 0.6$		
n		$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$
250	CV	95.49	96.89	97.70	82.16	82.87	79.80	56.41	47.80	37.70
	AL	93.69	95.09	95.60	76.35	75.15	71.10	47.39	37.78	28.60
	PI	85.37	82.76	81.90	67.23	60.62	54.50	41.38	32.06	21.80
	IL	94.09	93.99	92.00	90.18	86.77	83.10	79.66	71.64	60.00
1000	CV	99.60	100.00	100.00	97.20	98.10	99.00	82.20	78.90	72.10
	AL	98.90	99.90	100.00	93.20	95.20	96.40	71.50	66.40	57.80
	PI	98.50	98.60	99.00	92.60	94.30	93.10	73.30	67.50	58.40
	IL	100.00	100.00	99.70	99.60	99.30	98.90	96.30	93.10	89.60
4000	CV	100.00	100.00	100.00	99.80	100.00	100.00	96.20	98.30	98.60
	AL	99.80	100.00	100.00	98.50	99.50	100.00	86.60	90.70	91.80
	PI	99.80	100.00	100.00	99.10	99.90	100.00	94.50	95.20	94.50
	IL	100.00	100.00	100.00	100.00	100.00	100.00	99.80	100.00	99.30

Table 4: Average and standard error of the number of selected regressors

		$u = 0.0$			$u = 0.3$			$u = 0.6$		
n		$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$
250	CV	8.51 (7.69)	9.08 (7.74)	8.76 (7.62)	8.46 (7.43)	9.14 (7.60)	8.64 (7.05)	8.40 (7.42)	8.79 (7.54)	8.16 (7.07)
	AL	7.32 (7.10)	7.59 (6.94)	7.38 (6.68)	7.32 (7.01)	7.65 (7.00)	7.22 (6.55)	7.00 (6.84)	7.20 (6.80)	6.71 (6.45)
	PI	1.33 (0.66)	1.59 (0.84)	1.85 (1.08)	1.40 (0.72)	1.68 (0.86)	1.98 (1.10)	1.26 (0.60)	1.43 (0.73)	1.52 (0.80)
	IL	5.12 (5.32)	5.98 (2.29)	9.95 (4.19)	5.73 (7.18)	6.20 (2.31)	10.67 (4.22)	5.09 (8.02)	5.72 (2.19)	9.16 (3.70)
1000	CV	9.63 (9.31)	9.90 (9.25)	10.11 (10.16)	9.94 (9.39)	10.34 (9.34)	10.94 (10.32)	10.02 (9.36)	10.32 (9.26)	10.66 (9.52)
	AL	4.35 (8.13)	4.91 (9.32)	4.86 (8.86)	5.02 (8.60)	5.89 (9.63)	6.37 (9.70)	5.20 (8.63)	6.40 (9.98)	6.70 (9.61)
	PI	1.41 (0.61)	1.62 (0.81)	1.99 (1.16)	1.69 (0.71)	2.08 (0.99)	2.68 (1.37)	1.69 (0.67)	2.08 (0.88)	2.54 (1.13)
	IL	5.37 (5.27)	7.16 (1.98)	11.86 (4.11)	5.98 (7.20)	7.22 (2.04)	12.99 (4.03)	6.55 (11.25)	7.23 (1.91)	12.94 (3.66)
4000	CV	10.48 (11.52)	10.34 (11.13)	10.77 (11.15)	10.91 (11.57)	10.71 (11.06)	11.52 (11.59)	11.11 (11.54)	11.06 (11.12)	12.09 (11.57)
	AL	1.00 (0.04)	1.04 (1.10)	1.00 (0.00)	1.03 (0.50)	1.14 (1.98)	1.15 (2.27)	1.16 (1.84)	1.31 (2.76)	2.04 (6.01)
	PI	1.35 (0.57)	1.52 (0.74)	1.79 (1.03)	1.68 (0.74)	2.05 (1.01)	2.49 (1.30)	1.93 (0.80)	2.40 (1.10)	3.02 (1.30)
	IL	5.19 (5.65)	8.36 (1.70)	15.39 (3.44)	6.00 (11.38)	8.04 (1.89)	14.16 (3.88)	6.51 (10.51)	7.55 (1.96)	14.13 (3.54)

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 favour 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 iceberg lasso is a safe bet and is the best method. This is particularly the case if the relevant variable is highly correlated with other potential controls because in that case the iceberg lasso outperforms the adaptive lasso even for the larger samples considered in our experiments.

