Firm Level Pass Through: A Machine Learning Approach

Lu Han
University of Cambridge

August 3, 2018

Abstract

Understanding how exporters react to exchange rate shocks is important for evaluating international shock transmissions and setting optimal international monetary policy. Empirical studies have documented substantial heterogeneity in the degree to which different firms and products respond to exchange rate shocks. In addition, estimates of exchange rate pass through (ERPT) are time varying and depend on observed and unobserved variables in a nonlinear way. This paper proposes a machine learning algorithm that systematically detects the determinants of ERPT and estimates ERPT at the firm level in a large-scale custom dataset. The accuracy of the algorithm is tested on simulated data from an extended multi-country version of Atkeson and Burstein (2008). Applying the algorithm to China's custom data from 2000-2006, this paper estimates the ERPT of China's exporters and documents new evidence on the nonlinear relationships among market structures, unit value volatility and ERPT.

JEL classification: C50, E31, F41, F42
Keywords: firm-level, pass through, machine learning, exchange rates.

In the last decade, the increasing availability of large scale firm level datasets has greatly enlarged the ability to understand firm level heterogeneity and its implications at the aggregate level. Especially in the literature of ERPT, understanding why firms have different pricing behaviour in response to a common exchange rate shock has important implications in setting the optimal monetary policy[1].

Unlike micro studies in other fields, international trade firm-level datasets recently made available contain a significant proportion of firms in an economy and almost all custom transactions at firm product (8-digit) level in a given period. As the richness and the scale of micro dataset available to researchers develop rapidly, conventional methods applied extensively by empirical researchers, fixed effects related methods for example, are either less flexible in their functional assumptions or not very effective in gathering all possible aspects of heterogeneity[2] in a large scale dataset. Therefore, these methods may not be the best option to understand firm-level heterogeneities. Conventional ERPT estimation methods generate large standard errors when applied at sector/firm level due to unobserved variables, e.g. marginal cost, heterogeneity in product characteristics and market structures. Empirically, researchers trade off controlling for unobserved variables against the flexibility of functional forms[3].

On the other hand, recent researches in machine learning focus especially on large datasets and heterogeneities. It seems to be the natural alternative for trade problems. In spite of successful applications of these algorithms in various subjects, economists often stay alarmed with the usage of these algorithms for two reasons. First, a machine learning algorithm often involves extracting maximum amount of information in a certain dataset and thus the results are often data driven and may not necessarily reflect the true relationship between variables being studied. Second, a machine learning algorithm may be good in making predictions but does not identify causal relationships, nor does it enhance our economic understanding.

For causality, the mainstream empirical papers working with firm-level data take a restrictive approach. Methods include restricting the dataset so that the subgroup being studied is no longer subject to omitted variable bias or adding multiple fixed effects such that "irrelevant" and possible confounding variations can be partitioned out. However, adding restrictions and layers means dropping observations and losing information[4]. These restrictions can help us build an "ideal" environment to study the hypothetical relationship but also limit our vision to a particular hypothetical situation.

Thanks to persistent advocates of applying machine learning methods to Economics[5], pioneer works on adapting machine learning methods to make casual inferences and solve policy problems have made a significant progress[6]. However, existing studies work under the condition of unconfoundedness. This

[1]Seminal contributions include Dornbusch (1987), Corsetti and Pesenti (2005), Corsetti and Dedola (2005), Corsetti, Dedola and Leduc (2007), Corsetti, Dedola and Leduc (2008*a*), Corsetti, Dedola and Leduc (2008*b*), Atkeson and Burstein (2008).

[2]The ability to search for all possible aspects of heterogeneities is important for predicting the firm-level response to policy shocks.

[3]For example, by dividing data into several bins according to destinations, quantiles of market shares, etc.

[4]Not a problem if obtaining one (aggregate) coefficient is the main concern but costly in understanding firm-level heterogeneities.

[5]See Varian (2014) for an introduction of various machine learning algorithms and how they can be applied to study economic questions.

[6]See Athey and Imbens (2015), Bajari et al. (2015), Chernozhukov, Hansen and Spindler (2015), Kleinberg et al. (2015). Pioneer works on adapting machine learning methods to make casual inferences include Chernozhukov et al. (2016), Athey, Tibshirani and Wager (2016), Athey and Imbens (2017), Athey et al. (2017) and Wager and Athey (2017).

condition will not be satisfied for international trade related problems. The marginal cost of the product being sold and the prices of competitors are unobserved and endogenous to exchange rate movements. Estimates ignoring these confounding variables will lead to biased point estimates of individual treatment effects.

This paper follows recent work on making casual inferences and proposes an algorithm specifically designed to estimate firm-level heterogeneities in response to macro shocks in a multi-dimensional panel. The proposed algorithm features in two aspects: (a) it uses the high predicting power of the gradient boosting regression tree algorithm (GBRT)[7] to construct counterfactual environments; (b) it uses orthogonal variations across dimensions to control for unobservables. The proposed algorithm contributes to the machine learning literature in its awareness that the monotonic property of tree based algorithms can be used together with orthogonal variations across dimensions to control for unobserved components[8].

The central idea behind the proposed algorithm is that machine learning algorithms making causal inferences should be assisted with structural information implied by economic models. Most machine learning algorithms take an agnostic data driven approach. As in most estimation techniques, adding correct structural assumptions will increase the precision of estimation. However, how to add economic assumptions into a machine learning algorithm is still not clear[9]. This proposed algorithm presents a novel approach to feed structural information into a tree based machine learning algorithm in a multi-dimensional panel framework[10].

The proposed algorithm is designed to work directly with large scale custom datasets and identify the ERPT parameter for each exporter in an economy. The algorithm learns from reading records of trade patterns. It not only predicts export prices and quantities at the firm level conditioning on values of future environments of aggregate variables but also generates the genetic rules governing the data generating process[11]. To assess the performance of the proposed algorithm, I build the following multi-country trade model.

In order to understand how firms optimally price their products under a multi-sector multi-country environment, I extend the two-country model of Atkeson and Burstein (2008) to a multi-country framework and introduce heterogeneity in productivity distributions across sectors and countries. The main features of the model are as follows. First, there are N countries in the world and each country owns S

[7] GBRT is featured in its power of out-of-sample prediction accuracy due to its ability to capture nonlinearities. GBRT algorithm has been widely applied in frontier studies of a wide range of topics, e.g. global distribution and the risk of dengue [Bhatt et al. (2013)], effect of climate change [Cox et al. (2013), Randall and Van Woesik (2015)]. The algorithm is proved to be effective in solving practical classification and prediction problems and has been actively implemented in international computing and machine learning challenges [For details, click List of Winning Solutions]. An introduction to the GRBT algorithm can be found in the appendix.

[8] Most custom datasets have disaggregated and detailed firm level transaction records, but key elements such as the marginal costs are not observed and are difficult to be estimated even when one can complement the custom database with some industry surveys.

[9] In general, the range of structural assumptions is not limited to probabilistic assumptions and could include information on how variables interact in a structural model. This paper only works on a limited case and shows that adding the log-linearised version of the structural pricing equation will significantly improve firm level ERPT estimates of the proposed algorithm.

[10] As I will illustrate in the model section, adding structural information has a marginal effect on the predictive power of the dependent variable but is critical in getting the correct causal inference.

[11] For example, how ERPT depends on observed firm level characteristics.

sectors with heterogeneous productivity distributions. Second, within each sector there are local firms as well as exporters from other countries competing under Cournot competition with a demand elasticity structure similar to Atkeson and Burstein (2008). The result of the competition is determined by the productivity distribution of the participating firms as well as aggregate variables such as bilateral exchange rate shocks. Due to the Cournot competition structure, there is no closed form solution for the model[12]. I construct counterfactual environments to understand how ERPT differs under different productivity distributions of local firms and foreign exporters and different compositions of exchange rate shocks.

In terms of macroeconomic modeling, although such a framework may be helpful in understanding the pricing behavior of an exporter facing competition from local competitors and other exporters from other countries, the need to add extra levels of heterogeneity seems to be less justified. The main drawback of a micro-founded multi-sector multi-country trade model is the lack of ability to map into the real world due to its demanding requirement in calibration. In practice, estimating the productivity distribution for a particular country/sector, provided the existence of good data, is already a challenge[13].

However, such highly detailed micro-founded models may have much to offer in an alternative modeling strategy. The procedure is given as follows. First, notice what data are available in reality. Second, simulate the highly micro founded model with arbitrary calibration. Third, subset the simulated data and construct a dataset similar to what is observed in the reality. Fourth, write an algorithm or econometric method to estimate key parameters of interest[14] for each individual/firm of interest in the constructed dataset. Fifth, change the parameter value of the model, re-simulate and construct a new dataset. Test the algorithm's ability to estimate the key parameter of interest and revise the algorithm if not. Sixth, apply the algorithm to observed data. If the model is believed to be correct, the estimates from the algorithm are reliable.

The advantage of this approach is that we will have a structural model that enables us to understand how the mechanism works and figure out the key variable of interest in reality provided that the model is correctly specified[15]. This approach is useful in a scenario where we do not have full information to estimate the whole model but may have enough information to identify a subset of parameters implied by the structural model. The proposed approach is similar to the indirect inference approach but differs in that my proposed approach, particularly steps 4 and 5, puts emphasis on building an algorithm that provides the correct estimates of interest for all possible calibrations of the micro-founded model. The indirect inference approach emphasises on finding an auxiliary model such that estimates from the true model or data are as close as possible to the estimates from the auxiliary model.

The algorithm is proved to be successful in recovering the true ERPT parameter at firm level in the simulated model and is applied to the custom database of China from 2000-2006. Using this nonparametric approach, my finding confirms that ERPT is a nonlinear function of firm-level characteristics

[12] A market of N firms would give N simultaneous equations.

[13] For this particular question of interest, one would need to estimate the productivity distributions for each country in the world, which is difficult.

[14] Please note that these parameters are not necessarily the same as the parameters needed for calibration.

[15] If the model specification is in doubt, an additional loop between step 2 and 5 can be added to evaluate possible alternatives in model specification. I am still working on the proper way to evaluate and compare different model specifications under my proposed procedure.

depending on various measures of market structure. Consistent with theoretical and empirical works, the relationship between ERPT and several market share measures resembles a U shape[16].

The rest of the paper is organised as follows. Section 2 formalises the empirical question. Section 3 introduces the proposed algorithm and explains the mechanism behind it. Section 4 presents the theoretical model of ERPT and various exercises for recovering the true ERPT estimates using the proposed algorithm. Section 5 presents empirical results on China's custom data. Section 6 concludes.

1 Problem

This section gives a formal presentation of the empirical question that this paper tries to address. In addition, I construct two-dimensional numerical examples to illustrate how conventional fixed effects related methods may fail to capture the nonlinear features of ERPT estimates.

The pricing equation of exporters can be presented as follows.

$$p_{\mathcal{I}} = g\left(e_{\mathcal{I}_1}, X_{\mathcal{I}}, M_{\mathcal{I}_2}, \epsilon_{\mathcal{I}}\right)$$

where $\mathcal{I} = \{i, f, d, t\}$ represents the dimensions along which a variable varies, with i, f, d, t standing for product, firm, destination and time respectively; the missing variables vary along dimensions that satisfy $\mathcal{I}_2 \subset \mathcal{I}$ and $\mathcal{I}_1 \neq \mathcal{I}_2$; g is an unknown function; p is a scalar dependent variable representing the exporter's price; e is the key variable of interest, the bilateral nominal exchange rates; X is a vector of observed feature variables; M is a vector of unobserved variables that correlate with e; ϵ is an error term that does not correlate with e. The objective is to understand how changes in exchange rates affect the exporter's price conditioning on the set of observed firm level characteristics X, such as various market share measures, and unobserved variables M that are not varying along all dimensions, such as firm specific marginal costs.

$$\frac{\partial g(.)}{\partial e_{d,t}}$$

1.1 Conventional approaches in the literature are not informative about the underlying structure of ERPT

In this subsection, I construct examples to explain why conventional fixed effect methods are not very informative. In my examples, I restrict my focus to the same exporter selling the same product to different destinations d over a certain time period t . For simplicity, I assume a linear process of export prices $p_{d,t}$ that depends on bilateral exchange rates $e_{d,t}$, market shares $ms_{d,t}$, and marginal cost of the product mc_t. $\beta_{d,t}$ represents the ERPT coefficient which is assumed to be a nonlinear function of market shares

[16]Krugman (1986), Dornbusch (1987), Atkeson and Burstein (2008), Melitz and Ottaviano (2008), Chen, Imbs and Scott (2009), Berman, Martin and Mayer (2012), Amiti, Itskhoki and Konings (2014), Auer and Schoenle (2016)

and marginal costs.[17]

$$p_{d,t} = \mu + \beta_{d,t} e_{d,t} + ms_{d,t} + mc_t + \epsilon_{d,t}$$

In constructing the rest of series, I assume simple linear relationships under which all explanatory variables, $e_{d,t}, ms_{d,t}$ and mc_t, are correlated with each other. Specifically, market shares are constructed to be linear in a destination time specific factor and nominal exchange rates. Marginal cost is constructed similarly.

$$\begin{aligned} ms_{d,t} &= u_{d,t} + 0.1 e_{d,t} \\ mc_t &= u_t - 0.1\overline{e_t}; \quad \overline{e_t} := \frac{\sum_d e_{d,t}}{n_d} \\ e_{d,t} &\sim N(0,1),\ u_{d,t} \sim uniform(0,1),\ \epsilon_{d,t} \sim N(0, 0.01) \end{aligned}$$

Next, I simulate the model for three different underlying ERPT functions $\beta_{d,t}$ and compare results of applying the standard fixed effect estimator with destination and time fixed effects.

$$\begin{aligned} Spec1: &\quad \beta_{d,t} = (ms_{d,t} - 0.5)^2 + mc_t;\ u_t \sim uniform(0,1) \\ Spec2: &\quad \beta_{d,t} = 2(ms_{d,t} - 0.5)^2 * mc_t;\ u_t \sim uniform(0,1) \\ Spec3: &\quad \beta_{d,t} = 2(ms_{d,t} - 0.5)^2 * mc_t;\ u_t \sim N(0,1) \end{aligned}$$

The objective is to read simulated data records of $d, t, p_{d,t}, e_{d,t}, ms_{d,t}$ and estimate the ERPT $\beta_{d,t}$. There are two difficulties in estimating ERPT in this simulated example: (a) the ERPT is not a constant parameter but an unknown function of firm characteristics; (b) the marginal cost mc_t is not observed.

Table (1) presents results with each specification being simulated for 2000 destinations[18] and 40 time periods. Columns (1) - (5) resemble the empirical discovery process of the relationship between ERPT and market shares. The estimated coefficients in column (1) represent a general response of prices to exchange rates. Column (2) adds market share in levels and finds significant coefficients for both variables. Columns (3) and (4) try different interaction terms between exchange rates and market share but no significant result is found. This reflects the main drawback of fully specified structural equations compared to nonparametric approaches. The rejection of one specification is not informative about the alternative right specification. If the researcher stops at regression (4), the discovery that ERPT is U shaped in market share is likely to be delayed.

Even at the correct regression specification column (5)[19], results are not very informative about the

[17] A more realistic model is discussed in section 4.

[18] In my original experiment, I set this number to be significantly bigger than the number of time periods n_T. A more realistic example where $n_D = 200$ can be found in the appendix.

[19] Regression (5) is the specification closest to the true data generating process among these three specifications. For example, specification 1 can be rewritten as follows: $p_{d,t} = 10 + ms_{d,t}^2 e_{d,t} - ms_{d,t} e_{d,t} + (mc_t + 0.25) e_{d,t} + ms_{d,t} + mc_t + \epsilon_{d,t}$ where $\overline{mc_t} = \frac{\sum_t mc_t}{n_T} = 0.5$. Coefficient on $ms_{d,t}$ is close to the theoretical value of 1. The exchange rate interacting with market share has a significant coefficient close to -1 and the interaction term with exchange rate squared has a coefficient close to 1. The coefficient on $e_{d,t}$ is slightly downward biased as the mean of mc_t equals 0.5, which gives the theoretical value of 0.815.

Table 1: Estimates from the fixed effect method

	(1)	(2)	(3)	(4)	(5)
Specification 1: $p_{d,t} = \mu + \left[(ms_{d,t} - 0.5)^2 + mc_t\right] e_{d,t} + ms_{d,t} + mc_t + \epsilon_{d,t}$					
$e_{d,t}$	0.757***	0.635***	0.759***	0.638***	0.796***
	(0.002)	(0.001)	(0.003)	(0.002)	(0.003)
$ms_{d,t}$		1.198***		1.198***	0.995***
		(0.004)		(0.004)	(0.005)
$e_{d,t} * ms_{d,t}$			-0.006	-0.006	-1.019***
			(0.006)	(0.004)	(0.011)
$e_{d,t} * ms_{d,t}^2$					1.016***
					(0.011)
Adjusted R^2	0.663	0.865	0.722	0.865	0.879
Specification 2: $p_{d,t} = \mu + \left[2(ms_{d,t} - 0.5)^2 * mc_t\right] e_{d,t} + ms_{d,t} + mc_t + \epsilon_{d,t}$					
$e_{d,t}$	0.212***	0.090***	0.217***	0.093***	0.242***
	(0.002)	(0.001)	(0.003)	(0.001)	(0.001)
$ms_{d,t}$		1.197***		1.196***	1.002***
		(0.002)		(0.002)	(0.002)
$e_{d,t} * ms_{d,t}$			-0.009**	-0.005***	-0.962***
			(0.004)	(0.002)	(0.004)
$e_{d,t} * ms_{d,t}^2$					0.957***
					(0.004)
Adjusted R^2	0.168	0.825	0.227	0.825	0.885
Specification 3: $p_{d,t} = \mu + \left[2(ms_{d,t} - 0.5)^2 * mc_t\right] e_{d,t} + ms_{d,t} + mc_t + \epsilon_{d,t}; \quad u_t \sim N(0,1)$					
$e_{d,t}$	-0.177***	-0.275***	-0.180***	-0.279***	-0.088***
	(0.017)	(0.016)	(0.016)	(0.016)	(0.019)
$ms_{d,t}$		1.005***		1.003***	1.038***
		(0.016)		(0.016)	(0.016)
$e_{d,t} * ms_{d,t}$			0.302***	0.296***	0.301***
			(0.016)	(0.016)	(0.016)
$e_{d,t} * ms_{d,t}^2$					-0.190***
					(0.011)
Adjusted R^2	0.001	0.049	0.006	0.053	0.056
Time FE	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes
Observations	80,000	80,000	80,000	80,000	80,000

Note: This table presents estimation results after applying the conventional fixed effect estimator to Monte-Carlo simulated data from specification 1 to 3.

underlying structure driving the heterogeneity of ERPT due to the existence of the unobserved marginal cost. The estimated coefficients of the interaction terms from column (2) - (5) are very similar to the results under specification 1. From regression results under specification 1 and 2, it is difficult to make an inference on how $\beta_{d,t}$ depends on firm-level characteristics such as market share $ms_{d,t}$ and marginal cost mc_t. Specification 3 shows that the estimated coefficients can be very sensitive to the distribution of the unobserved variable mc_t where the random factor u_t is assumed to be standard normally distributed rather than uniformly distributed.

2 Algorithm

This section explains the proposed algorithm. The first part of this section introduces the general property which the proposed algorithm relies on under the framework of statistical learning theory. The second part explains how this property can be exploited to control for unobserved variables in tree based algorithms.

2.1 The proposed idea under statistical learning theory

A standard statistical learning problem can be formulated as follows. Consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$. $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are random variables with an unknown joint distribution P. We observe a sequence of n i.i.d. pairs of (X_i, y_i) sampled according to P. The goal of the learning problem is to construct a function $g : \mathcal{X} \to \mathcal{Y}$ such that this function minimises the risk of all possible measurable functions:

$$R(g) := \int h(g(X), Y) dP$$

where $h(.)$ is a criterion function[20]. Empirically, the optimal g is given by

$$\widehat{g}_n := \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} h(g(X_i), p_i)$$

where the expectation is taken over the distribution of P_{XY}. $\mathcal{G}$ is a space of allowed functions depending on the classification algorithm. $\widehat{g_n}$ stands for the estimated function f from data. The main concern of the statistical learning theory is to establish bounds for $R(\widehat{g_n}) - \inf_g R(g)$ so that we know when empirical error $R(\widehat{g_n})$ is a good representation of the true risk measure $\inf_g R(g)$. This measure can be further decomposed into two components, the estimation error and the approximation error.

$$R(\widehat{g_n}) - \inf_g R(g) = \underbrace{\left(R(\widehat{g_n}) - \inf_{g \in \mathcal{G}} R(g) \right)}_{\text{Estimation Error}} + \underbrace{\left(\inf_{g \in \mathcal{G}} R(g) - \inf_g R(g) \right)}_{\text{Approximation Error}} \tag{1}$$

[20] Empirically, $h(,)$ may take the form of $|y - g(X)|$ or $(y - g(X))^2$ depending on the assumptions of P.

Estimating the approximation error is usually hard since it requires knowledge about the target.[21] Important contribution on establishing the relationship between estimation error bounds and entropy measures of classifiers/algorithms has been made by Vladimir N. Vapnik[22].

The method introduced in this paper takes a different appraoch. Instead of assuming that the observed set of variables are given, I consider a parallel set of learning problems. The dependent variable will be the same and take the same value. Most feature variables will also be the same. But some feature variables may be different.

Suppose we have a set of parallel learning problems indexed by $1, ..., j$. For each j, we have an input space $\mathcal{X} \times \mathcal{M}^{(j)}$ and an output space $\mathcal{Y}$. $(X, M^{(j)}, Y) \in \mathcal{X} \times \mathcal{M}^{(j)} \times \mathcal{Y}$ are random variables with joint distribution $P^{(j)}$ unknown to us. We observe n i.i.d. pairs of $(X_i, M_i^{(0)}, y_i)$. We know that $M_i^{(j)}$ is a function of $M_i^{(0)}$. For each j, we have the conventional learning problem of constructing a function $g^{(j)} : \mathcal{X} \times \mathcal{M}^{(j)} \to \mathcal{Y}$ such that this function minimises the risk:

$$R^{(j)}(g) := \int h(g(X, M^{(j)}), Y) dP^{(j)}$$
$$\widehat{g}_n^{(j)} := \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} h(g(X_i, M_i^{(j)}), p_i)$$

Define the numerical measure of the partial derivative as

$$h_2(g, x_1, X_{-x_1}, M^{(j)}, \epsilon) := \frac{g(x_1 + \epsilon, X_{-x_1}, M^{(j)}) - g(x_1 - \epsilon, X_{-x_1}, M^{(j)})}{2\epsilon}$$

Suppose we are originally interested in the case $g^{(0)} : \mathcal{X} \times \mathcal{M}^{(0)} \to \mathcal{Y}$, the question is to what extent we can infer the answer of (0) from results from $g^{(j)} : \mathcal{X} \times \mathcal{M}^{(j)} \to \mathcal{Y}$. In this case, we can write the problem as an expression similar to equation (1):

$$\int h_3[h_2(\widehat{g}_n^{(j)}, x_1, X_{-x_1}, M^{(j)}, \epsilon) - h_2(\arg\inf_g R^{(0)}(g), x_1, X_{-x_1}, M^{(0)}, \epsilon)] dP^{(0)} =$$
$$\int h_3 \Bigg[\underbrace{\left(h_2(\widehat{g}_n^{(j)}, x_1, X_{-x_1}, M^{(j)}, \epsilon) - h_2(\arg\inf_{g \in \mathcal{G}} R^{(j)}(g), x_1, X_{-x_1}, M^{(j)}, \epsilon) \right)}_{\text{Estimation Error}}$$
$$+ \underbrace{\left(h_2(\arg\inf_{g \in \mathcal{G}} R^{(j)}(g), x_1, X_{-x_1}, M^{(j)}, \epsilon) - h_2(\arg\inf_{g \in \mathcal{G}} R^{(0)}(g), x_1, X_{-x_1}, M^{(0)}, \epsilon) \right)}_{\text{Substitution Error}}$$
$$+ \underbrace{\left(h_2(\arg\inf_{g \in \mathcal{G}} R^{(0)}(g), x_1, X_{-x_1}, M^{(0)}, \epsilon) - h_2(\arg\inf_g R^{(0)}(g), x_1, X_{-x_1}, M^{(0)}, \epsilon) \right)}_{\text{Approximation Error}} \Bigg] dP^{(0)}$$

[21] Most statistical learning algorithms take an agnostic approach and avoid making specific assumptions about the underlying distribution.

[22] See Vapnik (1999) for a literature review.

The first term, the estimation error, is where most frontier machine learning algorithms making casual inferences work on. The third term is a conventional term but very difficult to measure without prior assumptions.

The second term is new from this paper. It reflects the effect of substituting the learning problem from (0) to (j). Note that this substitution error depends on three elements, i.e. the group of allowed functions $\mathcal{G}$, the relationship between the variable of interest x_1 and $\mathcal{M}^{(0)}$, and the relationship between input spaces being substituted $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(j)}$.

The interesting part is that the linkage between the size of the substitution error and the set of allowed functional classes $\mathcal{G}$, and the relationship between input spaces being substituted $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(j)}$ can be exploited to control for unobserved variables by adding the third channel of optimisation.

Consider two cases. If two input spaces $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(j)}$ are very different from each other, the range of allowed functions $\mathcal{G}$ will have a considerable impact on the substitution error through affecting $\widehat{g}_n^{(j)}$ and $\widehat{g}_n^{(j)}$ being selected. On the contrary, if two input spaces $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(j)}$ are identical, the set of functions $\mathcal{G}$ can be of any range and do not change the substitution error.

Alternatively, for any given set of functions $\mathcal{G}$, it is possible to figure out the maximum "distance" between $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(j)}$ such that the substitution error is zero.

2.2 Tree based algorithms

This paper exploits a special case where the set of functions $\mathcal{G}$ are tree based algorithms and designs a procedure that can be applied to control for variables that do not vary along all dimensions in a multi-dimensional panel.

The next two subsections explain how tree based algorithms can be exploited to control for unobserved components and facilitate the identification of casual inferences. Subsection 2.2.1 starts with a simple example and outlines the condition (compared with the conventional monotonic transformation) that needs to be satisfied for a variable to be a good proxy for the unobserved variable. I will refer to this condition as weak monotonic transformation. I also discuss additional simulation results for general cases. These simulation results suggest that a more general form of the monotonic transformation condition exists. I will try my best to describe this condition and refer to it as the conditional monotonic transformation property.

In practice, even a conditional (weak) monotonic transformation of the unobserved variable is difficult to find. Subsection 2.2.2 shows that, in a multi-dimensional panel, parameters from structural estimations in a set of restricted dimensions can be used as a proxy for the weak monotonic transformation of unobserved variables not varying along all dimensions.

