TEXTO PARA DISCUSSÃO

No. 653

ARCO: an artificial counterfactual approach for high-dimensional panel time-series data

Carlos C. Carvalho
Ricardo Masini
Marcelo C. Medeiros

DEPARTAMENTO DE ECONOMIA
www.econ.puc-rio.br

ARCO: AN ARTIFICIAL COUNTERFACTUAL APPROACH FOR HIGH-DIMENSIONAL PANEL TIME-SERIES DATA

Carlos V. Carvalho
Department of Economics, Pontifical Catholic University of Rio de Janeiro
E-mail: cvianac@econ.puc-rio.br

Ricardo Masini
São Paulo School of Economics, Getulio Vargas Foundation
E-mail: ricardo.masini@fgv.br

Marcelo C. Medeiros
Department of Economics, Pontifical Catholic University of Rio de Janeiro
E-mail: mcm@econ.puc-rio.br

Abstract

We consider a new method to estimate causal effects when a treated unit suffers a shock or an intervention, such as a policy change, but there is not a readily available control group or counterfactual. We propose a two-step approach where in the first stage an artificial counterfactual is estimated from a large-dimensional set of variables from pool of untreated units ("donors pool") using shrinkage methods, such as the *Least Absolute Shrinkage Operator* (LASSO). In the second stage, we estimate the average intervention effect on a vector of variables belonging to the treated unit, which is consistent and asymptotically normal. Our results are valid uniformly over a wide class of probability laws. Furthermore, we show that these results still hold when the date of the intervention is unknown and must be estimated from the data. Tests for multiple interventions and for contamination effects are also derived. By a simple transformation of the variables of interest, it is also possible to test for intervention effects on several moments (such as the mean or the variance) of the variables of interest. Finally, we can disentangle the actual intervention effects from confounding factors that usually bias "before-and-after" estimators. A detailed Monte Carlo experiment evaluates the properties of the method in finite samples and compares our proposal with other alternatives such as the differences-in-differences, factor models and the synthetic control method. An empirical application to evaluate the effects on inflation of a new anti tax evasion program in Brazil is considered. Our methodology is inspired by different branches of the literature such as: the Synthetic Control method, the Global Vector Autoregressive models, the econometrics of structural breaks, and the counterfactual analysis based on macro-econometric and panel data models.

Keywords: counterfactual analysis, comparative studies, LASSO, ArCo, synthetic control, policy evaluation, intervention, structural break, panel data, factor models.

JEL Codes: C22, C23, C32, C33.

1. Introduction

We propose a method for counterfactual analysis to evaluate the impact of interventions such as regional policy changes, the start of a new government, or outbreaks of wars, just to name a few possible cases. Our approach is specially useful in situations where there is a single treated unity and no available "controls". Furthermore, it is robust to the presence of confounding effects, such as a global shock[1]. The idea is to construct an artificial counterfactual based on a large-dimensional panel of observed time-series data from a pool of untreated peers. The introduced methodology shares roots with the panel factor model, hereafter PF, of Hsiao, Ching, and Wan (2012) and Gobillon and Magnac (2016), the Synthetic Control method, hereafter SC, pioneered by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010), as well as with the work of Pesaran and Smith (2012). Nevertheless, the overall proposed procedure differs from prior methods in several dimensions as will become clear in the next paragraphs.

Causality is a research topic of major interest in empirical Economics. Usually, causal statements with respect of the adoption of a given treatment rely on the construction of counterfactuals based on the outcomes from a group of individuals not affected by the treatment. Notwithstanding, definitive cause-and-effect statements are usually hard to formulate given the constraints that economists face in finding sources of exogenous variation. However, in micro-econometrics there has been major advances in the literature and the estimation of treatment effects is part of the toolbox of applied economists; see Angrist and Imbens (1994), Angrist, Imbens, and Rubin (1996), Abadie and Gardeazabal (2003), Heckman and Vytlacil (2005), Conley and Taber (2011), Belloni, Chernozhukov, and Hansen (2014), Ferman and Pinto (2015), and Belloni, Chernozhukov, Fernández-Val, and Hansen (2016).

On the other hand, when there is not a natural control group which is usually the case when handling aggregated (macro) data, the econometric tools have evolved in a slower pace and much of the work has focused on simulating counterfactuals from structural models. However, in recent years, some authors have proposed new techniques inspired partially by the developments in micro-econometrics that are able, under some assumptions, to conduct counterfactual analysis with aggregate data; see, for instance, Hsiao, Ching, and Wan (2012), Pesaran and Smith (2012), and Gobillon and Magnac (2016).

1.1. **Contributions of the Paper.** This paper fits into the literature of dynamic treatment effects and counterfactual analysis when a control group is not available. We propose a two-step approach called the **Artificial Counterfactual (ArCo)** method to estimate the average treatment (intervention) effect on the treated (ATET). In the first step, we propose a model and use the data before the intervention to estimate it. Then, we combine the estimated model with the data after the intervention to create the artificial counterfactual.

[1] Although the results in the paper are derived under the assumption of single treated unit, they can be easily generalized to the case of multiple units suffering the treatment.

Our proxy control unit is built as a function of a high-dimensional set of observed variables from a pool of untreated peers and without any stringent assumption about the actual Data Generating Process (DGP). High-dimensionality is relevant as the dimension of the estimation problem in the first-step can grow very fast when the number of peers is large and/or the number of variables for each peer increases. We use the *Least Absolute Selection and Shrinkage Operator* (LASSO) put forward by Tibshirani (1996) to estimate the model parameters. Nonlinearities can be handled by including in the first-stage model some transformations of the explanatory variables, such as polynomials or splines. The method is able to simultaneously test for effects in different variables as well as in multiple moments of a set of variables such as the mean and the variance. Furthermore, we propose a test of no intervention effects with a standard limiting distribution uniformly in a wide class of DGPs.

In addition, we propose extensions of the basic estimator. First, we accommodate situations when the exact time of the intervention is unknown. This is of importance in the presence of anticipation effects, for instance. We propose (and tabulate for some values) a $\mathcal{L}_p$ test inspired by the literature on structural breaks (Bai, 1997; Bai and Perron, 1998; Hansen, 2000). Inspired by the same literature we derive a test for the case of multiple interventions. Finally, we construct a test to check for contamination effects among units.

The identification of the average intervention effect relies on the assumption of independence between the intervention and the treated peers but we allow for heterogeneous common, possibly nonlinear, deterministic time trends among units. Our results are derived under asymptotic limits on the time dimension (T). However, we allow the number of peers (n) and the number of observed variables for each peer to grow as a function of T.

A thorough Monte Carlo experiment is conducted in order to evaluate the small sample performance of the ArCo methodology and to compare it to well-established alternatives, namely: the before-and-after (BA) estimator, the differences-in-differences (DiD) estimator assuming each peer to be an individual in the control group, the PF approach of Gobillon and Magnac (2016), hereafter PF-GM, and the SC method. We show that the bias of the ArCo method is, in general, negligible and much smaller than competing alternatives.

Finally, we apply the ArCo methodology to evaluate the impacts on inflation of an anti tax-evasion program implemented in October 2007 by the state government of São Paulo, the large city in Brazil. The mechanism works by giving monetary incentives (tax rebates) for consumers who ask for sales receipts. Additionally, the registered sales receipts give the consumer the right to participate in monthly lotteries promoted by the government. Similar initiatives relying on consumer auditing schemes were proposed in the European Union and in China. Under the premisses that (i) a certain degree of tax evasion was occurring before the intervention, (ii) the sellers has some degree of market power and (iii) the penalty for tax-evasion is large enough to alter the seller behaviour, one is expected to see an upwards movements in prices due to an increase in marginal cost. Therefore, our goal is to investigate whether the program had an impact on consumer prices.

1.2. **Connections to the literature.** The paper most similar to ours is Hsiao, Ching, and Wan (2012). The authors considered a two-step method where in their first step the counterfactual for a single treated variable of interest is constructed as a linear combination of a low-dimensional set of observed covariates from pre-selected elements from a pool of peers. The model is estimated using data from the pre-intervention period. Their theoretical results have been derived under the hypothesis of correct specification of a linear panel data model with common factors and no covariates. The selection of the included peers in the linear combination is carried out by information criteria. Recently, several extensions of the above methods have been proposed. Ouyang and Peng (2015) relaxed the linear conditional expectation assumption by introducing a semi-parametric estimator. Du and Zhang (2015) made improvements on the selection mechanism for the constituents of the donors pool.

Contrary to Hsiao, Ching, and Wan (2012), Gobillon and Magnac (2016) consider directly the estimation of a correctly specified linear panel model with interactive fixed effects, strictly exogenous regressors and known number of common factors. The model is an extension of the usual DiD specification augmented by a known number of common factors and the estimation is carried out in the whole sample. Their theoretical results rely on double asymptotics when both T and n go to infinity. The number of untreated units must grow in order to guarantee the identification of the common factors. The authors allow the common confounding factors to have nonlinear deterministic trends with heterogenous loadings, which is an utmost generalization of the common linear parallel trend hypothesis assumed when DiD estimation is considered.

The ArCo method differs from the above mentioned works in several directions. First, we do not restrict the analysis to a single treated variable. We can, for instance, measure the impact of interventions in several variables of the treated unit simultaneously. We also allow for tests on several moments of the variable of interest. Second, contrary to Hsiao, Ching, and Wan (2012) and Gobillon and Magnac (2016), all our theoretical results are derived under no stringent assumptions about the DGP, which we assume to be unknown. We do not need to estimate either the common factors or the true conditional expectation. This is a nice feature of the ArCo methodology, as usually models are misspecified and, even more important, consistent estimation of factors needs that both the time-series and the cross-section dynamics diverge to infinity. On the other hand, we consider a flexible linear-in-parameters high-dimensional model and our asymptotic results holds uniformly on a wide class of probability laws when the first step is estimated by the LASSO and the number of parameters to be estimated diverge. Furthermore, we show how to construct asymptotically honest confidence intervals to the average intervention effect. Third, we also demonstrate that our methodology can still be applied when the intervention time is unknown. Finally, we also develop tests for multiple interventions and contamination effects.

When compared to DiD estimators, the advantages of the ArCo methodology are three-fold. First, we do not need the number of treated units to grow. In fact, the workhorse situation is when there is a single treated unit. The second, and most important difference, is that the ArCo methodology has been developed to situations where the $n-1$ untreated units differ substantially from the treated one and can not form a control group even after conditioning on a set of observables. Finally, the ArCo methodology works even without the parallel trends hypothesis. The first difference can be attenuated in light of the recent results of Conley and Taber (2011) and Ferman and Pinto (2015) who put forward inferential procedures when the number of treated groups is small.

Although, both the ArCo and the SC methods construct an artificial counterfactual as a function of observed variables from a pool of peers (untreated units), the two approaches are different in several dimensions. First, the SC method relies on a convex combination of peers to construct the counterfactual. The ArCo solution is a general, possibly nonlinear, function. Even in the case of linearity, the method does not impose any restriction on the parameters of the linear combination. Furthermore, the weights in the SC method are usually estimated using time averages of the observed variables for each peer. Therefore, all the time-series dynamics is removed and the weights are determined in a pure cross-sectional setting. In some applications of the SC method, the number of observations to estimate the weights is much lower than the number of parameters to be determined. For example, in Abadie and Gardeazabal (2003) the authors have 13 observations to estimate 16 parameters. A similar issue also appears in Abadie, Diamond, and Hainmueller (2010, 2014). In addition, the SC method was designed to evaluate the effects of the intervention in a single variable. In order to evaluate the effects in a vector of variables, the method has to be applied several times. The ArCo methodology can be directly applied to a vector of variables of interest. Finally, our inferential procedure is not based on permutation tests.

With respect to the methodology by Pesaran and Smith (2012), the major difference is that the authors construct the counterfactual based on variables that belong to the treated unit and they do not rely on a pool of untreated peers. Their key assumption is that a subset of variables of the treated unit is invariant to the intervention. Although, in some specific cases this could be a reasonable hypothesis, in a general framework this is quite restrictive.

Recently, Angrist, Jordá, and Kuersteiner (2013) propose a semiparametric method to evaluate the effects of monetary policy based on the so called policy propensity score. Similar to Pesaran and Smith (2012), the authors only rely on information on the treated unit and no donor pool is available. As before, this is a major difference from our approach. Furthermore, their methodology seems to be particularly appealing to monetary economics but hard to be applied in other settings without major modifications.

Finally, it is important to compare the ArCo methodology with the work of Belloni, Chernozhukov, and Hansen (2014) and Belloni, Chernozhukov, Fernández-Val, and Hansen (2016). Both papers consider the estimation of intervention effects in large dimensions. First, Belloni, Chernozhukov, and Hansen (2014) consider a pure cross-section setting where the intervention is correlated to a large set of regressors and the approach is to consider an instrumental variable estimator to recover the intervention effect, as there is no control group available. In the ArCo framework, on the other hand, the intervention is assumed to be exogenous with respect to the peers. Notwithstanding, the intervention may not be (and probably is not) independent of variables belonging to the treated unit. This key assumption enables us to construct honest confidence bands by using the LASSO in the first step to estimate the conditional model. Belloni, Chernozhukov, Fernández-Val, and Hansen (2016) proposed a general and flexible approach for program evaluation in high dimensions. They provide efficient estimators and honest confidence bands for a large number of treatment effects. However, they do not consider the case where there is no control group available.

1.3. **Potential Applications.** There has been a large body of studies that require the estimation of intervention effects with no group of controls.

Measuring the impacts of regional policies is a potential application. For example, Hsiao, Ching, and Wan (2012) measure the impact of economic and political integration of Hong Kong with mainland China on Hong Kong's economy whereas Abadie, Diamond, and Hainmueller (2014) estimate spillovers of the 1990 German reunification in West Germany. Pesaran, Smith, and Smith (2007) used the Global Vector Autoregressive (GVAR) framework of Pesaran, Schuermann, and Weiner (2004) and Dees, Mauro, Pesaran, and Smith (2007) to study the effects of the launching of the Euro. Gobillon and Magnac (2016) considered the impact on unemployment of a new police implemented in France in the 1990s. The effects of trade agreements and liberalization have been discussed in Billmeier and Nannicini (2013), and Jordan, Vivian, and Wohar (2014). The rise of a new government or new political regime are, as well, a relevant "intervention" to be studied. For example, Grier and Maynard (2013) considered the economic impacts of the Chavez era.

Other potential applications are new regulation on housing prices as in Bai, Li, and Ouyang (2014) and Du and Zhang (2015), new labor laws as considered in Du, Yin, and Zhang (2013), and macroeconomic effects of economic stimulus programs Ouyang and Peng (2015). The effects of different monetary policies have been discussed in Pesaran and Smith (2012) and Angrist, Jordá, and Kuersteiner (2013). Estimating the economic consequences of natural disasters, as in Belasen and Polachek (2008), Cavallo, Galiani, Noy, and Pantano (2013), Fujiki and Hsiao (2015), and Caruso and Miller (2015), is also a promising area of research.

The effects of market regulation or the introduction of new financial instruments on the risk and returns of stock markets has been considered in Chen, Han, and Li (2013) and Xie and Mo (2013). Testing the intervention effects in multiple moments of the data can be of special interest in Finance, where the goal could be the effects of different corporate governance policies in the returns and risk of the firms (Johnson, Boone, Breach, and Friedmand, 2000).

1.4. **Plan of the paper.** Apart from this introduction, the paper is organized as follows. In Section 2 we present the ArCo method and discuss the conditional model used in the first step of the methodology. In Section 3 we derive the asymptotic properties of the ArCo estimator and state our main result. Sub-section 3.3 deals with the test for the null hypothesis of no causal effect. Extensions for unknown intervention time, multiple interventions and possible contamination effects are described in Section 4. In Section 5 we discuss some potential sources of bias in the ArCo method. A detailed Monte Carlo study together with a horse race among competitor estimators is conducted in Section 6. Section 7 deals with the empirical exercise. Finally, Section 8 concludes. Tables and Figures together with all proofs are relegated to the Appendix.

2. The Artificial Counterfactual Estimator

2.1. **Definitions and Main Idea.** Suppose we have n units (countries, states, municipalities, firms, etc) indexed by $i = 1, \ldots, n$. For each unit and for every time period $t = 1, \ldots, T$, we observe a realization of $\boldsymbol{z}_{it} = (z_{it}^1, \ldots, z_{it}^{q_i})' \in \mathbb{R}^{q_i}$, $q_i \geq 1$. Furthermore, assume that an intervention took place in unit $i = 1$, and only in unit 1, at time $T_0 = \lfloor \lambda_0 T \rfloor$, where $\lambda_0 \in (0, 1)$.