These results, 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 283 trade agreements,[15] 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 emerges as providing a good compromise between parsimony and the ability to identify the relevant variables. It consequently is our preferred approach.

4 Lasso Results

In this section, we present our lasso results obtained using the methods described in the previous section. We first present results for the plug-in method before briefly discussing the results obtained using cross-validation. We then turn to the iceberg lasso results, which themselves are based on provisions selected by the plug-in method.

4.1 Plug-in Lasso Results

Table 5 presents results for the plug-in lasso and post-lasso regressions discussed before. In column 1, we start by presenting the results of a traditional PPML estimation with a dummy for the presence of a preferential trade agreement between the trading partners. This shows that we can replicate the usual finding that PTAs lead to a significant increase in trade flows in our data. Specifically, we find that the PTAs in our data increase trade by $\exp(0.130) - 1 = 13.8\%$. Column (2) then shows the results of our first-step lasso regression, showing only the coefficients that the lasso finds to be non-zero. In a subsequent step, we then estimate a “post-lasso” PPML regression—a standard PPML regression using only the provisions that were selected by the lasso in the first step.

[15] 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.

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				PPML
	PPML	Lasso	Post-lasso	PPML	PPML	Lasso	Post-lasso
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PTA	0.130***			−0.030	0.083**		
	(0.038)			(0.054)	(0.038)		
EU					0.688***	0.416	0.589***
					(0.065)		(0.084)
AD14. Anti-dumping –Material Injury		0.172	0.313***	0.303***		0.188	0.343***
			(0.114)	(0.116)			(0.105)
CP23. Competition Policy –Transparency / Coordination		0.031	0.075	0.078		0.011	0.046
			(0.056)	(0.056)			(0.054)
SUB12. Subsidies –Discipline (general)		0.008	0.099*	0.108**			
			(0.052)	(0.055)			
TBT provisions:							
TBT2. Mutual Recognition		0.084	0.073	0.068			
			(0.093)	(0.094)			
TBT7. Technical Reg's: use International Standards		0.034	0.111	0.121		0.055	0.106
			(0.080)	(0.082)			(0.077)
TBT33. Standards: use Regional Standards		0.067	0.046	0.050		0.039	0.039
			(0.066)	(0.066)			(0.051)
Trade Facilitation provisions:							
TF41. Harmonization and Common Legal Framework						0.038	0.550***
							(0.126)
TF42. Customs and Other Duties Collection		0.227	0.354***	0.352***			
			(0.121)	(0.121)			
TF45. Issuance of Proof of Origin		0.022	0.076***	0.096**		0.016	0.079***
			(0.029)	(0.043)			(0.028)

Notes: Gravity estimates are obtained using PPML with exporter-time, importer-time, and exporter-importer FEs. The number of observations is 194,092. Columns labelled “PPML Post-lasso” report PPML coefficients for all variables selected by a plugin lasso method in a prior step. The difference between columns 2-3 and 6-7 is that the latter includes the EU dummy in the lasso step as a possible predictor to be selected. All other columns report further experiments using PPML. PPML cluster-robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < .05$, *** $p < .01$.

Using the plug-in approach, the lasso selects a small number of trade agreement provisions related to anti-dumping, competition policy, domestic subsidies, 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, competition policy, and subsidy provisions all create more certainty as to how disciplinary investigations and proceedings will be carried out in these various policy areas. 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.

The corresponding post-lasso PPML results, shown in column (3), finds 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 36.8% ($\exp(0.313) - 1 = 0.368$). Even larger effects are found for having trade facilitation provisions that regulate customs and other duties collection (TF42), which has an estimated effect of 42.5% ($\exp(0.354) - 1 = 0.425$). 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 implications for model fit when a variable is not included, 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 8 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 estimated in column (1).