2.2.1 One-dimensional example

In this subsection, I temporarily abstract away from my ERPT question and discuss this one-dimensional example that helps to understand and clarify the mechanism of the proposed algorithm. Consider the

case of identifying the individual treatment effect.[23]

$$\begin{aligned} y_i &= \beta_i T_i + M_i \\ \beta_i(M_i) &:= M_i \\ T_i &\in \{0,1\},\ M_i \in \{0,1\} \end{aligned}$$

where T_i is a treatment indicator randomly drawn from $\{0,1\}$ with equal probability and β_i is the treatment effect for individual i. M_i is the unobserved variable. In this first example, I assume $\beta_i(M_i) = M_i$ for simplicity. More general cases are discussed in later sections. The objective is to find β_i given data of individual outcomes y_i and its treatment indicator T_i. The data generating process (the functional form of each variable) is unknown to economists. M_i is unobserved.

Suppose all explanatory variables are observed and the functional form is known, obtaining the causal inference is equivalent to estimating parameter values of the function and taking the partial derivative. Similarly, if all explanatory variables are observed but the functional form is unknown, one could fit a nonparametric function and then perform a numerical partial differentiation with the estimated model. That is, suppose M_i is observed, we can estimate the individual treatment effect β_i using the following two-step procedure:

1. Use a nonparametric econometric method or a machine learning algorithm to recognise the pattern of y_i using T_i and M_i. Obtain
$$model_1 : (T_i, M_i) \rightarrow p_i$$

2. Use $model_1$ to construct counterfactual predictions conditioning on the value of M_i and calculate individual treatment effect[24].

$$\beta_i^{Est} = model_1(1, M_i) - model_1(0, M_i)$$

In this procedure, the ability to make predictions conditioning on the value of M_i is important. If the explanatory variable M_i is unobserved, the individual treatment effect β_i will not be identifiable in general.

In many cases, economists do not observe M_i. But it may be possible to have/create a variable $\mathfrak{M}_i$ that preserves some structural information in M_i. If we could construct counterfactuals conditioning on the structural information provided by $\mathfrak{M}_i$, we will be able to recover β_i using the above procedure. In general, the structural information contained in the alternative variable $\mathfrak{M}_i$ could be highly nonlinear. I find that the tree based algorithms have a unique advantage in addressing this type of problems.

Consider the following data generating process of 200 individuals:

[23] You can also think of this setting in terms of the conventional framework characterising treatment effects: $y_i = [y_{1i}(M_i) - y_{0i}(M_i)]T_i + y_{0i}(M_i)$ with $y_{1i}(M_i) = 2M_i$ and $y_{0i}(M_i) = M_i$.

[24] In the case where T_i is continuous, a similar numerical derivation can be obtained by $\beta_i^{Est} = \frac{model_1(T_i+c,i) - model_1(T_i-c,i)}{2c}$

Table 2: Values of y_i

y_i	β_i	M_i	T_i
0	0	0	0
0	0	0	1
1	1	1	0
2	1	1	1

Table 3: Assignment of M_i

M_i	i
0	1-100
1	101-200

In this constructed example, the assignment of M_i happens to be ordered and I want to utilise the information provided by the index i to estimate the individual treatment effect β_i. To capture the highly nonlinear information contained in i, I will run a tree based algorithm with the supervisor/dependent variable y_i on explanatory variables T_i and i in step 1 and construct counterfactuals based on the estimated tree structure $model_1(T_i, i)$ in step 2.

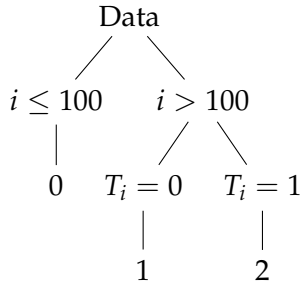
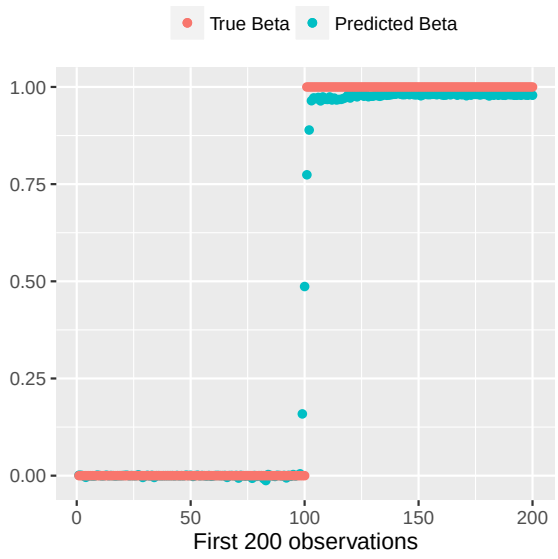


Figure 1: Fitted $model_1(T_i, i)$

The result of applying a decision tree is presented in figure 1. In producing $model_1(T_i, i)$, the algorithm compares the resulted MPE[25] between the best split of T_i and the best split of i. Splitting the sample into $T_i = 0$ and $T_i = 1$ will result in $MPE = 0.5$, while splitting into $i \leq 100$ and $i > 100$ (the best split for i) will result in $MPE = 0.25$. Therefore, the algorithm will choose to split i for its first split. After this split, the MPE for the subgroup $i \leq 100$ is 0 and no further split is needed. The algorithm will try to find the best split for the subgroup $i > 100$. In this case, the best split will be $T_i = 0$ and $T_i = 1$.

[25]I use $h(.) = |y_i - m_c|$ as the error criteria.

i	$\widehat{y_i}$ Evaluated at $T_i = 1$	$\widehat{y_i}$ Evaluated at $T_i = 0$	Estimated β_i	True β_i
1-50	0	0	0	0
51-100	0	0	0	0
101-150	2	1	1	1
151-200	2	1	1	1



The table on the left hand side illustrates the second step. The right hand side graph compares the true β_i with the estimated β_i in an environment closer to reality. First, a random noise is added to the data generating process, i.e. $y_i = \beta_i T_i + M_i + \epsilon_i$, where $\epsilon_i \sim N(0, 0.01)$. Second, in estimating the $model_1(.)$, I add a random variable $\zeta_i \sim N(0,1)$ as an additional explanatory variable which is independently generated from the data generating process of p_i. The idea is that the algorithm should be able to distinguish informative variables from uninformative ones by utilising additional structural information implied by the index i.

Figure 6 gives the estimated β_i under three different settings[26]. Sub-figure (a) represents the result when I use e_i and ζ_i as the explanatory variables. As the algorithm can no longer make predictions of β_i conditioning on useful information in M_i, the predicted β_i can be very different from the true β_i. Sub-figure (c) represents the estimates after adding $N-1$ dummies of the index i. Sub-figure (d) represents the estimates when the true $\beta_i = M_i$ is added as a feature variable.

In general, the assignment of M_i will not be ordered and adding index i will not provide relevant information. Figure 7 presents estimates of β_i where M_i is randomly drawn from $\{0,1\}$ with equal probability for each individual i.

To make the correct prediction of β_d, one needs to find a transformation of the unobserved variable M_i that satisfies a weak monotonic property $\urcorner$ defined below:

Definition 1. Let $\{a_n\}$ be a sequence of real numbers. $\urcorner : \{a_n\} \to \{b_n\} \in \mathbb{R}^N$ is a weak monotonic transformation if

$$a_j > a_i \Rightarrow b_j > b_i \quad \forall i, j \in \{1, ..., N\} \quad or$$
$$a_j > a_i \Rightarrow b_j < b_i \quad \forall i, j \in \{1, ..., N\}$$

[26]The additional variable ζ_i is included in all cases.

Proposition 1. *Let $\{a_n\}$ be a sequence of real numbers. Suppose that entering $\{a_n\}$ as an explanatory variable in a recursive binary splitting algorithm results in $k < n$ unique splitting points $a_{m_1}, ..., a_{m_k}$ with m_i indicating the index for the i th split, then entering any weak monotonic transformation of $\{a_n\}$ will result in the same splitting indices $m_1, ..., m_k$.*

The proof of proposition 1 is given in appendix B. Intuitively, since a binary splitting algorithm only uses ordinal information of an explanatory variable in its splitting criteria, any transformation that preserves the ordinal information of this explanatory variable should result in the same splitting points. Compared to proposition 1, a more interesting and useful property of tree based algorithms is the conditional monotonic transformation property as stated below:

Proposition 2. *Let X_n be a set of feature variables excluding a_n. If $var(a|X_n) \neq 0$ for some values of X_n and there is a large number of observations for these subsets of X_n, entering $\{a_n\}$ as a feature variable is equivalent to entering any $\urcorner(\{a_n\}|X_n)$ in a recursive binary splitting algorithm.*

I have not proved proposition 2 yet. I illustrate my idea using the following numerical example. Consider the case where the treatment effect β_i depends on another observable explanatory variable X_i. To keep the story simple, I assume that X_i is independently drawn from $\{0, 1\}$ with equal probability and β_i is linear in X_i and M_i.

$$
\begin{aligned}
&y_i = \beta_i T_i + M_i \\
&\beta_i = X_i + M_i \\
&T_i \in \{0,1\},\ M_i \in \{0,1\},\ X_i \in \{0,1\}
\end{aligned}
$$

I experiment on the following two formulations of M_t:

$$
\mathfrak{M}_d^1 = \begin{cases} -|\epsilon_i| & if\ M_i = 0 \\ |\epsilon_i| & if\ M_i = 1 \end{cases};\qquad \mathfrak{M}_i^2 = \begin{cases} -|\epsilon_i| + X_i & if\ M_i = 0 \\ |\epsilon_i| + X_i & if\ M_i = 1 \end{cases};\quad where\ \epsilon_i \sim N(0,1)
$$

These two formulations are (a) highly nonlinear but (b) satisfy the weak or conditional weak monotonic transformations[27]. Estimation results are given in Figure 8 and 9. Please note that the algorithm does not know the data generating process of $\mathfrak{M}_i^1$ and $\mathfrak{M}_i^2$ and thus cannot see the clear distinction between red and green points in sub-figures (a) and (b). It classifies by recognising patterns between p_i and $\mathfrak{M}_i^1, T_i$ (or $\mathfrak{M}_i^2, T_i, X_i$ in the second case).

To sum up, tree based algorithms only use the ordinal information in its classification process. Any transformation that contains the same ordinal information of the unobserved variable will produce the same tree structure. Therefore, if one can find a variable or a set of variables that contain approximately the same ordinal information of the unobserved variable, the casual inference can be made (approximately) as if we had observed the unobserved variable.

[27] Please note that $\mathfrak{M}_d^1$ and $\mathfrak{M}_i^2$ do not satisfy the conventional monotonic transformation definition which requires $a_j \leq a_i \Leftrightarrow b_j \leq b_i \quad \forall i, j \in \{1, ..., N\}$

2.2.2 Utilising orthogonal dimensions

In general, it is difficult to apply the conditional weak monotonic transformation property in a one dimensional data framework. However, in a multi-dimensional panel, this property can be exploited together with orthogonal variations across dimensions to control for unobserved variables. The proposed approach exploits the fact that certain dimensions are less influenced by certain types of errors. Conditioning on a particular dimension, the structural estimates may be biased. However, as long as the bias is "well structured" in other dimensions, the weak monotonic transformation property will apply and the individual treatment effect is identifiable[28]. The idea of using all possible combinations of subsets of dimensional-limited structural estimations to control for unobserved variables is first proposed in this paper.

In the context of the constructed two-dimensional examples in section 2, the procedure of the proposed algorithm can be applied as follows. First, run simple OLS regressions using the pricing equation implied by the structural model in all possible subsets of dimensions. Second, gather these estimates from regressions and enter them as variables in a tree based algorithm to predict the dependent variable[29]. In this process, only informative coefficients on predicting the dependent variable from the first step will be selected. Third, use the obtained non-parametric model to predict the dependent variable by changing the key variable of interest and keeping other explanatory variables and the obtained coefficients in step 1 fixed. Fourth, calculate the numerical partial derivative and perform a second algorithm mapping this numerical partial derivative on observed explanatory variables and estimated structural coefficients.

1. For $t = 1...n_t$, run OLS, and collect coefficients b_t^0, b_t^1

$$p_{d,t} = b_t^0 + b_t^1 e_{d,t}$$

   For $d = 1...n_d$, run OLS, and collect coefficients b_d^0, b_d^1

$$p_{d,t} = b_d^0 + b_d^1 e_{d,t}$$

2. Approximating p. Run GBRT entering coefficients $\{b_t^0, b_t^1, b_d^0, b_d^1\}$ as additional feature variables. Obtain

$$model_1 : (e_{d,t}, ms_{d,t}, b_t^0, b_t^1, b_d^0, b_d^1) \rightarrow p_{d,t}$$

3. Numerical differentiation. Use $model_1$ to construct counterfactual predictions conditioning on the

[28] In the context of the ERPT problem, well structured means that the covariance between the volatility (second moment) of bilateral exchange rates and the level (first moment) of marginal cost of the firm at the time dimension does not vary across destinations. This point is explained by analytical examples in the appendix.

[29] Unlike the fixed effect related methods, instead of partitioning out information, the proposed approach adds back these estimates to the main estimation question.

values of $ms_{d,t}, b_t^0, b_t^1, b_d^0, b_d^1$ and calculate[30]:

$$\begin{aligned} p_{d,t}^{Est1} =& model_1(e_{d,t} - \epsilon, b_t^0, b_t^1, b_d^0, b_d^1) \\ p_{d,t}^{Est2} =& model_1(e_{d,t} + \epsilon, ms_{d,t}, b_t^0, b_t^1, b_d^0, b_d^1) \\ \beta_{d,t}^{Est} =& \frac{p_{d,t}^{Est2} - p_{d,t}^{Est1}}{2\epsilon} \end{aligned}$$

4. Approximating β^{Est}. Run GBRT with the dependent variable $\beta_{d,t}^{Est}$ on $e_{d,t}, ms_{d,t}, b_t^0, b_t^1, b_d^0, b_d^1$, and get

$$model_2 : (e_{d,t}, ms_{d,t}, b_t^0, b_t^1, b_d^0, b_d^1) \rightarrow \beta_{d,t}^{Est}$$

Table 4 presents the estimated ERPT from applying the proposed algorithm to three examples constructed in the section 2.[31]

The algorithm is evaluated in two aspects, the ability to recover the key parameter of interest, $\beta_{d,t}$ and the ability to discover the underlying structure $\beta_{d,t}$. In contrast to conventional regression methods, the algorithm estimates β for each d and t, which generates $80,000$ estimates. I construct three measures to evaluate the algorithm's ability to recover the key parameter of interest, $\beta_{d,t}$: (a) the usual absolute measure of distances defined as the sum of squared residuals, SSR; (b) the measure of the number of outliers or extreme values defined as the number of estimated β^{Est} that lies outside one standard deviation of the true β over total number of estimated β^{Est},[32]; (c) visualisation plotting the first 50 observations.

$$\begin{aligned} SSR :=& \sum_d \sum_t (\beta_{d,t}^{Est} - \beta_{d,t})^2 \\ Error\ Rate :=& \frac{|\{\beta_{d,t}^{Est} : |\beta_{d,t}^{Est} - \beta_{d,t}| > \sigma_\beta\}|}{|\{\beta_{d,t}^{Est}\}|} \end{aligned}$$

For evaluating the ability to recover the underlying structure of the ERPT function, I construct the following three measures. Measure 1 and 2 will enable us to compare the true relationship between ERPT and market share with the algorithm estimated relationship. Measure 3 is helpful in understanding why the algorithm estimated relationship is different from the true relationship under some circumstances.

1. The true relationship between ERPT and market share evaluated at different quantiles of the marginal cost.
    - Calculate $mc^q := quantile(mc, q)$ from data; $q \in [0.3, 0.5, 0.7]$[33]
    - Plot $f(ms) = (ms - 0.5)^2 + mc^q$

[30] Throughout my analysis, I choose ϵ to be half standard deviation of the policy variable, i.e. $\epsilon = 0.5std(e_d)$

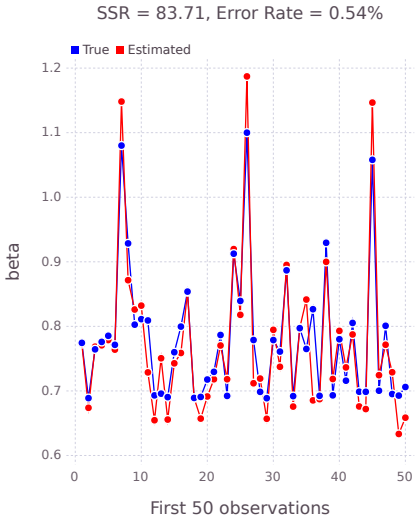
[31] In evaluating the algorithm, I construct two different datasets of the same size generated by the data generating process specified in section 2. The algorithm is first trained in one dataset. The fitted model is then tested in the second dataset.

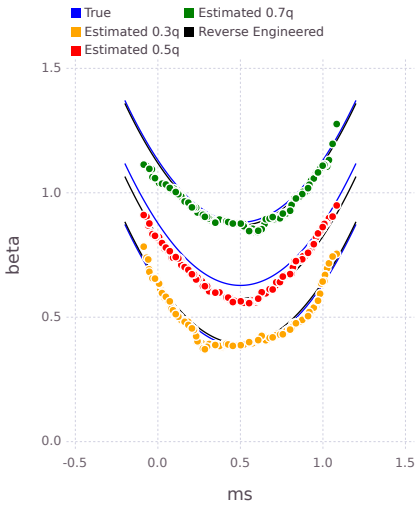
[32] In my examples, the panel is balanced, $|\{\beta_{d,t}^{Est}\}| = n_D n_T$

[33] In my initial experiments, I arbitrarily chose these three quantiles $0.3, 0.5, 0.7$. A more common choice may be $0.25, 0.5, 0.75$.

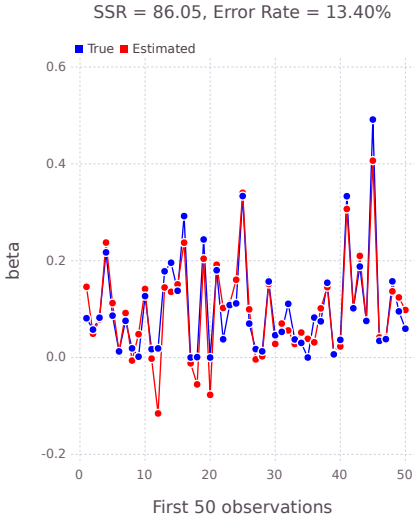
Table 4: Estimates of the proposed algorithm

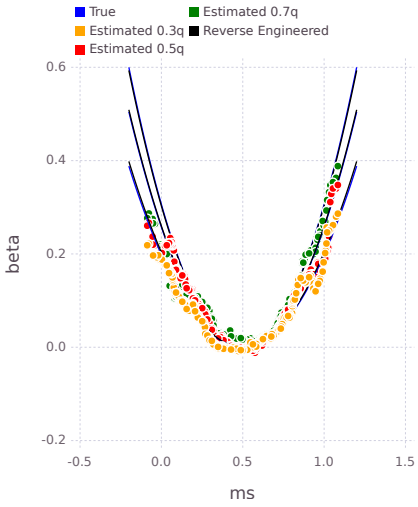
Specification 1:



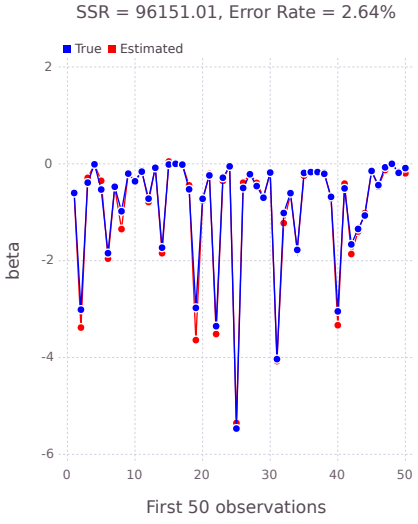


Specification 2:





Specification 3:



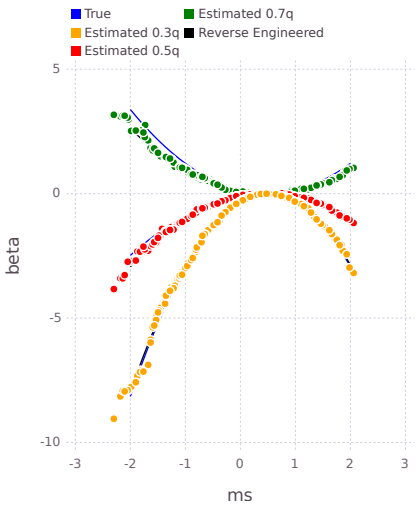
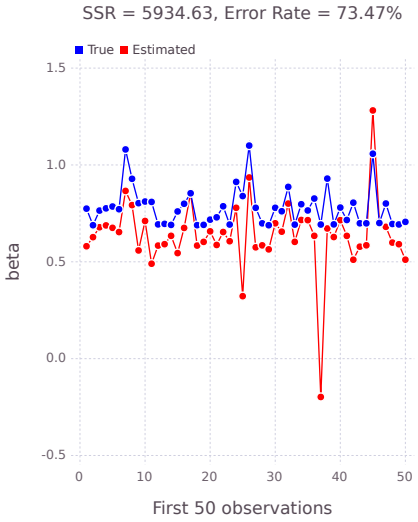
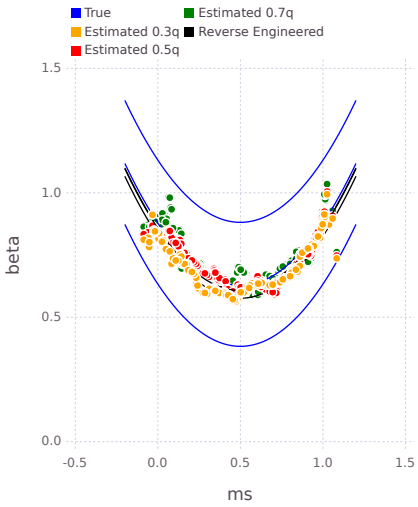


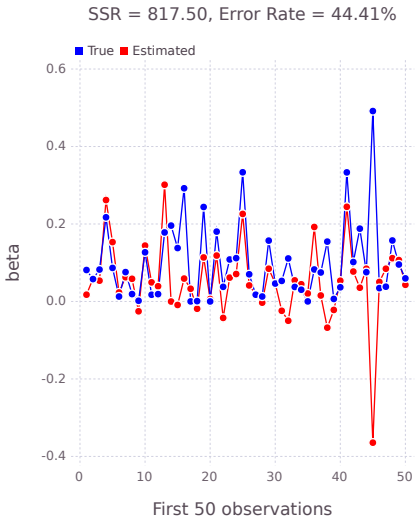
Table 5: No additional information

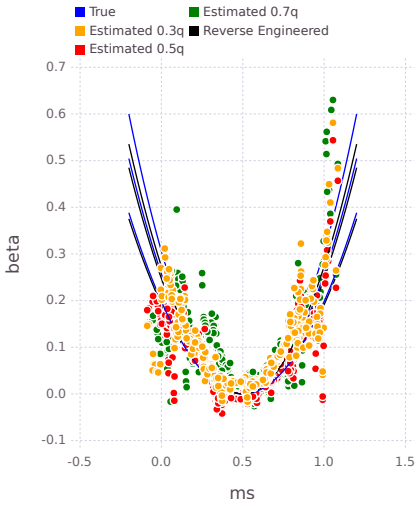
Specification 1:



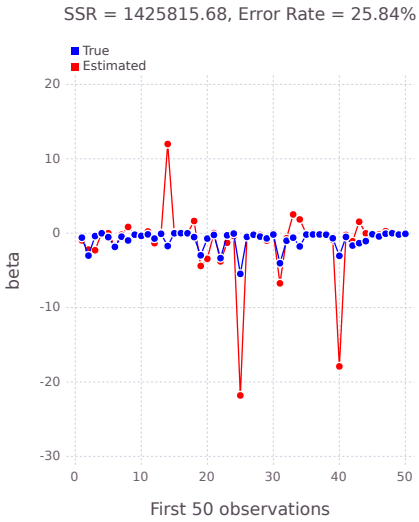


Specification 2:





Specification 3:



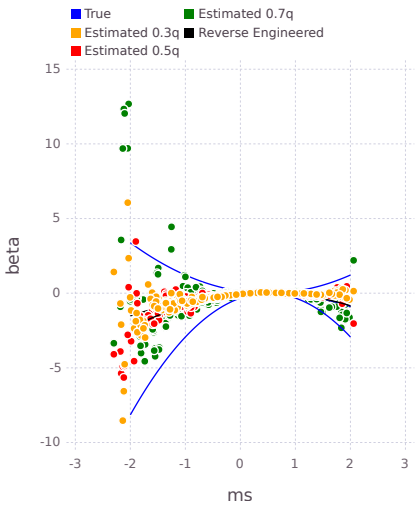
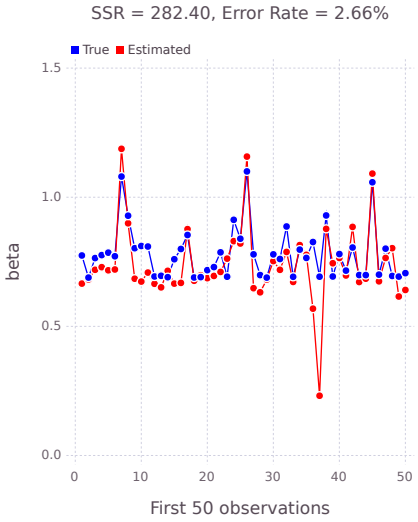
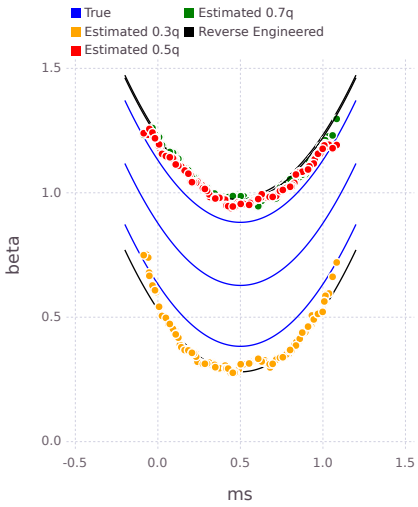


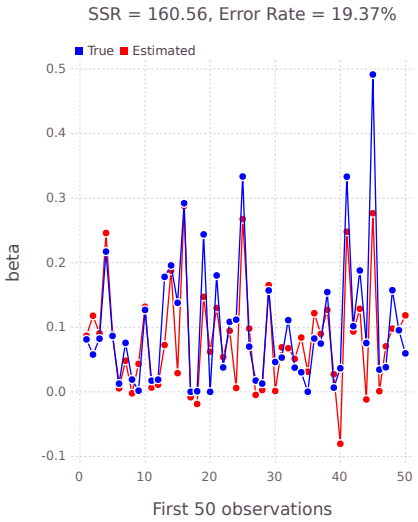
Table 6: Adding dimensional index

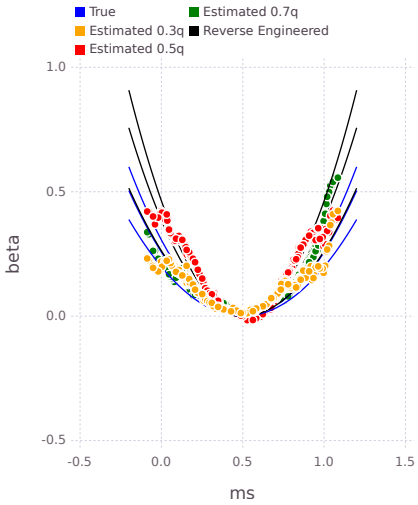
Specification 1:



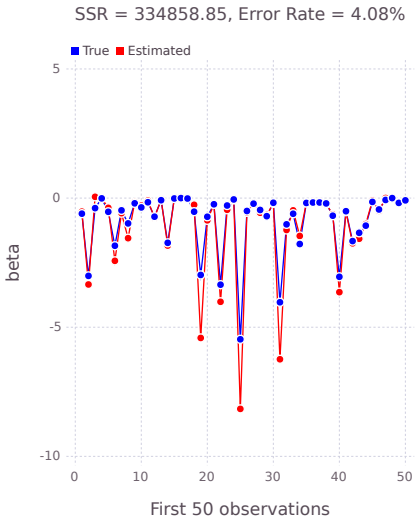


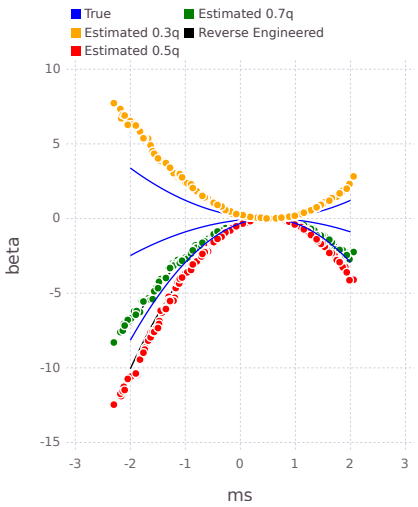
Specification 2:





Specification 3:





2. The estimated relationship between ERPT and market share evaluated at different quantiles of feature variables excluding market share ms, i.e. $\mathcal{X}_{-ms} := e_{d,t}, b_t^0, b_t^1, b_d^0, b_d^1$.

