Determining which trade agreement provisions matter for trade

Online Appendix – Not for publication

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

24 May 2026

This Appendix is divided into 5 parts:

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

Appendix A: More details on provision variables

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

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

Table A1: Main Provisions

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

Table A1 (continued): Main Provisions

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

Table A1 (continued): Main Provisions

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

Appendix B: More details on implementation

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

Details on HDFE-PPML-lasso estimation

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

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

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

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

$$\widehat{\alpha}_{it} : \frac{1}{n}\sum_{j} y_{ijt} - \widehat{\mu}_{ijt} = 0, \qquad \forall i,t,$$

$$\widehat{\gamma}_{jt} : \frac{1}{n}\sum_{i} y_{ijt} - \widehat{\mu}_{ijt} = 0, \qquad \forall j,t,$$

$$\widehat{\eta}_{ij} : \frac{1}{n}\sum_{t} y_{ijt} - \widehat{\mu}_{ijt} = 0, \qquad \forall i,j,$$

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

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

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

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

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

where

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

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

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

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

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

where

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

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

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

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

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

Details on cross-validation

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

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

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

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

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

Details on plug-in lasso

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

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

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

Intuitively,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Bootstrap lasso and iceberg lasso

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

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

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

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

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

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

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

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

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

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

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

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

A simple example

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof of Theorem 1: Let

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

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

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

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

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

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

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

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

so that

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

and by Lemma 1,

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

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

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

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

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

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

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

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

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

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

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

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

Hence, since the arg min is a continuous operation,

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

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

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

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

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

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

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

which establishes (22).

Note that, (22) implies that

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

and

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

This completes the proof of (20).

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

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

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

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

where $\phi^*_k=\lim_{r\to\infty}\phi^{*(r-1)}_k$. Note that minimizing $Q^*_n(\beta)$ with respect to β results in $\beta^*_n=\lim_{r\to\infty}\beta^{*(r)}_n$.

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

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

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

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

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

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

These expressions can then be combined as

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

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

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

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

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

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

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

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

To extend this to the full coefficient vector, note that

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

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

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

Appendix D: Further empirical results

Additional boostrap-lasso results

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

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

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

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

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

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

Cross-validation lasso results

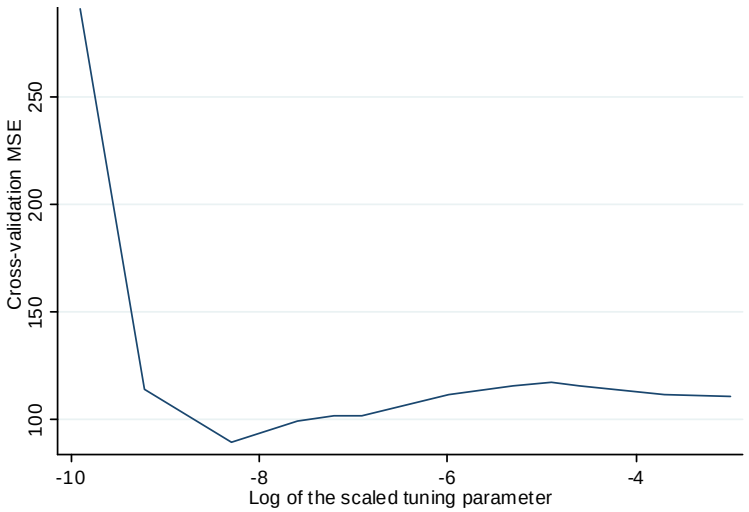
As discussed in the paper, the plug-in approach to choosing penalty parameter tends to choose a relatively small set of regressors and may fail to pick the "correct" regressors. For comparison, we now discuss the choice of regressors when we apply the cross-validation approach to our data.[9]