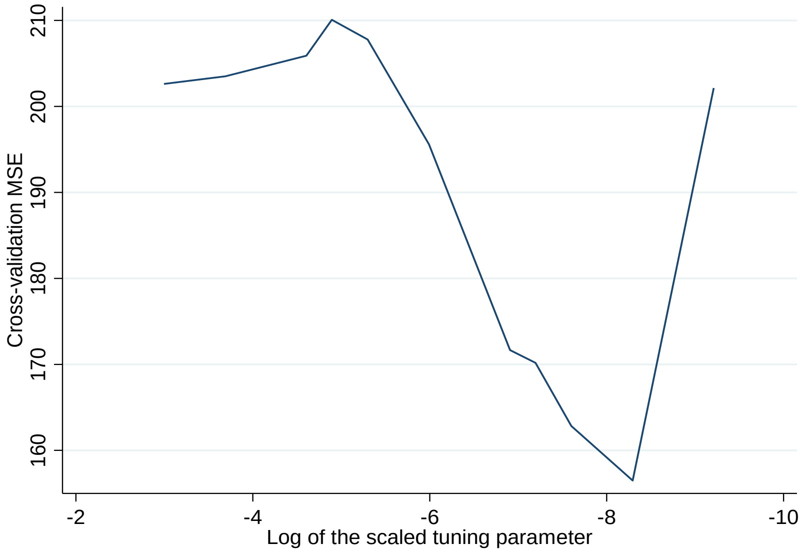
Next, column (5) returns to our original simple model from column (1) but adds a second dummy variable for the EU agreement. Our reasons for treating the EU separately from other agreements are three-fold. First, we suspect that not all of the EU's efforts to promote trade are captured in how their provisions 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 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 EU 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, the more important exercise is in column (6), where we now treat the EU as a possible predictor in the lasso. Because the EU is indeed selected as being an important predictor of changes in trade flows, the value of this exercise is that the selection of other predictors is solely based on information from other agreements aside from the EU. Consequently, the set of provision variables selected by the lasso is now slightly different than in column (2), adding TF41 (which calls for harmonization of customs procedures) but losing TBT2, SUB12, and TF42. Notably, the post-lasso estimates in column (7) find TF41 to be highly significant both statistically and economically, with an estimated effect of $\exp(0.550) - 1 = 73.3\%$. Given the possible issues with the EU we have outlined, this last set of provision variables is our preferred set to work with in the subsequent iceberg lasso analysis.

4.2 Cross-Validation Lasso Results

As discussed previously, the plug-in approach to choosing λ is conservative, in the sense that it 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. 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. The out-of-sample MSE initially decreases as λ is increased and then increases again, with a minimum reached at $\lambda / \sum_{ijt} y_{ijt} = 0.00025$.

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



For more illustration, Figures 3 and 4 show the corresponding regularization paths for selected provisions. 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. At the optimal value of $\lambda/\sum_{ijt} y_{ijt} = 0.00025$, our cross-validation approach selects 124 provisions to have non-zero effects, which is many more than what we found using our plug-in approach.[16]

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 our plug-in approach, is only selected for higher values for λ. Intuitively, as we lower λ, more provisions are selected and some of these are correlated with provision AD14. This then implies that adding AD14 itself does not lead to significant improvements in 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 that AD14 on its own gains predictive power and is included. That said, for higher values of λ, we generally see a close correspondence between the results along the regularization path indicated in Figures 3 and 4 and those that we found earlier using the plug-in method.

Overall, Figures 3 and 4 show that our two approaches to selecting λ lead to very different sets of trade agreement provisions being selected. While some provision, such as CP23 or SUB12 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 provision AD05. Furthermore, we also see in Figures 3 and 4 that many of the estimated effects for the provisions that are selected are too large in absolute magnitude to be 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.

4.3 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 last 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. 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 instead relying on cross-validation. Furthermore, as we have seen from our simulations, it should also give us more confidence in the results.

[16] In each panel of the figure, the second-to-last set of estimates corresponds to the 124 variables selected by the cross-validation method.