   - Calculate the q-th quantile of each variable in $\mathcal{X}_{-ms}$;
   - Plot $f(ms) = model_2[ms, (\mathcal{X}_{-ms})^q]$ where $(\mathcal{X}_{-ms})^q := (e_{d,t})^q, (b_t^0)^q, (b_t^1)^q, (b_d^0)^q, (b_d^1)^q$

3. The estimated relationship between ERPT and market share evaluated at the reverse engineered quantiles of the marginal cost.

   - Estimate mc^q implied by $(\mathcal{X}_{-ms})^q$; $q \in [0.3, 0.5, 0.7]$
     (a) Run GBRT with mc_t as the dependent variable on $\mathcal{X}_{-ms}$ and get $model_{mc}$
     (b) Estimate $mc^q = model_{mc}[(\mathcal{X}_{-ms})^q]$
   - Plot $f(ms) = (ms - 0.5)^2 + mc^q$

I compare results for the proposed method with two alternative settings. Table 5 presents results when no additional information is added. $model_1$ will be a function mapping $(e_{d,t}, ms_{d,t}) \rightarrow p_{d,t}$ and $\mathcal{X}_{-ms} = e_{d,t}$. The estimation procedure includes step 2-4 only. Table 6 presents results using indices d and t as controls. $model_1$ will be a function mapping $(e_{d,t}, ms_{d,t}, d, t) \rightarrow p_{d,t}$ with $\mathcal{X}_{-ms} = (e_{d,t}, d, t)$.

Comparing results of three tables, the proposed method is significantly better at estimating $\beta_{d,t}$ and approximating the underlying structure of $\beta_{d,t}$ in all three specifications. The method without adding any additional information generates large errors in the point estimate of $\beta_{d,t}$ due to alignments of the unobserved variable mc_t. Given that, the graph on the right hand side shows that the estimated relationship represents the ERPT function evaluated at the median of the unobserved variable mc_t. Adding dimensional indices as additional feature variables will improve the accuracy of point estimates (by a smaller amount compared to the proposed approach) but does not provide additional information on the quantile of the unobserved variable mc_t. As a result, the resulting underlying structure of $\beta_{d,t}$ can be very different from the true structure.

The key to improve the estimates relies on feeding the correct additional structural information about the functional forms to the machine learning algorithm. This type of algorithm has not been explored by existing machine learning approaches because adding such structural information is not possible for prediction problems[34]. A formal presentation of the algorithm can be found in the appendix.

3 Model and Recovering ERPT from Simulated Exporters

The previous section tests the algorithm using simple numerical examples. This section tests the performance of the algorithm in a workhorse international macroeconomic model with heterogeneous firms.

[34]They require information of the dependent variable to estimate structural coefficients.

3.1 Model

I take the seminal contribution of Atkeson and Burstein (2008) as the benchmark model. The model is designed to understand how strategic competition due to different market structures (productivity distributions) could reach different equilibria after an exchange rate shock. There are N countries in the world trading with each other. Within each country, there are a large but limited number of sectors S. As in Atkeson and Burstein (2008), these sectors can be interpreted as "the lowest level of disaggregation of commodities used in economic censuses and price index construction". Within each sector, there are a limited number of firms producing goods. Each firm produces a distinct product with the elasticity of substitution within the sector being ρ_s.

To make the model tractable, I will stick to the following two simplifications made in the original model. First, the model starts with an equilibrium and firms do not make entry and exit decisions[35]. Second, firms only use labour in their production and no imported inputs are needed.

To customise the model to fit the purpose of this paper, I extend the model in three aspects. First, I allow asymmetries in industry structures. To achieve this, I assume a large but limited number of sectors. Second, I extend the original two-country framework to an N-country trading system. This modification allows the model to study the effect of asymmetric exchange rate shocks on trade pattens. Third, to ensure a unique equilibrium in this multi-country world, I assume that only the best domestic firm in each sector exports.[36] This setting can be tough as there exists a hidden sector specific trading barrier such that only the best firm in each sector finds it profitable to export. Technically, this simplification makes this multi-firm multi-sector multi-country model stable and avoids multiple equilibria. In an N-country framework, it generates a very nice market structure with productive N-1 firms from trade partners and a bunch of domestic firms that may be less productive but large in numbers [see figure 18].

3.1.1 Firm's Problem

Variables in this model have five dimensions with f, s, o, d, t standing for firm, sector, origin, destination, time respectively. The final consumption $D_{d,t}$ in destination d is aggregated across sectors using the CES production function with the elasticity of substitution across sectors being equal to η. The price index for final consumption $P_{d,t}$ can be derived as follows.

$$D_{d,t} \equiv \left[\sum_s (D_{s,d,t})^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}}, \quad P_{d,t} \equiv \left[\sum_s (p_{s,d,t})^{1-\eta}\right]^{\frac{1}{1-\eta}} \tag{2}$$

Within a sector, there are foreign firms in this sector $\mathbb{1}_{s,o}$ winning the exporting games I_E from each origin o and all domestic firms in this sector $\mathbb{1}_{s,d}$ competing together with the within-sector elasticity of substitution ρ_s. The sectoral demand $D_{s,d,t}$ and price $P_{s,d,t}$ are given by:

[35]The rationale is that firms' decisions depend on long-run sum of expectations of all future profits. As this model aims to study short-run effects of exchange rate fluctuations, this is a relatively safe condition.

[36]When the best firm is determined, it exports to all countries.

$$D_{s,d,t} \equiv \left[\sum_{o} \sum_{f \in \mathbb{1}_{s,o} \cap \mathbb{1}_E} \left(q_{f,s,o,d,t}\right)^{\frac{\rho_s - 1}{\rho_s}} + \sum_{f \in \mathbb{1}_{s,d}} \left(q_{f,s,o,d,t}\right)^{\frac{\rho_s - 1}{\rho_s}} \right]^{\frac{\rho_s}{\rho_s - 1}}$$

$$P_{s,d,t} \equiv \left[\sum_{o} \sum_{f \in \mathbb{1}_{s,o} \cap \mathbb{1}_E} \left(p_{f,s,o,d,t}\right)^{1-\rho_s} + \sum_{f \in \mathbb{1}_{s,d}} \left(p_{f,s,o,d,t}\right)^{1-\rho_s} \right]^{\frac{1}{1-\rho_s}}$$

Firms compete in quantities $q_{f,s,o,d,t}$ under Cournot competition within each sector s[37]:

$$\max_{q_{f,s,o,d,t}} q_{f,s,o,d,t}(p_{f,s,o,d,t}e_{o,d,t} - mc_{f,s,o,t})$$

subject to

$$q_{f,s,o,d,t} = \left(\frac{p_{f,s,o,d,t}}{P_{s,d,t}}\right)^{-\rho_s} \left(\frac{P_{s,d,t}}{P_{d,t}}\right)^{-\eta} D_{d,t} \tag{3}$$

where $mc_{f,s,o,t}$ is the marginal cost of firm f from sector s and origin o at time t.

3.1.2 Price, Market Share and Demand Elasticity

The optimal price $p_{f,s,o,d,t}$ for an exporter from origin o to destination d can be expressed as a function of price elasticity of demand $\varepsilon_{f,s,o,d,t}$, marginal cost $ms_{f,s,o,d,t}$ and bilateral exchange rate $e_{o,d}$, which is defined as units of currency o per unit of currency d at time t.

$$p_{f,s,o,d,t} = \frac{\varepsilon_{f,s,o,d,t}(ms_{f,s,o,d,t})}{\varepsilon_{f,s,o,d,t}(ms_{f,s,o,d,t}) - 1} \frac{mc_{f,s,o,t}}{e_{o,d,t}} \tag{4}$$

The price elasticity of demand $\varepsilon_{f,s,o,d,t}$ can be expressed as a function of the market share and the elasticity of substitution. Specifically, under the assumption that $\rho > \eta$, the price elasticity of demand is a strictly decreasing function of market share, i.e. bigger firms face a less elastic demand and charge a higher markup.

$$\varepsilon_{f,s,o,d,t} = \frac{1}{\frac{1}{\rho}(1 - ms_{f,s,o,d,t}) + \frac{1}{\eta} ms_{f,s,o,d,t}} \tag{5}$$

where market share is defined as

$$ms_{f,s,o,d,t} = \frac{p_{f,s,o,d,t} q_{f,s,o,d,t}}{\sum_f p_{f,s,o,d,t} q_{f,s,o,d,t}} = \frac{p_{f,s,o,d,t}^{1-\rho}}{\sum_f (p_{f,s,o,d,t})^{1-\rho}} \tag{6}$$

Substituting (9) into (8), we can express elasticity of demand as relative prices. ERPT is less than one as a decrease in $e_{d,t}$ leads to an increase in optimal price, which in turn leads to a lower market share and increases the optimal markup. A log-linearised version of the above description can be derived as follows:

[37] In this nested CES structure, the main theoretical result is not sensitive to whether firms compete in prices or quantities. Atkeson and Burstein (2008) show that similar expressions can be derived if firms are competing in prices.

Log-linearising equation (4), deviations of optimal price can be expressed as a function of deviations of its own market share, its own marginal cost and the bilateral exchange rate between the origin country and the destination country.

$$\widehat{p}_{k,s,o,d,t} = \kappa_{k,s,o,d,t}\widehat{ms}_{k,s,o,d,t} + \widehat{mc}_{k,s,o,t} - \widehat{e}_{o,d,t} \tag{7}$$

where $\kappa_{f,s,o,d,t}$ is the price elasticity with respect to a firm's own market shares, which equals the desired markup times a multiplier due to differences in elasticity of substitution across sectors and within sectors.

$$\kappa_{f,s,o,d,t} \equiv \left(\frac{\varepsilon_{f,s,o,d,t}}{\varepsilon_{f,s,o,d,t}-1}\right)\left(-\frac{1}{\rho_s}+\frac{1}{\eta}\right) \tag{8}$$

Note that both $\widehat{mc}_{k,s,o,t}$ and $\widehat{e}_{o,d,t}$ are state variables and exogenous to firms. After a shock, firms reach the new equilibrium through Cournot competition. The deviation of market share $\widehat{ms}_{k,s,o,d,t}$ for firm k depends on ex ante market structure, i.e. market share distributions $\{ms_{k,s,o',d,t}\}_{k\in\mathbb{1}_f,o'\in\mathbb{1}_o}$, marginal cost shocks $\{\widehat{mc}_{k,s,o',t}\}_{k\in\mathbb{1}_f,o'\in\mathbb{1}_o}$, and the bilateral exchange rate movements of all trade partners from country d, $\{\widehat{e}_{o',d,t}\}_{o'\in\mathbb{1}_o}$.

$$\begin{aligned}
&\widehat{ms}_{k,s,o,d,t}\left[1-(1-ms_{k,s,o,d,t})(1-\rho_s)\kappa_{k,s,o,d,t}\right] \\
&= (1-ms_{k,s,o,d,t})\left\{(1-\rho_s)\left[\widehat{mc}_{k,s,o,t}-\widehat{e}_{o,d,t}\right]\right\} \\
&-\sum_{o'}\sum_{f\neq k} ms_{f,s,o',d,t}\left\{(1-\rho_s)\left[\widehat{mc}_{f,s,o',t}-\widehat{e}_{o',d,t}-\kappa_{f,s,o',d,t}\widehat{ms}_{f,s,o',d,t}\right]\right\}
\end{aligned} \tag{9}$$

It is worth stressing that even under a firm specific shock, the equilibrium effect of changing market shares for other firms $\sum_{o'}\sum_{f\neq k} ms_{f,s,o',d,t}\kappa_{f,s,o',d,t}\widehat{ms}_{f,s,o',d,t}$ will not be zero in most cases[38]. The importance of competitors' market share reactions is weighed by the market share with its importance strictly increasing in the market share of the competitor[39].

Substituting (9) into (7), we can obtain a general equation for price deviations in a multi-country environment.

$$\widehat{p}_{k,s,o,d,t} = \lambda_{k,s,o,d,t}\left[\widehat{mc}_{k,s,o,t}-\widehat{e}_{o,d,t}-\kappa_{k,s,o,d,t}\widehat{CE}_{k,s,o,d,t}\right] \tag{10}$$

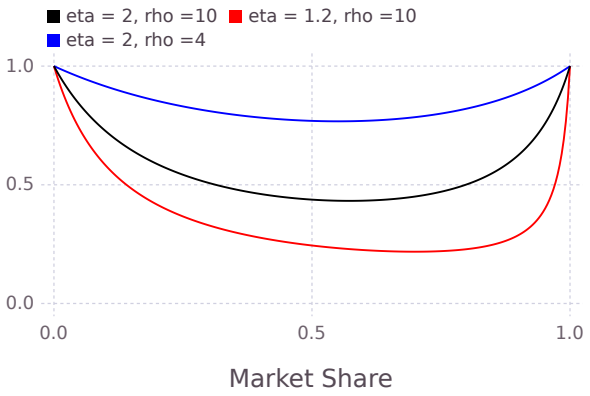
where $\lambda_{k,s,o,d,t}$ is the theoretical ERPT and it is U-shaped in market share as derived in most ERPT literature,

$$\lambda_{f,s,o,d,t} = \frac{1}{1-(1-ms_{f,s,o,d,t})(1-\rho_s)\kappa_{f,s,o,d,t}} \tag{11}$$

[38]In the presence of $\widehat{ms}_{f,s,o',d,t}$, there is no simple analytical solution for the optimal market share change after a shock even after log-linearisation. Given a set of realised shocks and prior market structure, market share conditions (9) will formulate a system of f nonlinear equations and can be solved numerically. As I will show in later simulations, reaction from other firms will make ERPT fail to present the U-shaped response in market share.

[39]Note that the expression $\kappa_{f,s,o',d,t}$ is strictly increasing in market share $ms_{f,s,o',d,t}$.

Figure 2: Plot of $\lambda_{f,s,o,d,t}$



and $\widehat{CE}_{k,s,o,d,t}$ is the total effect of competitors' reactions.

$$\widehat{CE}_{k,s,o,d,t} = \sum_{o'} \sum_{f \neq k} ms_{f,s,o',d,t}(1-\rho_s)\left[\widehat{mc}_{f,s,o',t} - \widehat{e}_{o',d,t} - \kappa_{f,s,o',d,t}\widehat{ms}_{f,s,o',d,t}\right] \tag{12}$$

In a multi-country setting, the optimal price response of an exporter is a function of origin specific exchange rate shock minus bilateral exchange rate shocks of all other trade partners weighted by a non-linear function of corresponding competitor's market share.

The household's problem follows closely with Atkeson and Burstein (2008). There is a representative household in each destination d maximising its expected utility by choosing optimal final consumption $C_{d,t}$ and optimal labour supply $L_{d,t}$. The representative consumer can trade a complete set of international assets from all trade partners.

$$\max_{C_{d,t},L_{d,t}} E_0 \sum_{t=0}^{\infty} \beta^t U(C_{d,t}, L_{d,t})$$

subject to

$$U_{d,t} = log[C_{d,t}^{\mu}(1-L_{d,t})^{1-\mu}]$$

$$P_{d,t}C_{d,t} + \sum_{o}\left[\sum_{\nu} p_{o,t}^{B}(\nu)B_{o,t}(\nu) - (1+i_{o,t-1})B_{o,t-1}\right] * e_{o,d,t} = W_{d,t}L_{d,t} + \Pi_{d,t}$$

where holding $B_{o,t}(\nu)$ will earn $B_{o,t}$ unit of currency o at $t+1$ if state ν happens. $p_{o,t}^{B}(\nu)$ is the price of bond from origin o with state ν. $i_{o,t-1}$ represents the interest paid in the unit of currency o from $t-1$ to t. $\Pi_{d,t}$ is the lump-sum profit transfer from all domestic firms and exporters in country d.

The optimal solution of household's problem is given by

$$\frac{1-\mu}{\mu}\frac{C_{d,t}}{1-L_{d,t}} = \frac{W_{d,t}}{P_{d,t}} \tag{13}$$

$$\frac{C_{o,t}P_{o,t}}{e_{o,d,t}C_{d,t}P_{d,t}} = \frac{C_{o,t+1}(\nu)P_{o,t+1}(\nu)}{e_{o,d,t+1}(\nu)C_{d,t+1}(\nu)P_{d,t+1}(\nu)} \tag{14}$$

where (13) represents the optimal division of consumption and labor and (14) stands for the conventional international risk sharing condition.

3.1.3 Other equilibrium conditions

The production function is assumed to be linear in labour where the marginal cost $mc_{f,s,o,t}$ of a firm is calculated by dividing the nominal wage of the origin country $W_{o,t}$ by its productivity $\Omega_{f,s,o,t}$. For each firm, the total quantity of products sold $\sum_d q_{f,s,o,d,t}$ equals the quantity produced $\Omega_{f,s,o,t}l_{f,s,o,t}$. The last equation is the labor market clearing condition.

$$mc_{f,s,o,t} = \frac{W_{o,t}}{\Omega_{f,s,o,t}}$$
$$\Omega_{f,s,o,t}l_{f,s,o,t} = \sum_d q_{f,s,o,d,t}$$
$$\sum_{f,s} l_{f,s,o,t} = L_{o,t}$$

I select the nominal wage $W_{o,t}$ in each origin as the numeraire and set it equal to one. In this model, the productivity distribution can be asymmetric across sectors and countries. As a result, the bilateral nominal exchange rate is not necessarily equal to one. In my simulation the steady state bilateral exchange rate is determined by the bilateral balance of trade condition, i.e.

$$\sum_{f,s} p_{f,s,d,o,t}q_{f,s,d,o,t} = \sum_{f,s} p_{f,s,o,d,t}q_{f,s,o,d,t} * e_{o,d,t} \quad for\ o \neq d$$

3.2 Recovering ERPT from simulated exporters

In the following subsection, I use the model to test the proposed algorithm. Specifically, I simulate the model under different scenarios, calculate the model implied ERPT at firm level by constructing counterfactual environments, run the proposed algorithm using simulated data and compare estimated pass through with its theoretical value.

3.2.1 Model Simulation

I use the same calibration for the elasticity of substitution across sectors η and within sectors ρ as in Atkeson and Burstein (2008). In the benchmark case, I choose a model of three countries. The number of sectors is chosen to be 25, consistent with the classification of popular industry coding standards[40]. For a given prior productivity distribution, the number of domestic firms in each county determines the degree of home bias in the sector. For a model of three countries, I set the number of domestic firms to

[40]Increasing the number of countries and sectors will exponentially increase the number of nonlinear equations needed to solve for each period.

be 3. As a result, there will be 5 firms in each sector, including two relatively more competitive foreign firms and three domestic firms. This setting gives a reasonable median home market share around 50% depending on the productivity distribution of the sector in other countries[41]. Firm level productivity shocks are assumed to follow a simple AR(1) process with persistence equal to 0.95.

	Countries	S	N	ρ	η	$\Phi(\Omega)$
Benchmark	3	25	3+2	10	2	Uniform
Robustness	4,5	10-35	3 to 10 + Countries - 1	10	2	Uniform

To ensure the existence of a unique equilibrium in each period, I consider a financial autarky case and give exogenous exchange rate shocks to the model[42]. I further assume that the no financial market exchange rate arbitrage condition holds, i.e.

$$e_{1,2,t} = \frac{e_{1,3,t}}{e_{2,3,t}}$$

which implies a maximum of 2 exchange rate shocks in this three country world[43].

$$e_{1,2,t} = \xi_{1,t} e_{1,2,ss}, \quad e_{3,2,t} = \xi_{3,t} e_{3,2,ss}, \quad \xi_{i,t} \sim uniform(0.8, 1.2)$$

There are two sets of state variables influenced by two sets of shocks, i.e. the set of productivity shocks $\Omega_{f,s,o,t}$ for each firm, each sector and each country and the set of bilateral exchange rate shocks $e_{o,d,t}$. In each period, productivity shocks and exchange rate shocks are first realised and the corresponding values of state variables are then calculated. The most productive domestic firm in each sector wins the exporting game and exports to all trade partners. Collecting all equilibrium conditions for all countries, solving the model is equivalent to solving a large-scale constrained system of nonlinear equations[44].

In the following exercise, I will take country 1 as the home country and try to recover ERPT of country 1's exporters. Counterfactual macro state is constructed as follows[45]

$$e^c_{1,2,t} = e_{1,2,t-1}, \quad e^c_{3,2,t} = \xi_{3,t} e_{3,2,ss}, \quad e^c_{1,3,t} = \frac{e^c_{1,2,t}}{e^c_{3,2,t}}$$

[41] In this model, increasing the number of domestic firms will not always lead to a greater home bias as it will also make foreign firms surviving from the exporting game more competitive. Home firms increase in numbers but foreign competitors increase in quality. The equilibrium result depends on the assumption of productivity distributions.

[42] The international risk sharing condition (14) no longer applies.

[43] The third bilateral exchange rate is determined by the no arbitrage condition.

[44] The setting that only the most productive firm in a sector exports avoids potential multiple equilibria and returns a unique solution in most calibrations. The model is built using Julia JUMP module and solved using Ipopt solver.

[45] The algorithm is tested on various settings of exchange rate shocks and a world with maximum 5 countries. Related results are available upon request.

3.2.2 Estimation procedure

After the model is simulated, an artificial dataset is constructed to resemble those observable variables in China's custom dataset and test the proposed algorithm. The objective of the algorithm is to use only information from the constructed dataset to (a) learn from trade patterns, (b) estimate price changes under a bilateral exchange rate shock at period t given market conditions at $t-1$ and (c) recover the model implied ERPT for estimated firms.

The estimation procedure is given as follows:

1. Simulate the model for 240 periods (20 years). Record variables that are accessible from a common custom database including $f, s, o, d, t, p_{f,s,o,d,t}, q_{f,s,o,d,t}$, plus some observable macroeconomic indices including $D_{s,d,t}, P_{s,d,t}, e_{d,t}, P_{d,t}, L_{d,t}, C_{d,t}$.

2. Re-simulate the model to recover the model implied counterfactual ERPT. To calculate the model implied theoretical ERPT, I load all variables including the productivity shock from the simulated model. I then construct the counterfactual equilibrium using the productivity distribution at period t-1 and bilateral exchange rate at period t. The price difference between the counterfactual and the original equilibrium reflects the equilibrium effect of a pure exchange rate shock.

3. Identify a simple regression relationship between the dependent variable and the observable independent variables based on economic theory.

$$log(p_{f,s,d,t}) = a + b * log(e_{d,t}) + c * log(p_{f,s,d,t-1})$$

4. Identify dimensions to be fixed as $\{s, t, d, sd\}$. Run regressions and collect coefficients $a_s, b_s, c_s, a_t, b_t, c_t, ...$

5. Create market share measure $ms_{fsdt,sdt} = \frac{p_{f,s,d,t} q_{f,s,d,t}}{P_{s,d,t} D_{s,d,t}}$

6. Run GBRT with supervisor $log(p_{f,s,d,t})$ on the policy variable $log(e_{d,t})$ and the feature variables

$$\begin{aligned}\mathcal{X} = \; & log(e_{d,t-1}), log(e_{d',t}), log(e_{d',t-1}), log(ms_{fsdt-1,sdt-1}), \\ & log(D_{s,d,t-1}), log(P_{s,d,t-1}), log(P_{d,t-1}), log(L_{d,t-1}), log(C_{d,t-1}), \\ & a_s, b_s, c_s, a_t, b_t, c_t, ...\end{aligned}$$

   and obtain $model_1$

7. Numerical differentiation using the predicted price at current exchange rate and the predicted price

if the exchange rate was the same as in the previous period[46].

$$p^{Est1}_{f,s,d,t} = model_1(e_{d,t}, \mathcal{X})$$
$$p^{Est2}_{f,s,d,t} = model_1(e_{d,t-1}, \mathcal{X})$$
$$ERPT^{Est}_{f,s,d,t} = \frac{log(p^{Est1}_{f,s,d,t}) - log(p^{Est2}_{f,s,d,t})}{e_{d,t} - e_{d,t-1}}$$

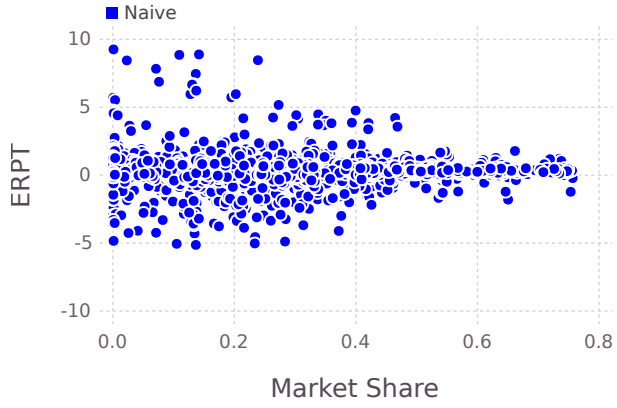
8. Run GBRT again with supervisor $ERPT^{Est}_{f,s,d,t}$ on $log(e_{d,t})$, $\mathcal{X}$ and obtain $model_2$.

3.2.3 Results

The performance of the algorithm is tested in two cases. Case 1 shuts down the idiosyncratic productivity shocks of firms in all countries[47], leaving only multilateral exchange rate shocks. Case 2 represents the world with both the idiosyncratic productivity shocks and multilateral exchange rate shocks.