Let $\mathcal{D}_t$ be a binary variable flagging the periods when the intervention was in place. We can express the observable variables of unit 1 as

$$\boldsymbol{z}_{1t} = \mathcal{D}_t \boldsymbol{z}_{1t}^{(1)} + (1 - \mathcal{D}_t) \boldsymbol{z}_{1t}^{(0)},$$

where $\mathcal{D}_t = I(t \geq T_0)$, $I(A)$ is an indicator function that equals 1 if the event A is true, and $\boldsymbol{z}_{1t}^{(1)}$ denotes the outcome when the unit 1 is exposed to the intervention and $\boldsymbol{z}_{1t}^{(0)}$ is the potential outcome of unit 1 when there is no intervention.

We are ultimately concerned in testing hypothesis on the effects of the intervention in unit 1 for $t \geq T_0$. In particular, we consider interventions of the form

$$\boldsymbol{y}_t^{(1)} = \begin{cases} \boldsymbol{y}_t^{(0)}, & t = 1, \ldots, T_0 - 1, \\ \boldsymbol{\delta}_t + \boldsymbol{y}_t^{(0)}, & t = T_0 \ldots, T, \end{cases} \tag{1}$$

where $\boldsymbol{y}_t^{(j)} \equiv \boldsymbol{h}(\boldsymbol{z}_{1t}^{(j)})$ for $j \in \{0, 1\}$, $\boldsymbol{h} : \mathbb{R}^{q_1} \mapsto \mathbb{R}^q$ is a measurable function of $\boldsymbol{z}_{1t}$ that will be defined latter, and $\{\boldsymbol{\delta}_t\}_{t=T_0}^T$ is a deterministic sequence. Due to the flexibility of the mapping $\boldsymbol{h}(\cdot)$, interventions modeled as (1) are quite general. It includes, for instance, interventions affecting the mean, variance, covariances or any combination of moments of $\boldsymbol{z}_{1t}$. The null hypothesis of interest is

$$\mathcal{H}_0 : \boldsymbol{\Delta}_T = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \boldsymbol{\delta}_t = \boldsymbol{0}. \tag{2}$$

The quantity $\boldsymbol{\Delta}_T$ in (2) is quite similar to the traditional *average treatment effect on the treated* (ATET) vastly discussed in the literature. Furthermore, the null hypothesis (2) encompasses the case where the intervention is a sequence $\{\boldsymbol{\delta}_t\}_{t=T_0}^T$ under the alternative, which obviously is a special case of uniform treatments by setting $\boldsymbol{\delta}_t = \boldsymbol{\delta}, \forall t \geq T_0$.

The particular choice of the transformation $\boldsymbol{h}(\cdot)$ will depend on which moments of the data the econometrician is interested in testing for effects of the intervention. In other words, the goal will be to test for a break in a set of unconditional moments of the data and check if this break is solely due to the intervention or has other (global) causes (confounding effects). Typical choices for $\boldsymbol{h}(\cdot)$ are presented as examples below.

Example 1. *For the univariate case ($q_1 = 1$), we can use the identity function $h(a) = a$ for testing changes in the mean. In fact, provided that the p-th moment of the data is finite, we can use $h(a) = a^p$ to test any change in the unconditional p-th moment.*

Example 2. *In the multivariate case ($q_1 > 1$) we can consider*

$$\boldsymbol{h}(\boldsymbol{z}_{1t}) = \begin{cases} \boldsymbol{z}_{1t} & \textit{for testing changes in the mean,} \\ \mathsf{vech}\ (\boldsymbol{z}_{1t}, \boldsymbol{z}_{1t}') & \textit{for testing changes in the second moments.} \end{cases}$$

Example 3. *We can also conduct joint tests by combining the different choices for $\boldsymbol{h}$. For example, for testing simultaneously a change in the mean and variance we can set $\boldsymbol{h}(a) = (a, a^2)'$. In the multivariate we can set $\boldsymbol{y}_t = \mathsf{diag}\ (\boldsymbol{z}_{1t}, \boldsymbol{z}_{1t}')$.*

Set $\boldsymbol{y}_t = \mathcal{D}_t \boldsymbol{y}_t^{(1)} + (1 - \mathcal{D}_t)\boldsymbol{y}_{yt}^{(0)}$. The exact dimension of $\boldsymbol{y}_t$ depends on the chosen $\boldsymbol{h}(\cdot)$. However, regardless of the choice of $\boldsymbol{h}(\cdot)$, we will consider, without loss of generality, that $\boldsymbol{y}_t \in \mathcal{Y} \subset \mathbb{R}^q$, $q > 0$, and that we have a sample $\{\boldsymbol{y}_t\}_{t=1}^T$, being the first $T_0 - 1$ observations before the intervention and the $T - T_0 + 1$ remaining observations after the intervention.

Clearly we do not observe $\boldsymbol{y}_t^{(0)}$ after $T_0 - 1$. For that reason, we call thereafter the *counterfactual*, i.e., what would $\boldsymbol{y}_t$ have been like had there been no intervention (potential outcome). In order to construct the counterfactual, let $\boldsymbol{z}_{0t} = (\boldsymbol{z}_{2t}', \dots, \boldsymbol{z}_{nt}')'$ and $\boldsymbol{Z}_{0t} = \left(\boldsymbol{z}_{0t}', \dots, \boldsymbol{z}_{0t-p}'\right)'$ be the collection of all the untreated units observables up to an arbitrary lag $p \geq 0$. The exact dimension of $\boldsymbol{Z}_{0t}$ depends upon the number of peers $(n-1)$, the number of variables per peer, $q_i, i = 2, \dots, n$, and the choice of p. However, without loss of generality, we assume that $\boldsymbol{Z}_{0t} \in \mathcal{Z}_0 \subseteq \mathbb{R}^d$, $d > 0$.

Consider the following model

$$\boldsymbol{y}_t^{(0)} = \mathcal{M}_t + \boldsymbol{\nu}_t,\ t = 1, \dots, T, \tag{3}$$

where $\mathcal{M}_t \equiv \mathcal{M}(\boldsymbol{Z}_{0t})$, $\mathcal{M} : \mathcal{Z}_0 \to \mathcal{Y}$ is a measurable mapping, and $\mathbb{E}(\boldsymbol{\nu}_t) = \boldsymbol{0}$.[2]

Set $T_1 \equiv T_0 - 1$ and $T_2 \equiv T - T_0 + 1$ as the number of observations before and after the intervention, respectively. One can estimate the model above using the first T_1 observations since, in that case, $\boldsymbol{y}_t^{(0)} = \boldsymbol{y}_t$. Then, the estimate $\widehat{\mathcal{M}}_{t,T_1} \equiv \widehat{\mathcal{M}}_{T_1}(\boldsymbol{Z}_{0t})$ can be used to construct the estimated counterfactual as:

$$\widehat{\boldsymbol{y}}_t^{(0)} = \begin{cases} \boldsymbol{y}_t^{(0)}, & t = 1, \dots, T_0 - 1, \\ \widehat{\mathcal{M}}_{t,T_1}, & t = T_0, \dots, T. \end{cases} \tag{4}$$

Consequently, we can define:

[2] Which can be ensured by either including a constant in the model $\mathcal{M}$ or by centering the variables in a linear specification.

Definition 1. *The Artificial Counterfactual (ArCo) estimator is*

$$\widehat{\boldsymbol{\Delta}}_T = \frac{1}{T-T_0+1}\sum_{t=T_0}^{T}\widehat{\boldsymbol{\delta}}_t, \tag{5}$$

where $\widehat{\boldsymbol{\delta}}_t \equiv \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t^{(0)}$, *for* $t = T_0, \dots, T$.

Therefore, the ArCo is a two-stage estimator where in the first stage we choose and estimate the model $\mathcal{M}$ using the pre-intervention sample and in the second we compute $\widehat{\boldsymbol{\Delta}}_T$ defined by (5). At this point the following remarks are in order.

Remark 1. *The ArCo estimator in (5) is defined under the assumption that* λ_0 *(consequently* T_0*) is known. However, in some cases the exact time of the intervention might be unknown due to, for example, anticipation effects. On the other hand, the effects of a policy change may take some time to be noticed. Although the main results are derived under the assumption of known* λ_0*, we later show they are still valid when* λ_0 *is unknown.*

Remark 2. *In most applications the intervention exists for sure (the outbreak of a war, a new government, a different policy, a new law, etc) and, differently from comparative studies with micro data, it is usually an idiosyncratic change in unit 1. For example, only in unit 1 the new law was enforced or a new government began.*

2.2. A Key Assumption and Motivations.

In order to recover the effects of the intervention by the ArCo we need the following key assumption.

Assumption 1. $\boldsymbol{z}_{0t} \perp\!\!\!\perp \mathcal{D}_s$, *for all* t, s.

Roughly speaking the assumption above is sufficient for the peers to be unaffected by intervention on the unit of interest. The independence is actually stronger than necessary. Technically, what is necessary for the results is the mean independence of the chosen model as in $\mathbb{E}(\mathcal{M}_t|\mathcal{D}_t) = \mathbb{E}(\mathcal{M}_t)$. Nevertheless, the latter is implied by Assumption 1 regardless of the choice of $\mathcal{M}$. It is worth mentioning that since we allow $\mathbb{E}(\boldsymbol{z}_{1t}|\mathcal{D}_t) \neq \mathbb{E}(\boldsymbol{z}_{1t})$ we might have some sort of selection on observables and/or non-observables belonging to the treated unit. Of course, selection on features of the untreated units is ruled out by Assumption 1.

Even though we do not impose any specific DGP, the link between the treated unit and its peers can be easily motivated by a very simple, but general, common factor model:

$$\boldsymbol{z}_{it}^{(0)} = \boldsymbol{\mu}_i + \boldsymbol{\Psi}_{\infty,i}(L)\boldsymbol{\varepsilon}_{it}, \quad i = 1, \dots, n;\ t \geq 1 \tag{6}$$

$$\boldsymbol{\varepsilon}_{it} = \boldsymbol{\Lambda}_i \boldsymbol{f}_t + \boldsymbol{\eta}_{it}, \tag{7}$$

where $\boldsymbol{f}_t \in \mathbb{R}^f$ is a vector of common unobserved factors such that $\mathsf{sup}_t\,\mathbb{E}(\boldsymbol{f}_t\boldsymbol{f}_t') < \infty$ and $\boldsymbol{\Lambda}_i$, is a $(q_i \times f)$ matrix of factor loadings. Therefore, we allow for heterogeneous determinist trends of the form $\zeta(t/T)$, where ζ is a integrable function on $[0, 1]$ as in Bai (2009). $\{\boldsymbol{\eta}_{it}\}$,$i = 1, \dots, n$, $t = 1, \dots, T$, is a sequence of uncorrelated zero mean random variables. Finally, L is the lag operator and the polynomial matrix $\boldsymbol{\Psi}_{\infty,i}(L) = (\boldsymbol{I}_{q_i} + \boldsymbol{\psi}_{1i}L + \boldsymbol{\psi}_{2i}L^2 + \cdots)$ is such that $\sum_{j=0}^{\infty} \boldsymbol{\psi}_{ji}^2 < \infty$ for all $i = 1, \ldots, n$. $\boldsymbol{I}$ is the identity matrix. Usually we have $f < n$ thus, as long as we have a "truly common" factor in the sense of having some rows of $\boldsymbol{\Lambda}_i$ non zero, we expect correlation among the units.

The DGP originated by (6) is fairly general and nests several models as by the multivariate Wold decomposition and under mild conditions, any second-order stationary vector process can be written as an infinite order vector moving average process; see Niemi (1979). Furthermore, under a modern macroeconomics perspective, reduced-forms for Dynamic Stochastic General Equilibrium (DSGE) models are written as vector autoregressive moving average (VARMA) processes, which, in turn, are nested in the general specification in (6) (Fernández-Villaverde, Rubio-Ramírez, Sargent, and Watson, 2007; An and Schorfheide, 2007). Gobillon and Magnac (2016) is a special case of the general model described above.

In case of Gaussian errors, the above model will imply that $\mathbb{E}[\boldsymbol{y}_t^{(0)} | \boldsymbol{Z}_{0t}] = \boldsymbol{\Pi} \boldsymbol{Z}_{0t}$. Otherwise, we can choose model $\mathcal{M}$ to be a linear approximation of the conditional expectation. The strategy is to define $\boldsymbol{x}_t$ as a set of transformations of $\boldsymbol{Z}_{0t}$, such as, for instance, polynomials or splines, and write $\boldsymbol{y}_t^{(0)}$ as a linear function of $\boldsymbol{x}_t$.

There are at least two major advantages of applying the ArCo estimator instead of just computing a simple difference in the mean of $\boldsymbol{y}_t$ before and after the intervention as a estimator for the intervention effect. The first is an efficiency argument. Note that the "before and after" estimator defined as $\widehat{\boldsymbol{\Delta}}_T^{BA} \equiv \frac{1}{T-T_0+1} \sum\limits_{t=T_0}^{T} \boldsymbol{y}_t - \frac{1}{T_0-1} \sum\limits_{t=1}^{T_0-1} \boldsymbol{y}_t$ is a particular case of our estimator when you have "bad peers", in the sense they are uncorrelated to the unit of interest. In this case, $\mathcal{M}(\cdot) =$ constant and $\widehat{\boldsymbol{\Delta}}_T = \widehat{\boldsymbol{\Delta}}_T^{BA}$. In fact, the additional information provided by the peers helps to reduce the ArCo estimator variance.

The second, and more important, argument in favor of the ArCo method is related to its capability of isolate the intervention of interest from aggregated shocks. When attempting to measure the effect of a particular intervention we are usually in a scenario that other aggregated shocks took place at the same time. The ability to disentangle these two effects is vital if one intends to provide a meaningful estimation of the intervention effect. A simple mental experiment illustrates the point: suppose *all* units at time T_0 are hit by a (aggregated) shock that changes all the means by the same amount. If we apply the BA estimator we will eventually encounter this mean break and would erroneously attribute it to the intervention of interest[3]. On the other hand, if we use the ArCo approach, since all the units have changed equally, the estimated effect will probably be insignificant.

Finally, it is important to stress that the validity of the ArCo procedure does not rely on the traditional parallel trend assumption such as the one usually considered in DiD techniques nor does it assume the trend to be the same for all the units at a given time, as for instance in the SC framework. The necessary assumption for our methodology to work properly is some sort of combination of peers (model $\mathcal{M}$) that can generate an artificial counterfactual

[3] Unless the intervention of interest is the aggregated shock but in that case we have invalid peers since they were treated.

whose difference from the real counterfactual is well behaved (in the sense of admitting a Law of Large Numbers and Central Limit theorems). This is usually possible with deterministic trends that do not dominate the stationary stochastic component asymptotically as well as when there is some common structure among units.

3. Asymptotic Properties and Inference

3.1. Choice of the Pre-intervention Model and a General Result.

The first stage of the ArCo method requires the choice of the model $\mathcal{M}$. One should aim for a model that captures most of the information from the available peers. Once the choice is made, the model must be estimated using the pre-intervention sample.

It is important to recognise that we do not consider that the model choice is actually the true model. We can consider that the $\boldsymbol{z}_{it}$ is generated by a DGP such as (6) irrespective of the choice of $\mathcal{M}$. Ideally, in the mean square error sense, we would like to set $\mathcal{M}$ as the conditional expectation model $\boldsymbol{m}(\boldsymbol{a}) = \mathbb{E}(\boldsymbol{y}_t|\boldsymbol{Z}_{0t} = \boldsymbol{a})$.

Motivated by the fact the dimension of $\boldsymbol{Z}_{0t}$ can grow quite fast in any simple application (by either including more peers, more covariates, or by simply considering more lags) we propose a fully parametric specification in order to approximate $\boldsymbol{m}(\cdot)$ as opposed to try to estimate it non-parametrically. In particular, we approximate it by a linear model (q linear models to be precise) of some transformation of $\boldsymbol{Z}_{0t}$. Consequently, the model is linear in $\boldsymbol{x}_t = \boldsymbol{h}_x(\boldsymbol{Z}_{0t})$, where in $\boldsymbol{x}_t$ we include a constant term. In particular, $\boldsymbol{h}_x$ could be a dictionary of functions such as polynomials, splines, interactions, dummies or any another family of elementary transformations the $\boldsymbol{Z}_{0t}$, in the spirit of sieve estimation (Chen, 2007). The same approach has been adopted in Belloni, Chernozhukov, and Hansen (2014) and Belloni, Chernozhukov, Fernández-Val, and Hansen (2016).