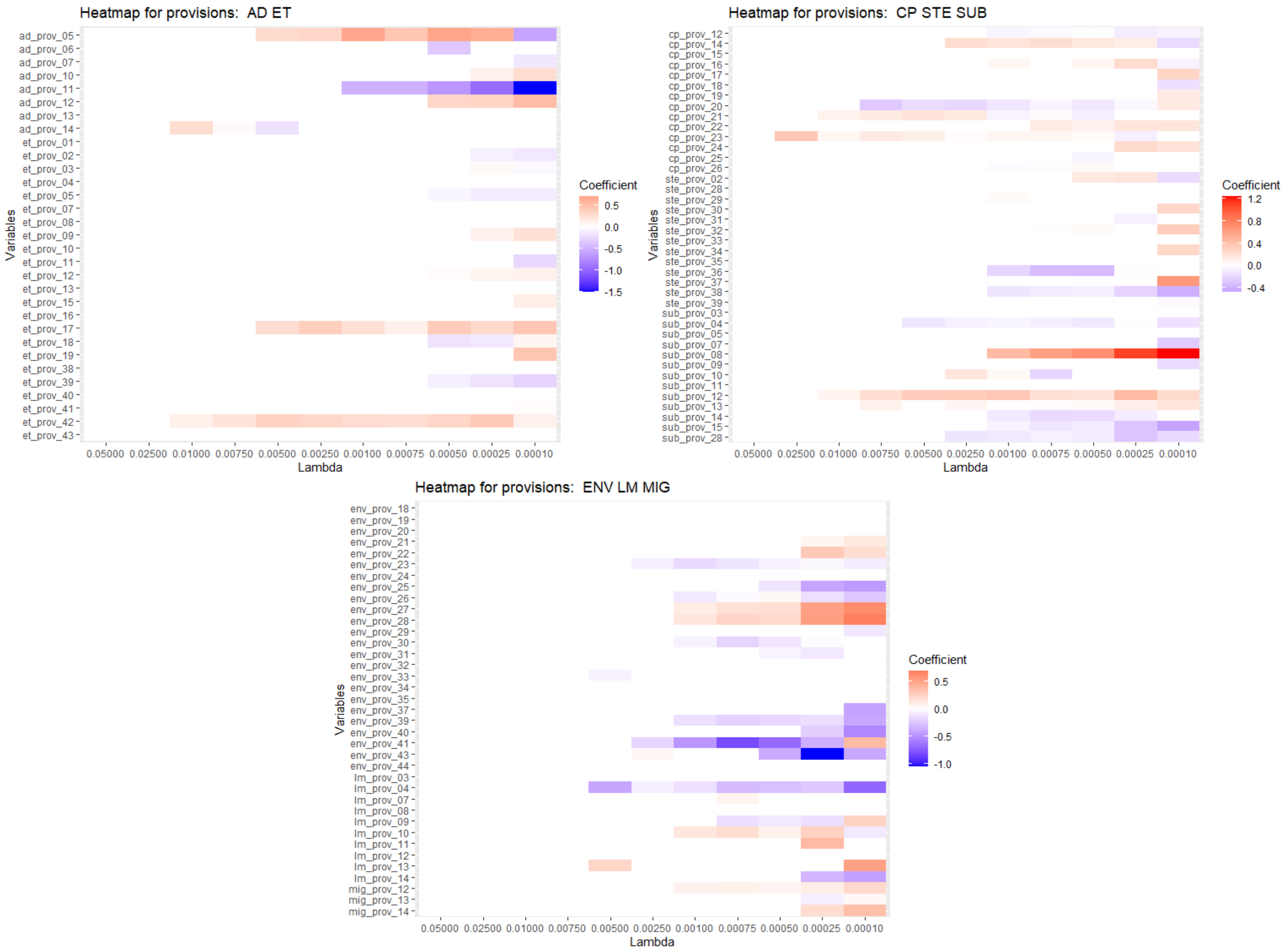


Figure 3: Regularization path for selected provisions (AD, ET, CM, STE, SUB, ENV, LM, and MIG)