In a multi-country world, multilateral rather than bilateral exchange rate movements matter. As derived in (9) and (12), the multilateral exchange rate shocks transmit into exporter's prices through the competition channel. However, controlling for the effect of multilateral exchange rate movements is not straightforward and most empirical works estimating the ERPT from the exporters' perspective only focus on bilateral movements.

Figure 3: Naive ERPT estimates

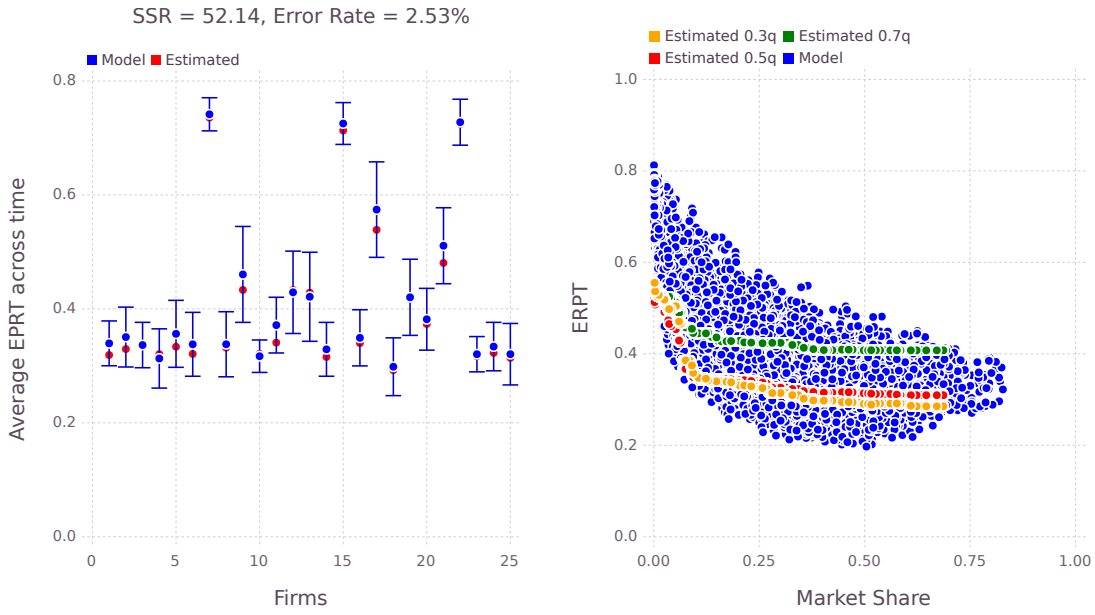


Note: The blue dots represent the model implied firm-level ERPT without accounting for the exchange rate movements of other trade partners. The model is simulated under case 1 where there is no productivity shock. If there is no exchange rate shock, the price $p_{f,s,1,2,t}$ will be the same across all time periods.

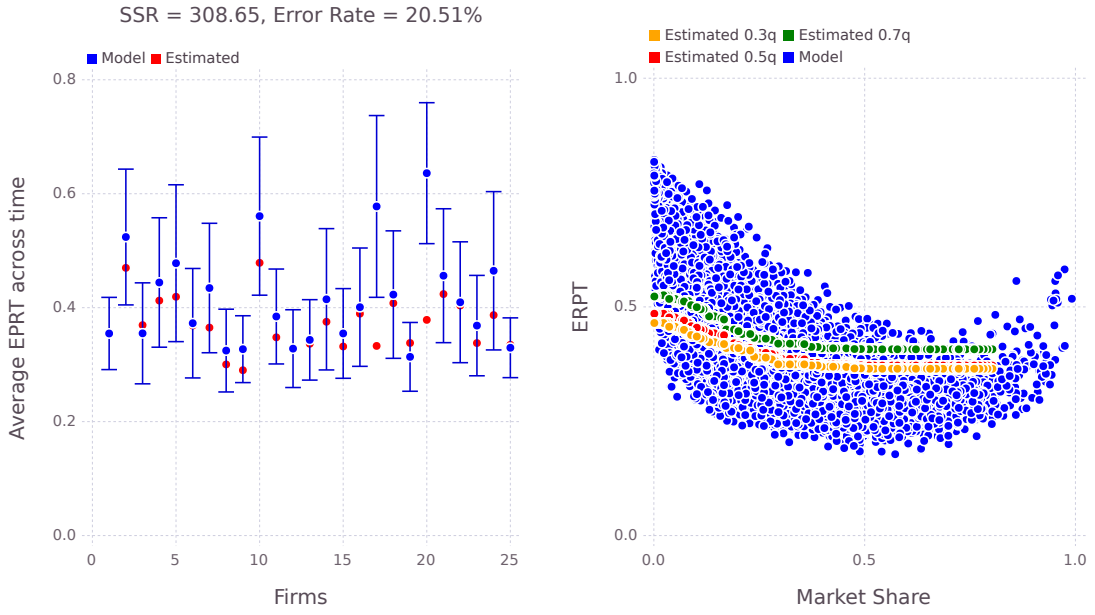
Ignoring this competition effect will potentially lead to seemingly unacceptable ERPT estimates. Figure 3 shows calculation of firm level ERPT for exporters in country 1 selling in country 2 without controlling for the exchange rate movement between country 2 and 3. In the simulated model, the price of

[46] All other variables in $\mathcal{X}$ take their current value at time t.
[47] Firms are still different in their productivity drawn.

Figure 4: ERPT estimates of the proposed algorithm versus model implied counterfactuals



Case 1: only exchange rate shocks



Case 2: add productivity shocks

Note: The left graph represents the time-average of ERPT for exporters originating from country 1 exporting to country 2. The x-axis of the left graph represents the index of exporters. The red dots represent estimates of the proposed algorithm. The blue dots represent the time-average of model implied firm-level ERPT backed up through counterfactual analysis. The blue bars reflect the time fluctuation of model implied ERPT for each firm. Error rate and SSR are calculated based on point estimates of ERPT for each firm-time combination (i.e. not time-averages).
The right graph represents the estimated relationship between market share and ERPT. The blue dots plot the model implied ERPT. The coloured lines provide the algorithm estimated relationship evaluated at different quantiles of the feature variables excluding the market share, $\mathcal{X}_{-ms}$.

exporters from country 1 at country 2, $p_{f,s,1,2,t}$, reacts to both $e_{1,2,t}$ and $e_{2,3,t}$. The bilateral exchange rate movements of other trade partners of the destination country could potentially magnify or mitigate the effect of the bilateral exchange rate movement from the origin country. If the ERPT is calculated without accounting for this effect, calculated results can be significantly greater than 1 or smaller than 0.

Figure 4 shows the results of the time-averaged estimated ERPT[48] from the proposed algorithm. The proposed algorithm performs extremely well under case 1. All time-averaged estimates lie within one standard deviation of the model implied estimate and are very close to the mean value of the model implied estimates. The error rate on point estimates is only 2.53%. The right graph shows that the estimated relationship between ERPT and market share is well aligned with the true relationship of the model implied estimates[49].

For the second case, adding productivity shocks increases the error rate. However, out of 25 firms, only two firms' time-average ERPT lie outside one standard deviation of the true value. The right figure is relatively weak in identifying the correct quantiles but still well aligned with the true relationship implied by the model.

4 Empirical Results

This section presents three empirical contributions on understanding firms' pricing behaviour. First, with the proposed algorithm, this paper presents estimates of ERPT for each firm-product-destination combination of China's exporters during the sampling period 2000-2006.[50] These estimates can be later used to construct effective exchange rate measures using a bottom-up approach based on firm-level ERPT; to identify the most and the least influenced commodity, industry and trade partner by exchange rate shocks; or to plot distributions of ERPT for different types of firms, industries and destinations, etc.

Second, this paper takes an agnostic approach to study the relationship between ERPT and various market share measures. With a four-dimensional panel (firm-product-destination-time), 12 market share measures can be constructed. Among these 12 market share measures, 9 measures are economically meaningful. Although there has been increasing attention in the trade and international literature on how different market share measures capture different aspects of firms' pricing decision and international shock transmissions, most studies work on a subset of the market share framework presented in table 7. My estimates contribute to the literature by assessing the relative statistical importance of market shares in explaining variations of ERPT and the unit value volatility. In addition, this paper empirically documents the nonlinear relationship between various market share measures and confirms

[48] i.e. $\frac{1}{n_t}\sum_t \beta^{Est}_{f,s,1,2,t}$. As only the best firm exports, the 25 firms in the figure stand for 25 sectors in the model. Graphs for detailed point estimates and the comparison with alternative methods can be found in the appendix.

[49] Note that for a given market share, the same exchange rate shock may have different impacts on each sector of an economy, which depends on two factors: (a) the underlying distribution of productivity for this particular sector of exporters from all countries and (b) the general equilibrium effect due to the change in aggregate environments of a local destination and countries exporting to this destination.

[50] I use China Import and Export Custom Database funded by Cambridge Endowment for Research in Finance. Data are available at the monthly frequency from 2000 to 2006. I aggregate these monthly series into quarterly frequency to accommodate the availability of macro series such as CPI index. Details of the database and its related descriptive statistics can be found in the second chapter.

various theoretical predictions.

Table 7: Dimensions of market shares

Measure	Construction	Abbreviation	Notation
Classical Market Share	$\frac{V_{fdi}}{\sum_f V_{fdi}}$	`fdi_di`	Firm Share (DI)
Local Core Product Measure	$\frac{V_{fdi}}{\sum_i V_{fdi}}$	`fdi_fd`	Product Share (FD)
Destination Importance at Firm-product Level	$\frac{V_{fdi}}{\sum_d V_{fdi}}$	`fdi_fi`	Destination Share (FI)
Global Firm Competitiveness	$\frac{\sum_d V_{fdi}}{\sum_f \sum_d V_{fdi}}$	`fi_i`	Firm Share (I)
Global Core Product Measure	$\frac{\sum_d V_{fdi}}{\sum_d \sum_i V_{fdi}}$	`fi_f`	Product Share (F)
Destination Importance at Product Level	$\frac{\sum_f V_{fdi}}{\sum_f \sum_d V_{fdi}}$	`di_i`	Destination Share (I)
Local Firm Competitiveness	$\frac{\sum_i V_{fdi}}{\sum_f \sum_d V_{fdi}}$	`fd_d`	Firm Share (D)
Local Taste Preference	$\frac{\sum_f V_{fdi}}{\sum_f \sum_i V_{fdi}}$	`di_d`	Product Share (D)
Destination Importance within Firm	$\frac{\sum_i V_{fdi}}{\sum_d \sum_i V_{fdi}}$	`fd_f`	Destination Share (F)

Third, this paper provides the first evidence that the underlying factors explaining unit value volatility and ERPT may be different. The price volatility is strictly decreasing in all market share measures, while the relationship between ERPT and the market structure is nonlinear and varies depending on the specific share measure. In addition, increasing the volatility of bilateral and multilateral exchange rates, the volatility of destination CPI and the frequency of trade have ambiguous positive effects on unit value volatility. The effects of these variables on ERPT are heterogeneous and highly nonlinear. Interestingly, I find that both EPRT and unit value volatility are hump-shaped[51] in a number of observed trading periods.

4.1 Estimation procedure

The first two stages of the empirical procedure follow closely with the one introduced in section 3.1. In addition, I explore and estimate the factors explaining the volatility of unit values and compare them to the results on ERPT obtained in stage 2.

1. **Stage 1:**

   (a) Identify a simple regression relationship between the dependent variable and observable independent variables based on economic theory.

$$log(p_{i,f,d,t}) = a + b * log(p_{i,f,d,t-1}) + c * log(e_{d,t})$$

[51] U-shaped from the importers' perspective.

(b) Identify dimensions to be fixed as $\{i, f, d, t, if, id, fd, ft\}$. Run regressions and collect coefficients $\{a_i, b_i, c_i, a_f, b_f, c_f, ...\}$

(c) Create and select combinations of market share measures.[52]

Table 8: Classification of market share measures

	Destination Specific	Global Counterparts
Firm	`fdi_di`	`fd_d & fi_i`
Product	`fdi_fd`	`fi_f` & `di_d`
Destination	`fdi_fi`	`di_i` & `fd_f`

(d) Estimate GBRT model with supervisor $log(p_{f,s,d,t})$ on the policy variable $log(e_{d,t})$ and feature variables[53]

$$\begin{aligned}\mathcal{X}_1 := \; & log(oneer_{d,t}), log(cpi_{d,t}), log(p_{i,f,d,t-1}), \\ & \texttt{fdi\_di, fdi\_fd, fdi\_fi, fd\_d, fi\_f, di\_i,} \\ & a_i, b_i, c_i, a_f, b_f, c_f, ...\end{aligned}$$

and obtain fitted $model_1$

(e) Numerical differentiation on predicted counterfactual bilateral exchange rates

$$\begin{aligned} p^{Est1}_{f,s,d,t} &= model_1(e_{d,t} + 0.5\sigma_{e_d}, \mathcal{X}_1) \\ p^{Est2}_{f,s,d,t} &= model_1(e_{d,t} - 0.5\sigma_{e_d}, \mathcal{X}_1) \\ ERPT^{Est}_{f,s,d,t} &= \frac{log(p^{Est1}_{f,s,d,t}) - log(p^{Est2}_{f,s,d,t})}{\sigma_{e_d}} \end{aligned}$$

2. **Stage 2:**
Estimate GBRT model with supervisor $ERPT^{Est}_{i,f,d,t}$ on feature variables including volatilities of unit values and three macro price indicators (bilateral nominal exchange rates, ONEER, Destination CPI), two measures of firm-product-destination level characteristics (frequency of trade and observed trading periods), 6 market share measures and controlling coefficients.

$$\begin{aligned}\mathcal{X}_2 := \; & \sigma_{p_{i,f,d}}, \sigma_{e_d}, \sigma_{oneer_d}, \sigma_{cpi_d}, \text{Frequency of Trades}_{i,f,d}, \text{Observed Trading Periods}_{i,f,d} \\ & \texttt{fdi\_di, fdi\_fd, fdi\_fi, fd\_d, fi\_f, di\_i,} \\ & a_i, b_i, c_i, a_f, b_f, c_f, ...\end{aligned}$$

[52] Table 8 reclassifies the 9 economically meaningful measures. As these market share measures are interdependent, one can find the minimal set of variables to represent the information of these 9 statistics. It can be shown that it is sufficient to include three destination specific measures (`fdi_di, fdi_fd, fdi_fi`) and the first column of global measures (`fd_d, fi_f, di_i`).

[53] $oneer_{d,t}$ indicates the orthogonal destination NEER which is constructed using quarterly data by the same method introduced in the second chapter.

and obtain fitted $model_2$.

3. **Stage Volatility:**
   Estimate GBRT model with supervisor volatility of unit values, $\sigma_{p_{i,f,d}}$, on the same set of feature variables as in stage 2 excluding $\sigma_{p_{i,f,d}}$

$$\begin{aligned}\mathcal{X}_{Vola} := \; & \sigma_{e_d}, \sigma_{oneer_d}, \sigma_{cpi_d}, \text{Frequency of Trades}_{i,f,d}, \text{Observed Trading Periods}_{i,f,d} \\ & \texttt{fdi\_di, fdi\_fd, fdi\_fi, fd\_d, fi\_f, di\_i,} \\ & a_i, b_i, c_i, a_f, b_f, c_f, ...\end{aligned}$$

   and obtain fitted $model_{Vola}$.

4.2 Main results

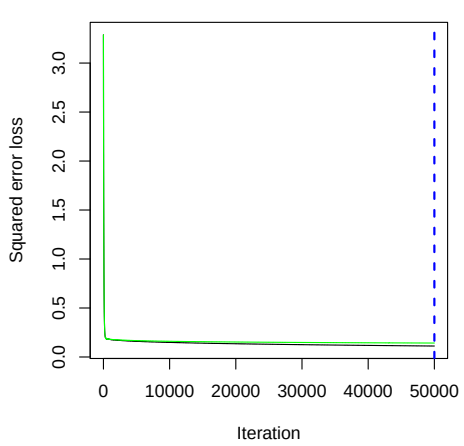
Deploying the algorithm on a real custom dataset is computationally demanding.[54] At this stage, the graphs are still sensitive to the economic equation relationship being assumed in step (a) of the first stage and the feature variables entering the first and the second stages of the algorithm. The following graphs summarise my preliminary findings.

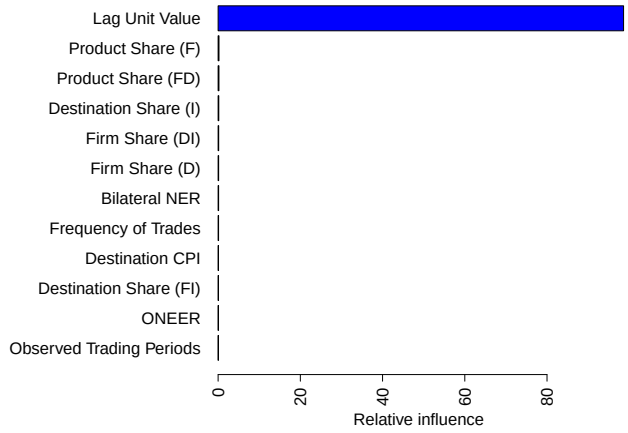
[54]The cleaned dataset has a size around 5 Gigabytes. In the proposed method, a large number of estimated structural parameters need to be stored in the memory. As a result, it currently requires around 100 times the memory of the original dataset. My codes are running on a computational cluster CamGrid [see http://help.uis.cam.ac.uk/supporting-research/research-support/camgrid/camgrid] which allows me to have maximum 128 Gigabytes memory. The following result is based on a sample of 5% randomly selected firms in the China's Custom dataset.

The second practical issue is that the amount of computational resources needed increases exponentially with the size of the dataset and the number of iterations to run. The computing time is mainly consumed in running cross validation simulations. Ideally, the optimal number of iterations needs to be determined by the cross validation simulations. By increasing the number of iterations, the within sample prediction error will always decrease but the cross validation error may or may not decrease depending on whether the additional iteration improves the fit for all parallel sub-samples. The optimal number of iteration is defined at the iteration where cross validation errors stop decreasing. However, due to computational time limits, I force the program to stop at 50,000 iterations before the optimal iteration is reached. With a 5% sample and 50,000 iterations, the program takes around 1 week to complete. As can be seen in table 9, the rate of the decreasing squared error loss is sufficiently low at the 50,000th iteration.

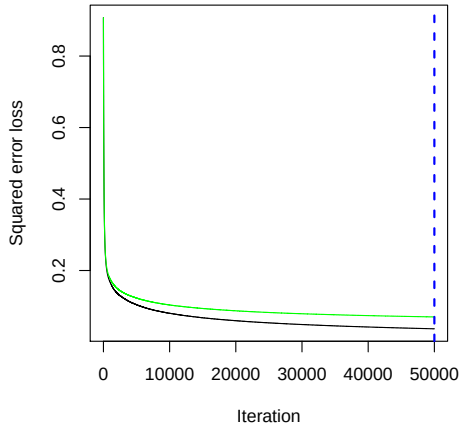
Table 9: Cross validation and relative importance

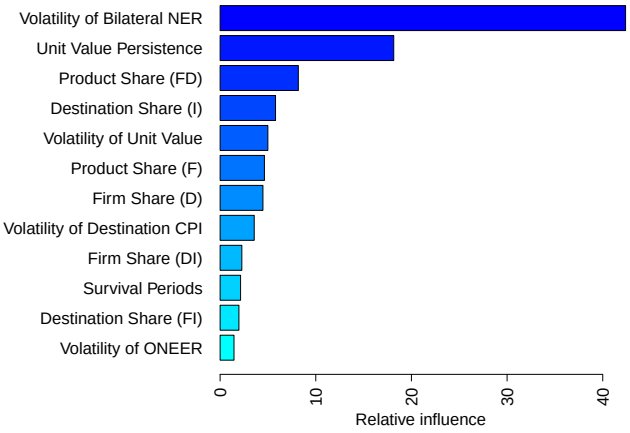
Stage 1:



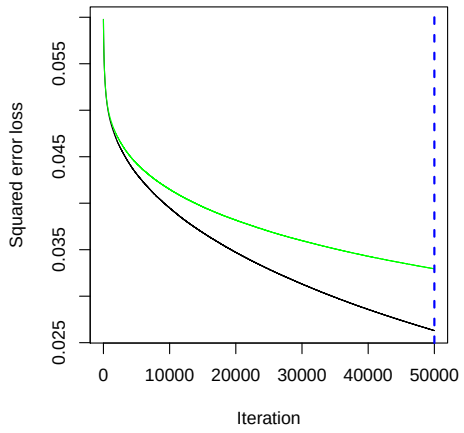


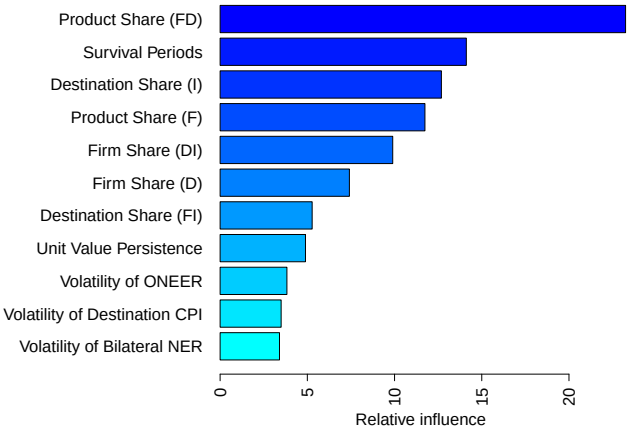
Stage 2:





Volatility:

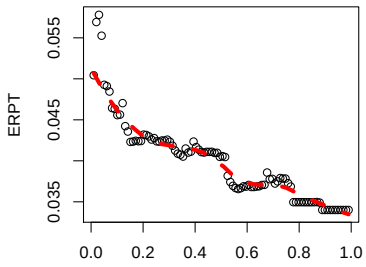


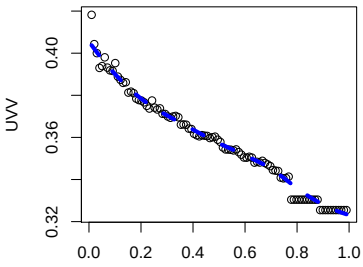


Note: The left panel presents cross validations of 3 models. The green and black line represent the cross-validation prediction error and within-sample prediction error respectively. The blue dashed line shows the optimal iteration indicated by cross validation errors. The right panel presents feature variables' contribution in error reduction. The supervisors in these three models are logged unit value, point estimate of ERPT and unit value volatility respectively. Unit value persistence and survival periods are measured at firm-product-destination level proxied by the frequency of trades and the number of observed trading periods respectively.

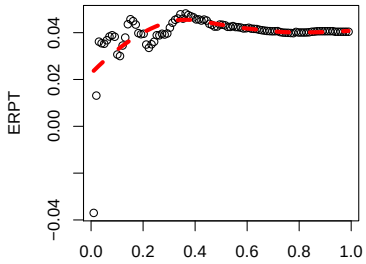
Table 10: Mapping firm-product-destination characteristics to ERPT (red) and unit value volatility (blue)

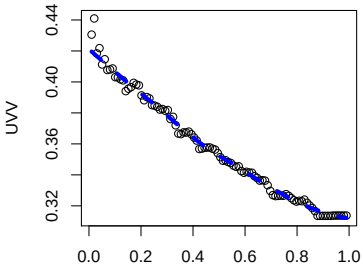
Classical Market Share (`fdi_di`)



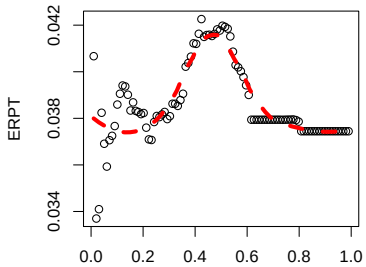


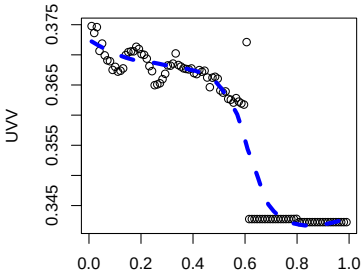
Local Core Product Measure (`fdi_fd`)



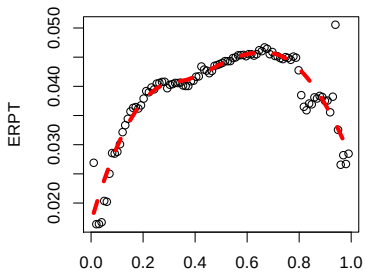


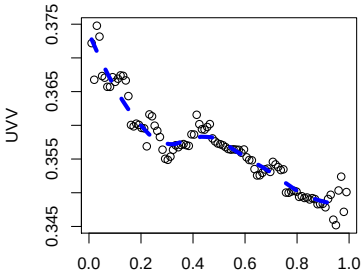
Destination Importance at Firm-Product Level (`fdi_fi`)



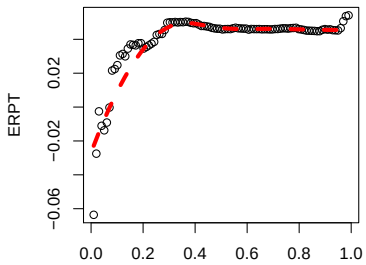


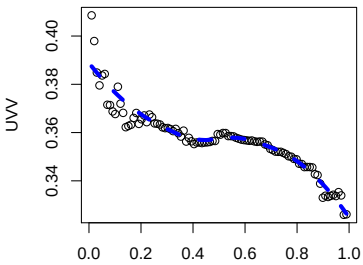
Local Firm Competitiveness (`fd_d`)



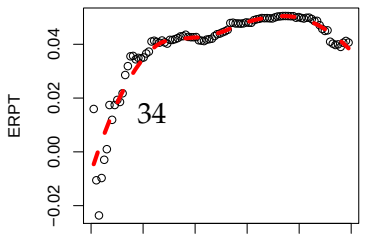


Global Core Product Measure (`fi_f`)





Destination Importance at Product Level (`di_i`)



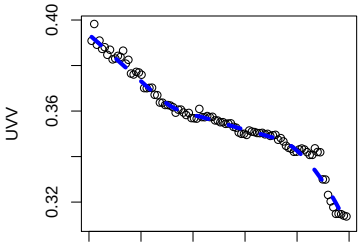
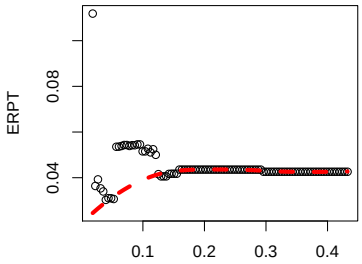
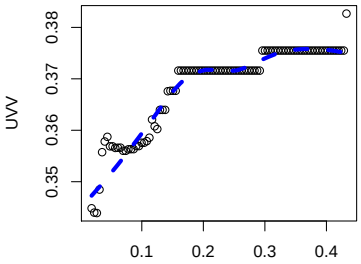


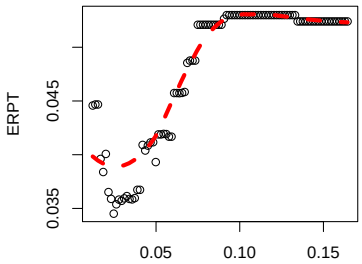
Table 11: Mapping firm-product-destination characteristics to ERPT (red) and unit value volatility (blue) (cont.)