Hence, $\mathcal{M}_t = \mathsf{diag}\,(\boldsymbol{\theta}_{0,1}', \ldots, \boldsymbol{\theta}_{0,q}')\boldsymbol{x}_t$, where both $\boldsymbol{x}_t$ and $\boldsymbol{\theta}_{0,j}$, $j = 1, \ldots, q$, are d-dimensional vectors for $j = 1, \ldots, q$. We allow d to be a function of T. Hence, $\boldsymbol{x}_t$ and $\boldsymbol{\theta}_{0,j}$ depend on T but the subscript T will be omitted in what follows. Set $\boldsymbol{r}_t \equiv \boldsymbol{m}_t - \mathcal{M}_t$ as the approximation error and $\boldsymbol{\varepsilon}_t \equiv \boldsymbol{y}_t - \boldsymbol{m}_t$ as the projection error. We can write the model as in (3), with $\boldsymbol{\nu}_t = \boldsymbol{r}_t + \boldsymbol{\varepsilon}_t$. The model is then compromised of q linear regressions:

$$y_{jt}^{(0)} = \boldsymbol{x}_t'\boldsymbol{\theta}_{0,j} + \nu_{jt}, \quad j = 1, \ldots, q, \tag{8}$$

where $\boldsymbol{\theta}_{0,j}$ are the best (in the MSE sense) linear projection parameters which are properly identified as long as we rule out multicollinearity among $\boldsymbol{x}_t$ (Assumption 2).

We consider the sample (in the absence of intervention) as a single realization of the random process $\{\boldsymbol{z}_t^{(0)}\}_{t=1}^T$ defined on a common measurable space $(\Omega, \mathcal{F})$ with a probability law (joint distribution) $P_T \in \mathcal{P}_T$, where $\mathcal{P}_T$ is (for now) an arbitrary class of probability laws. The subscript T makes it explicit the dependence of the joint distribution on the sample size T, but we omit it in what follows. We write $\mathbb{P}_P$ and $\mathbb{E}_P$ to denote the probability and expectation with respect to the probability law $P \in \mathcal{P}$, respectively.

We establish the asymptotic properties of the ArCo estimator by considering the whole sample increasing, while the proportion between the pre-intervention to the post-intervention sample size is constant. The limits of the summations are from 1 to T whenever left unspecified. Recall that $T_1 \equiv T_0 - 1$ and $T_2 \equiv T - T_0 + 1$ are the number of pre and post intervention periods, respectively and $T_0 = \lfloor \lambda_0 T \rfloor$. Hence, for fixed $\lambda_0 \in (0,1)$ we have $T_0 \equiv T_0(T)$. Consequently, $T_1 \equiv T_1(T)$ and $T_2 \equiv T_2(T)$. All the asymptotics are taken as $T \to \infty$. We denote convergence in probability and in distribution by "$\xrightarrow{p}$" and "$\xrightarrow{d}$", respectively.

First, we state a general result under very high level assumptions from which all the other subsequent results rely on. Let $\widehat{\mathcal{M}}_{t,T_1} = (\boldsymbol{x}_t'\widehat{\boldsymbol{\theta}}_{1,T_1}, \dots, \boldsymbol{x}_t{}'\widehat{\boldsymbol{\theta}}_{q,T_1})'$, for $t \geq T_0$, where $\widehat{\boldsymbol{\theta}}_{j,T_1}$, $j = 1, \dots, q$, is estimated with only the first T_1 pre-intervention observations, and define $\boldsymbol{\eta}_{t,T_1} \equiv \widehat{\mathcal{M}}_{t,T_1} - \mathcal{M}_t$, $t \geq T_0$.

PROPOSITION 1. *Under Assumption 1, consider further that, uniformly in $P \in \mathcal{P}$ (an arbitrary class of probability laws):*

(a) $\sqrt{T}\left(\frac{1}{T_2}\sum_{t\geq T_0} \boldsymbol{\eta}_{t,T_1} - \frac{1}{T_1}\sum_{t\leq T_1} \boldsymbol{\nu}_t\right) \xrightarrow{p} \mathbf{0}$

(b) $\frac{1}{\sqrt{T_1}}\boldsymbol{\Gamma}_{T_1}^{-1/2}\sum_{t\leq T_1} \boldsymbol{\nu}_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$, *where* $\boldsymbol{\Gamma}_{T_1} = \mathbb{E}_P\left[\frac{1}{T_1}(\sum_{t\leq T_1} \boldsymbol{\nu}_t)(\sum_{t\leq T_1} \boldsymbol{\nu}_t')\right]$.

(c) $\frac{1}{\sqrt{T_2}}\boldsymbol{\Gamma}_{T_2}^{-1/2}\sum_{t\geq T_0} \boldsymbol{\nu}_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$, *where* $\boldsymbol{\Gamma}_{T_2} = \mathbb{E}_P\left[\frac{1}{T_2}(\sum_{t\geq T_0} \boldsymbol{\nu}_t)(\sum_{t\geq T_0} \boldsymbol{\nu}_t')\right]$.

Then, uniformly in $P \in \mathcal{P}$, $\sqrt{T}\boldsymbol{\Omega}_T^{-1/2}\left(\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$, where $\mathcal{N}(\cdot,\cdot)$ is the multivariate normal distribution and $\boldsymbol{\Omega}_T \equiv \frac{\boldsymbol{\Gamma}_{T_1}}{T_1/T} + \frac{\boldsymbol{\Gamma}_{T_2}}{T_2/T}$.

Condition (a) above sets a limit for the estimation error to be asymptotic negligible, ensuring the $\sqrt{T}$ rate of convergence of the estimator. Under condition (a) we can write:

$$\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T = \frac{1}{T_2}\sum_{t\geq T_0} \boldsymbol{\nu}_t - \frac{1}{T_1}\sum_{t\leq T_1} \boldsymbol{\nu}_t + o_p(T^{-1/2}).$$

Finally, conditions (b) and (c) ensure the asymptotic normality of the terms above after appropriate normalization. From the asymptotic variance $\boldsymbol{\Omega}_T$ it becomes evident that an intervention at the middle of the sample, $\lambda_0 = 0.5$, is desirable when $\lim_{T\to\infty} \boldsymbol{\Gamma}_{T_1} = \lim_{T\to\infty} \boldsymbol{\Gamma}_{T_2} \equiv \boldsymbol{\Gamma}$, which happens for instance when $\{\boldsymbol{\nu}_t\}$ is a stationary process. In this case, $\lim_{T\to\infty} \boldsymbol{\Omega}_T = \boldsymbol{\Gamma}/\lambda_0(1-\lambda_0)$.

Recall that if $\mathcal{M} = \boldsymbol{\alpha}_0$, the estimator is equivalent to the BA estimator. Therefore, one advantage of the ArCo is to provide a systematic way to extract as most information as possible from the peers in order to reduce the asymptotic variance of the prediction error. We can make more explicit the peers' contribution in reducing the asymptotic variance of the average ArCo estimator by the following matrix inequality (in term of positive definiteness)

$$\mathbf{0} \leq \lim_{T\to\infty} \boldsymbol{\Omega}_T \equiv \boldsymbol{\Omega} \leq \lim_{T\to\infty} T\mathbb{V}\left(\frac{1}{T_2}\sum_{t\geq T_0} \boldsymbol{y}_t^{(0)} - \frac{1}{T_1}\sum_{t\leq T_1} \boldsymbol{y}_t^{(0)}\right) \equiv \widetilde{\boldsymbol{\Omega}},$$

where $\mathbb{V}$ is the variance operator defined for any random vector $\boldsymbol{v}$ as $\mathbb{V}(\boldsymbol{v}) = \mathbb{E}(\boldsymbol{v}\boldsymbol{v}') - \mathbb{E}(\boldsymbol{v})\mathbb{E}(\boldsymbol{v}')$.

The upper bound $\widetilde{\boldsymbol{\Omega}}$ is the long run variance of the variables of the unit of interest (unit 1) weighted by the intervention fraction time λ_0. As a consequence, our estimator variance for any given λ_0, lies in between those two polar cases. One polar case is when there is a perfect artificial counterfactual and the other one is when the peers contribute with no information. Thus, the peer's contribution in reducing the ArCo estimator asymptotic variance could be represented by a R^2-type statistic measuring the "ratio" between the explained long-run variance $\boldsymbol{\Omega}$ to the total long-run variance $\widetilde{\boldsymbol{\Omega}}$.

3.2. **Assumptions and Asymptotic Theory in High-Dimensions.** The dimension d of $\boldsymbol{x}_t$ can be potentially very large, even larger than the sample size T, whenever the number of peers and/or the number of variables per peer is large. In these cases it is standard to allow d, and consequently $\boldsymbol{\theta}_j$, $j = 1\ldots,q$, to be function of the sample size, such that $d \equiv d_T$ and $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j,T}$. In order to make estimation feasible, regularization (shrinkage) is usually adopted, which is justified by some sparsity assumption on the vector $\boldsymbol{\theta}_{0,j}$, $j = 1\ldots,q$, in the sense that only a small portion of its entries are different from zero.

We propose the estimation of (8), equation by equation, by the LASSO approach proposed by Tibshirani (1996) and we allow that dimension $d > T$ to grow faster than the sample size[4]. Also, since each equation in the model is the same, we drop the subscript j from now on to focus on a generic equation. Therefore, we estimate $\boldsymbol{\theta}_0$ via

$$\widehat{\boldsymbol{\theta}} = \mathsf{arg\,min}\left\{\frac{1}{T_0-1}\sum_{t<T_0}(y_t - \boldsymbol{x}_t'\boldsymbol{\theta})^2 + \varsigma\|\boldsymbol{\theta}\|_1\right\}, \tag{9}$$

where $\varsigma > 0$ is a penalty term and $\|\cdot\|_1$ denotes the ℓ_1 norm.

Let $\boldsymbol{\theta}[A]$ denote the vector of parameters indexed by A and S_0 the index set of the non-zero (relevant) parameters $S_0 = \{i : \theta_{0,i} \neq 0\}$ with cardinality s_0. We consider the following set of assumptions.[5]

Assumption 2. *(DESIGN) Let $\boldsymbol{\Sigma} \equiv \frac{1}{T_1}\sum_{t=1}^{T_1}\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$. There exists a constant $\psi_0 > 0$ such that*

$$\|\boldsymbol{\theta}[S_0]\|_1^2 \leq \frac{\boldsymbol{\theta}\boldsymbol{\Sigma}\boldsymbol{\theta}s_0}{\psi_0^2},$$

for all $\|\boldsymbol{\theta}[S_0^c]\|_1 \leq 3\|\boldsymbol{\theta}[S_0]\|_1$.

Assumption 3. *(HETEROGENEITY AND DEPENDENCY) Let $\boldsymbol{w}_t \equiv (\nu_t, \boldsymbol{x}_t')'$, then:*

(a) $\{\boldsymbol{w}_t\}$ is strong mixing with $\alpha(m) = \exp(-cm)$ for some $c \geq \underline{c} > 0$

(b) $\mathbb{E}|w_{it}|^{2\gamma+\delta} \leq c_\gamma$ for some $\gamma > 2$ and $\delta > 0$ for all $1 \leq i \leq d, 1 \leq t \leq T$ and $T \geq 1$,

[4] Some efficiency gain could be potentially obtain by a joint estimation, for instance, a SUR (seemly unrelated regression) setting if the regressors of each equation are the not the same. We do not pursue this route in here.

[5] Recall that since we drop the equation subscript j, the assumptions below must understood for each equation $j = 1,\ldots,q$ separately.

(c) $\mathbb{E}(\nu_t^2) \geq \epsilon > 0$*, for all* $1 \leq t \leq T$ *and* $T \geq 1$.

Assumption 4. *(REGULARITY)*
(a) $\varsigma = O\left(\frac{d^{1/\gamma}}{\sqrt{T}}\right)$
(b) $s_0 \frac{d^{2/\gamma}}{\sqrt{T}} = o(1)$

Assumption 2 is known as the compatibility condition, which is extensively discussed on Bülhmann and van der Geer (2011). It is quite similar to the restriction of the smallest eigenvalue of $\mathbf{\Sigma}$, when one replace $\|\boldsymbol{\theta}[S_0]\|_1^2$ by its upper bound $s_0\|\boldsymbol{\theta}[S_0]\|_2^2$. Notice that we make no compatibility assumption regarding the sample counterpart $\widehat{\mathbf{\Sigma}} \equiv \frac{1}{T_1}\sum_{t=1}^{T_1} \boldsymbol{x}_t\boldsymbol{x}_t'$.

Assumption 3 controls for the heterogeneity and the dependence structure of the process that generates the sample. In particular Assumption 3(a) requires $\{\boldsymbol{w}_t\}$ to be an α-mixing process with exponential decay. It could be replaced by more flexible forms of dependence such as near epoch dependence or $\mathcal{L}_p$-approximability on an α-mixing process as long as we control for the approximation error term. Assumption 3(b) bounds uniformly some higher moment which ensures an appropriate Law of Large Numbers, and Assumption 3(c) is sufficient for the Central Limit Theorem. The latter bounds the variance of the regression error away from zero, which is plausible if we consider that the fit will never be perfect regardless of how much relevant variables we have in (8).

Assumption 4(a) and (b) are regularity conditions on the growth rate of the penalty parameter and the number of (relevant/total) parameters, respectively. They are obviously smaller than the analogous results found in the literature for the case of fix design and normality of the error term.[6]

We can now define $\mathcal{P}$ as the class of probability law that satisfies Assumptions 2,3 and 4(b). However, for convenience we explicitly state all those assumptions underlying the results that follows. Here is our main result.

Theorem 1. (MAIN) *Let* $\mathcal{M}$ *be the model defined by* (8)*, whose parameters are estimated by* (9)*, then under Assumptions 1-4:*

$$\sup_{P\in\mathcal{P}}\sup_{\boldsymbol{a}\in\mathbb{R}^q}\left|\mathbb{P}_P\left[\sqrt{T}\mathbf{\Omega}_T^{-1/2}(\widehat{\mathbf{\Delta}}_T - \mathbf{\Delta}_T) \leq \boldsymbol{a}\right] - \Phi(\boldsymbol{a})\right| \to 0, \quad \textit{as } T \to \infty,$$

where $\mathbf{\Omega}_T$ *is defined in Proposition 1,the event* $\{\boldsymbol{a} \leq \boldsymbol{b}\} \equiv \{a_i \leq b_i, \forall i\}$ *and* $\Phi(\cdot)$ *is the cumulative distribution function of a zero-mean identity covariance normal random vector.*

The results above are uniform with respect to the class of probability laws $\mathcal{P}$, which we believe to be large enough to be of some interest. Notice that we do *not* require any strong separation of the parameters away from zero, which is usually accomplished in the literature by imposing a $\theta_{\min}$ which is uniformly bounded away from zero. The uniformity convergence above is possible, in our case, as consequence of Assumption 1, which translate into the treatment $\mathcal{D}_t$ be uncorrelated to the regressors $\boldsymbol{x}_t$. In other words, the potential

[6] Under those condition, 4(a) and (b) become $\varsigma = O\left(\sqrt{\frac{\log d}{T}}\right)$ and $s_0\frac{\log d}{\sqrt{T}} = o(1)$, respectively.

non-uniformity issues regarding the estimation of the parameters of $\boldsymbol{\theta}_0$ do not contaminate the estimation of $\boldsymbol{\Delta}_T$, even if the coefficients of the conditional model are of order $O(T^{-1/2})$ as discussed in Leeb and Pötscher (2005,2008,2009).

In a different set-up, Belloni, Chernozhukov, and Hansen (2014) consider the case where the treatment is correlated with the set of regressors. Consequently, they propose the estimation via a moment condition with the so called *orthogonality property* in order to achieve uniform convergence. Further, Belloni, Chernozhukov, Chetverikov, and Wei (2016) generalize this idea to conduct uniform inference in a broad class of Z-estimators. As a parallel, our framework ensures a moment condition with the orthogonality property as a consequence of Assumption 1.