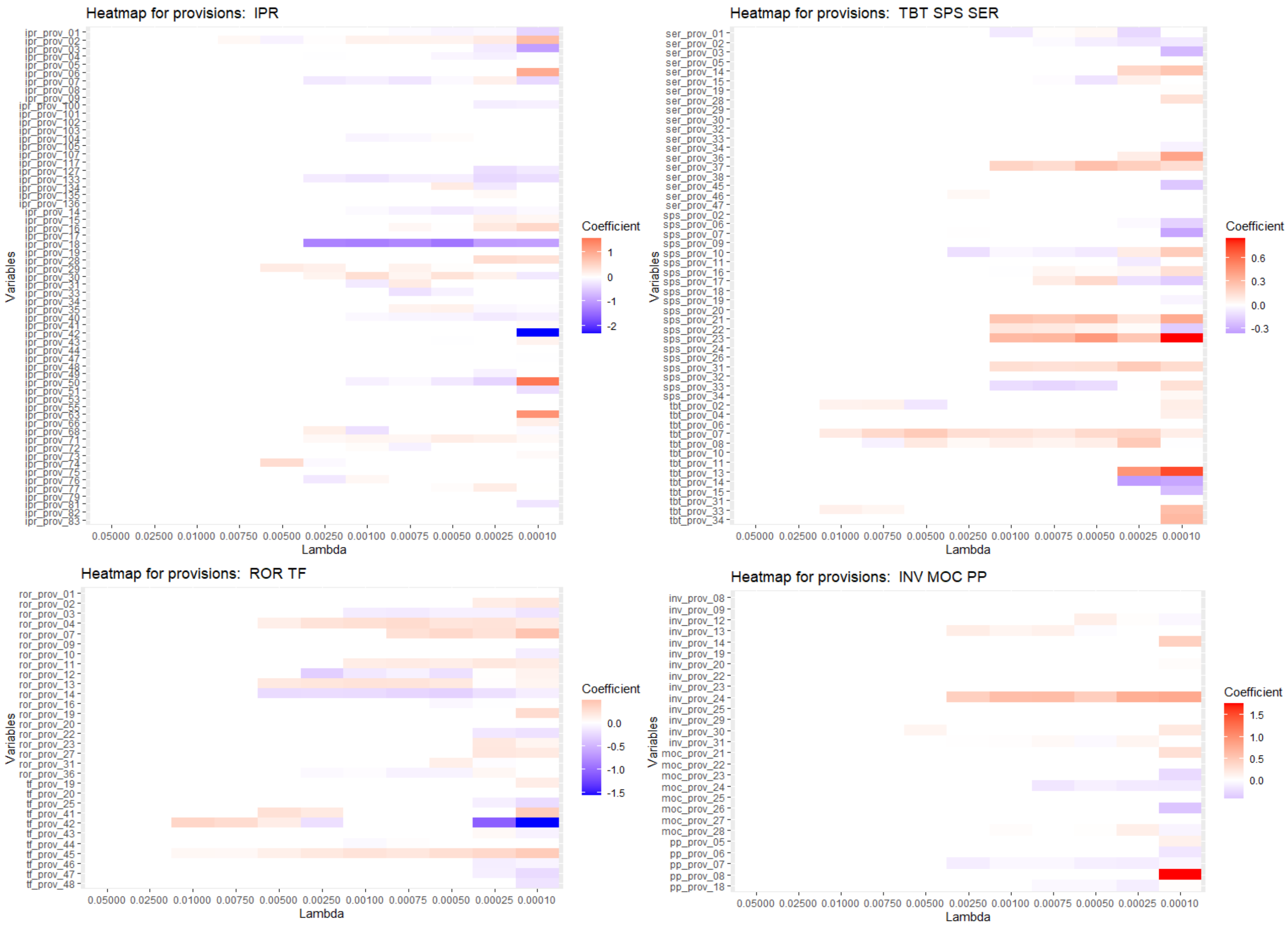


Figure 4: Regularization path for selected provisions (IPR, TBT, SPS, SER, ROR, TF, INV,MOC, and PP)

Table 6 presents the results of our iceberg lasso analysis. The first two rows of Table 6 list each of the six provisions selected by the first-stage plug-in lasso when the EU dummy is included, as well as their estimated impact on trade flows from column (6) 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 antidumping provision AD14 is highly correlated with two further antidumping provisions (AD06 and AD08) as well as with one provision on environmental protection (ENV42);[17] the R^2 of the regression of AD14 on these three provisions is 0.95.

Table 6: Iceberg lasso results

(1)	(2)	(3)	(4)	(5)	(6)
AD14	CP23	TBT07	TBT33	TF41	TF45
(+41%)	(+4.7%)	(+11.2%)	(+4%)	(+73.3%)	(+8.2%)
AD06 (0.97)	AD06 (0.46)	AD06 (0.54)	AD06 (0.48)	AD05 (0.89)	AD11 (0.09)
AD08 (0.97)	AD08 (0.46)	AD08 (0.54)	AD08 (0.48)		CP15 (0.73)
ENV42 (0.97)	CP22 (0.78)	ENV42 (0.54)	AD12 (-0.11)		ET03 (0.51)
	CP24 (0.89)	ENV44 (0.06)	ENV42 (0.48)		SUB10 (0.25)
	ENV42 (0.46)	SPS21 (0.23)	ENV44 (-0.01)		SUB11 (0.28)
	ET41 (0.16)	SUB07 (0.08)	INV24 (0.11)		TF44 (0.98)
	IPR42 (-0.00)	TBT15 (0.73)	IPR71 (-0.08)		
	IPR55 (-0.01)	TBT34 (0.94)	IPR103 (-0.11)		
	IPR63 (-0.00)		IPR107 (-0.12)		
	IPR74 (-0.01)		MOC26 (-0.10)		
	PP08 (0.08)		SPS21 (0.19)		
	SPS21 (0.17)		SUB04 (-0.11)		
	STE31 (0.57)		SUB07 (0.07)		
	TBT02 (0.56)		TBT05 (0.61)		
	TBT15 (0.37)		TBT06 (0.98)		
	TBT29 (0.56)		TBT15 (0.69)		
	TF42 (0.56)		TBT32 (0.61)		
	TF44 (0.38)		TBT34 (0.53)		
0.95	0.83	0.89	0.97	0.80	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 3-20; 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.