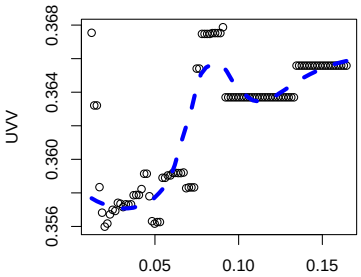
Volatility of Nominal Bilateral Exchange Rates



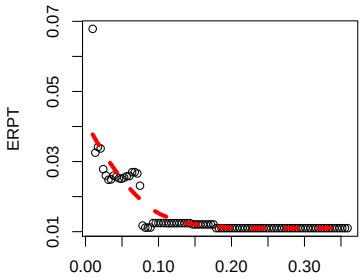


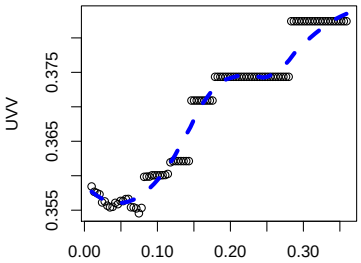
Volatility of Orthogonal Destination NEER



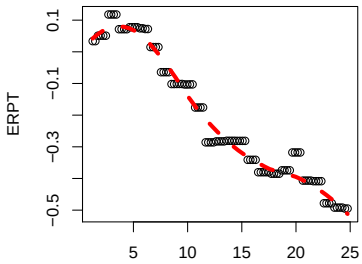


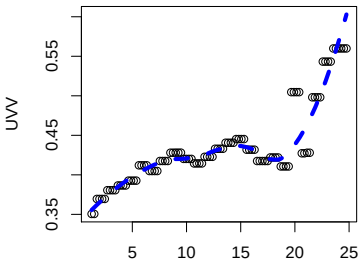
Volatility of Destination CPI



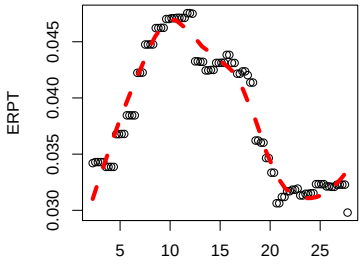


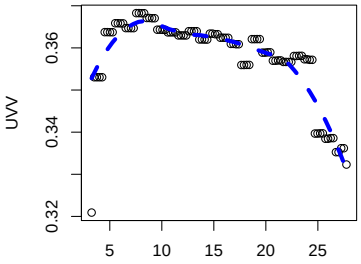
Frequency of Trades





Observed Trading Periods





Note: The x-axis of the graphs for the three volatility measures represents the standard deviation of logged macro price indicators. The x-axis of the group for the frequency of trade represents the period gap (in quarters) between two observations at the firm-product-destination level. The x-axis of the bottom two graphs represents the total number trade records observed in the sampling period at the firm-product-destination level. The circled-dots represent the estimated ERPT and unit value volatility respectively. The dashed coloured line represents the smoothed version using second order polynomials. A pass through value of 0.05 means that the RMB price goes up by 0.05% in reaction to a 1% bilateral exchange rate shock, i.e. a 95% destination country pass through. The median of the standard deviation of logged unit values at firm-product-destination level is around 0.36.

5 Conclusion

This paper differs from existing methodologies in emphasizing a holistic approach to estimating ERPT and proposes a machine learning algorithm to study the heterogeneity in ERPT at firm-level.

The core of the proposed algorithm consists of two elements. First, I find that the fact that tree based algorithms are robust to monotonic transformations of its feature variables can be exploited to control for unobserved components. Second, in a multi-dimensional panel, estimates from structural estimations in a range of limited-dimensional spaces can help to restrain the behaviour of unobserved components.

This paper extends Atkeson and Burstein (2008) and builds a multi-sector multi-country model to study how markets reach equilibrium under Cournot competition. From the simulated model, I construct a dataset that resembles available information in the real custom database to test the performance of the algorithm under complicated scenarios. The proposed method shows an extremely high accuracy rate on estimating firm-level ERPT and approximating the relationship between ERPT and the destination market share.

Applying the algorithm to China's custom data from 2000-2006, this paper documents new evidence on the relationships among various market share measures, firm-product-destination characteristics, unit value volatility and ERPT.

A Introduction to Classification And Regression Tree and Gradient Boosting Models

Classification And Regression Tree (CART)[55], a method of supervised learning, is a recursive binary splitting algorithm producing nonparametric mapping functions from independent variables (feature variables) to the dependent variable. Depending on the type of the dependent variables, tree based models are divided into classification trees (discrete dependent variable) and regressions trees (continuous dependent variable). Tree based methods are excellent at accommodating interactions between variables and complex nonlinear structures as well as handling outliers and missing observations. Modern decision tree algorithms are introduced by Breiman et al. (1984) and Friedman, Hastie and Tibshirani (2001).

In a decision tree algorithm, the dataset is binary partitioned sequentially until certain stop criterion has been met. At each partition, the algorithm will search all possible splits for all feature variables and select the split that minimises the prediction error. The procedure of a basic decision tree algorithm is given as follows:

$$MPE = \sum_{\tau \in leaves(T)} \sum_{i \in \tau} h(y - m_c)$$
$$m_c = \frac{1}{n_c} \sum_{i \in \tau} p_i$$

A decision tree algorithm recursively binary splits/partitions data at the point which minimises the mean prediction error (MPE) measured by criteria $h(.)$[56]. (1) The algorithm starts a tree of single node containing all points. If all the points in the node have the same value for all the input variables, stop. (2) Search over all binary splits of all variables for the one which reduces MPE as much as possible. If the largest decrease in MPE is below some threshold, or one of the resulting nodes contains fewer than q points, stop. Otherwise, take that split, creating two new nodes. (3) In each new node, go back to step 1.

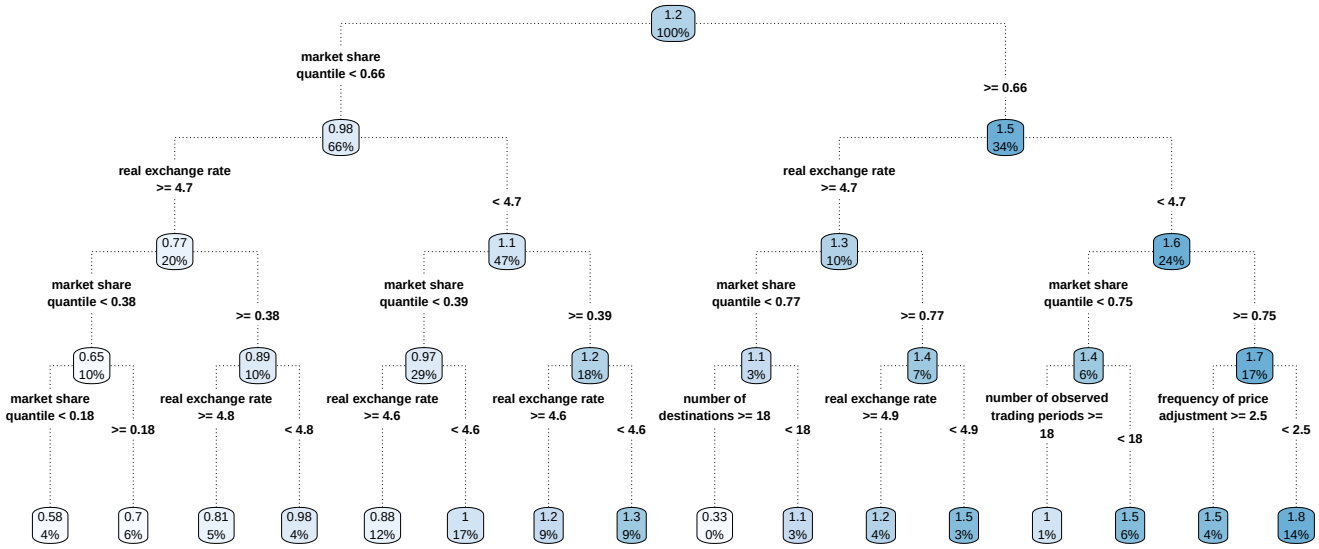
Figure 5 shows the results from applying CART to analyse the factors explaining the variation in export prices (unit values at firm-product level) of China's exporters. Entered feature/explanatory variables include the quantiles of market shares[57], logged real bilateral exchange rate, frequency of unit value adjustment at firm-product-destination level, number of observed trading periods and number of exporting destinations (during the period 2000-2006) at firm level. As can be seen from figure 5, the first split is made at the quantile of market shares. The algorithm predicts a higher average unit value for firms with high market shares among Chinese exporters. After the first split, several more splits are made sequentially in each subgroup based on other feature/explanatory variables. There is an interesting pattern for the last set of splits made for the left branch (market share quantile < 0.66) and the right

[55] A commonly used alias is "decision tree" algorithm. In my following discussions, I will refer to this algorithm as "decision tree" algorithm or tree based algorithm.

[56] Commonly used functions include $h(.) = |p_i - m_c|$ and $h(.) = (p_i - m_c)^2$.

[57] I use the classical market share measure at firm-product-destination level $\equiv \frac{\sum_t V_{i,f,d,t}}{\sum_f \sum_t V_{i,f,d,t}}$ calculated among China's exporters.

Figure 5: Predicting export unit values



Note: Calculation is based on quarterly data of China's import and export database 2000-2006. Unit values are measured in US dollars. The number in the circled note represents the average unit value of the classified group. The percentage below the number shows the proportion of data (counted by number of observations) located in this classification. Light (dark) blue indicates low (high) average unit values.

branch (market share quantile ≥ 0.66). The left branch suggests that the unit value variation is mostly explained by market share and real exchange rates variations for those firm-product combinations with small market shares, whereas the right branch suggests that other firm-product characteristics, such as the number of exported destinations and the frequency of price adjustments, start to play a role after the first few splits for those firm-product combinations with large market shares[58].

There are three advantages of tree based algorithms.

First, the binary splitting rule represents a natural decision-making process and the resulting tree structure is easy to understand and interpret.

Second, the recursive binary splitting feature makes decision tree methods a natural nonlinear estimator. Interactions between variables are accounted from the sequential feature of the partition process as the next partition depends on the previous partitions being made. "Trees tend to work well for problems where there are important nonlinearities and interactions." Tree based algorithms can discover nonlinear patterns that conventional econometric methods may fail to detect. More discussions can be found in Varian (2014).

Third, tree based models are robust to certain types of outliers and irregularities of data. Due to the binary splitting structure, only ordinal information of explanatory variables is used. Therefore, the resulting tree structure is robust to monotonic transformation of the explanatory variables. As discussed

[58]Please note that these results represent statistical relationships between variables only. As these classifications are not conditional on the characteristics of firms, products, destination competition environments, no further economic inference should be made based on these results.

in subsection 2.2, this property can be exploited to control for unobserved variables.

Given these advantages, empirically applying a decision tree algorithm also has various problems. First, finding the optimal decision tree in a large dataset is computationally difficult[59]. Second, practical decision tree solutions often lead to a local rather than global optimisation. Third, the algorithm is sensitive to small changes of the ordinal structure of explanatory variables. The resulted tree structure is often sensitive to its initial splits. More details and discussions of tree based algorithms can be found in Rokach and Maimon (2005).

The above problems can be overcome by various machine learning techniques including bagging, stacking, model averaging, random forest and boosting, etc.[60] The gradient boosting model, introduced by Friedman (2002)[61], is one of the most effective algorithms.

The boosting algorithm is based on the idea that adaptively integrating many small models can achieve and even outperform the predictive power of a single big model. Gradient boosting regression tree (GBRT) algorithm combines elements of gradient boosting and decision tree algorithms. In GBRT, trees are grown sequentially: each tree is grown conditional on the classification from previously grown trees. Adding the boosting procedure makes tree based models more robust, less path dependent and easy to work with large datasets.

The procedure of a workhorse GBRT algorithm is given as follows. A GBRT algorithm is a numerical optimisation technique with the objective to find the mapping $f(\mathbf{x})$ to minimise the expected loss function Ψ by sequentially adding a new tree that best reduces the gradient of the loss function:

$$\widehat{f}(\mathbf{x}) = arg\min_{f(\mathbf{x})} E_{y,\mathbf{x}}\Psi(y, f(\mathbf{x})) \tag{15}$$

The algorithm starts by initialising $\widehat{f}(\mathbf{x})$ to be a constant and iterating the following steps until reaching the specified $Iter_{max}$.

1. Compute the negative gradient as the working response

$$h_i = -\frac{\partial}{\partial f(\mathbf{x_i})}\Psi(p_i, f(\mathbf{x_i}))\Big|_{f(\mathbf{x_i})=\widehat{f}(\mathbf{x_i})} \tag{16}$$

2. Randomly select a fraction `bf` from the dataset (Random Forest/Bagging)

3. Fit a regression tree with `inter.depth` splits , $g(\mathbf{x})$, predicting h_i from the covariates $\mathbf{x_i}$.

4. Update the estimate of $f(\mathbf{x})$ as

$$\widehat{f}(\mathbf{x}) \to \widehat{f}(\mathbf{x}) + \texttt{lr} * g(\mathbf{x}) \tag{17}$$

5. Repeat step 1-4 until $Iter_{max}$

[59]There have been papers proving that finding an optimal decision tree from a given data is NP-hard or NP-complete under different scenarios. See Hyafil and Rivest (1976) and Hancock et al. (1996)

[60]See Breiman (1996) and Breiman (2001)

[61]Freund and Schapire (1996) developed the first two-class boosting classification algorithm called AdaBoost.

	Φ	iter	inter.depth	lr	bf
Benchmark	Normal	Cross Validation	8	0.01	0.5
Robustness	-	50,000	1-10	0.005, 0.001	0.3, 1

A GBRT model is calibrated with four parameters.[62] First, the bagging fraction, bf.[63] Second, a parameter controls the depth of interactions between variables, inter.depth. Third, the shrinkage or the learning rate, lr controls the weight of each iteration and a higher value means a quicker convergence rate. Fourth, the distribution Φ of the error term defines the loss function Ψ.[64]

The optimal number of iterations is often selected by the cross validation process. The model is first run with large number of iterations and the best iteration iter is then selected with k-fold cross validations.

It is worth noticing that the difference in ideology of modeling between machine learning and economic models. Economic and most econometric models start with structural assumptions reflecting economists understanding of how the world operates. Machine learning models, on the other side, assume that the true data-generating process is infinitely complex and all variables in the model are correlated in a nonlinear manner. Machine learning approaches try to maximally recover the ground truth by appointing a learning algorithm (a classier) to learn the relationship between variables. More specifically, the objective of a learning algorithm is to recover patterns among variables with the performance evaluated by prediction/classification errors in a given dataset. The cost of such non-parametric ideology of machine learning approaches is data driven, i.e. the ability of an algorithm to describe the ground truth of the world depends critically on the quality of data being supplied. If an important variable is not observed in the dataset, conventional machine learning approaches fail to capture the information in this variable and the resulting model is less satisfactory. In this aspect, it is worth designing an approach to integrate structural economic models and machine learning algorithms.

[62] Elith, Leathwick and Hastie (2008) provide a good introduction for modeling tuning practices. Ridgeway (2007) provides a good guidance on modeling tuning for the gbm package in R.

[63] This parameter helps to ensure the robustness of the model and prevents overfitting. The conventional value is $0.3 - 0.5$.

[64] For example, Gaussian implies a squared error loss function.

B Proof of Proposition 1 and Simulations of Example 1

Proof of Proposition 1. The first split is made at the point m_1. This implies

$$\frac{1}{|\{i: a_i \le a_{m_1}\}|}\sum_{i:a_i\le a_{m_1}} g[f(a_i) - \frac{\sum_{i:a_i\le a_{m_1}} f(a_i)}{|\{i: a_i \le a_{m_1}\}|}] + \frac{1}{|\{i: a_i > a_{m_1}\}|}\sum_{i:a_i> a_{m_1}} g[f(a_i) - \frac{\sum_{i:a_i> a_{m_1}} f(a_i)}{|\{i: a_i > a_{m_1}\}|}]$$
$$< \frac{1}{|\{i: a_i \le a_q\}|}\sum_{i:a_i\le a_q} g[f(a_i) - \frac{\sum_{i:a_i\le a_q} f(a_i)}{|\{i: a_i \le a_q\}|}] + \frac{1}{|\{i: a_i > a_q\}|}\sum_{i:a_i> a_q} g[f(a_i) - \frac{\sum_{i:a_i> a_q} f(a_i)}{|\{i: a_i > a_q\}|}] \quad \forall q \ne m_1 \tag{18}$$

where $g(.)$ is a loss function. Let $\{b_n\} = \urcorner(\{a_n\})$.

$a_j > a_i \Rightarrow b_j > b_i \quad \forall i, j \in \{1, ..., N\}$ implies

$$\{i: a_i < a_j\} \subseteq \{i: b_i < b_j\} \quad and \quad \{i: a_i > a_j\} \subseteq \{i: b_i > b_j\} \quad \forall j \in \{1, ..., N\} \tag{19}$$

Suppose that the transformation $\{b_n\}$ falls into the first category that $a_j > a_i \Rightarrow b_j > b_i \quad \forall i, j \in \{1, ..., N\}$ and the first optimal splitting point of $\{b_n\}$ is $b_{m_1^*}$. As the splitting criterion only uses the ordinal information, it can be written as

$$\frac{1}{|\{i: b_i \le b_{m_1^*}\}|}\sum_{i:b_i\le b_{m_1^*}} g[f(a_i) - \frac{\sum_{i:b_i\le b_{m_1^*}} f(a_i)}{|\{i: b_i \le b_{m_1^*}\}|}] + \frac{1}{|\{i: b_i > b_{m_1^*}\}|}\sum_{i:b_i> b_{m_1^*}} g[f(a_i) - \frac{\sum_{i:b_i> b_{m_1^*}} f(a_i)}{|\{i: b_i > b_{m_1^*}\}|}]$$
$$< \frac{1}{|\{i: b_i \le b_q\}|}\sum_{i:b_i\le b_q} g[f(a_i) - \frac{\sum_{i:b_i\le b_q} f(a_i)}{|\{i: b_i \le b_q\}|}] + \frac{1}{|\{i: b_i > b_q\}|}\sum_{i:b_i> b_q} g[f(a_i) - \frac{\sum_{i:b_i> b_q} f(a_i)}{|\{i: b_i > b_q\}|}] \quad \forall q \ne m_1^* \tag{20}$$

I want to prove $m_1 = m_1^*$.
(18) and (20) imply

$$\{i: a_i < a_{m_1}\} = \{i: b_i < b_{m_1^*}\} \tag{21}$$
$$\{i: a_i > a_{m_1}\} = \{i: b_i > b_{m_1^*}\} \tag{22}$$

(19) implies

$$\{i: a_i < a_{m_1}\} \subseteq \{i: b_i < b_{m_1}\} \quad and \quad \{i: a_i < a_{m_1^*}\} \subseteq \{i: b_i < b_{m_1^*}\} \tag{23}$$
$$\{i: a_i > a_{m_1}\} \subseteq \{i: b_i > b_{m_1}\} \quad and \quad \{i: a_i > a_{m_1^*}\} \subseteq \{i: b_i > b_{m_1^*}\} \tag{24}$$

(21) and (23), and (22) and (24) imply

$$\{i : a_i < a_{m_1^*}\} \subseteq \{i : b_i < b_{m_1^*}\} = \{i : a_i < a_{m_1}\} \subseteq \{i : b_i < b_{m_1}\} \tag{25}$$

$$\{i : a_i > a_{m_1^*}\} \subseteq \{i : b_i > b_{m_1^*}\} = \{i : a_i > a_{m_1}\} \subseteq \{i : b_i > b_{m_1}\} \tag{26}$$

From which, it can be derived that $a_{m_1} = a_{m_1^*}$ and $b_{m_1} = b_{m_1^*}$. Because

$$\{i : a_i < a_{m_1^*}\} \subseteq \{i : a_i < a_{m_1}\} \Rightarrow a_{m_1^*} \leq a_{m_1} \tag{27}$$

$$\{i : a_i > a_{m_1^*}\} \subseteq \{i : a_i > a_{m_1}\} \Rightarrow a_{m_1^*} \geq a_{m_1} \tag{28}$$

(25) and (26) can be simplified as

$$\{i : a_i < a_{m_1^*}\} = \{i : b_i < b_{m_1^*}\} = \{i : a_i < a_{m_1}\} = \{i : b_i < b_{m_1}\} \tag{29}$$

$$\{i : a_i > a_{m_1^*}\} = \{i : b_i > b_{m_1^*}\} = \{i : a_i > a_{m_1}\} = \{i : b_i < b_{m_1}\} \tag{30}$$

which implies

$$\{i : a_i = a_{m_1}\} = \{i : a_i = a_{m_1^*}\} \tag{31}$$

By the uniqueness of m_1, we have (from 18)

$$\{i : a_i > a_{m_1}\} \neq \{i : a_i > a_q\} \quad \forall q \in \{1, ..., N\} \neq m_1 \tag{32}$$

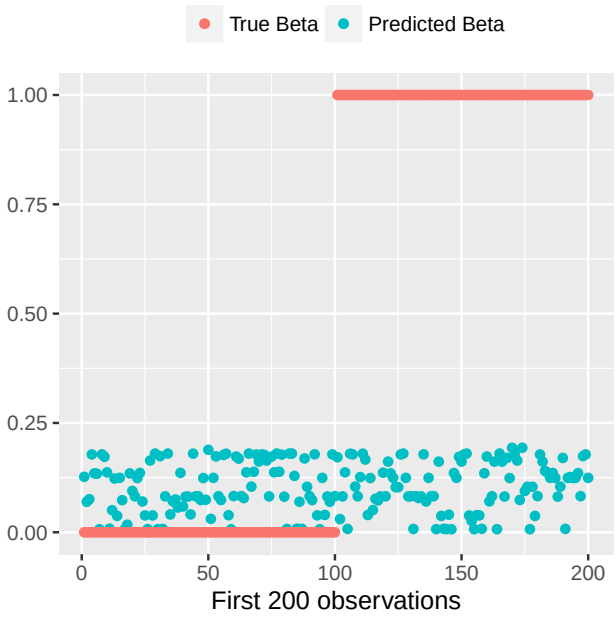
Therefore,

$$m_1 = m_1^* \tag{33}$$

The case where the transformation $\{b_n\}$ falls into the second category that $a_j > a_i \Rightarrow b_j < b_i \quad \forall i, j \in \{1, ..., N\}$ can be proved following a similar procedure. Recursively applying this argument to the rest of splitting points $a_{m_2}, ..., a_{m_k}$ completes the proof.

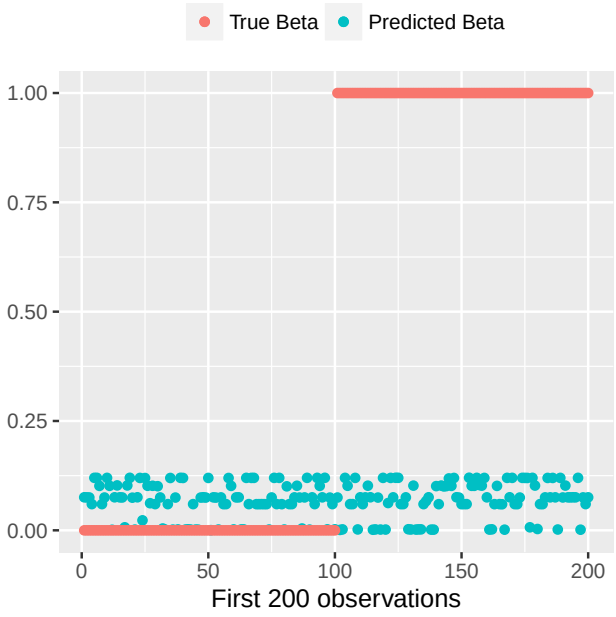
Remark. Intuitively, since a binary splitting algorithm only uses ordinal information of an explanatory variable in its splitting criteria, any transformation that preserves the ordinal information of this explanatory variable should result in the same splitting points.