3.3. **Hypothesis Testing under Asymptotic Results.** Given the asymptotic normality of $\widehat{\boldsymbol{\Delta}}_T$, it is straightforward to conduct hypothesis testing. It is important, however, to remember the dependence of the results upon knowing the exact point of a possible break and the assurance that the peers are in fact untreated. Fortunately, both conditions can be tested which is the topic of the next sections. For now will we consider that the unit 1 is the only one potentially treated and the moment of the intervention, T_0, is known for certain.

First we need a consistent estimator for the variance $\boldsymbol{\Omega}_T$. More precisely, we need estimators for both $\boldsymbol{\Gamma}_{T_1}$ and $\boldsymbol{\Gamma}_{T_2}$. If we expect to have uncorrelated residuals and given the consistency of $\widehat{\boldsymbol{\theta}}$, we can simply estimate it by the average of the sum of squares of residuals in the pre-intervention model. A popular choice for serial correlated residuals are presented in Andrews (1991) and Newey and West (1987). Both have a similar structure given by the weighted autocovariance estimator as

$$\widehat{\boldsymbol{\Gamma}}_{T_i} = \widehat{\boldsymbol{\Gamma}}_{0_i} + \sum_{k=1}^{M} \phi(k) \left(\widehat{\boldsymbol{\Gamma}}_{k_i} + \widehat{\boldsymbol{\Gamma}}'_{k_i} \right), \quad i = \{1, 2\}, \tag{10}$$

where $\widehat{\boldsymbol{\Gamma}}_{k_1} \equiv \frac{1}{T_1-k}\sum_{t=1}^{T_1-k} \widehat{\boldsymbol{\nu}}_t \widehat{\boldsymbol{\nu}}'_{t+k}$, $\widehat{\boldsymbol{\Gamma}}_{k_2} \equiv \frac{1}{T_2-k}\sum_{t=T_0}^{T-k} \widehat{\boldsymbol{\nu}}_t \widehat{\boldsymbol{\nu}}'_{t+k}$, $k = 0, \ldots, M$, and $\widehat{\boldsymbol{\nu}}_t = \boldsymbol{y}_t - \widehat{\mathcal{M}}_{T_0}(\boldsymbol{x}_t) - \widehat{\boldsymbol{\Delta}}_T I(t \geq T_0)$.

In practice, we still need to specify the maximum number of lags/bandwidth to consider and the weight function. Usually, the later is a kernel function centered at zero. A common choice is a Bartlett kernel where the weights are given simply by $\phi(k) = 1 - \frac{k}{M+1}$. Theorem 2 of Newey and West (1987) and Proposition 1 of Andrews (1991) give general conditions under which the estimator is consistent. Moreover, Andrews (1991) discusses what kind of kernels are allowed and present a sizeable list of options. It also describes a data-driven procedure for bandwidth selection.

Therefore, if we replace $\boldsymbol{\Omega}_T$ by $\widehat{\boldsymbol{\Omega}}_T \equiv \frac{\widehat{\boldsymbol{\Gamma}}_{T_1}}{T_1/T} + \frac{\widehat{\boldsymbol{\Gamma}}_{T_2}}{T_2/T}$, we can construct honest (uniform) asymptotic confidence intervals and hypothesis testing as follows:

Proposition 2. *(Uniform Confidence Interval) Let $\widehat{\boldsymbol{\Omega}}_T$ be a consistent estimator for $\boldsymbol{\Omega}_T$ uniformly in $P \in \mathcal{P}$. Under the same conditions of Theorem 1, for any given significance* *level α:*

$$\mathcal{I}_\alpha \equiv \left[\widehat{\Delta}_{j,T} \pm \frac{\widehat{\omega}_j}{\sqrt{T}}\Phi^{-1}(1-\alpha/2)\right]$$

for each $j = 1, \ldots, q$, where $\widehat{\omega}_j = \sqrt{[\widehat{\mathbf{\Omega}}]_{jj}}$ and $\Phi^{-1}(\cdot)$ is the quantile function of a standard normal distribution. The confidence interval $\mathcal{I}_\alpha$ is uniformly valid (honest) in the sense that for a given $\epsilon > 0$, there exists a T_ϵ such that for all $T > T_\epsilon$:

$$\sup_{P \in \mathcal{P}} \left|\mathbb{P}_P\left(\Delta_{j,T} \in \mathcal{I}_\alpha\right) - (1-\alpha)\right| < \epsilon,$$

Proposition 3. *(Uniform Hypothesis Test) Let $\widehat{\mathbf{\Omega}}_T$ be a consistent estimator for $\mathbf{\Omega}_T$ uniformly in $P \in \mathcal{P}$. Under the same conditions of Theorem 1, for a given $\epsilon > 0$, there exists a T_ϵ such that for all $T > T_\epsilon$:*

$$\sup_{P \in \mathcal{P}} \left|\mathbb{P}_P\left(W_T \leq c_\alpha\right) - (1-\alpha)\right| < \epsilon,$$

where $W_T \equiv T\widehat{\mathbf{\Delta}}_T'\widehat{\mathbf{\Omega}}_T^{-1}\widehat{\mathbf{\Delta}}_T$, $\mathbb{P}(\chi^2_q \leq c_\alpha) = 1-\alpha$ and χ^2_q is a chi-square distributed random variable with q degrees of freedom.

4. Extensions

We consider extensions of the framework developed previously. In Section 4.1 we deal with the problem of an unknown intervention time and propose a procedure to account for that and develop a consistent estimator for the most likely intervention time. The case of multiple intervention points is treated in Section 4.2 and, finally, Section 4.3 investigates the presence of treated unit among the controls, which is particularly useful for testing for spillover effects.

4.1. Unknown Intervention Timing.

There are reasons why the intervention timing might not be known for certainty. It could be due to anticipation effects related to rational expectations regarding an announced change in future policy. Or, on the other hand, a simple delay in the response of the variable of interest. Regardless of the cause of uncertainty, we propose a way to apply the methodology even when T_0 is unknown.

We start by reinterpreting our estimator as a function of λ (or $T_\lambda \equiv \lfloor \lambda T \rfloor$), where $\lambda \in \Lambda$, a compact subset of $(0, 1)$:

$$\widehat{\mathbf{\Delta}}_T(\lambda) = \frac{1}{T - T_\lambda + 1}\sum_{t \geq T_\lambda} \widehat{\boldsymbol{\delta}}_{t,T}(\lambda), \quad \forall \lambda \in \Lambda \tag{11}$$

where $\widehat{\boldsymbol{\delta}}_{t,T}(\lambda) = \boldsymbol{y}_t - \widehat{\mathcal{M}}_T(\lambda)(\boldsymbol{x}_t)$, for $t = T_\lambda, \ldots, T$, and $\widehat{\mathcal{M}}_T(\lambda)$ is the estimate of the model $\mathcal{M}$ based on the first $T_\lambda - 1$ observations. Also, consider a λ-dependent version of our average treatment effect, given by

$$\mathbf{\Delta}_T(\lambda) = \frac{1}{T - T_\lambda + 1}\sum_{t=T_\lambda}^{T} \boldsymbol{\delta}_t.$$

For fix λ provided that the condition of Proposition 1 are satisfied for T_λ (as opposed to just $T_0 \equiv T_{\lambda_0}$) we have the convergence in distribution to a Gaussian. Hence, it is sufficient to consider the following extra assumption.

ASSUMPTION 5. $\{(\boldsymbol{y}_t', \boldsymbol{x}_t')'\}$ *is a strictly stationary process.*

Assumption 5 above is clearly stronger than necessary. For instance, it would be enough to have $\{\boldsymbol{\nu}_t\}$ as a weakly stationary process. However, in order to avoid assumptions that are model dependent (via the choice of $\mathcal{M}$) we state Assumption 5 as it is. It follows for instance if the process that generates the observable data in the absence of the intervention $\{\boldsymbol{z}_t^{(0)}\}$ is strictly stationary and both transformations $\boldsymbol{h}(\cdot)$ and $\boldsymbol{h}_x(\cdot)$ are measurable.

In order to analize the properties of the estimator (11) it is convenient to define the stochastic process $\{\boldsymbol{S}_T\}$ index by $\lambda \in \Lambda$, such that for each $\lambda \in \Lambda$, we have $\boldsymbol{S}_T(\lambda) \equiv \sqrt{T}\boldsymbol{\Gamma}_T^{-1/2}\left[\boldsymbol{\Delta}_T(\lambda) - \boldsymbol{\Delta}_T(\lambda)\right]$. Note that unlike the notation used in Proposition 1, we do not include the factors T_1/T and T_2/T inside the asymptotic variance term also since all the results will be under stationarity (Assumption 5) we replace $\boldsymbol{\Gamma}_{T_1}$ and $\boldsymbol{\Gamma}_{T_2}$ by its asymptotic equivalent $\boldsymbol{\Gamma}_T$, which is independent of $\lambda \in \Lambda$.

Therefore, the convergence in distribution of $\boldsymbol{S}_T(\boldsymbol{\lambda})$ to a Gaussian for any finite dimension $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_k)'$ follows directly from Theorem 1 combined with Assumption 5 and the Cramèr-Wold device. Furthermore the next theorem shows that $\boldsymbol{S}_T$ converges uniformly in $\lambda \in \Lambda$.

THEOREM 2. *Under the conditions of Proposition 1 and Assumption 5:*

$$\boldsymbol{S}_T(\lambda) \equiv \sqrt{T}\boldsymbol{\Gamma}_T^{-1/2}\left[\boldsymbol{\Delta}_T(\lambda) - \boldsymbol{\Delta}_T(\lambda)\right] \xrightarrow{d} \boldsymbol{S} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}(\lambda, \lambda') = \frac{I_q}{(\lambda \vee \lambda')(1 - \lambda \wedge \lambda')}$, $\forall (\lambda, \lambda') \in \Lambda^2$. *For* $p \in [1, \infty]$, $\|\boldsymbol{S}_T\|_p \xrightarrow{d} \|\boldsymbol{S}\|_p$, *where* $\|f\|_p = \left(\int |f(x)|^p dx\right)^{1/p}$ *if* $1 \leq p \leq \infty$ *and* $\|f\|_\infty = \mathsf{sup}_{x \in \mathcal{X}} |f(x)|$.

The second part of Theorem 2 gives us a direct approach to conduct inference in the case of unknown intervention time. We can replace $\boldsymbol{\Gamma}_T$ by a consistent estimator $\widehat{\boldsymbol{\Gamma}}_T$ (as for instance the one discussed in in Section 3.3) and conduct inference on $\|\widehat{\boldsymbol{S}}_T\|_p$ under a slightly stronger version of $\mathcal{H}_0$, (which clearly implies $\mathcal{H}_0$):

$$\mathcal{H}_0^\lambda : \boldsymbol{\delta}_t = \boldsymbol{0}, \quad \forall t \geq 1.$$

In practice, as it is the case for the structural breaks tests, we trim the sample to avoid finite sample bias close to the boundaries and select $\Lambda = [\underline{\lambda}, \bar{\lambda}]$. Table 1 presents the critical values for common choices of $p = \{1, 2, \infty\}$ and trimming values.

The procedure above suggests a natural estimator for the unknown intervention time, which might be useful in situations such as the one discussed in Section 4.2 where we treat multiple unknown intervention times.

We assume a constant intervention such as

ASSUMPTION 6. $\boldsymbol{\delta}_t = \boldsymbol{\Delta}$, *for* $t = T_0, \dots, T$, *where* $\boldsymbol{\Delta} \in \mathbb{R}^q$ *is non-random.*

Remark 3. *Recall that Assumption 6 is not overly restrictive due to the flexibility provided by the transformation $h(.)$. The mean of $\boldsymbol{y}_t$ might as well represent the variance, covariances or any other moment of interest of the original $\boldsymbol{z}_{1t}$ variable.*

Remark 4. *Assumption 6 implies an instantaneous treatment effect (step function) at $t = T_0$. In most cases, however, we might encounter a continuous intervention effect, possibly reaching a distinguishable new steady state value. We could accommodate these cases by trimming this transitory part of the sample, provided we have enough data, and then apply the methodology in the trimmed sample where the Assumption 6 holds.*

Proposition 4. *Under the conditions of Proposition 1 and Assumptions 5 and 6, $\widehat{\boldsymbol{\Delta}}_T(\lambda) \xrightarrow{p} \phi(\lambda)\boldsymbol{\Delta}$, where*

$$\phi(\lambda) = \begin{cases} \frac{1-\lambda_0}{1-\lambda} & \text{if } \lambda \leq \lambda_0, \\ \frac{\lambda_0}{\lambda} & \text{if } \lambda > \lambda_0. \end{cases}$$

Since both $\frac{1-\lambda_0}{1-\lambda}$ and $\frac{\lambda_0}{\lambda}$ are bounded between 0 and 1, we have that $\|\mathsf{plim}\, \widehat{\boldsymbol{\Delta}}_T(\lambda)\|_p \leq \|\boldsymbol{\Delta}\|_p$ for all $\lambda \in \Lambda$, where $\|\cdot\|_p$ denotes the ℓ_p norm. Under the maintained hypothesis that $\boldsymbol{\Delta} \neq 0$, we can establish the identification result that $\mathsf{plim}\, \widehat{\boldsymbol{\Delta}}_T(\lambda) = \boldsymbol{\Delta}$ if and only if $\lambda = \lambda_0$. The result above naturally suggests an estimator for λ_0:

$$\widehat{\lambda}_{0,p} = \underset{\lambda \in \Lambda}{\arg\max}\, J_{T,p}(\lambda) \quad \text{and} \quad J_{T,p}(\lambda) \equiv \|\widehat{\boldsymbol{\Delta}}_T(\lambda)\|_p. \tag{12}$$

Theorem 3. *Let $p \in [1, \infty]$. Under the conditions of Proposition 1 and Assumptions 5 and 6, for $\boldsymbol{\Delta} \neq \boldsymbol{0}$, $\widehat{\lambda}_{0,p} = \lambda_0 + o_p(1)$. If $\boldsymbol{\Delta} = \boldsymbol{0}$, $\widehat{\lambda}_{0,p}$ converges in probability to any $\lambda \in \Lambda$ with equal probability.*

4.2. **Multiple Intervention Points.** We can readily extend our analysis to the case of more than one intervention taking place in the unit of interest as long as, in each of them, Assumption 6 is valid. Suppose that we have S ordered known intervention points corresponding to the fractions of the sample given by $\lambda_0 \equiv 0 < \lambda_1 < \cdots < \lambda_S < 1 \equiv \lambda_{S+1}$.

For each of the intervention points $s = \{1, \ldots, S\}$ we can define the time of each intervention by $T_s \equiv \lfloor \lambda_s T \rfloor$ and construct our estimator the same way we did for the single intervention case. To simplify notation we define the set of all periods after intervention s but before the intervention $s+1$ by $\tau_s = \{T_s, T_s+1, \ldots, T_{s+1}-1\}$ and set $\#\{A\}$ the number of elements in the set A. Then we have S estimators given by:

$$\widehat{\boldsymbol{\Delta}}_T^s \equiv \widehat{\boldsymbol{\Delta}}_T(\lambda_s, \widehat{\boldsymbol{\theta}}_s) = \frac{1}{\#\{\tau_s\}} \sum_{t \in \tau_s} \left[\boldsymbol{y}_t - \mathcal{M}_p(\boldsymbol{x}_t, \widehat{\boldsymbol{\theta}}_{s,T})\right], \qquad s = 1, \ldots, S,$$

where once again $\widehat{\boldsymbol{\theta}}_{s,T}$ is the LASSO estimator using the sample index by $t \in \tau_{s-1}$. Note that we could allow the linear model to depend on s, i.e., differ from one intervention point to another. However, a much more parsimonious estimation could be obtained by choosing the same model to all intervention periods.

Under the same set of assumptions for the single intervention case plus Assumption 6, we have the sequence of estimators $\{\widehat{\boldsymbol{\Delta}}_T^s\}_{s=1}^S$ consistent to their respective intervention effects $\{\boldsymbol{\Delta}^s\}_{s=1}^S$ and also asymptotically normal. However, we need to make a minor adjustment in the asymptotic covariance matrix to reflect the intervention timing as:

$$\sqrt{T}\boldsymbol{\Gamma}_T^{-1/2}\left(\widehat{\boldsymbol{\Delta}}_T^s - \boldsymbol{\Delta}^s\right) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \frac{1}{(\lambda_s - \lambda_{s-1})(\lambda_{s+1} - \lambda_s)}\right], \quad s = 1, \ldots, S.$$

Since under Assumption 6 all the interventions are constant, we have that the asymptotic variance $\boldsymbol{\Gamma}$ is the same across all intervention points. Therefore, we can apply the inference for each breaking point as we have described for the single intervention case.