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

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

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

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



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

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

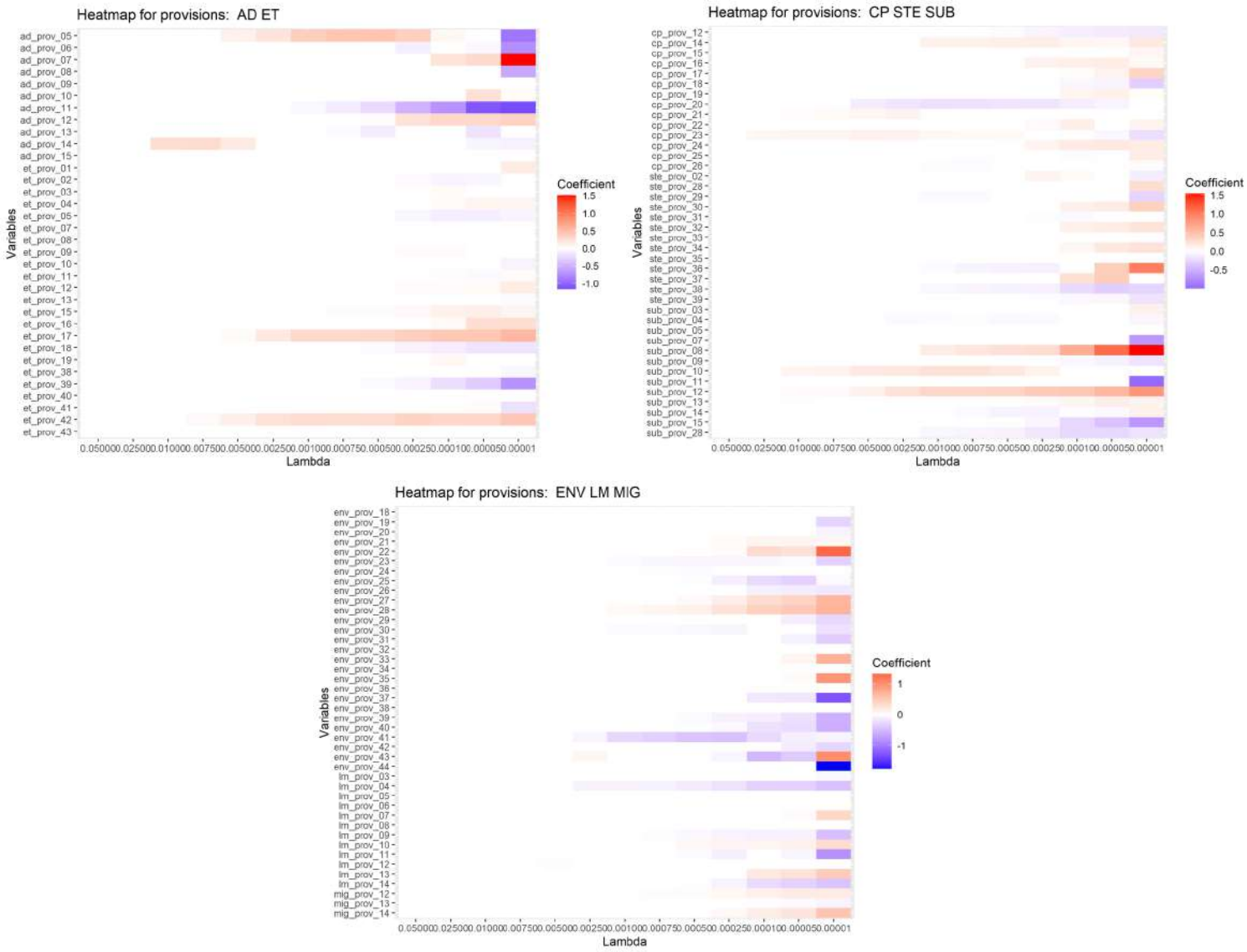


Figure A2: Regularization path for selected provisions (AD, ET, CP, STE, SUB, ENV, LM, and MIG)