Figure 6: Simulation of example 1: the ordered case

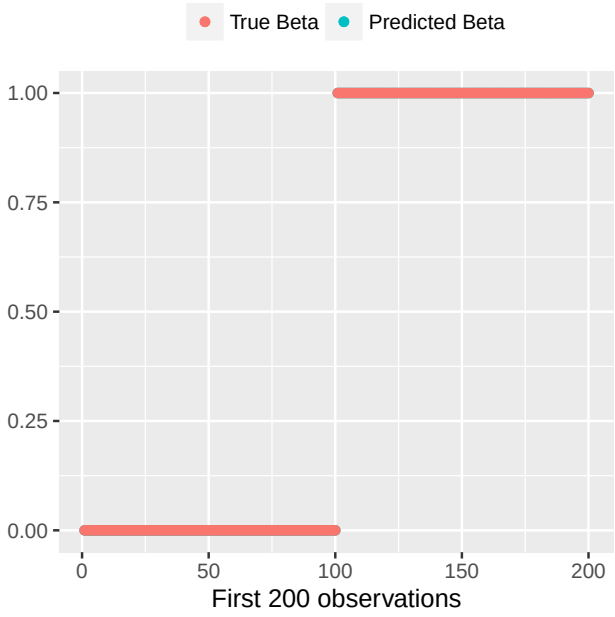


(a) No index variable

(b) Add index i

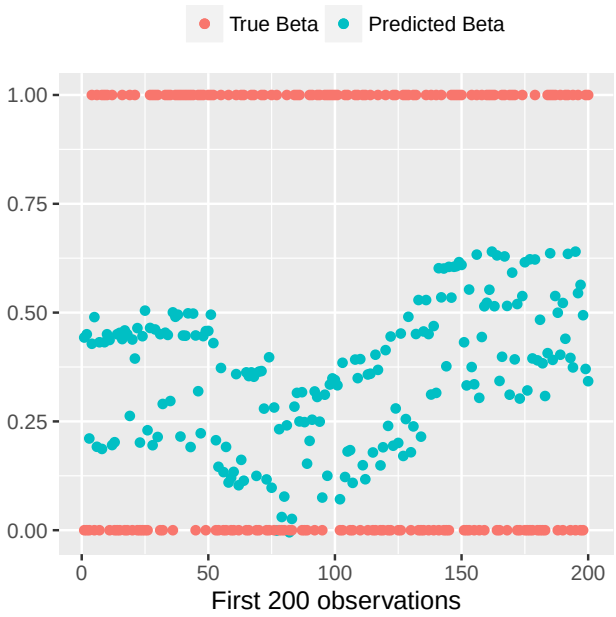


(c) Add dummy variables of i

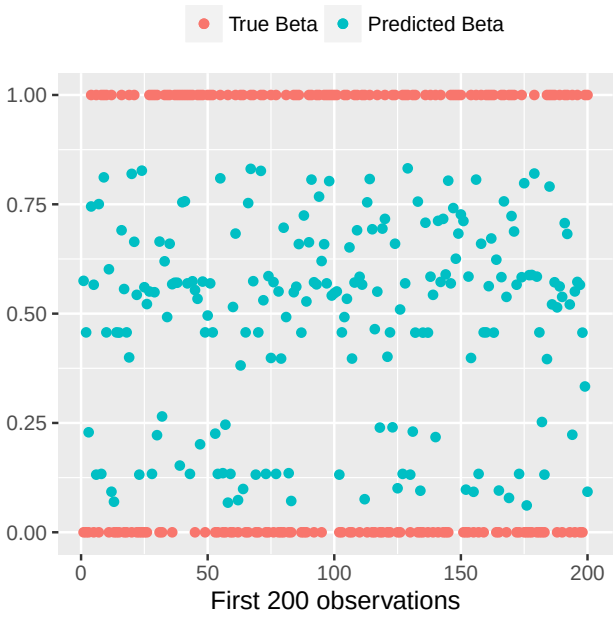


(d) Add true β_d

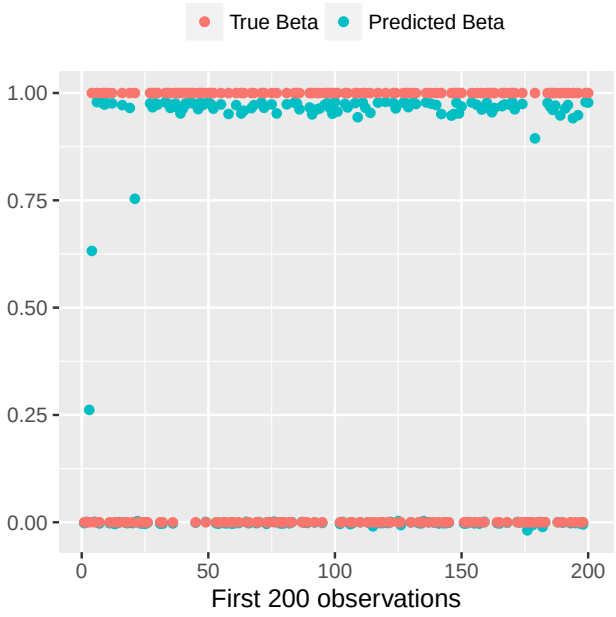
Figure 7: Simulation of example 1: the randomly assigned case



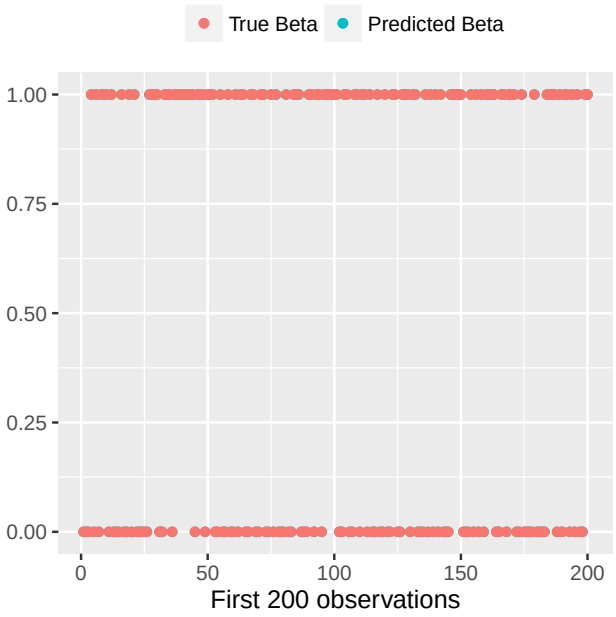
(a) Add index variable i



(b) Add dummy variables of i



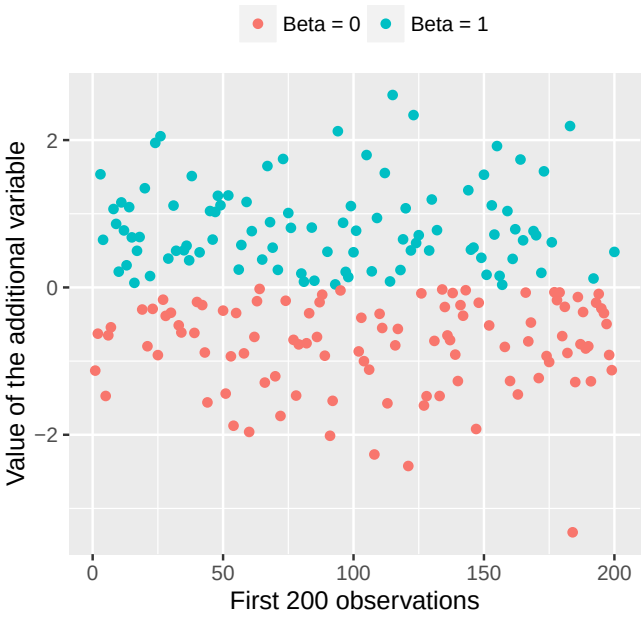
(c) Add $\mathfrak{M}_i$



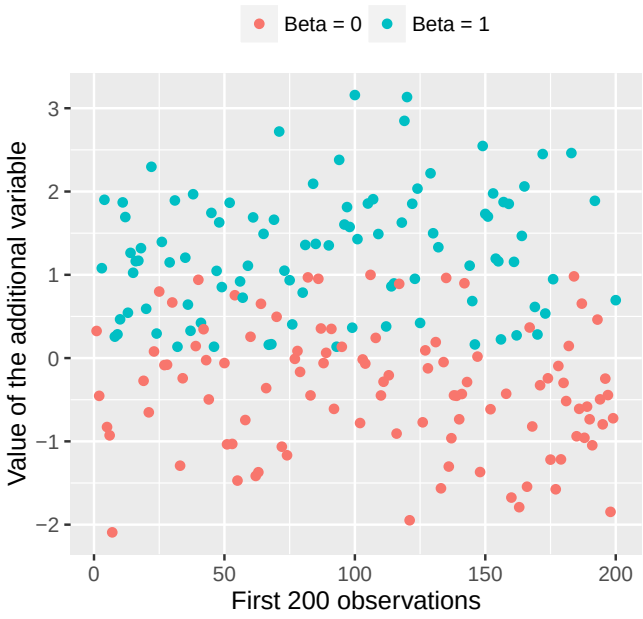
(d) Add true β_d

where $\mathfrak{M}_i = -|\epsilon|$ if $\beta_d = 0$ and $\mathfrak{M}_i = |\epsilon|$ if $\beta_d = 1$ $\epsilon; \sim N(0,1)$

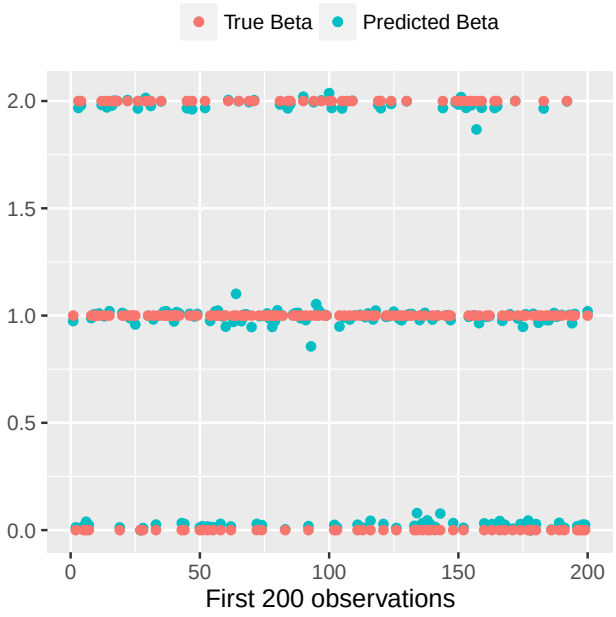
Figure 8: Simulation of example 1: the property of weak monotonic transformation



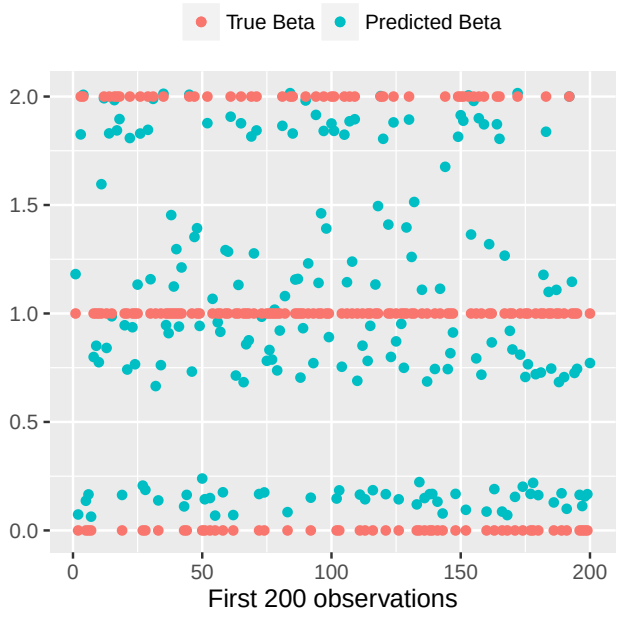
(a) Value of $\mathfrak{M}_d^1$



(b) Value of $\mathfrak{M}_i^2$



(c) Add $\mathfrak{M}_d^1$ as control

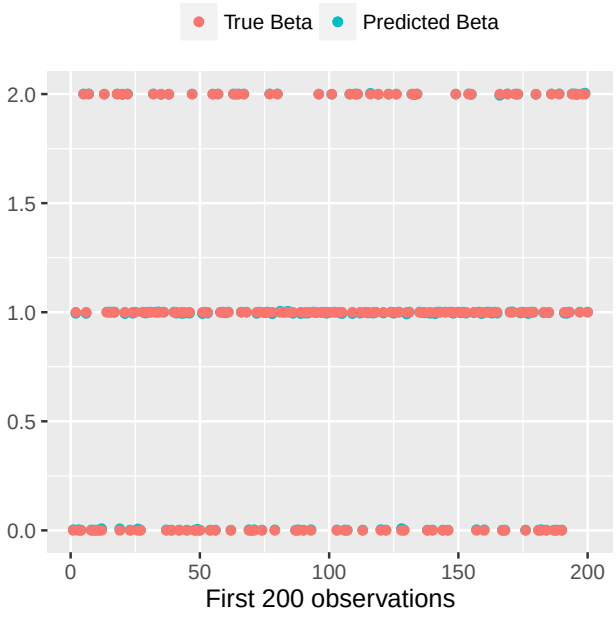


(d) Add $\mathfrak{M}_i^2$ as control

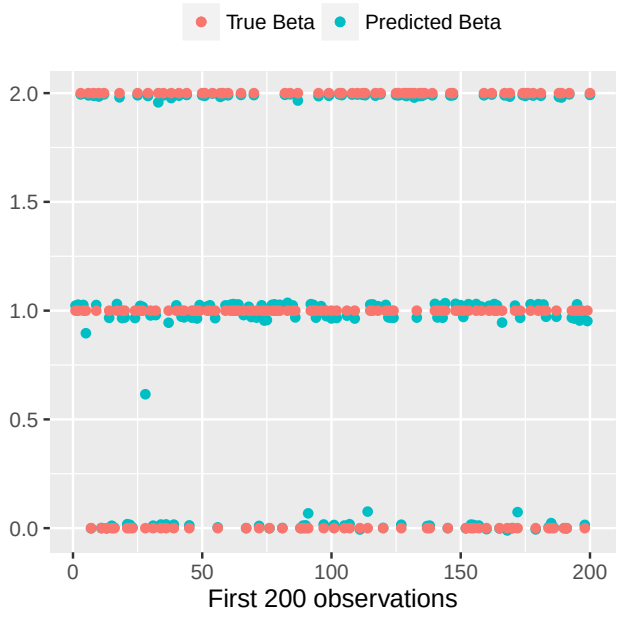
Figure 9: Simulation of example 1: the property of weak monotonic transformation; increase n to 2,000

(a) Value of $\mathfrak{M}_d^1$

(b) Value of $\mathfrak{M}_i^2$



(c) Add $\mathfrak{M}_d^1$ as control



(d) Add $\mathfrak{M}_i^2$ as control

C An Analytical Discussion of Simple Cases of the Two Dimensional Example

The key issue here is to understand how and when adding estimated parameters from structural estimations can help to control for the unobserved variables and estimate $\beta_{d,t}$. In the following subsection, I will discuss a simple case where the unobserved variable does not appear in the outer part of the equation of $p_{d,t}$.[65] For the data generating process given in equation (34) and (35), I want to estimate how the price $p_{d,t}$ changes under an exchange rate $e_{d,t}$ shock conditioning on values of M_t and $X_{d,t}$. However, M_t is unobserved. The objective is to understand how and when conditioning on values of coefficients $b_t^0, b_t^1, b_d^0, b_d^1$ and $X_{d,t}$ could achieve the same result as conditioning on values of M_t and $X_{d,t}$.

$$p_{d,t} = \beta_{d,t} e_{d,t} \tag{34}$$

$$\beta_{d,t} = f(X_{d,t}, M_t) \tag{35}$$

In the following discussion, I will try to express the true $\beta_{d,t}$ as a function of regression estimated coefficients.[66] In order to be clear about the information contained in estimated regression coefficients, I take first order approximations to decompose the data generating process into factors that only vary in one dimension. With this approximation, I will be able to obtain analytical solutions expressing values of b_t^1 and b_d^1 as $\beta_{d,t}$.[67] The gain in efficiency by adding coefficients versus using the indices will depend on the complexity of the hidden function implied by estimated coefficients versus the complexity of the hidden function implied by indices.

The coefficients of obtained from step 1 can be written as:

$$b_t^1 = \sum_i \beta_{d,t} w_{d,t} \quad b_d^1 = \sum_t \beta_{d,t} \omega_{d,t} \tag{36}$$

$$w_{d,t} := \frac{e_{d,t}(e_{d,t} - \overline{e_t})}{\sum_i (e_{d,t} - \overline{e_t})^2} \quad \omega_{d,t} := \frac{e_{d,t}(e_{d,t} - \overline{e_i})}{\sum_t (e_{d,t} - \overline{e_i})^2} \tag{37}$$

where $\overline{e_t} = \sum_i e_{d,t}/n_I$ with n_I being the number of observations at dimension i.[68] As I will illustrate below, by conditioning on b_t^1 and b_d^1, I am essentially conditioning on values of $\beta_{d,t}$ with a particular weight.

Case A: If weights $w_{d,t} = 1/n_I$ and $\omega_{d,t} = 1/n_T$ and $\beta_{d,t}$ can be approximated by $\beta_{d,t} = v_d + v_t + v_d * v_t$, then $\overline{\beta_t} = \overline{v^d}(1 + v_t)$ and $\overline{\beta_d} = (v_d + 1)\overline{v^t}$, where $\overline{v^t} := \sum_t v_t/n_T$.

[65] In this simple case, I only need to make counterfactual predictions of $p_{d,t}$ changing $e_{d,t}$ conditioning on the value of $\beta_{d,t}$. When the unobserved variable M_t^{outer} does appear in the outer part of the equation of $p_{d,t}$, I will need to make predictions conditioning on the values of both $\beta_{d,t}$ and M_t^{outer}. In general, the missing component in the inner part of $\beta_{d,t}$ and the outer part M_t^{outer} need not take the same functional form nor the same value.

[66] More formally, I should define a multi-dimensional monotonic transformation measure. I leave this task for my future work.

[67] In this simple case, I only need to use information provided by b_t^1 and b_d^1. b_t^0 and b_d^0 will be used in more complicated cases as in subsection C.1.

[68] I assume that the panel is balanced. I use notations n_I and n_T rather than the conventional N and T.

$$\frac{b_t^1}{\overline{v^d}} * \frac{b_d^1}{\overline{v^t}} - 1 = v_d + v_t + v_d * v_t = \beta_{d,t} \tag{38}$$

Case B: If $\beta_{d,t} = v_d + v_t + v_d * v_t$ but $w_{d,t} \neq 1/n_I$ and $\omega_{d,t} \neq 1/n_T$, then

$$b_t^1 = \sum_i (v_d + v_t + v_d * v_t) w_{d,t} = \sum_i v_d w_{d,t} + v_t (\sum_i w_{d,t} + \sum_i v_d w_{d,t}) \tag{39}$$

$$b_d^1 = \sum_t (v_t + v_d + v_d * v_t) \omega_{d,t} = \sum_t v_t \omega_{d,t} + v_d (\sum_t \omega_{d,t} + \sum_t v_t \omega_{d,t}) \tag{40}$$

Notice $\sum_t v_t \omega_{d,t} = 0$ and $\sum_i v_d w_{d,t} = 0$

This is still too abstract. Let $e_{d,t} = k_t + k_d + k_d * k_t$. Then $\overline{e_t} = \sum_i (k_t + k_d + k_d * k_t)/n_I$ and $(e_{d,t} - \overline{e_t})^2 = [k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t]^2$.

$$\begin{aligned} b_t^1 &= \frac{\sum_i (v_d + v_t + v_d * v_t)(k_t + k_d + k_d * k_t)[k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t]}{\sum_i [k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t]^2} \\ &= v_t + (1 + v_t) \frac{\sum_i v_d (k_t + k_d + k_d * k_t)[k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t]}{\sum_i [k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t]^2} \end{aligned} \tag{41}$$

$$\begin{aligned} b_d^1 &= \frac{\sum_t (v_d + v_t + v_d * v_t)(k_t + k_d + k_d * k_t)[k_t - \overline{k^T} + (k_t - \overline{k^T}) * k_d]}{\sum_t [k_t - \overline{k^T} + (k_t - \overline{k^T}) * k_d]^2} \\ &= v_d + (1 + v_d) \frac{\sum_t v_t (k_t + k_d + k_d * k_t)[k_t - \overline{k^T} + (k_t - \overline{k^T}) * k_d]}{\sum_t [k_t - \overline{k^T} + (k_t - \overline{k^T}) * k_d]^2} \end{aligned} \tag{42}$$

Define sample conditional covariance measures as[69]:

$$Cov_d(x_{d,t}, z_{d,t}) := \sum_i (x_{d,t} - \overline{x_t})(z_{d,t} - \overline{z_t})/n_I \tag{43}$$

$$Var_i(x_{d,t}) := Cov_d(x_{d,t}, x_{d,t}) \tag{44}$$

From the relationship $\sum_i x_{d,t} z_{d,t}/n_I = Cov_d(x_{d,t}, z_{d,t}) + \sum_i x_{d,t} \sum_i z_{d,t}/n_I^2$, b_t^1 and b_d^1 can be rewritten as:

$$b_t^1 = [v_t + (1 + v_t)\overline{v^d}] + (1 + v_t) \frac{Cov_d\{v_d, (k_t + k_d + k_d * k_t)[k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t]\}}{Var_i(k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t)} \tag{45}$$

$$b_d^1 = [v_d + (1 + v_d)\overline{v^t}] + (1 + v_d) \frac{Cov_t\{v_t, (k_t + k_d + k_d * k_t)[k_t - \overline{k^T} + (k_t - \overline{k^T}) * k_d]\}}{Var_t[k_t - \overline{k^T} + (k_t - \overline{k^T}) * k_d]} \tag{46}$$

[69] These definitions are used to simplify my notation only and may not be a consistent measure of conditional covariance.

Equations (45) and (46) can be expressed as[70]:

$$b_t^1 = [v_t + (1+v_t)\overline{v^d}] + (1+v_t)\frac{Cov_d[v_d, (e_{d,t} - \overline{e_t})^2]}{Var_i(e_{d,t} - \overline{e_t})^2} \tag{47}$$

$$b_d^1 = [v_d + (1+v_d)\overline{v^t}] + (1+v_d)\frac{Cov_t[v_t, (e_{d,t} - \overline{e_i})^2]}{Var_t(e_{d,t} - \overline{e_i})^2} \tag{48}$$

The last part can be further simplified as

$$\frac{Cov_d[v_d, (e_{d,t} - \overline{e_t})^2]}{Var_i(e_{d,t} - \overline{e_t})^2} = \frac{Cov_d[v_d, (k_d - \overline{k^I})^2]}{Var_i(k_d)} =: c_1 \tag{49}$$

$$\frac{Cov_t[v_t, (e_{d,t} - \overline{e_i})^2]}{Var_t(e_{d,t} - \overline{e_i})^2} = \frac{Cov_t[v_t, (k_t - \overline{k^T})^2]}{Var_t(k_t)} =: c_2 \tag{50}$$

In this example, due to the simple factorisation of $e_{d,t}$, I assume that the last part is a constant and does not vary along the other dimensions. This property no longer holds in a more general factorisation process, e.g. $e_{d,t} = k_t^1 + k_d^1 + k_d^1 k_t^2$.

In general, I discuss three possibilities here:

Case B.1: If $Cov_d[v_d, (e_{d,t} - \overline{e_t})^2] = 0$ and $Cov_t[v_t, (e_{d,t} - \overline{e_i})^2] = 0$, the true $\beta_{d,t}$ can be expressed as a simple nonlinear equation of b_t^1 and b_d^1 as in Case A.

Case B.2: If $Cov_d[v_d, (e_{d,t} - \overline{e_t})^2] \neq 0$ and $Cov_t[v_t, (e_{d,t} - \overline{e_i})^2] \neq 0$, but $Var_t\left\{\frac{Cov_d[v_d, (e_{d,t} - \overline{e_i})^2]}{Var_i(e_{d,t} - \overline{e_t})^2}\right\} = 0$ and $Var_i\left\{\frac{Cov_t[v_t, (e_{d,t} - \overline{e_i})^2]}{Var_t(e_{d,t} - \overline{e_i})^2}\right\} = 0$, v_d and v_t can be written as follows:

$$b_t^1 = v_t + (1+v_t)(\overline{v^d} + c_1) \tag{51}$$

$$b_d^1 = v_d + (1+v_d)(\overline{v^t} + c_2) \tag{52}$$

$$v_t = \frac{b_t^1 - (\overline{v^d} + c_1)}{1 + (\overline{v^d} + c_1)} \tag{53}$$

$$v_d = \frac{b_d^1 - (\overline{v^t} + c_2)}{1 + (\overline{v^t} + c_2)} \tag{54}$$

It is clear that $\beta_{d,t}$ can be expressed as a nonlinear function of b_t^1 and b_d^1.

Case B.3: If $\frac{Cov_d[v_d, (e_{d,t} - \overline{e_t})^2]}{Var_i(e_{d,t} - \overline{e_t})^2} \neq c_1$ and $\frac{Cov_t[v_t, (e_{d,t} - \overline{e_i})^2]}{Var_t(e_{d,t} - \overline{e_i})^2} \neq c_2$ but $Var_t\left\{\frac{Cov_d[v_d, (e_{d,t} - \overline{e_t})^2]}{Var_i(e_{d,t} - \overline{e_t})^2}\right\}$ and $Var_i\left\{\frac{Cov_t[v_t, (e_{d,t} - \overline{e_i})^2]}{Var_t(e_{d,t} - \overline{e_i})^2}\right\}$ are very small, the weak monotonic property will make it work.

This exercise gives me two interesting insights. First, variances of the bias at the other dimension matter. As long as the bias of estimated parameters at one dimension is "well-structured" in the other dimension, adding these estimated parameters will help to estimate the desired $\beta_{d,t}$.

Second, it is the covariance between elements driving $\beta_{d,t}$ and a local measure of second moments

[70] Hint: rewrite $k_t + k_d + k_d * k_t = [k_d - \overline{k^I} + (k_d - \overline{k^I}) * k_t] + [k_t + \overline{k^I} + \overline{k^I} k_t]$

of the policy variable $(e_{d,t} - \overline{e_t})^2$ and $(e_{d,t} - \overline{e_i})^2$ that matters. Due to the linear regression structure, first order terms are filtered out and only second order terms will influence the bias. This is a very useful property for economics studies. Taking this property into the context of my empirical ERPT question, exchange rates may be correlated with the marginal cost of the exporter[71], but it is less likely for the volatility of exchange rates to be correlated with the level movement of the marginal cost of the exporter. Even if these two terms are correlated, as long as the correlations (as a function of d) do not change systematically across destinations, estimated parameters b_t^1 and b_d^1 will provide useful information which can be analysed through a tree based machine learning algorithm.

In general, adding estimated parameters from regressions and/or other structural estimations from a range of dimension-limited partition spaces should always be more efficient than adding indices i and t or dummies related to the indices provided that the assumed structural equation is not very far from the true specification.

C.1 The case where M_t appears in the outer part of the linear form

In this subsection, I discuss the case where M_t appears in the outer part of the linear form. As not only $\beta_{d,t}$ but also M_t need to be backed up from the estimated parameters, I also need to use information from b_t^0 and b_d^0.

$$
\begin{aligned}
p_{d,t} &= \beta_{d,t} e_{d,t} + M_t \\
\beta_{d,t} &= f(X_{d,t}, M_t)
\end{aligned}
$$

Regression estimated parameters can be written as

$$
\begin{aligned}
b_t^1 &= \frac{\sum_i [f(X_{d,t}, M_t) e_{d,t} + M_t](e_{d,t} - \overline{e_t})}{\sum_i (e_{d,t} - \overline{e_t})^2} \\
&= \sum_i \beta_{d,t} w_{d,t} + M_t \\
b_t^0 &= \sum_i (p_{d,t} - b_t^1 * e_{d,t})/n_I \\
&= \sum_i (\beta_{d,t} - \sum_i \beta_{d,t} w_{d,t}) e_{d,t}/n_I + M_t(1 - \sum_i e_{d,t}/n_I)
\end{aligned}
$$

$$
\begin{aligned}
b_d^1 &= \frac{\sum_t [f(X_{d,t}, M_t) e_{d,t} + M_t](e_{d,t} - \overline{e_i})}{\sum_t (e_{d,t} - \overline{e_i})^2} \\
&= \sum_t \beta_{d,t} \omega_{d,t} + \sum_t M_t \frac{\omega_{d,t}}{e_{d,t}} \\
b_d^0 &= \sum_t (p_{d,t} - b_d^1 * e_{d,t})/n_T \\
&= \sum_t (\beta_{d,t} - \sum_t \beta_{d,t} \omega_{d,t}) e_{d,t}/n_T + \sum_t M_t(1 - \omega_{d,t})/n_T
\end{aligned}
$$

[71] For example, the exporter may use imported inputs.