On the other hand, if the intervention points are unknown, we need to first estimate their location as in the single intervention case. Since the intervention points are assumed to be distinct, i.e. $\lambda_i \neq \lambda_j$, $\forall i, j$, it follows from Proposition 4 that there exists an interval of size $\epsilon > 0$ around every intervention point such that

$$\widehat{\boldsymbol{\Delta}}_T^p(\lambda) \xrightarrow{p} \begin{cases} \frac{1-\lambda_p}{1-\lambda}\boldsymbol{\Delta} & \text{if } \lambda \in [\lambda_p - \epsilon/2, \lambda_p], \\ \frac{\lambda_p}{\lambda}\boldsymbol{\Delta} & \text{if } \lambda \in (\lambda_p, \lambda_p + \epsilon/2]. \end{cases}$$

Nonetheless, in contrast to the single intervention scenario, in the case of multiple intervention points we need first to estimate how many are they and their respective location to construct $\{\widehat{\boldsymbol{\Delta}}_T^p\}_{p=1}^P$. One approach is to start with the null hypothesis of no intervention $(s = 0)$ against the alternative of a single one. We can then compute $\widehat{\lambda}_1$ as in (12) and test the null using $\widehat{\boldsymbol{\Delta}}_T^0(\widehat{\lambda}_1)$. In case we are able to reject the null, we split the sample at $\widehat{\lambda}_1$ and repeat the procedure in each of the two subsample. Every time we reject the null we split the sample in $\widehat{\lambda}_s$ and proceed sequentially until we no longer reject the null in any subsample.

The sequential procedure described above was advocated by Bai and Perron (1998). It in based on the observation that given a non-zero number of true intervention points, the first loop will encounter the most significant one (in terms of SSR reduction) and proceed sequentially until it finds the last one of them. In case we have multiple intervention points with the same magnitude the method would converge to any of them with equal probability.

Formally, starting from an arbitrary number of $s \geq 0$ intervention points and for a given significance level α we test for each of the $s + 1$ subsamples as:

$$\begin{aligned} \mathcal{H}_0^{(s)} &: \boldsymbol{\Delta} = \mathbf{0} \quad \text{for all } \lambda \in [\lambda_j, \lambda_{j+1})_{j=0}^s, \\ \mathcal{H}_1^{(s+1)} &: \boldsymbol{\Delta} \neq \mathbf{0} \quad \text{for any } \lambda \in [\lambda_j, \lambda_{j+1})_{j=0}^s. \end{aligned}$$

Note that the overall significance level of the test is no longer the individual significance level and it has to be adjusted to account for the sequential nature of the procedure.

4.3. **Testing for the unknown treated unit/Untreated peers.** All the analysis carried on so far relies on the knowledge of which unit is the treated one an also, more importantly, on the premisses that the remaining are in fact untreated during the sample period (Assumption 1). Yet, there might be cases where we are either unsure or would like to test for those conditions. Given any *finite* subset $\mathcal{I}$ of the available units we would like to test the following hypothesis

$$\mathcal{H}_0^n : \boldsymbol{\Delta}_T^{(i)} = \mathbf{0} \quad \forall i \in \mathcal{I} \subseteq \{1, \dots, n\}$$
$$\mathcal{H}_1^n : \boldsymbol{\Delta}_T^{(i)} \neq \mathbf{0} \quad \text{for some } i \in \mathcal{I}$$

Nothing prevent us from running the same procedure considering each unit $i \in \mathcal{I}$ to be the treated one to obtain $\widehat{\boldsymbol{\Delta}}_T^{(i)}$ as in (5) for $i = 1, \dots, n_{\mathcal{I}}$, where $n_{\mathcal{I}} < \infty$ is the cardinality of the set $\mathcal{I}$. We can then stack all of them in a vector as $\widehat{\mathbf{\Pi}}_T(\mathcal{I}) \equiv \left(\widehat{\boldsymbol{\Delta}}_T^{(1)'} \dots \widehat{\boldsymbol{\Delta}}_T^{(n_{\mathcal{I}})'}\right)'$ as an average estimator for the true average intervention effect vector $\mathbf{\Pi}_T(\mathcal{I}) \equiv \left(\boldsymbol{\Delta}_T^{(1)'} \dots \boldsymbol{\Delta}_T^{((\mathcal{I}))'}\right)'$ where $\boldsymbol{\Delta}_T^{(i)}$ is defined for each unit. Hence,

Proposition 5. *Under the conditions of Proposition 1, for any* finite *subset* $\mathcal{I} \subseteq \{1, \dots, n\}$

$$\sqrt{T}\mathbf{\Sigma}_{\mathcal{I}}^{-1/2}\left[\widehat{\mathbf{\Pi}}_T(\mathcal{I}) - \mathbf{\Pi}_T(\mathcal{I})\right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}),$$

where $\mathbf{\Sigma}_{\mathcal{I}}$ *is a covariance matrix with typical (matrix) element* $(i,j) \in \mathcal{I}^2$ *given by:*

$$\mathbf{\Omega}_T^{ij} \equiv T\mathbb{E}\left[\left(\widehat{\boldsymbol{\Delta}}_T^{(i)} - \boldsymbol{\Delta}_T^{(i)}\right)\left(\widehat{\boldsymbol{\Delta}}_T^{(j)} - \boldsymbol{\Delta}_T^{(j)}\right)'\right],$$

with $\mathbf{\Omega}_T^{ij} = \frac{\mathbf{\Gamma}_{T_1}^{ij}}{T_1/T} + \frac{\mathbf{\Gamma}_{T_2}^{ij}}{T_2/T}$, $\mathbf{\Gamma}_{T_1}^{ij} = \mathbb{E}\left[\frac{(\sum_{t \leq T_1} \boldsymbol{\nu}_t^i)(\sum_{t \leq T_1} \boldsymbol{\nu}_t^{j'})}{T_1}\right]$, *and* $\mathbf{\Gamma}_{T_2}^{ij} = \mathbb{E}\left[\frac{(\sum_{t \geq T_0} \boldsymbol{\nu}_t^i)(\sum_{t \geq T_0} \boldsymbol{\nu}_t^{j'})}{T_2}\right]$.

Therefore, for a given consistent estimator $\widehat{\mathbf{\Sigma}}$ *we have under* $\mathcal{H}_0^n$*:*

$$W_T^{\pi} \equiv T\widehat{\mathbf{\Pi}}_T'\widehat{\mathbf{\Sigma}}_{\mathcal{I}}^{-1}\widehat{\mathbf{\Pi}}_T \xrightarrow{d} \chi_{nq}^2.$$

We can obtain a consistent estimator for $\mathbf{\Sigma}_{\mathcal{I}}$ repeating the same procedure described in Section 3.3 for each pair $(ij) \in \mathcal{I}^2$ to obtain $\widehat{\mathbf{\Omega}}^{ij}$ and finally construct the matrix $\widehat{\mathbf{\Sigma}}_{\mathcal{I}}$. Hence for a desirable significance level, we can then use W_T^{π} to test $\mathcal{H}_0^n$. Once you remove the (likely) treated unit and re-test it again with the remanning units (peers) the test becomes yet more useful. In case we fail to reject the null, we can interpreted this result as a direct evidence in favour of the hypothesis that the peers are in fact untreated considering the sample at hand. Which ultimately provides support to our key Assumption 1.

5. Selection Bias, Contamination, Nonstationarity and Other Issues

In this section we discuss some possible sources of bias in the ArCo method. In particular, we consider the potential effects when the intervention does not affect only the outcome of the variable of the unit 1. Equivalently, we investigate the consequences whenever Assumption 1(b) fails and we expect to have $\mathbb{E}(\boldsymbol{z}_{0t}|\mathcal{D}_t) \neq 0$.

We consider without loss of generality a simpler version of the DGP described in Section 2. Each unit $i = 1, \dots, n$ under no intervention is represented by $z_{it}^{(0)} = l_i f_t + \eta_{it}$, where η_{it} is an zero mean independent and identically distributed (iid) idiosyncratic shock with variance

$\sigma^2_{\eta_i}$. Furthermore, $\mathbb{E}(\eta_{it}\eta_{jt}) = 0$, for all $i \neq j$. Also, the common factor vector f_t is an iid random variables with zero mean and variance σ^2_f.

Set $y_t = z_{1t}$, $\boldsymbol{x}_t = (z_{2t}, \ldots, z_{nt})'$, $\boldsymbol{l}_0 = (l_2, \ldots, l_n)'$ and $\boldsymbol{\sigma}^2_{\eta_0} = (\sigma^2_{\eta_2}, \ldots, \sigma^2_{\eta_n})'$. In this setup we can write

$$\begin{pmatrix} \boldsymbol{y}_t \\ \boldsymbol{x}_t \end{pmatrix} \sim \left[\boldsymbol{0}, \sigma^2_f \begin{pmatrix} l_1^2 + r_1 & l_1 \boldsymbol{l}_0' \\ l_1 \boldsymbol{l}_0 & \boldsymbol{l}_0 \boldsymbol{l}_0' + \mathsf{diag}\,(\boldsymbol{r}_0) \end{pmatrix} \right],$$

where $r_i \equiv \frac{\sigma^2_{\eta_i}}{\sigma^2_f}$ is the noise to signal ratio of unit $i = 1, \ldots, n$ and $\boldsymbol{r}_0 = (r_2, \ldots, r_n)'$.

As a consequence, the best linear projection model is given by $\mathbb{L}(\boldsymbol{y}_t|\boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0 = [\boldsymbol{l}_0\boldsymbol{l}_0' + \mathsf{diag}\,(\boldsymbol{r}_0)]^{-1}(l_1\boldsymbol{l}_0)$. Furthermore, $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_0 + \nu_t$, where $\mathbb{E}(\boldsymbol{x}_t\nu_t) = \boldsymbol{0}$ by definition, and $\sigma^2_\nu \equiv \mathbb{E}(\nu_t^2) = \sigma^2_f\,(l_1^2 + r_1 - \boldsymbol{\beta}_0' l_1 \boldsymbol{l}_0)$.

Therefore, we have that $\boldsymbol{\beta}_0 \equiv \boldsymbol{\beta}_0(\boldsymbol{l}, \boldsymbol{r})$ and $\sigma^2_\nu \equiv \sigma^2_\nu(\boldsymbol{l}, \boldsymbol{r}, \sigma^2_f)$, where $\boldsymbol{r} = (r_1, \boldsymbol{r}_0')'$ and $\boldsymbol{l} = (l_1, \ldots, l_n)'$.

Suppose now that we have an intervention affecting all units from T_0 onwards, i.e. Assumption 1(b) does *not* hold. We consider both situations, where the intervention is change in the common factor given by a deterministic sequence $\{c^f_t\}_{t \geq T_0}$ and one completely idiosyncratic $\{c^i_t\}_{t \geq T_0}$ for $i = 1, \ldots, n$, $z^{(1)}_{it} = z^{(0)}_{it} + 1\{t \geq T_0\}\left(c^i_t + l_i c^f_t\right)$.

Consequently, for $t = T_0, \ldots, T$:

$$\delta_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_0 = y^{(0)}_t + c^1_t + l_1 c^f_t - \left(\boldsymbol{x}^{(0)}_t + \boldsymbol{c}^0_t + \boldsymbol{l}_0 c^f_t\right)' \boldsymbol{\beta}_0 = c^1_t + \nu_t - \boldsymbol{c}^{0\prime}_t \boldsymbol{\beta}_0 + (l_1 - \boldsymbol{l}_0'\boldsymbol{\beta}_0)\, c^f_t.$$

Clearly, under Assumption 1(b), we have that $\boldsymbol{c}^{(0)}_t = c^f_t = 0, \forall t$, thus $\mathbb{E}(\delta_t) = c^1_t$ and, ignoring the sampling error of estimating $\boldsymbol{\beta}_0$, the ArCo estimator will be unbiased for the average of c^1_t for the post intervention period. On the other hand, without those assumptions we have the following bias in normalized statistic

$$b_t \equiv \mathbb{E}\left(\frac{\delta_t - c^1_t}{\sigma_\nu}\right) = \underbrace{\left(\frac{l_1 - \boldsymbol{l}_0'\boldsymbol{\beta}_0}{\sigma_\nu}\right)}_{\equiv \phi_f} c^f_t - \frac{\boldsymbol{c}^{0\prime}_t \boldsymbol{\beta}_0}{\sigma_\nu} \tag{13}$$

The factor in the first term of the bias $\phi_f = \phi_f(\boldsymbol{l}, \boldsymbol{r}, \sigma^2_f)$ is a non-linear expression which is hard to express in closed form. However, regardless of the choice of the factor loads $\boldsymbol{l}$ and idiosyncratic shock variances $\boldsymbol{\sigma}^2_\eta = (\sigma^2_{\eta_1}, \ldots, \sigma^2_{\eta_n})'$, we have that as $\sigma^2_f \to \infty$, $r \to 0$ and consequently $R^2 \to 1$. Hence we write $\phi_f = \phi_f(R^2)$. Moreover, $\phi_f(R^2)$ is strictly decreasing in R^2 and approach zero quite fast as it can be seen in the left scale of Figure 1. Also $\phi_f = \phi(s_0)$ is also decreasing in the number of relevant variables s_0 for fix R^2.

Hence, if $\boldsymbol{c}^0_t = \boldsymbol{0}$ but $c^f_t \neq 0$, even with moderate R^2, we have a reasonable small bias which cause the inference to be valid with minor overejection. This is in contrast to the case where we do not include relevant peers in our analysis . In fact, as mention previous in the Introduction, that is the main motivation for using the present methodology as oppose to an alternative that does not involve peers (a simple before-and-after estimation of averages for instance). ArCo can effectively isolate the intervention of interest even in the case of partially fulfilment of Assumption 1. In the limit of a perfect counterfactual, the bias is zero and the higher is the correlation among the treated unit and the peers, the smaller is the bias.

The second bias term in (13) can be seen as a result for instance of global shock that induce breaks in peers in non-systematic way, which makes this source of bias difficult to handle. To get a better sense, consider for instance the case where idiosyncratic shock is a fix proportion of the standard deviation of each unit, i.e. $c_t^i = k\sigma_i, \forall i$ for some $k \in \mathbb{R}$. In that case, $\phi_g = (\boldsymbol{\sigma}'\boldsymbol{\beta}_0/\sigma_\nu)k$, where $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)'$. Here the opposite happens, namely $\phi_g(R^2)$ is zero when $R^2 = 0$ and increases in the overall fit of the model. The bias increase is quite sharp as can been seen in the right scale of Figure 1.

Therefore, whenever one expect $\boldsymbol{c}_t^0 \neq \mathbf{0}$, the ArCo methodology does not work properly but the BA estimator does as it can be seen as a particular case of the ArCo estimator with $R^2 = 0$ (for instance by not including any peers) and hence the bias is zero. In general, the ArCo estimator gives the difference between the actual break in the treated unit and what is expected from the peers. A standard solution is to assume that the "treatment assignment" is independent of $\boldsymbol{z}_{0t} = (z_{2t}, \dots, z_{nt})'$, which is our Assumption 1 and the ArCo approach is not subject to selection bias. However, it is important to stress that the "treatment assignment" might be dependent on z_{1t} and our approach is still valid.[7] One way to check if there is no "treatment contamination" is to test the peers for possible breaks after T_0 as discussed in Section 4.3.

Other possible source of problems is the use of "non-stationary" processes, leading to spurious results. In this paper we focus solely in the case the variables of interest have some sort of "fading memory" behaviour. Thus, if one or more variables are found to be integrated, they must be differenced first in order to achieve stationarity. A full discussion of integrated process of order 1 can be seen in Masini and Medeiros (2016).

6. Monte Carlo Simulation

We conducted two sets of Monte Carlo simulations. First, we conduct size and power simulations in order to investigate the finite sample properties of the test. We consider a broad range of cases by combining different innovation distributions, sample sizes, number of peers, number of relevant peers, dependence structure, trends and intervention types. Second, a "horse race" is proposed in order to compare the ArCo estimator with potential alternatives. We consider the SC method of Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010), the PF estimator suggested in Gobillon and Magnac (2016) and the DiD and BA estimators.