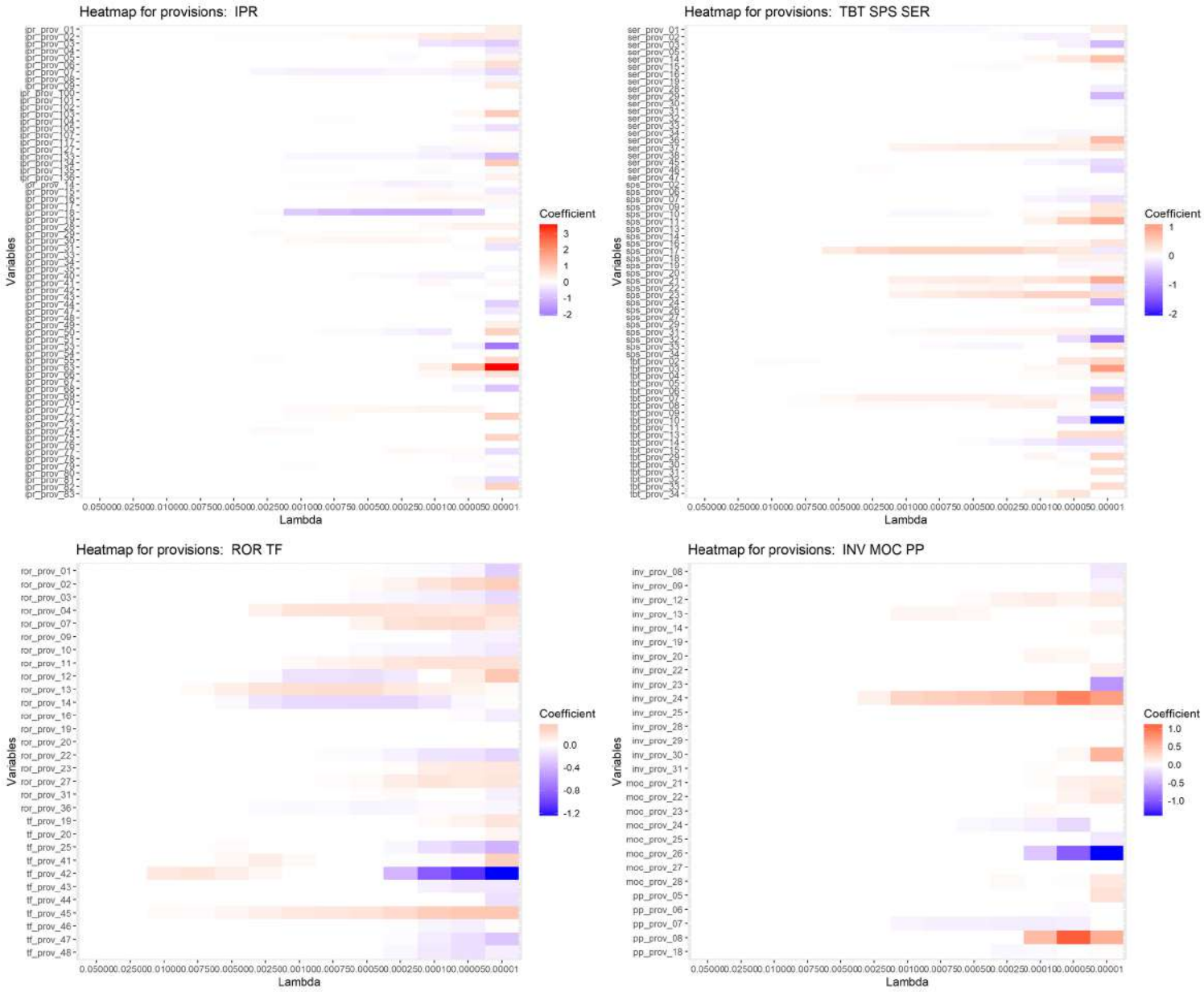


Figure A3: Regularization path for selected provisions (IPR, TBT, SPS, SER, ROR, TF, INV, MOC, and PP)

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

Appendix E: Prediction

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

Simulation results

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

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

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

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

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

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

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

Table A4: MSE for prediction errors

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

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

Empirical results

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

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

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

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

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

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

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

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

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

Table A5: Summarizing Estimates of Heterogeneous PTA Effects

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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