The results in Table 6 show that the iceberg-lasso identifies 43 provisions that are likely to be associated with increased trade. This finding contrasts with the 124 provisions identified by the cross-validation lasso, and the 6 provisions selected by the plug-in lasso. Therefore, as in the simulations in the preceding section, the iceberg lasso appears

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

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.

As noted above, we find that provision AD14 is correlated other antidumping provisions; this correlation is not surprising because all these provisions fulfill a similar purpose, which is to increase transparency in the use of antidumping duties. In that sense, one conclusion to be drawn from this exercise is that antidumping 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 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 antidumping and the environment, hence leading to a high correlation between the corresponding provisions in our data.

Template effects may also be important for understanding the variables highly correlated with the selected TBT provisions, TBT07 and TBT33. Indeed, some of the same anti-dumping and environmental provisions that were found to be correlated with AD14 show up here as well (AD6, AD8, ENV42). That said, the strongest correlations in these cases are with other TBT provision such as TBT06, TBT15 and TBT34. This is not surprising as these provisions also relate to the use of international standards. Thus, it seems likely 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 provisions TBT07 and TBT33, although we cannot say which of the individual TBT provisions is driving the observed effect.

The lasso also selects two provisions that reduce the administrative burden resulting from compliance with rules of origin and other customs procedures (TF41 and TF45), which are estimated to have a very large trade increasing effect (over 70% for TF41). Table 6 also indicates that other trade facilitation provisions are correlated with some of the provisions selected by the lasso; this is true both for TF45 and CP23. Thus, our results suggest that trade facilitation procedures are likely to be associated with significant trade flow increases.

Finally, we find that provision CP23, which serves to promote transparency in competition policy, is correlated with some of the previously identified types of provisions, as well as with two further provisions on competition policy (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.

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 intellectual property rights and sanitary and phytosanitary measures are related both to CP23 and 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 CP23 and 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 bundle of provisions that are likely to have the desired effect. In particular, provisions related to TBTs, antidumping, trade facilitation, and competition policy are likely to enhance the trade-increasing effect of trade agreements.

5 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 which of the many provisions in our data set should be treated as relevant for affecting trade flows. Using our preferred method, a two-step "iceberg lasso" approach, we identify a relatively parsimonious set of 43 provisions that are most likely to impact trade. While these 43 provisions span a range of policy areas, our results generally support the conclusion that a select number of provisions related to anti-dumping, competition policy, technical barriers to trade, and trade facilitation are most effective at promoting trade as compared to other types of provisions that appear in PTAs.

We need to be clear that interpreting these results requires some important caveats. We know that it is possible that our preferred method may fail to discover important trade-promoting provisions, and it is almost certain to lead to the inclusion of provisions that are not relevant. At present, we are not able to quantify either type of uncertainty. Developing metrics that can be used to guide researcher confidence represents an important avenue for future research.

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 method we have introduced could be a helpful tool for any researcher 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 variable selection and other related machine learning approaches 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	
AD05	Export price less than comparable price when destined for consumption in the exporting country
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
AD12	The consequent impact of dumped imports on the domestic industry
AD14	Requirement to establish material injury to domestic producers
Competition Policy	
CP15	Does the agreement prohibit/regulate cartels/concerted practices?
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	
ENV42	Does the agreement require states to comply with the UN Conference on Environment and Development?
ENV44	Does the agreement require states to comply with the International Energy Program?
Export Taxes	
ET03	Prohibits new export quotas/quantitative restrictions between the parties
ET41	Prohibits non-tariff measures related to export of goods
Investment	
INV24	Does the FET clause prohibit arbitrary, unreasonable or discriminatory measures?
Intellectual Property Rights	
IPR42	Prohibits requiring the recording of a trade mark license to establish license validity or as a condition for use
IPR55	Requires patent be made available for new processes of a known product
IPR63	Requires a period of sui generis protection for patents
IPR71	Requires system for protection of industrial designs
IPR74	Seek to improve industrial design systems
IPR103	Stipulates practices to be followed by collective management organizations
IPR107	Patent Law Treaty (2000)