To visualise the underlying structure of these estimated parameters, I take the following first order factorisation. Let

$$
\begin{aligned}
e_{d,t} &= k_t + k_d + k_d * k_t \\
\beta_{d,t} &= v_d + v_t + v_d * v_t \\
M_t &= m_t
\end{aligned}
$$

Therefore,

$$b_t^1 = [v_t + (1+v_t)\overline{v^d}] + (1+v_t)\frac{Cov_d[v_d, (e_{d,t} - \overline{e_t})^2]}{Var_i(e_{d,t} - \overline{e_t})^2} + m_t \tag{55}$$

$$b_d^1 = [v_d + (1+v_d)\overline{v^t}] + (1+v_d)\frac{Cov_t[v_t, (e_{d,t} - \overline{e_i})^2]}{Var_t(e_{d,t} - \overline{e_i})^2} + \frac{Cov_t(m_t, e_{d,t})}{Var_t(e_{d,t} - \overline{e_i})^2} \tag{56}$$

$$b_t^0 = Cov_d[\beta_{d,t}, e_{d,t}] - (\frac{Cov_d[\beta_{d,t}, (e_{d,t} - \overline{e_t})e_{d,t}]}{Var_i(e_{d,t} - \overline{e_t})} + M_t)\overline{e_t} + M_t \tag{57}$$

where I have used the following relationship in deriving the expression of b_t^0.

$$\sum_i [\beta_{d,t}(e_{d,t} - \overline{e_t})e_{d,t}]/n_I = Cov_d[\beta_{d,t}, (e_{d,t} - \overline{e_t})e_{d,t}] + \sum_i \beta_{d,t} \sum_i [(e_{d,t} - \overline{e_t})^2]/n_I^2 \tag{58}$$

With the assumed factorisation,

$$Cov_d[\beta_{d,t}, e_{d,t}] = (1+v_t)(1+k_t)Cov_d(v_d, k_d) \tag{59}$$

$$Cov_d[\beta_{d,t}, (e_{d,t} - \overline{e_t})e_{d,t}] = (1+v_t)(1+k_t)^2 Cov_d[v_d, (k_d - \overline{k^I})^2] \tag{60}$$

$$\frac{Cov_d[\beta_{d,t}, (e_{d,t} - \overline{e_t})e_{d,t}]}{Var_i(e_{d,t} - \overline{e_t})} = (1+v_t)\frac{Cov_d[v_d, (k_d - \overline{k^I})^2]}{Var_i(k_d)} =: (1+v_t)c_1 \tag{61}$$

This is where it becomes complicated. The expression of b_t^0 now involves a time-varying factor k_t of the observed policy variable $e_{d,t}$.

$$b_t^0 = (1+v_t)(1+k_t)Cov_d(v_d, k_d) - [(1+v_t)c_1 + m_t](k_t + k_t\overline{k^I} + \overline{k^I}) + m_t \tag{62}$$

$Cov_d(v_d, k_d)$, c_1, $\overline{k^I}$, $\overline{v^d}$, $\overline{v^t}$ are constants. This leaves v_t, v_d, m_t, k_t to be solved in 3 equations (55), (56) and (62). The tricky part to figure out is how and when the conditional weak monotonic transformation property in proportion 2 works in this case.

If I derive the expression of b_d^0, it would involve v_d, k_d. Together with $e_{d,t}$, (55), (56) and (62), there are 5 equations with 5 unknowns v_t, v_d, m_t, k_d, k_t. If this problem can be solved approximately, the efficiency gain in adding these estimated parameters should depend on the complexity of these equations compared to the complexity of hidden functions of using indices or related dummies.

Note that I discussed a general case here where the unobserved variable M_t needs not necessarily be

correlated with $\beta_{d,t}$.[72] If $\beta_{d,t}$ can be expressed as an explicit function of M_t and some observed variables, the derivation will be easier.

[72] I did not impose any restrictions on m_t and v_t.

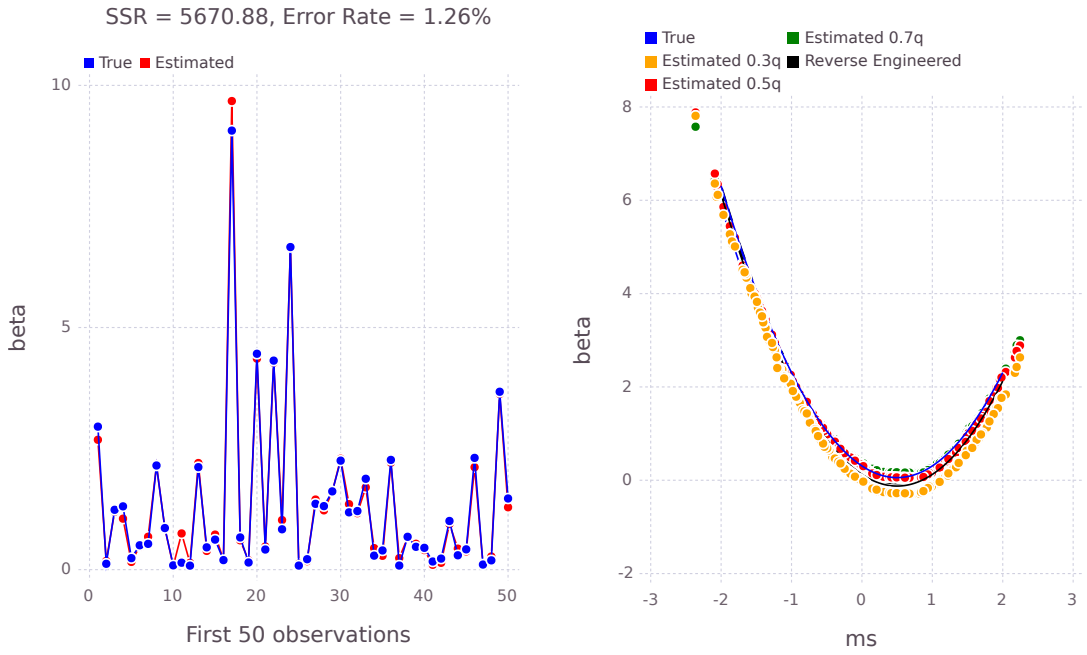
C.2 Tests on alternative specifications

C.2.1 High Nonlinearity

Setting:

$$
\begin{aligned}
p_{d,t} &= 10 + \beta_{d,t} e_{d,t} + ms_{d,t} - mc_t + \epsilon_{d,t} \\
\beta_{d,t} &= (ms_{d,t} - 0.5)^2 + sin(1000 mc_t) mc_t \\
ms_{d,t} &= u_{d,t} + 0.1 e_{d,t} \\
mc_t &= u_t - 0.1 \overline{e_t} \\
\overline{e_t} &= \frac{\sum_d e_{d,t}}{n_d} \\
n_d &= 2000;\ n_t = 40 \\
u_{d,t} &\sim N(0,1),\ u_t \sim N(0,1),\ \epsilon_{d,t} \sim N(0, 0.01)
\end{aligned}
$$

Figure 10: High nonlinearity: the proposed algorithm

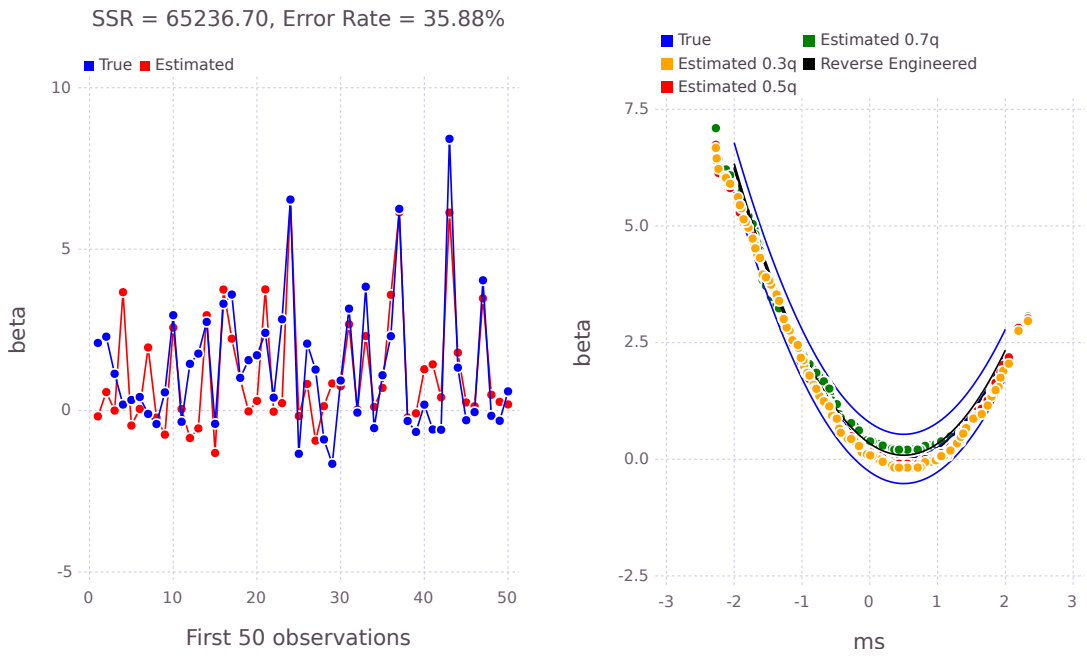


C.2.2 Not Identifiable

Setting:

$$
\begin{aligned}
p_{d,t} &= 10 + \beta_{d,t} e_{d,t} + ms_{d,t} - mc_{d,t} + \epsilon_{d,t} \\
\beta_{d,t} &= (ms_{d,t} - 0.5)^2 + mc_{d,t} \\
ms_{d,t} &= u_{d,t} + 0.1 e_{d,t} \\
mc_{d,t} &= u_{d,t} - 0.1 e_{d,t} \\
n_d &= 2000;\ n_t = 40 \\
u_{d,t} \sim & N(0,1),\ u_t \sim N(0,1),\ \epsilon_{d,t} \sim N(0, 0.01)
\end{aligned}
$$

Figure 11: Not identifiable: the proposed algorithm

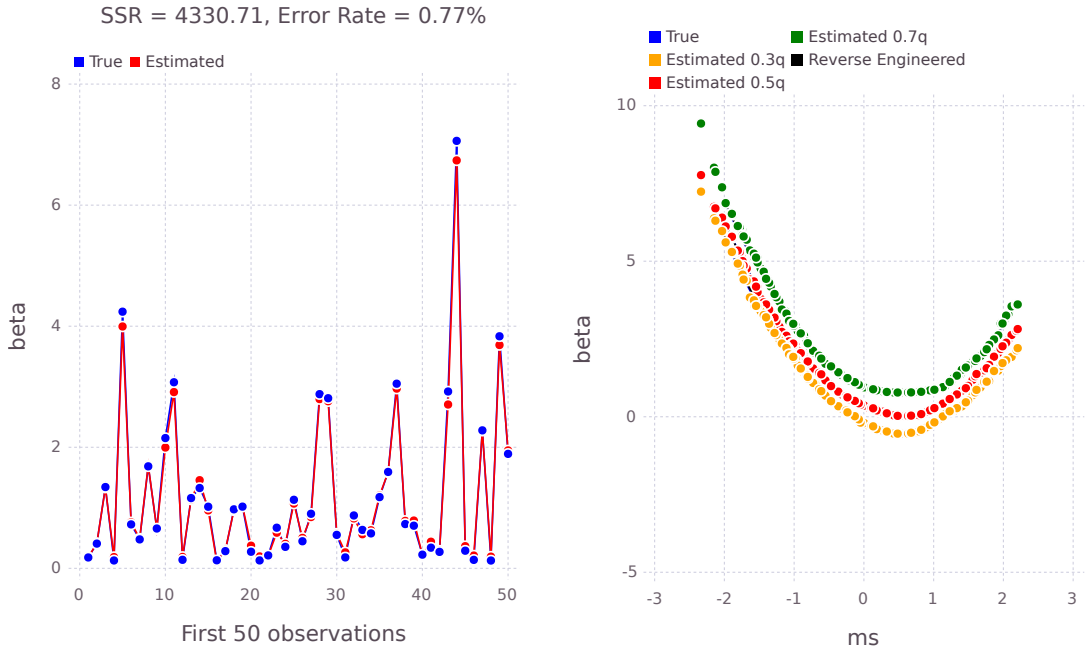


C.2.3 Larger Correlation

Setting:

$$
\begin{aligned}
p_{d,t} &= 10 + \beta_{d,t} e_{d,t} + ms_{d,t} - mc_t + \epsilon_{d,t} \\
\beta_{d,t} &= (ms_{d,t} - 0.5)^2 + mc_t \\
ms_{d,t} &= u_{d,t} + 0.1 e_{d,t} \\
mc_t &= u_t - 1\overline{e_t} \\
\overline{e_t} &= \frac{\sum_d e_{d,t}}{n_d} \\
n_d &= 2000;\ n_t = 40 \\
u_{d,t} \sim & N(0,1),\ u_t \sim N(0,1),\ \epsilon_{d,t} \sim N(0, 0.01)
\end{aligned}
$$

Figure 12: Larger correlation: the proposed algorithm

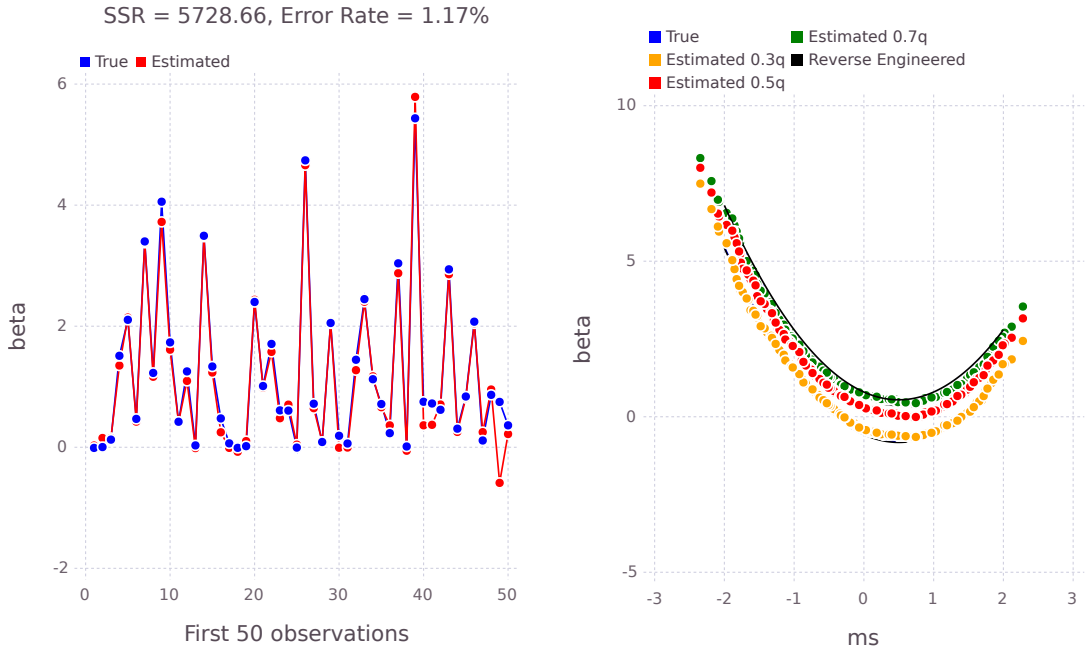


C.2.4 Different Function of the Outer Part

Setting:

$$
\begin{aligned}
p_{d,t} &= 10 + \beta_{d,t} e_{d,t} + ms_{d,t} - mc_t^2 + \epsilon_{d,t} \\
\beta_{d,t} &= (ms_{d,t} - 0.5)^2 + mc_t \\
ms_{d,t} &= u_{d,t} + 0.1 e_{d,t} \\
mc_t &= u_t - 0.1\overline{e_t} \\
\overline{e_t} &= \frac{\sum_d e_{d,t}}{n_d} \\
n_d &= 2000;\ n_t = 40 \\
u_{d,t} \sim & N(0,1),\ u_t \sim N(0,1),\ \epsilon_{d,t} \sim N(0, 0.01)
\end{aligned}
$$

Figure 13: Different function of the outer part: the proposed algorithm

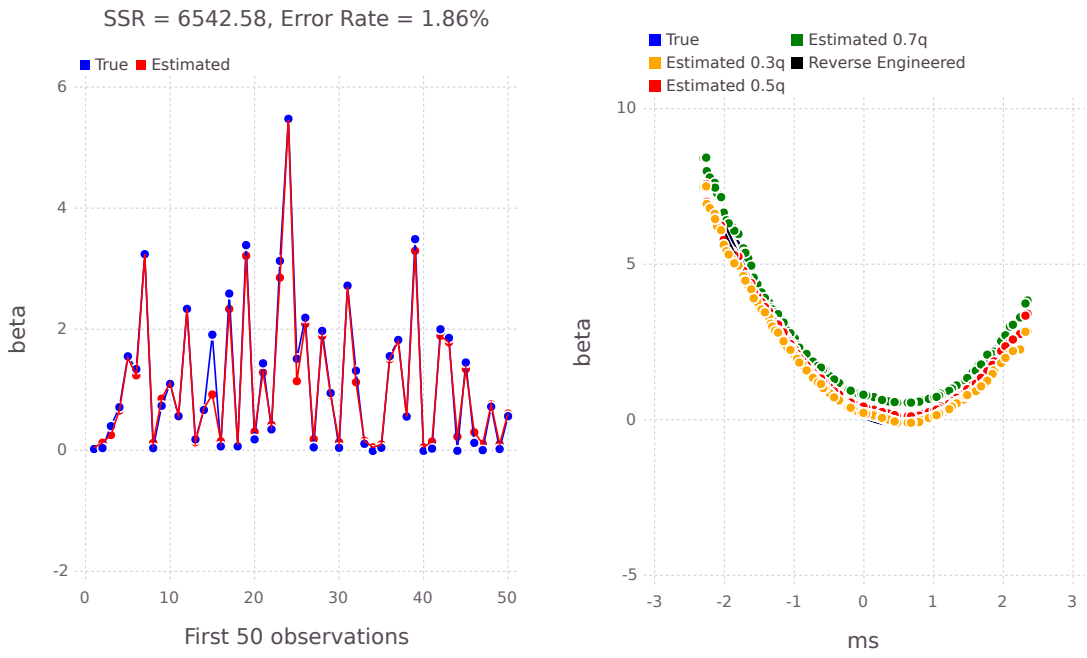


C.2.5 Arellano and Bond

Setting:

$$
\begin{aligned}
p_{d,t} &= 0.95p_{d,t-1} + \beta_{d,t}e_{d,t} + ms_{d,t} - mc_t + \epsilon_{d,t} \\
\beta_{d,t} &= (ms_{d,t} - 0.5)^2 + mc_t \\
ms_{d,t} &= u_{d,t} + 0.1e_{d,t} \\
mc_t &= u_t - 0.1\overline{e_t} \\
\overline{e_t} &= \frac{\sum_d e_{d,t}}{n_d} \\
n_d &= 2000;\ n_t = 40 \\
u_{d,t} &\sim N(0,1),\ u_t \sim N(0,1),\ \epsilon_{d,t} \sim N(0,0.01)
\end{aligned}
$$

Figure 14: Arellano and Bond: the proposed algorithm



C.2.6 Reduce Sample Size

Setting:

$$
\begin{aligned}
p_{d,t} &= 10 + \beta_{d,t} e_{d,t} + ms_{d,t} - mc_t + \epsilon_{d,t} \\
\beta_{d,t} &= (ms_{d,t} - 0.5)^2 + mc_t \\
ms_{d,t} &= u_{d,t} + 0.1 e_{d,t} \\
mc_t &= u_t - 0.1\overline{e_t} \\
\overline{e_t} &= \frac{\sum_d e_{d,t}}{n_d} \\
n_d &= 200;\ n_t = 40 \\
u_{d,t} &\sim uniform,\ u_t \sim uniform,\ \epsilon_{d,t} \sim N(0, 0.01)
\end{aligned}
$$

Figure 15: Reduce sample size: dummies

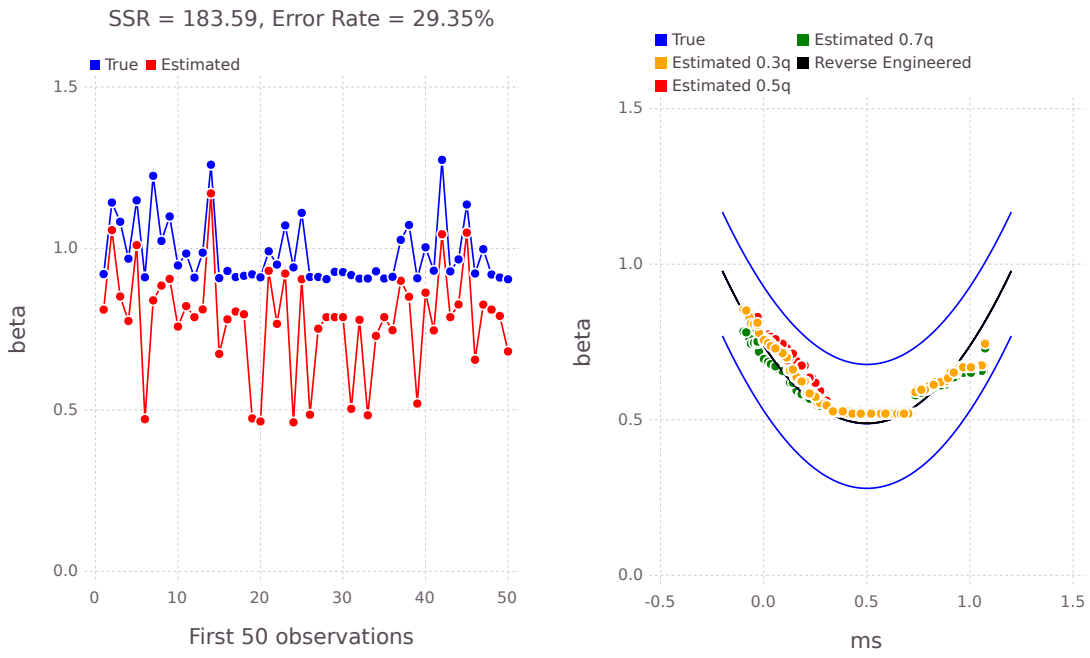
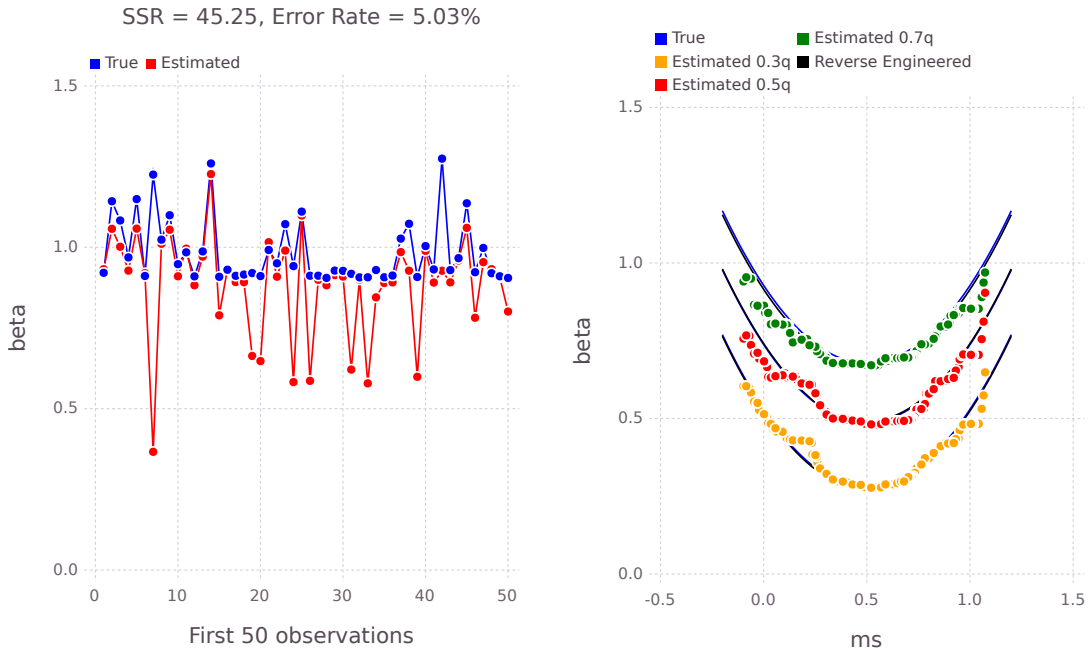
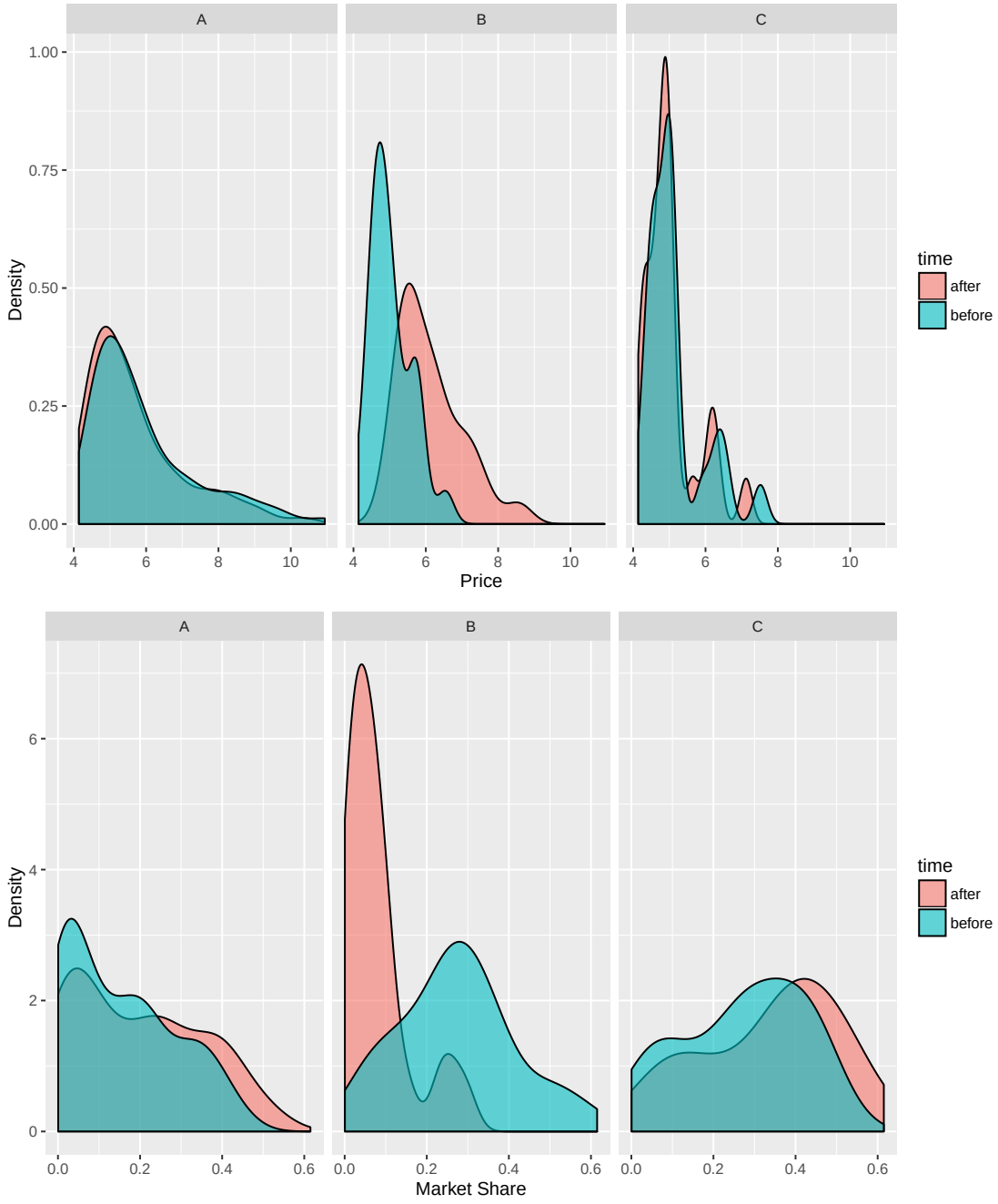