6.1. **Size and Power Simulations.** The DGP considered is a version of the common factor model (6) with the following baseline scenario: $T = 100$ observations, $n = 100$ units, $q = 1$

[7] The result is analogous to the average treatment effect on the treated not being biased by selection on (un)observables

one variable per unit, $\lambda_0 = 0.5$ (intervention at the middle of the sample), $s_0 = 5$ relevant (non-zero) parameters with loading factor equal to 1 and $f = 1$ common factor. The common factor and all idiosyncratic shocks are independent and identically normally distributed with zero mean and unit variance. We perform 10,000 simulations.

First, we analyze the influence of the underlying distribution on the test size by holding all the other parameters above fixed and performing the simulation for a chi-square distribution with 1 degree of freedom for asymmetry issues, t-Student distribution with 3 degrees of freedom for fat-tails and a mixed normal distribution for bimodality.[8] As shown in first panel of Table 2, little influence in the overall size of the test is perceived.

Next we analyze the influence of the number of observations in the test size. We consider $T = \{25, 50, 75, 100\}$. Surprisingly, the size distortions are small even with only 50 observations as shown in the second panel of Table 2. We stress that since we deal with the intervention at the middle of the sample we have less than $T/2$ observations to fit the high dimensional model.

We now investigate the influence of increasing the number of covariates (by increasing either the number of lags or the number of peers)[9]. We set $d = \{100, 200, 500, 1000\}$. The third panel of Table 2 shows that the test size seems to be unaffected by the increase in the model complexity. This should come with no surprise since consistent model selection is not an issue for the methodology to work.

We consider a change of relevant (non-zero) covariates (units) in the pre-intervention model. We consider a case where all the regressors are irrelevant, which reduces (asymptotically) the ArCo to the BA estimator, and we further increase s_0. In the last scenario we consider all regressors non-zero but with decreasing magnitude $1/\sqrt{j}, j = 1, \ldots, 100$. In all cases the LASSO do not overfit the pre-intervention data and the size distortions are small as displayed in Table 2.

Finally, we consider the case where each unit follows a first-order autoregressive process in order to investigate issues that arise in the presence of serial correlation. In this scenario we include lags of the relevant covariates instead of new peers. The results are shown in the last panel of Table 2. We note a persistent oversized test, which is more pronounced as the autoregressive coefficient (ρ) becomes closer to 1. The empirical distribution of the estimator (not shown) is, however, very close to normal, and the distortion is a sole consequence of the poor finite sample properties of the variance estimator . In particular it underestimate $\boldsymbol{\Omega}$. We tried several alternatives for $\widehat{\boldsymbol{\Omega}}_T$, including Newey and West (1987), Andrews (1991), Andrews and Monahan (1992), and Haan and Levin (1996). We obtain the best results (last panel of Table 2) using the procedure proposed in Andrews and Monahan (1992).

It is worth mention that the slightly oversized test are a direct consequence of the persistence of $\{\nu_t\}$ and not necessarily from the persistence of $\{(y_t, \boldsymbol{x}_t')\}$ per se. The problem

[8] All innovations are standardized to zero mean and unit variance.

[9] The difference is not completely innocuous since we loose one observation to each included lag. Therefore, we include new (uncorrelated) peers and deal with the lag inclusion in the serial correlation scenario

is attenuated, for instance, when enough lags are included to make $\{\nu_t\}$ closer to a white noise process, or when a linear combination of (potentially highly persistent) $\{(y_t, \boldsymbol{x}_t')\}$ is almost uncorrelated. For pure finite MA process the usual kernel HAC estimator are known to perform well and the tests are not oversized.

6.2. **Estimator Comparison.** In order to conduct the "horse race" among competitors for counterfactual analysis we consider the following DGP:

$$\boldsymbol{z}_{it}^{(0)} = \rho \boldsymbol{A}_i \boldsymbol{z}_{it-1}^{(0)} + \boldsymbol{\varepsilon}_{it}, \quad i = 1, \ldots, n, ; t = 1, \ldots, T, \tag{14}$$

where $\boldsymbol{\varepsilon}_{it} = \boldsymbol{\Lambda}_i \boldsymbol{f}_t + \boldsymbol{\eta}_{it}$, $\boldsymbol{f}_t = [1, (t/T)^{\varphi}, v_t]$, $\boldsymbol{z}_{it} \in \mathbb{R}^q$, $\rho \in [0, 1)$, $\varphi > 0$, $\boldsymbol{A}_i(q \times q)$ is a diagonal matrix with diagonal elements strictly between -1 and 1, $\{v_t\}$ is a sequence of iid standardized normal random variables, $\{\boldsymbol{\eta}_{it}\}$ is a sequence of iid normal random vectors with zero mean and covariance matrix $r_f^2 \boldsymbol{I}_{nq}$ where $r_f > 0$ can be interpreted as the noise-to-signal ratio which controls the overall correlation among the units, and $\boldsymbol{\Lambda}_i$ is a $(q \times 3)$ matrix of factor loadings.

Let $\boldsymbol{z}_t$ be the nq dimensional vector obtaining by stacking all the $\boldsymbol{z}_{it}^{(0)}$ and $\boldsymbol{\Lambda}$ is the $(nq \times 3)$ matrix after stacking all the $\boldsymbol{\Lambda}_i$. Similarly, define $\boldsymbol{\varepsilon}_t$ by stacking $\boldsymbol{\varepsilon}_{it}$ and $\boldsymbol{A}$ is the $(nq \times nq)$ diagonal matrix composed by the block diagonals $\boldsymbol{A}_i$. We use the notation $\boldsymbol{\Lambda}(j)$ to denote the jth column of $\boldsymbol{\Lambda}$, thus $\boldsymbol{\mu}_{\varepsilon,t} \equiv \mathbb{E}(\boldsymbol{\varepsilon}_t) = \boldsymbol{\Lambda}(1) + \boldsymbol{\Lambda}(2)(t/T)^{\varphi}$, $\boldsymbol{\Omega} \equiv \mathbb{V}(\boldsymbol{\varepsilon}_t) = \boldsymbol{\Lambda}(3)\boldsymbol{\Lambda}(3)' + r_f^2 \boldsymbol{I}_{nq}$, $\boldsymbol{\mu}_t \equiv \mathbb{E}(\boldsymbol{z}_t) = (\boldsymbol{I}_{nq} - \rho \boldsymbol{A})^{-1} \boldsymbol{\mu}_{\varepsilon,t}$, and $\mathsf{vec}\,(\boldsymbol{\Sigma}) \equiv \mathsf{vec}\,[(\mathbb{V}\boldsymbol{z}_t)] = [\boldsymbol{I}_{(nq)^2} - \rho^2 \boldsymbol{A} \otimes \boldsymbol{A}]^{-1} \mathsf{vec}\,(\boldsymbol{\Omega})$.

We set $y_{it}^{(1)} = y_{it}^{(0)} + \delta_t 1\{t \geq T_0 \text{ and } i = 1\}$, for simplicity we set $\delta_t = \delta$ constant and equal to one standard deviation from the unit of interest (unit 1). We are interested in estimating the average treatment effect

$$\Delta = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \delta_t = \delta.$$

We now briefly state the estimators considered in the Monte Carlo study. Whenever is convenient we use the following partition scheme: $\boldsymbol{z}_{it} = (y_{it}, \boldsymbol{x}_{it}')'$ and $\boldsymbol{z}_{0t} = (\boldsymbol{z}_{2t}', \ldots \boldsymbol{z}_{nt}')$.

Before-and-After (BA). The difference between the average of the y_{1t} before and after the intervention:

$$\widehat{\Delta}_{BA} = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} y_{1t} - \frac{1}{T_0 - 1} \sum_{t=1}^{T_0 - 1} y_{1t}.$$

Differences-in-Differences (DiD). The ordinary least squares (OLS) estimator of the dummy coefficient in the following regression models. For the case with covariates,

$$y_{it} = \alpha_0 + \boldsymbol{x}_{it}' \boldsymbol{\beta} + \alpha_1 I(i = 1) + \alpha_2 I(t \geq T_0) + \Delta_{DD^*} I(i = 1, t \geq T_0) + \varepsilon_{it},$$

or, for the case without covariates,

$$y_{it} = \alpha_0 + \alpha_1 I(i = 1) + \alpha_2 I(t \geq T_0) + \Delta_{DD} I(i = 1, t \geq T_0) + \varepsilon_{it}.$$

Gobillon and Magnac (GM). The estimator is defined as per Gobillon and Magnac (2016):

$$\widehat{\Delta}_{GM} = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} (y_{1t} - \widehat{y}_{1t}),$$

where $\widehat{y}^*_{1t} = \boldsymbol{x}_{1t}\widehat{\boldsymbol{\beta}} + \widehat{f}_t\widehat{\Lambda}_1$ or without including the covariates $\widehat{y}_{1t} = \widehat{f}_t\widehat{\Lambda}_1$. Ee choose r the number of factors to be 2 (or 3 if a trend is included).

Synthetic Control (SC). For the simulation purposes we use the algorithm *Synth*[10]. We choose on top of all covariates ($\boldsymbol{x}_{it}$), the average of the dependent variable ($\boldsymbol{y}_{it}$) during the pre-intervention period as a matching variables.

$$\widehat{\Delta}_{SC} = \tfrac{1}{T-T_0+1} \sum_{t=T_0}^{T} (y_{1t} - \widehat{y}_{1t}),$$

where $\widehat{y}_{1t} = \boldsymbol{w}^{*\prime}\boldsymbol{y}_{0t}$. The weight vector $\boldsymbol{w}$ must be non-negative entries that sum to one. It comes from a minimization process involving only values of the selected variables prior to the intervention. In our particular case, we take the pre-intervention average $\bar{\boldsymbol{z}} = \frac{1}{T_0-1}\sum_{t=1}^{T_0-1} \boldsymbol{z}_t$, partition as $\bar{\boldsymbol{z}} = (\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_0{}')'$ and reshape $\bar{\boldsymbol{z}}_0$ to a matrix $\bar{Z}_0(n-1 \times q)$ where each row are the variables of each of the remaining $n-1$ units

$$\boldsymbol{w}^*(\boldsymbol{V}) = \underset{\boldsymbol{w}\geq 0, \|\boldsymbol{w}\|_1 = 1}{\operatorname{arg\,min}} \|\bar{\boldsymbol{z}}_1 - \boldsymbol{w}'\bar{\boldsymbol{z}}_0\|_{\boldsymbol{V}},$$

where $\|\cdot\|_V$ is the norm induced by a positive definite matrix $\boldsymbol{V}$.

Finally, $\boldsymbol{V}$ is chosen as

$$\boldsymbol{V}^* = \operatorname{arg\,min} \frac{1}{T_0 - 1} \sum_{t=1}^{T0-1} \left[y_{1t} - \boldsymbol{w}^*(\boldsymbol{V})'\boldsymbol{y}_{0t}\right]^2, \tag{15}$$

and we set $\boldsymbol{w}^* \equiv \boldsymbol{w}^*(\boldsymbol{V}^*)$.

The results are presented in Table 4. The smoothed histograms can be found in Figures 2–7. Overall, the SC and the GM are heavily biased in most cases considered. For the former, this might well be a consequence of the instability of algorithm to find the minimizer of (15), since the bias persist even in the absence of time trends, where any fix linear combination of the peers should give us an unbiased estimator. For the former it is most likely a consequence of the poor finite sample properties of common factor estimator. It is well understood from Bai (2009) that the consistency depends on the double asymptotics on n and T. On the other hand, BA, DiD and the ArCo seems to have comparable small bias at least in absence of deterministic trends regardless of the presence of serial correlation. The ArCo seems to have better MSE performance. This comes with no surprise since by definition our estimator in the first stage searches for the linear combination that minimizes the MSE.

For the trended cases, first note the BA estimator is severely biased since since without using the information of the peers it cannot take into account the time trend effect. For

[10] R package maintained by Jens Hainmueller

the common trend cases, the DiD estimators have relatively small bias for both the linear and quadratic term. For the former it is excepted since a common linear time trend the exactly the kind of DGP that the DiD estimator was designed for. Once again, the ArCo estimators have comparable bias to the DD estimators for the common trend cases but with significant smaller variance (ranging from 6-16 times smaller). The clear advantage of the ArCo estimation can be seem in the idiosyncratic time trend cases. Even though some small (in finite sample) bias start to show up, it is clear much smaller than all other alternatives.

7. The Effects of an Anti Tax Evasion Program on Inflation

In this section we apply the ArCo methodology to estimate the effects of an anti tax evasion program in Brazil on inflation. Although, the causes of business non-compliance and tax evasion has been extensively studied in the literature, see, for example, Slemrod (2010), little attention has been devoted to measure the indirect effects from enforcing tax compliance.

In Brazil, tax evasion is a major fiscal concern and both the federal and local governments have been proposing new strategies to reduce evasion. Early in 1996, the federal government introduced the SIMPLES[11] system which drastically simplified the tax payments process and helped in reducing the tax burden on small enterprises. Later in 2005, the federal government launched the electronic sales receipt program (*Nota Fiscal Eletrônica*), to further reduce compliance costs to firms and, even more important, to standardize the disclosure of taxable.

In October 2007, the state government of São Paulo in Brazil implemented an anti tax evasion scheme called *Nota Fiscal Paulista* (NFP) program. The NFP program consists of a tax rebate from a state tax named ICMS (tax on circulation of products and services). ICMS is similar to the European VAT and the Canadian GST. However, unlike VAT and GST, ICMS does not apply to services other than those corresponding to interstate and intercity transportation and communication services. The NFP program works as an incentive to the consumer to ask for electronic sales receipts. The registered sales receipts give the consumer the right to participate in monthly lotteries promoted by the government. Furthermore, according to the rules of the program, registered consumers have the right to receive part of the ICMS paid by the seller, as tax rebate, when their security numbers (CPF) are included in the electronic sales receipts. Similar initiatives relying on consumer auditing schemes were proposed in the European Union and in China; see, for example, Wan (2010). The effectiveness of such programs have been discussed in Fatas, Nosenzo, Sefton, and Zizzo (2015) and Brockmann, Genschel, and Seelkopf (2016). In the Brazilian state of São Paulo, the NFP program has received extensive support from the population. In January 2008, 413 thousand people were registered in program while in October 2013 there were more than 15 million participants. The amount in Brazilian Reais distributed as rebates also grew

[11] Integrated System of Tax Payments for Micro and Small Enterprises

rapidly from 44 thousand Reais in January 2008 to an average of 70 million Reais distributed monthly by the end of the same year. Figure 8 illustrates the NFP participation as well as the value distributed as tax rebates.

Souza (2014) was the first author to discuss whether retailers increased prices in response to the NFP program and consequently whether the program impacted negatively consumers' purchasing power. By using the SC method to construct a counterfactual to the State of São Paulo, Souza (2014) showed that one year after the launching of the NFP program, the accumulated inflation on food outside home (FOH) was 5% higher in the state of São Paulo when compared to the synthetic control. In September 2009, the differences raised to 6.5%. We extend the analysis of Souza (2014) by considering the ArCo methodology as an alternative to the SC method. We also consider the BA, GM, and DiD estimators.

Under the premisses that (i) a certain degree of tax evasion was occurring before the intervention, (ii) the sellers has some degree of market power and (iii) the penalty for tax evasion is large enough to alter the seller behaviour, one is expected to see an upwards movements in prices due to an increase in marginal cost. Therefore, we would like to investigate whether the NFP had an impact on consumer prices in São Paulo. We test this hypothesis below as an empirical illustration of the ArCo methodology. The answer to this kind of question has important implications regarding social welfare effects that are usually neglected in the fiscal debate whenever the aim is to enforce tax compliance

The NFP was not implemented throughout the sectors in the economy at once. The first sector were restaurants, followed by bakeries, bars and other food service retailers. We do not possess a perfect match for a general consumer price index (IPCA - IBGE) and the sector where the NFP was implemented. However, we can take the IPCA component of food outside home (FOH) as a good indicator for price levels in those sectors. The sample then consists of monthly FOH index for 10 metropolitan areas[12] including São Paulo from January 1995 to September 2009. As a matter of comparison, Souza (2014) estimated a counterfactual by the SC method with assigning the following weights to Belo Horizonte, Recife, Goiânia, and Porto Alegre, respectively: 0.40, 0.27, 0.19, and 0.14. All other donors were assigned zero weights.