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

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	
SPS21	B. Risk Assessment: Is there reference to international standards/procedures?
State-Owned Enterprises	
STE31	Does the agreement prohibit anti-competitive behavior of state enterprises?
Subsidies	
SUB04	Does the agreement prohibit or regulate local-content subsidies?
SUB07	Does the agreement introduce any ceiling to permitted subsidies?
SUB10	Does the agreement include any specific regulation of fisheries subsidies?
SUB11	Does the agreement include any specific discipline for public services?
SUB12	Does the agreement include any other specific discipline for certain sectors or objectives?
Technical Barriers to Trade	
TBT02	B. Technical Regulations - Is mutual recognition in force?
TBT05	B. Technical Regulations - Are there specified existing standards to which countries shall harmonize?
TBT06	B. Technical Regulations - Is the use or creation of regional standards promoted?
TBT07	B. Technical Regulations - Is the use of international standards promoted?
TBT15	C. Conformity Assessment - Is the use of international standards promoted?
TBT29	A. Standards - Is mutual recognition in force?
TBT32	A. Standards - Are there specified existing standards to which countries shall harmonize?
TBT33	A. Standards - Is the use or creation of regional standards promoted?
TBT34	A. Standards - Is the use of international standards promoted?
Trade Facilitation and Customs	
TF41	Does the agreement require customs harmonization and a common legal framework?
TF42	Does the agreement regulate customs and other duties collection?
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

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

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

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

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 $\widehat{\alpha}_{it}(\widehat{\beta})$, $\widehat{\gamma}_{it}(\widehat{\beta})$, and $\widehat{\eta}_{it}(\widehat{\beta})$ as functions of $\widehat{\beta}$ that are implicitly defined by their FOCs. The resulting "concentrated" minimization problem is

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

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:

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

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

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

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

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

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, set

$$\widehat{\phi}_k^2 = \frac{1}{n}\sum_{i,j}\left(\sum_t \widetilde{x}_{ijt,k}\widehat{\epsilon}_{ijt}\right)^2 = \frac{1}{n}\sum_{i,j}\sum_t\sum_{t'}\widetilde{x}_{ijt,k}\widetilde{x}_{ijt',k}\widehat{\epsilon}_{ijt}\widehat{\epsilon}_{ijt'},$$

where $\widehat{\epsilon}_{ijt} = y_{ijt} - \exp(x'_{ijt}\widehat{\beta} + \widehat{\alpha}_{it} + \widehat{\gamma}_{jt} + \widehat{\eta}_{ij})$, but can also be obtained as $\widehat{\epsilon}_{ijt} = \widehat{\mu}_{ijt}(\widetilde{z}_{ijt} - \widetilde{x}'_{ijt}\widehat{\beta})$. By inspection, this expression provides an estimate of the variance of the score for β_k under the assumption that errors are correlated over time within the same pair, as is commonly assumed in this context. Provided there is weak temporal dependence (in the sense described by Hansen, 2007), $\widehat{\phi}_k^2 - \phi_k^2 = o_p(1)$ 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 $\widehat{\beta}_{PL}$,

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

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

In practice, the plug-in lasso method only 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 $\widehat{\epsilon}_{ijt}$ in each iteration to construct new values for $\widehat{\phi}_k$.

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 j. Define the CV bandwidth as

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

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, 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 (2014). "Inference on treatment effects after selection among high-dimensional controls," *Review of Economics Studies,* 81, 608-650.

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.

Fu, W. and K. Knight (2000). "Asymptotics for lasso-type estimators," *Annals of Statistics*, 28, 1356-1378.

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.

Hansen, C. (2007). "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics,* 141, 597-620.

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," *Handbook of International Economics*, Vol. 4: 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.

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.

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.

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.

Weidner, M., T. Zylkin (2020). “Bias and consistency in three-way gravity models,” arXiv:1909.01327.

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.