Figure 16: Reduce sample size: the proposed algorithm



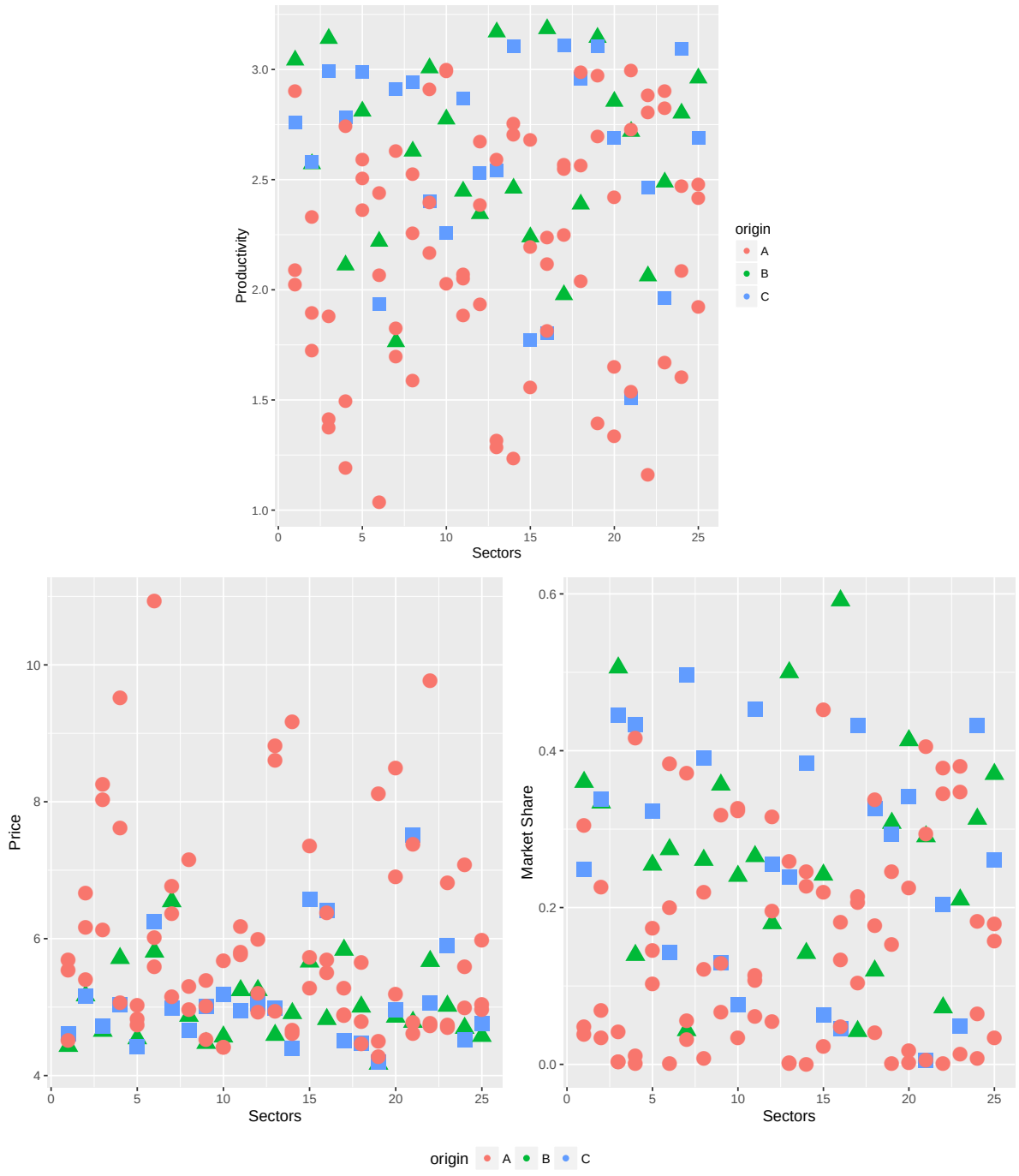
D Details of the Simulated Model

Figure 17: Responses of firms in country A to an appreciation in country B



Note: The left column presents the change of prices and market shares for domestic firms in country A. The middle and right columns present the reactions of exporters from country B and C respectively.

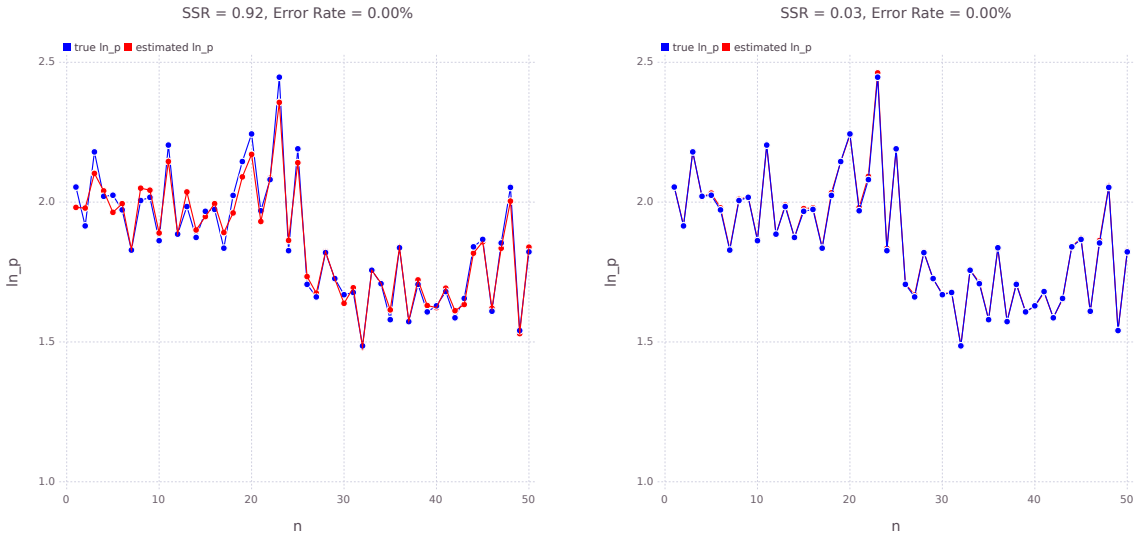
Figure 18: Visualisation of simulated firms in country A



Note: The top graph depicts the realised productivity of firms in country A. In each sector, there are three domestic firms and two foreign firms from country B and C respectively (only the best firm in each sector exports). The bottom two graphs depict the price and market shares of firms in country A. Exporters are firms with relatively high productivity and charge relatively low prices and own larger market shares. The assumption that only the best firms export gives a realistic market structure in this multi-country world.

D.1 Case 1: only exchange rate shocks

Figure 19: Case 1: Precision on price predictions



(a) GBRT without adding regression coefficients

(b) the proposed algorithm

Figure 20: Case 1: Without adding regression coefficients

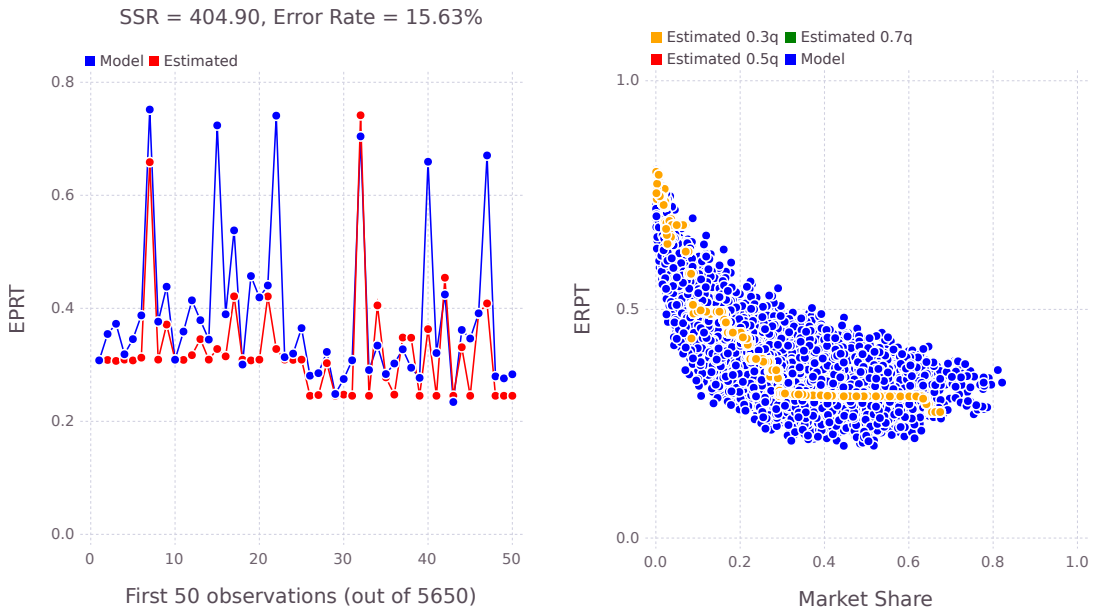
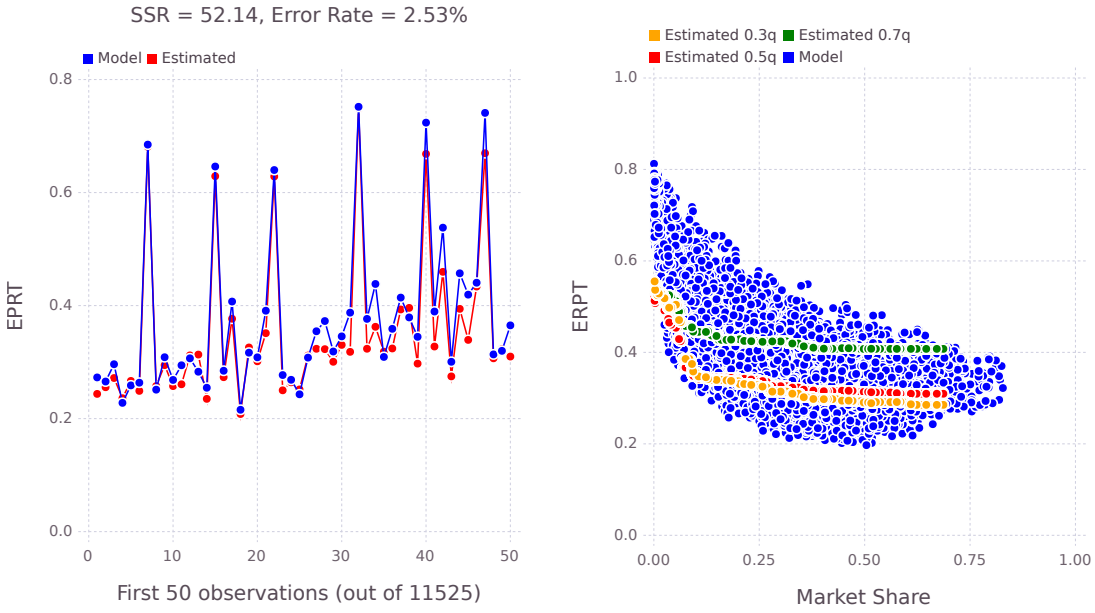
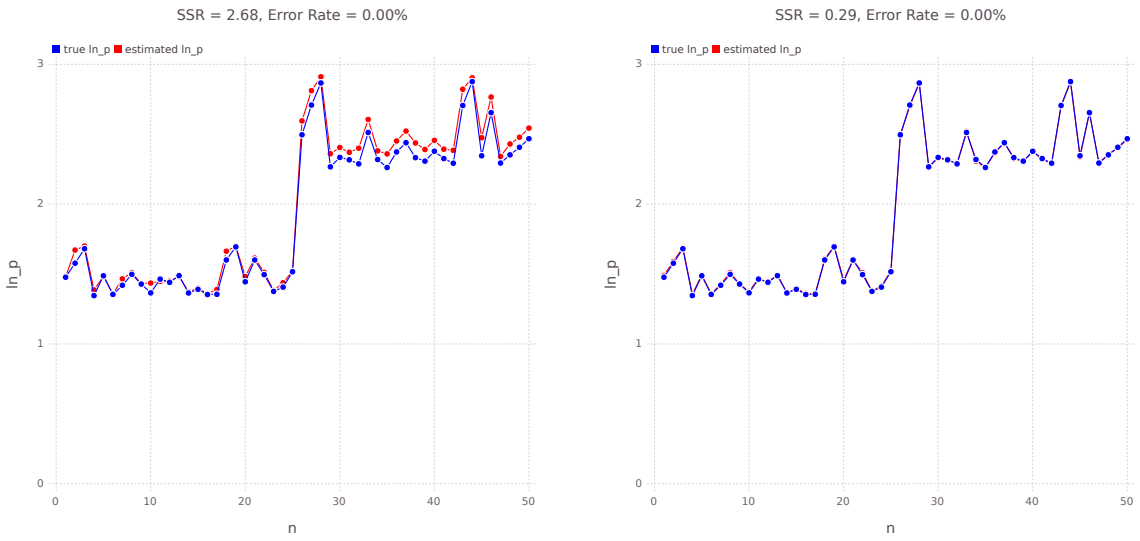


Figure 21: Case 1: Point estimates of the proposed algorithm compared to true counterfactual environments



D.2 Case 2: adding productivity shocks

Figure 22: Case 2: Precision on price predictions



(a) GBRT without adding regression coefficients

(b) the proposed algorithm

Figure 23: Case 2: Point estimates of the proposed algorithm compared to true counterfactual environments

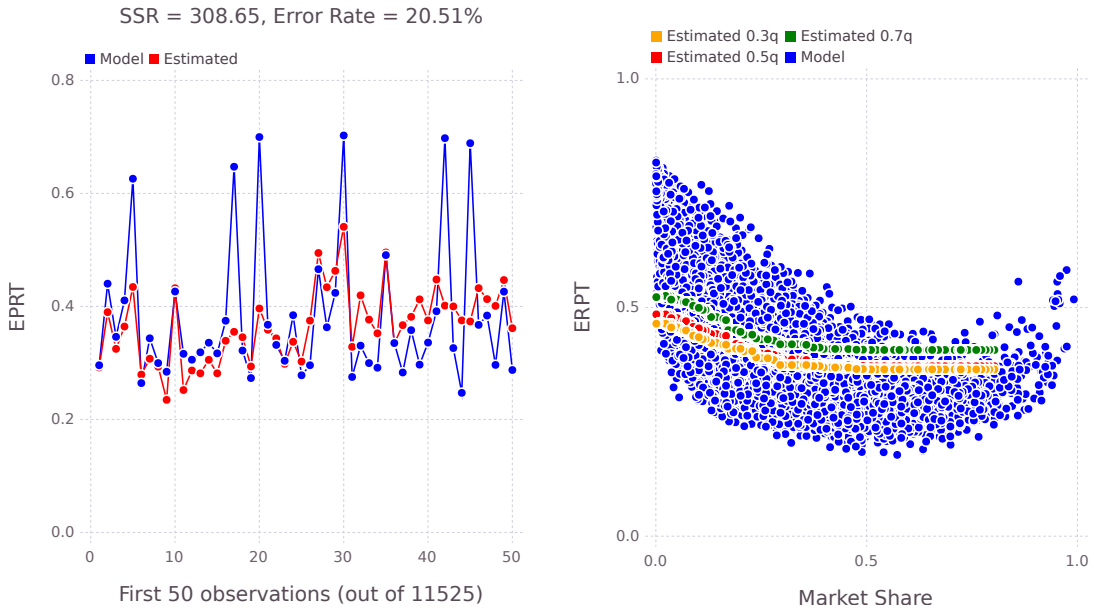
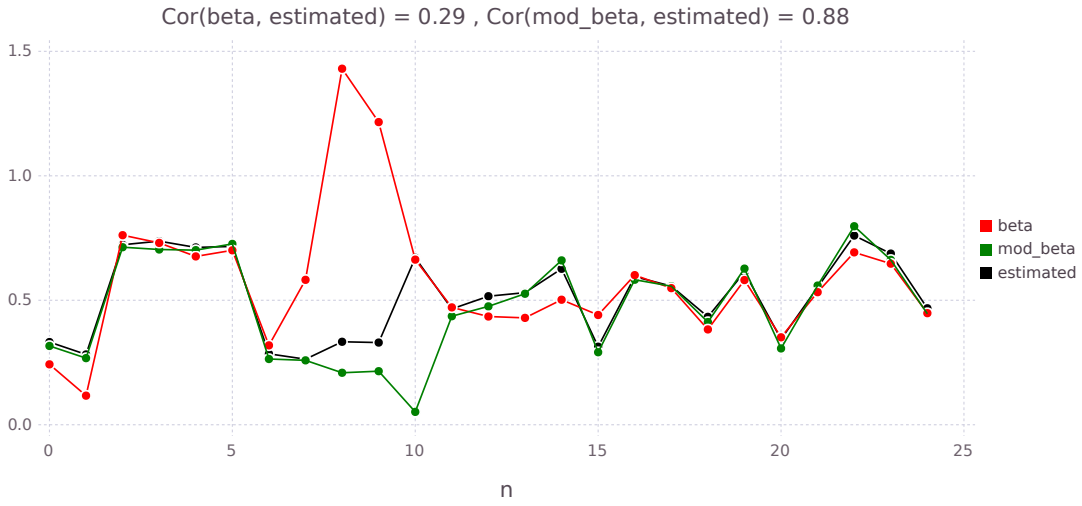


Figure 24: Comparing naive, counterfactual and algorithm predicted ERPT estimates



Note: Firm's productivity is assumed to follow an AR(1) process with a persistence of 0.95. The red line presents the ERPT estimates calculated using actual price changes of the simulated model. The green line represents the model implied ERPT estimates in a counterfactual equilibrium where there is no productivity shock in the next period. The black line represents ERPT estimates predicted by the proposed algorithm.

Algorithm 1 The Proposed Algorithm

Input data $\mathbf{I}, \mathbf{y}, \mathbf{X}, \mathbf{e}$

Obtain variable names of the index matrix $\mathbf{I}$ and the feature variable matrix $\mathbf{X}$ and save them as i_{names} and x_{names} respectively.
Calculate all non-repetitive combinations of dimension indices in i_{names} and save as S_i.
for s in S_i **do**
    $\mathbf{I_s} \leftarrow \mathbf{I}[i_{names} \in s]$
    $\widetilde{\mathbf{I}}_s \leftarrow unique(\mathbf{I_s})$
    for x in x_{names} **do**
        $x_s \leftarrow \mathbf{0}$
        for i_s in 1 to nrow($\widetilde{\mathbf{I}}_s$) **do**
            $x_s[\mathbf{I_s} = \widetilde{\mathbf{I}}[i_s]] \leftarrow mean(x|\mathbf{I_s} = \widetilde{\mathbf{I}}[i_s])$
        end for
    end for
end for
Calculate all non-repetitive binary combinations of S_i and save as S_{share}.
for s in S_{share} **do**
    $(s_a, s_b) \leftarrow s[sort(length(s[1], s[2]))]$
    for x in x_{names} **do**
        $x_{s_a,s_b} \leftarrow \frac{x_{s_a}}{x_{s_b}}$
    end for
end for
Observe dimensions in which the supervisor $\mathbf{y}$ and the policy/treatment variable $\mathbf{e}$ vary. Identify a subset available for controlling unobserved variables and save as S_{id}.
for s in S_{id} **do**
    Assume a possible (linear) structural equation based on economic rationale.
    for j in 1:(number of parameters in the structural model) **do**
        $coef_s^j \leftarrow \mathbf{0}$
    end for
    for d_s in 1 to nrow($\widetilde{\mathbf{I}}_s$) **do**
        Estimate the structural regression for the subset of data where $\mathbf{I_s} = \widetilde{\mathbf{I}}[i_s]$
        for j in 1:(number of parameters in the structural model) **do**
            $coef_s^j[\mathbf{I_s} = \widetilde{\mathbf{I}}[i_s]] \leftarrow parameter^j$
        end for
    end for
end for
Run GBRT with supervisor $\mathbf{y}$ on $\mathbf{e}, \mathbf{X}, \mathbf{X_{s_a,s_b}}, \mathbf{coef_{id}^j}$ and obtain $model_1$.
$\mathbf{y^{Est1}} \leftarrow model_1(\mathbf{e} - 0.5std(\mathbf{e}, \mathbf{X}, \mathbf{X_s}, \mathbf{coef_s^j})$
$\mathbf{y^{Est2}} \leftarrow model_1(\mathbf{e} + 0.5std(\mathbf{e}), \mathbf{X}, \mathbf{X_s}, \mathbf{coef_s^j})$
$\mathbf{beta^{Est}} \leftarrow \frac{\mathbf{y^{Est2}} - \mathbf{y^{Est1}}}{std(\mathbf{e})}$
Run GBRT again with supervisor $\mathbf{beta^{Est}}$ on $\mathbf{e}, \mathbf{X}, \mathbf{X_{s_a,s_b}}, \mathbf{coef_{id}^j}$ and obtain $model_2$.

Output: $model_1, model_2, \mathbf{beta^{Est}}$

References

Amiti, Mary, Oleg Itskhoki, and Jozef Konings. 2014. "Importers, Exporters, and Exchange Rate Disconnect." *The American Economic Review*, 104(7): 1942–1978.

Athey, Susan, and Guido Imbens. 2015. "A Measure of Robustness to Misspecification." *The American Economic Review*, 105(5): 476–480.

Athey, Susan, and Guido Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *The Journal of Economic Perspectives*.

Athey, Susan, Guido Imbens, Thai Pham, and Stefan Wager. 2017. "Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges." *The American Economic Review*, 107(5): 278–81.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2016. "Solving Heterogeneous Estimating Equations with Gradient Forests." *Arxiv Preprint Arxiv:1610.01271*.

Atkeson, Andrew, and Ariel Burstein. 2008. "Pricing-to-Market, Trade Costs, and International Relative Prices." *The American Economic Review*, 98(5): 1998–2031.

Auer, Raphael A., and Raphael S. Schoenle. 2016. "Market Structure and Exchange Rate Pass-Through." *Journal of International Economics*, 98: 60–77.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. "Machine Learning Methods for Demand Estimation." *The American Economic Review*, 105(5): 481–485.

Berman, Nicolas, Philippe Martin, and Thierry Mayer. 2012. "How Do Different Exporters React to Exchange Rate Changes?" *The Quarterly Journal of Economics*, 127(1): 437–492.

Bhatt, Samir, Peter W. Gething, Oliver J. Brady, Jane P. Messina, Andrew W. Farlow, Catherine L. Moyes, John M. Drake, John S. Brownstein, Anne G. Hoen, Osman Sankoh, et al. 2013. "The Global Distribution and Burden of Dengue." *Nature*, 496(7446): 504–507.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning*, 24(2): 123–140.

Breiman, Leo. 2001. "Random Forests." *Machine Learning*, 45(1): 5–32.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees.* CRC press.

Chen, Natalie, Jean Imbs, and Andrew Scott. 2009. "The Dynamics of Trade and Competition." *Journal of International Economics*, 77(1): 50–62.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Post-selection and Post-regularization Inference in Linear Models with Many Controls and Instruments." *The American Economic Review*, 105(5): 486–90.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, et al. 2016. "Double Machine Learning for Treatment and Causal Parameters." *Arxiv Preprint Arxiv:1608.00060*.

Corsetti, Giancarlo, and Luca Dedola. 2005. "A Macroeconomic Model of International Price Discrimination." *Journal of International Economics*, 67(1): 129–155.

Corsetti, Giancarlo, and Paolo Pesenti. 2005. "International Dimensions of Optimal Monetary Policy." *Journal of Monetary Economics*, 52(2): 281–305.

Corsetti, Giancarlo, Luca Dedola, and Sylvain Leduc. 2007. "Optimal Monetary Policy and the Sources of Local-currency Price Stability." In *International Dimensions of Monetary Policy*. 319–367. University of Chicago Press.

Corsetti, Giancarlo, Luca Dedola, and Sylvain Leduc. 2008*a*. "High Exchange-Rate Volatility and Low Pass-Through." *Journal of Monetary Economics*, 55(6): 1113–1128.

Corsetti, Giancarlo, Luca Dedola, and Sylvain Leduc. 2008*b*. "International Risk Sharing and the Transmission of Productivity Shocks." *Review of Economic Studies*, 75(2): 443–473.

Cox, Peter M., David Pearson, Ben B. Booth, Pierre Friedlingstein, Chris Huntingford, Chris D. Jones, and Catherine M. Luke. 2013. "Sensitivity of Tropical Carbon to Climate Change Constrained by Carbon Dioxide Variability." *Nature*, 494(7437): 341–344.

Dornbusch, Rudiger. 1987. "Exchange Rates and Prices." *The American Economic Review*, 77(1): 93–106.

Elith, Jane, John R Leathwick, and Trevor Hastie. 2008. "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology*, 77(4): 802–813.

Freund, Yoav, and Robert E. Schapire. 1996. "Experiments with a New Boosting Algorithm." 148–156. Morgan Kaufmann.

Friedman, Jerome H. 2002. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis*, 38(4): 367–378.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning.* Vol. 1, Springer series in statistics Springer, Berlin.

Hancock, Thomas, Tao Jiang, Ming Li, and John Tromp. 1996. "Lower Bounds on Learning Decision Lists and Trees." *Information and Computation*, 126(2): 114–122.

Hyafil, Laurent, and Ronald L Rivest. 1976. "Constructing Optimal Binary Decision Trees Is NP-complete." *Information Processing Letters*, 5(1): 15–17.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *The American Economic Review*, 105(5): 491–495.

Krugman, Paul. 1986. "Pricing to Market When the Exchange Rate Changes." National Bureau of Economic Research Working Paper 1926.

Melitz, Marc J., and Gianmarco I. Ottaviano. 2008. "Market Size, Trade, and Productivity." *The Review of Economic Studies*, 75(1): 295–316.

Randall, CJ, and R Van Woesik. 2015. "Contemporary White-band Disease in Caribbean Corals Driven by Climate Change." *Nature Climate Change*, 5(4): 375–379.

Ridgeway, Greg. 2007. "Generalized Boosted Models: A Guide to the Gbm Package." *R "gbm" Package USer Manual*.

Rokach, Lior, and Oded Maimon. 2005. "Top-down Induction of Decision Trees Classifiers-a Survey." *Ieee Transactions on Systems, Man, and Cybernetics, Part C (applications and Reviews)*, 35(4): 476–487.

Vapnik, Vladimir N. 1999. "An Overview of Statistical Learning Theory." *IEEE Transactions on Neural Networks*, 10(5): 988–999.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *The Journal of Economic Perspectives*, 28(2): 3–27.

Wager, Stefan, and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association*.