In order to compute the counterfactual by the ArCo methodology we consider the following variables from the pool of donors: monthly inflation (FOH), monthly GDP growth, monthly retail sales growth and monthly credit growth. All variables are stationary and no lags or additional transformations are considered. The conditional model is linear and is estimated by LASSO, where the penalty parameter is selected by the Hannan and Quinn (HQ) criterium. The choice of the HQ instead of the BIC, for example, is that the latter delivers conditional models with no variables in most of the cases. The in-sample period (pre-intervention) consists of 33 observations while the size of the out-of-sample period is 23.

[12] Goiânia-GO, Fortaleza-CE, Recife-PE, Salvador-BA, Rio de Janeiro-RJ, São Paulo-SP, Porto Alegre-RS, Curitiba-PR, Belém-PA, Belo Horizonte-MG

The factor in the GM methodology is computed from the monthly growth in GDP, retail sales and credit by principal component methods. The number of factors are selected as to explain 80% of the total variance in the data. The BA estimator considers only variables from the treated unit.

The results are depicted in Table 5. The upper panel in the table reports, for different choices of conditioning variables, the estimated ATET after the adoption of the NFP. The standard errors are reported between parenthesis. Diagnostic tests do not evidence any residual autocorrelation and the standard errors are computed without any correction. The table also shows the R-squared of the first stage estimation, the number of included regressors in each case as well as the number of selected regressors by the LASSO. In all cases, the ATET is significant at the 1% level. The highest R-squared is achieved when inflation and GDP are used as conditioning variables, followed by a model with inflation, GDP and retail sales. In the first case, column (5) of Table 5, the monthly ATET is 0.4478%. The aggregated effect during the out-of-sample period is 10.72%. In the second case, column (6) of Table 5, the monthly ATET is 0.3796% and the aggregated effect is 9.04%. Two facts worth discussion. The first one is the much higher estimated effect when only credit variables are included. This is due to huge outliers (huge increase) observed in credit series in the out-of-sample period for the states of Pernambuco and Rio de Janeiro. If these two states are removed from the donors pool, the monthly ATET drops to 0.5768%. The second point that deserves attention is the much lower effect when only inflation is considered, although the in-sample fit is reasonably good.

Figure 9 and 10 show the actual and counterfactual data, both in-sample and out-of-sample. Figure 9 considers the case where only inflation and GDP growth are considered as conditioning variables while the plots in Figure 10 consider the case where retail sales growth are also included as a potential regressor in the first stage model.

The lower panel of Table 5 presents some alternative measures of the ATET, namely the BA, GM and DiD estimators. In all cases the estimated effects are smaller than the ones estimated with the ArCo. The DiD estimators are closer to the SC. The GM falls somehow in between the SC/DiD and the ArCo.

We also run a placebo ArCo estimator to check the robustness of the method. When we do this we find that Porto Alegre seems to have nontrivial breaks after October 2007; see Table 6. For this reason we re-run the analysis without Porto Alegre in the donor pool. The results are reported in Table 7. The overall picture seems unchanged.

8. Conclusions and Future Research

In this paper we proposed a new method to conduct counterfactual analysis with aggregated data, specially in situations where there is a single treatedunit and not "controls" are readily available. Our proposal called the Artificial Control (ArCo) share some common roots to Hsiao, Ching, and Wan (2012), the synthetic control method of Abadie and Gardeazabal

(2003) and Abadie, Diamond, and Hainmueller (2010). Comparing to other alternatives to conduct counterfactual analysis, we believe the ArCo method has several advantages: (1) It accommodates both high-dimensional covariates and multivariate unit of interest; (2) Possess a complete asymptotic theory which can be used to jointly test for intervention effects in a group of variables; (3) The counterfactual model can be written as a misspecified nonlinear function of observed variables for peers (untreated units); (4) The methodology can be applied even if the time of the intervention is not known for certain, which gives us a consistent estimator for the time of the intervention; (5) Multiple interventions can be handled; (6) We also propose a test for the presence of spillover effects among the units.

The current research can be extended in several directions as, for example, the case where the variables are nonstationary (either with cointegrated or not). A non-parametric or semiparametric estimation in the pre-intervention model. A Bayesian approach can also be easily accommodated in the present framework with the advantage of incorporating any pre-knowledge of the researcher about the pre-intervention model directly as priors.

Acknowledgemnts

We thank participants at the 2014 NBER-NSF Time Series Conference, the 2014 Meeting of the Brazilian Econometric Society (SBE), the 2015 World Congress of the Econometric Society in Montréal, the 2015 conference of the European Economic Association in Mannheim, the 2014 Workshop on Measuring and Modeling Financial Risk with High Frequency Data, the Workshop Rio-São Paulo of Econometrics and at seminars at Emory University, Ohio State University, Erasmus University and at the PUC-Rio Student Workshop for valuable suggestions. We are grateful to James Heckman and Peter C. B. Phillips for fruitful discussions on an earlier version of this paper presented during the conference in honor of Essie Maasoumi. We are also in debt to Alexandre Belloni for a decisive contribution towards the closure of this paper. We are also thankful to Bruno Ferman, Pedro Souza, Marcelo J. Moreira and Leonardo Rezende for great comments and discussions. Finally, we are specially thankful to Eduardo Mendes for hours of rich discussions as well as Fernando Friaça de Souza for rich discussions on the application of the method to the "Nota Fiscal Paulista". Any mistakes are our own.

Appendix A: Proofs

We begin by proving an uniform version for the Continuous Mapping Theorem (UCMT) and the Slutsky Theorem (UST). For the next 2 Lemmas, $\boldsymbol{X}_T$, $\boldsymbol{Y}_T$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random elements taking values on a subset $\mathcal{D}$ of the Euclidean space (real-valued scalar, vector or matrix) defined over the same probabilistic space with distribution P index by $\mathcal{P}$.

Lemma 1. ***(Uniform Continuous Mapping Theorem)*** *Let $\boldsymbol{g} : \mathcal{D} \to \mathcal{E}$ be uniformly continuous at every point of a set $\mathcal{C} \subseteq \mathcal{D}$ where $\mathbb{P}_P(\boldsymbol{X} \in \mathcal{C}) = 1$ for all $P \in \mathcal{P}$.*

(a) If $\boldsymbol{X}_T \xrightarrow{p} \boldsymbol{X}$ uniformly in $P \in \mathcal{P}$, then $\boldsymbol{g}(\boldsymbol{X}_T) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{X})$ uniformly in $P \in \mathcal{P}$.
(b) If $\boldsymbol{X}_T \xrightarrow{d} \boldsymbol{X}$ uniformly in $P \in \mathcal{P}$, then $\boldsymbol{g}(\boldsymbol{X}_T) \xrightarrow{d} \boldsymbol{g}(\boldsymbol{X})$ uniformly in $P \in \mathcal{P}$.

Proof. The proof is similar to the classical Continuous Mapping Theorem proof but with continuity replaced by uniform continuity. For (a), by the definition of uniform continuity, for any $\epsilon > 0$, there is a $\delta > 0$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ if $d_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{y}) \leq \delta \Rightarrow d_{\mathcal{E}}\left[\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{y})\right] \leq \epsilon$ for some metric $d_{\mathcal{D}}$ and $d_{\mathcal{E}}$, defined on $\mathcal{D}$ and $\mathcal{E}$ respectively. Therefore,

$$\mathbb{P}_P\left\{d_{\mathcal{E}}\left[\boldsymbol{g}(\boldsymbol{X}_T), \boldsymbol{g}(\boldsymbol{X})\right] > \epsilon\right\} \leq \mathbb{P}_P[d_{\mathcal{D}}(\boldsymbol{X}_T, \boldsymbol{X}) > \delta] + \mathbb{P}_P(\boldsymbol{X} \notin \mathcal{C}).$$

The result follows since the first term on the right hand side converges to zero uniformly in $P \in \mathcal{P}$ by assumption and the second is zero for all $P \in \mathcal{P}$ also by assumption.

For (b), given a set $E \in \mathcal{E}$ we have the preimage of $\boldsymbol{g}$ denoted by $\boldsymbol{g}^{-1}(E) \equiv \{\boldsymbol{x} \in \mathcal{D} : \boldsymbol{g}(\boldsymbol{x}) \in E\}$. For close $F \in \mathcal{E}$ we have that $\boldsymbol{g}^{-1}(F) \subset \overline{\boldsymbol{g}^{-1}(F)} \subset g^{-1}(F) \cup \mathcal{C}^c$ due to the continuity of $\boldsymbol{g}$ on $\mathcal{C}$. Clearly, the event $\{\boldsymbol{g}(\boldsymbol{X}_T) \in F\}$ is the same of $\{\boldsymbol{X}_T \in \boldsymbol{g}^{-1}(F)\}$, then we can write

$$\begin{aligned}\limsup \sup_{P\in\mathcal{P}} \mathbb{P}[\boldsymbol{X}_T \in \boldsymbol{g}^{-1}(F)] &\leq \limsup \sup_{P\in\mathcal{P}} \mathbb{P}[\boldsymbol{X}_T \in \overline{\boldsymbol{g}^{-1}(F)}] \\ &\leq \sup_{P\in\mathcal{P}} \mathbb{P}[\boldsymbol{X} \in \overline{\boldsymbol{g}^{-1}(F)}] \leq \sup_{P\in\mathcal{P}} \mathbb{P}[\boldsymbol{X} \in \boldsymbol{g}^{-1}(F)] + \underbrace{\sup_{P\in\mathcal{P}} \mathbb{P}(\boldsymbol{X} \notin \mathcal{C}\}}_{=0},\end{aligned}$$

where the second inequality is a consequence of the uniform convergence in distribution of $\boldsymbol{X}_T$ to $\boldsymbol{X}$ and the Portmanteau Lemma (Lemma 2.2 Van der Vaart, 2000). The result follows again by the Portmanteau Lemma in the other direction. □

Lemma 2. ***(Uniform Slutsky Theorem)*** *Let $\boldsymbol{X}_T \xrightarrow{p} \boldsymbol{C}$ uniformly in $P \in \mathcal{P}$, where $\boldsymbol{C} \equiv \boldsymbol{C}(P)$ is a non random conformable matrix and $\boldsymbol{Y}_T \xrightarrow{d} \boldsymbol{Y}$ uniformly in $P \in \mathcal{P}$, then*

(a) $\boldsymbol{X}_T + \boldsymbol{Y}_T \xrightarrow{d} \boldsymbol{C} + \boldsymbol{Y}$ uniformly in $P \in \mathcal{P}$
(b) $\boldsymbol{X}_T\boldsymbol{Y}_T \xrightarrow{d} \boldsymbol{C}\boldsymbol{Y}$ uniformly in $P \in \mathcal{P}$, if $\boldsymbol{C}$ is bounded uniformly in $P \in \mathcal{P}$.
(c) $\boldsymbol{X}_T^{-1}\boldsymbol{Y}_T \xrightarrow{d} \boldsymbol{C}^{-1}\boldsymbol{Y}$ uniformly in $P \in \mathcal{P}$, if $\det(\boldsymbol{C})$ is bounded away from zero uniformly in $P \in \mathcal{P}$.

Proof. If $\boldsymbol{X}_T \xrightarrow{p} \boldsymbol{C}$ uniformly in $P \in \mathcal{P}$, then $\boldsymbol{X}_T \xrightarrow{d} \boldsymbol{C}$ uniformly in $P \in \mathcal{P}$ Let $\boldsymbol{Z}_T \equiv (\operatorname{vec} \boldsymbol{X}_T, \operatorname{vec} \boldsymbol{Y}_T)'$, then $\boldsymbol{Z}_T \xrightarrow{d} \boldsymbol{Z} \equiv (\operatorname{vec} \boldsymbol{C}', \operatorname{vec} \boldsymbol{Y}')'$ uniformly in $P \in \mathcal{P}$. Now the sum of two real number seen as the mapping $(x, y) \mapsto x + y$ is uniformly continuous. The product mapping $(x, y) \mapsto x.y$ is also uniformly continuous provided that the domain of one of the arguments is bounded. The inverse mapping $x \mapsto 1/x$ can also be made uniformly continuous if the argument is bounded away for zero. Since all the transformations above applied to $\boldsymbol{Z}_T$ are (entrywise) compositions of uniform continuous mapping (hence uniformly continuous), the results follow from Lemma 1(b). □

Proof of Proposition 1.

Proof. Recall that $\mathcal{M}_t \equiv \mathcal{M}(\boldsymbol{x}_t)$, $\boldsymbol{\nu}_t \equiv \boldsymbol{y}_t^{(0)} - \mathcal{M}_t$ for $t \geq 1$ and $\boldsymbol{\eta}_{t,T} \equiv \widehat{\mathcal{M}}_t - \mathcal{M}_t$ for $t \geq T_0$. From the definition of our estimator we have:

$$\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T = \frac{1}{T_2} \sum_{t \geq T_0} \left[\boldsymbol{y}_t - \boldsymbol{\Delta}_T - \widehat{\mathcal{M}}(\boldsymbol{x}_t)\right] = \frac{1}{T_2} \sum_{t \geq T_0} \left[\boldsymbol{y}_t^{(0)} - \widehat{\mathcal{M}}(\boldsymbol{x}_t)\right] = \frac{1}{T_2} \sum_{t \geq T_0} \left[\boldsymbol{\nu}_t - \boldsymbol{\eta}_{t,T}\right].$$

After multiplying the last expression by $\sqrt{T}$ we can rewrite it as:

$$\sqrt{T}\left(\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T\right) = \underbrace{\frac{\sqrt{T}}{T_2} \sum_{t \geq T_0} \boldsymbol{\nu}_t}_{\equiv \boldsymbol{V}_{2,T}} - \underbrace{\frac{\sqrt{T}}{T_1} \sum_{t \leq T_1} \boldsymbol{\nu}_t}_{\equiv \boldsymbol{V}_{1,T}} - \sqrt{T}\left(\tfrac{1}{T_2} \sum_{t \geq T_0} \boldsymbol{\eta}_{t,T} - \tfrac{1}{T_1} \sum_{t \leq T_1} \boldsymbol{\nu}_t\right) \tag{16}$$

By condition (a) in the proposition, the last term in the right hand side converges to zero uniformly in $P \in \mathcal{P}$. Under condition (b), each one of the first two terms individually converges in distribution to a Gaussian random variable uniformly in $P \in \mathcal{P}$, which is not enough to ensure that the joint distribution is also Gaussian. However, notice that both $\boldsymbol{V}_{1,T}$ and $\boldsymbol{V}_{2,T}$ are defined with respect to the same random sequence. Hence, not only they are jointly Gaussian but also they are also asymptotically independent since they are summed over non-overlapping intervals:

$$\boldsymbol{V}_T \equiv (\boldsymbol{V}_{1,T}, \boldsymbol{V}_{2,T})' \xrightarrow{d} (\boldsymbol{Z}_1, \boldsymbol{Z}_2)' \equiv \boldsymbol{Z} \sim \mathcal{N}\left\{\boldsymbol{0}, \begin{bmatrix} \lambda_0^{-1}\boldsymbol{\Gamma} & \boldsymbol{0} \\ \boldsymbol{0} & (1-\lambda_0)^{-1}\boldsymbol{\Gamma} \end{bmatrix}\right\},$$

uniformly in $P \in \mathcal{P}$, where $\boldsymbol{\Gamma} \equiv \lim_{T \to \infty} \boldsymbol{\Gamma}_T$.

It follows from Lemma 1(a) that $\boldsymbol{V}_{2,T} - \boldsymbol{V}_{1,T} \xrightarrow{d} \boldsymbol{Z}_2 - \boldsymbol{Z}_1$, uniformly in $P \in \mathcal{P}$. By Lemma 2(a), $\sqrt{T}\left(\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T\right) \xrightarrow{d} \mathcal{N}\left[\boldsymbol{0}, \frac{\boldsymbol{\Gamma}}{\lambda_0(1-\lambda_0)}\right]$, uniformly in $P \in \mathcal{P}$. □

We now state some auxiliary lemmas that will provide bounds in probability used throughout the proof of the main theorem:

LEMMA 3. *Let $\{u_t\}_{t \in \mathbb{N}}$ be strong mixing sequence of centered random variables with mixing coefficient with exponential decay. Also for some real $r > 2$, $\mathsf{sup}_t \mathbb{E}|u_t|^{r+\delta} < \infty$ for some $\delta > 0$, then there exist a positive constant C_r (not depending on n) such that*

$$\mathbb{E}|u_1 + \cdots + u_T|^r \leq C_r T^{r/2}.$$

Proof. See Doukhan and Louhichi (1999) and Rio (1994). □

LEMMA 4. *Under Assumptions 2-4, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 = O_P\left(s_0 \frac{d^{1/\gamma}}{\sqrt{T}}\right)$.*

Proof. For real $a, b > 0$ define:

$$\mathscr{A}(a) = \left\{ \left\| \frac{2}{T_1} \sum_{t=1}^{T_1} \boldsymbol{x}_t \nu_t \right\|_{\max} \leq a \right\}, \quad \boldsymbol{p}_t(d \times 1) \equiv \boldsymbol{x}_t \nu_t;$$

$$\mathscr{B}(b) = \left\{ \left\| \frac{1}{T_1} \sum_{t=1}^{T_1} \boldsymbol{M}_t \right\|_{\max} \leq b \right\}, \quad \boldsymbol{M}_t(d \times d) \equiv \boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t'),$$

where $\| \cdot \|_{\max}$ is the maximum entry-wise norm.

Following Corollary 6.10 of Bülhmann and van der Geer (2011) on $\mathscr{A}(a) \cap \mathscr{B}(b)$, we have that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \leq \frac{32\varsigma s_0}{\psi_0^2}$, provided that $\varsigma \geq 8a$, $b \leq \frac{\psi_0^2}{32 s_0}$ and the compatibility constraint is satisfied for $\boldsymbol{\Sigma} \equiv \mathbb{E}\left(\frac{1}{T_1}\sum_{t=1}^{T_1} \boldsymbol{x}_t \boldsymbol{x}_t'\right)$ with constant $\psi_0 > 0$ (Assumption 2). For convenience set $a = \frac{\varsigma}{8}$ and $b = \frac{\psi_0^2}{32 s_0}$. Then, we can write

$$\begin{aligned}
\mathbb{P}\left(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 > \frac{32\varsigma s_0}{\psi_0^2} \right) &\leq \mathbb{P}\left(\left\| \frac{2}{T} \sum_{t=1}^{T_1} \boldsymbol{p}_t \right\|_{\max} > \frac{\varsigma}{8} \right) + \mathbb{P}\left(\left\| \frac{1}{T_1} \sum_{t=1}^{T_1} \boldsymbol{M}_t \right\|_{\max} > \frac{\psi_0^2}{32 s_0} \right) \\
&\leq d \max_{1 \leq i \leq d} \mathbb{P}\left(\left| \sum_{t=1}^{T_1} p_{i,t} \right| > \frac{\varsigma T_1}{16} \right) + d^2 \max_{1 \leq i,j \leq d} \mathbb{P}\left(\left| \sum_{t=1}^{T_1} m_{ij,t} \right| > \frac{\psi_0^2 T_1}{32 s_0} \right) \\
&\leq d \left(\frac{16}{\varsigma T_1} \right)^{\gamma} \max_{1 \leq i \leq d} \mathbb{E} \left| \sum_{t=1}^{T_1} p_{i,t} \right|^{\gamma} + d^2 \left(\frac{32 s_0}{\psi_0^2 T_1} \right)^{\gamma} \max_{1 \leq i,j \leq d} \mathbb{E} \left| \sum_{t=1}^{T_1} m_{ij,t} \right|^{\gamma} \\
&\leq C_1(\gamma) \frac{d}{T_1^{\gamma/2} \varsigma^{\gamma}} + C_2(\gamma, \psi_0) \frac{d^2 s_0^{\gamma}}{T_1^{\gamma/2}},
\end{aligned}$$

where the second inequality follows from the union bound. The third inequality follows from the Markov inequality applied for some $\gamma > 2$. The last inequality is a consequence of Lemma 3, since (i) by Assumption 3(a) both $\{\boldsymbol{p}_t\}$ and $\{\boldsymbol{M}_t\}$ are strong mixing sequences with exponential decay as measurable functions of $\{\boldsymbol{w}_t\}$; and (ii) by Cauchy-Schwartz inequality combined with Assumption 3(b) we have for some $\delta > 0$:

$$\mathbb{E}|p_{j,t}|^{\gamma+\delta/2} \leq \left(\mathbb{E}|x_{j,t}|^{2\gamma+\delta} \mathbb{E}|\nu_t|^{2\gamma+\delta} \right)^{\frac{\gamma+\delta/2}{2\gamma+\delta}} \leq c_\gamma, \quad 1 \leq i \leq d;\ t \geq 1$$

$$\mathbb{E}|m_{ij,t} - \mathbb{E}(x_{i,t}x_{j,t})|^{\gamma+\delta/2} \leq \left(\mathbb{E}|x_{i,t}|^{2\gamma+\delta} \mathbb{E}|x_{j,t}t|^{2\gamma+\delta} \right)^{\frac{\gamma+\delta/2}{2\gamma+\delta}} \leq c_\gamma, \quad 1 \leq i,j \leq d;\ t \geq 1.$$

The result follows since, by Assumption 4(a) $\varsigma = O\left(\frac{d^{1/\gamma}}{\sqrt{T}}\right)$ and by Assumption 4(b), $s_0 \frac{d^{2/\gamma}}{\sqrt{T}} = o_P(1)$. □

Lemma 5. *Let* $\boldsymbol{S}_T \equiv \sum_{t=1}^{T} \boldsymbol{u}_t$ *where* $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{dt})' \in \mathcal{U} \subset \mathbb{R}^d$ *is a zero mean random vector, such that the process* $(u_{j,t})$ *fulfils the conditions of Lemma 3 for some real* $r > 2$ *for all* $j \in \{1, \ldots, d\}$. *Then,* $\|\boldsymbol{S}_T\|_{\max} = O_P(d^{1/r}\sqrt{T})$.

Proof. For a given $\epsilon > 0$, By the union bound, followed by Markov inequality we have:

$$\mathbb{P}\left(\frac{\|\boldsymbol{S}_T\|_{\max}}{d^{1/r}\sqrt{T}} > \epsilon \right) \leq d \max_{1 \leq i \leq d} \mathbb{P}\left(\frac{|S_{i,T}|}{d^{1/r}\sqrt{T}} > \epsilon \right) \leq \frac{\max_{1 \leq i \leq d} \mathbb{E}|S_{i,T}|^r}{T^{r/2}\epsilon^r} \leq \frac{C_r}{\epsilon^r},$$

where the last inequality follows from Lemma 3. □

Proof of Theorem 1.

Proof. Recall that $\eta_{t,T} = \boldsymbol{x}_t'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ for $t \geq T_0$, and let $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0')'$, where α is the parameter of the intercept while $\boldsymbol{\beta}$ is the vector of remaining parameters. Similar, let $\boldsymbol{x}_t = (1, \widetilde{\boldsymbol{x}}_t)$. From the definition of the estimator, $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = \frac{1}{T_1}\sum_{t\leq T_1} \boldsymbol{\nu}_t - \frac{1}{T_1}\sum_{t\leq T_1} \widetilde{\boldsymbol{x}}_t \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$. Combining the last two expressions we can rewrite the estimation error as

$$\begin{aligned}\eta_{t,T} &= \frac{1}{T_1}\sum_{s\leq T_1} \boldsymbol{\nu}_s - \frac{1}{T_1}\sum_{s\leq T_1} \widetilde{\boldsymbol{x}}_s \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + \widetilde{\boldsymbol{x}}_t \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \\ &= \frac{1}{T_1}\sum_{s\leq T_1} \boldsymbol{\nu}_s - \left[\frac{1}{T_1}\sum_{s\leq T_1} \widetilde{\boldsymbol{x}}_s - \widetilde{\boldsymbol{x}}_t\right] \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right).\end{aligned}$$

Taking the average over $t = T_0, \ldots, T$, multiplying by $\sqrt{T}$ and rearranging yields:

$$\sqrt{T}\left(\frac{1}{T_2}\sum_{t\geq T_0} \boldsymbol{\eta}_{t,T} - \frac{1}{T_1}\sum_{t\leq T_1} \boldsymbol{\nu}_t\right) = \left(\frac{\sqrt{T}}{T_2}\sum_{t\geq T_0} \widetilde{\boldsymbol{x}}_t - \frac{\sqrt{T}}{T_1}\sum_{t\leq T_1} \widetilde{\boldsymbol{x}}_t\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right).$$

We now show that the last expression is $o_P(1)$ uniformly in $P \in \mathcal{P}$. First, we bound it in absolute term by:

$$\left\|\frac{\sqrt{T}}{T_2}\sum_{t\geq T_0} \widetilde{\boldsymbol{x}}_t - \frac{\sqrt{T}}{T_1}\sum_{t\leq T_1} \widetilde{\boldsymbol{x}}_t\right\|_{\max} \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_1.$$

Adding and subtracting the mean, the first term is the sum of two $O_P\left(d^{1/\gamma}\right)$ terms by Lemma 5 combined with Assumption 3(a)-(b). The second term is $O_P\left(s_0\frac{d^{1/\gamma}}{\sqrt{T}}\right)$ by Lemma 4. Hence, the last term in the above display is $O_P\left(s_0\frac{d^{2/\gamma}}{\sqrt{T}}\right) = o_P(1)$ by Assumption 4(b), which verifies condition (a) of Proposition 1.

Now $\{\nu_t\}$ is a strong mixing process with mixing coefficient with exponential decay and $\sup_t \mathbb{E}|\nu_t|^r < \infty$ for some $r > 4$ by Assumption 3(a) and (b). Also, $\mathbb{E}(\nu_t^2)$ is bounded by below uniformly by Assumption 3(c). Hence, we have a Central Limit Theorem as per Theorem 10.2 of Pötscher and Prucha (1997). Therefore, conditions (b) and (c) of Proposition 1 are verified and the result follows directly from Proposition 1.

□

Proof of Propositions 2 and 3.

Proof. Both follows directly from Theorem 1 combined with Lemma 2(c) □

Proof of Theorem 2.

Proof. From (16) in the Proof of Proposition 1, we have for $T_\lambda = \lfloor \lambda T \rfloor$, $\lambda \in \Lambda$

$$\boldsymbol{\Gamma}^{1/2}\boldsymbol{S}_T(\lambda) = \frac{\sqrt{T}}{T - T_\lambda + 1}\sum_{t\geq T_\lambda} \boldsymbol{\nu}_t - \frac{\sqrt{T}}{T_\lambda - 1}\sum_{t< T_\lambda} \boldsymbol{\nu}_t - \frac{\sqrt{T}}{T - T_\lambda + 1}\sum_{t\geq T_\lambda} \boldsymbol{\eta}_{t,T} + \frac{\sqrt{T}}{T_\lambda - 1}\sum_{t< T_\lambda} \boldsymbol{\eta}_{t,T}.$$

The last two terms are $o_p(1)$ uniformly in $\lambda \in \Lambda$, under the conditions of Proposition 1, Assumption 5 and the fact that Λ is compact.

For fix $\lambda \in \Lambda$ the pointwise convergence in distribution follows under the conditions of from Proposition 1 (for instance under the assumptions of Theorem 1). The uniform convergence result then follows from the invariance principle in McLeish (1974) applied to $\boldsymbol{V}_T(\lambda) \equiv \frac{1}{\sqrt{T}} \sum_{t \geq T_\lambda} \boldsymbol{\nu}_t$ and the Continuous Mapping Theorem.

To obtain the covariance structure let $\boldsymbol{\Gamma}_{s-t} = \mathbb{E}(\boldsymbol{\nu}_t \boldsymbol{\nu}'_s)$ for all s, t and note that for any pair $(\lambda, \lambda') \in \Lambda^2$ we have that

$$\frac{1}{T} \sum_{t \geq T_\lambda} \sum_{s \geq T_{\lambda'}} \boldsymbol{\Gamma}_{s-t} = \frac{T - T_{\lambda \vee \lambda'} + 1}{T} \left[\frac{1}{T - T_{\lambda \vee \lambda'} + 1} \sum_{t \geq T_\lambda} \sum_{s \geq T_{\lambda'}} \boldsymbol{\Gamma}_{s-t} \right] = (1 - \lambda \vee \lambda') \frac{\boldsymbol{\Gamma}}{\lambda \vee \lambda} + o_p(1),$$

where $\lambda \vee \lambda' = \max(\lambda, \lambda')$ and $\lambda \wedge \lambda' = \min(\lambda, \lambda')$. Finally, we have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{S}_T(\lambda) \boldsymbol{S}'_t(\lambda')] &= \boldsymbol{\Gamma}^{-1/2} \left[\frac{T^2}{(T - T_\lambda + 1)(T - T\lambda' + 1)} \frac{1}{T} \sum_{t \leq T_\lambda} \sum_{s \leq T_{\lambda'}} \boldsymbol{\Gamma}_{s-t} \right] \boldsymbol{\Gamma}^{-1/2} + o_p(1) \\
&= \left[\frac{1}{(1-\lambda)(1-\lambda')} \right] \frac{(1 - \lambda \vee \lambda')}{\lambda \vee \lambda} + o_p(1) \\
&= \frac{1}{(\lambda \vee \lambda)(1 - \lambda \wedge \lambda')} + o_p(1) \equiv \boldsymbol{\Sigma}_{\boldsymbol{\lambda}} + o_p(1)
\end{aligned}$$

□

Proof of Proposition 4.

Proof. Below we write T_λ we mean $\lfloor \lambda T \rfloor$. All the convergence in probability are a direct consequence of the Weak Law of Large Numbers ensured by the conditions of Proposition 1 combined with Assumption 5: Let $\lambda \leq \lambda_0$:

$$\begin{aligned}
\widehat{\boldsymbol{\Delta}}_T(\lambda) &\equiv \frac{1}{T - T_\lambda + 1} \sum_{t=T_\lambda}^{T} \widehat{\boldsymbol{\delta}}_t(\lambda) = \left(\frac{T_0 - T_\lambda}{T - T_\lambda + 1} \right) \frac{\sum_{t=T_\lambda}^{T_0 - 1} \widehat{\boldsymbol{\Delta}}_t(\lambda)}{T_0 - T_\lambda} + \left(\frac{T - T_0 + 1}{T - T_\lambda + 1} \right) \frac{\sum_{t=T_0}^{T} \widehat{\boldsymbol{\delta}}_t(\lambda)}{T - T_0 + 1} \\
&= o_p(1) + \left(\frac{1 - \lambda_0}{1 - \lambda} \right) \boldsymbol{\Delta}.
\end{aligned}$$

Similarly, consider a guess after the true value, $\lambda > \lambda_0$. Then:

$$\begin{aligned}
\widehat{\boldsymbol{\Delta}}_T(\lambda) &\equiv \frac{1}{T - T_\lambda + 1} \sum_{t=T_\lambda}^{T} \widehat{\boldsymbol{\delta}}_t(\lambda) = \frac{1}{T - T_\lambda + 1} \sum_{t=T_\lambda}^{T} \left[\boldsymbol{y}_t - \widehat{\mathcal{M}}(\boldsymbol{x}_t) \right] \\
&= \frac{1}{T - T_\lambda + 1} \sum_{t=T_\lambda}^{T} [\boldsymbol{y}_t - \mathcal{M}(\boldsymbol{x}_t)] - \frac{\lambda - \lambda_0}{\lambda} \boldsymbol{\Delta} + o_p(1) \\
&= \frac{1}{T - T_\lambda + 1} \sum_{t=T_\lambda}^{T} \left[\boldsymbol{y}_t^{(0)} - \boldsymbol{\alpha}_0 - g(\boldsymbol{\theta}_0) \right] + \frac{\lambda_0}{\lambda} \boldsymbol{\Delta} + o_p(1) = \frac{\lambda_0}{\lambda} \boldsymbol{\Delta} + o_p(1),
\end{aligned}$$

where the second equality follows from Assumption 6, since a step intervention will only affect (asymptotically) the constant regressor estimation of the model $\mathcal{M}$ by a factor of $\frac{\lambda-\lambda_0}{\lambda_0}$ times the intervention size $\boldsymbol{\Delta}$. To see this let $\boldsymbol{\alpha}_0$ be the constant and $\boldsymbol{\beta}_0$ the remaining parameters. Then,

$$\widehat{\boldsymbol{\alpha}} = \frac{1}{T_\lambda}\sum_{t\leq T_\lambda}\boldsymbol{y}_t^{(0)} + \frac{1}{T_\lambda}\sum_{t\leq T_\lambda}\boldsymbol{\Delta}I(t \geq T_0) - \frac{1}{T_\lambda}\sum_{t\leq T_\lambda}\widetilde{\mathcal{M}}(\widehat{\boldsymbol{\beta}}),$$

where $\mathcal{M}(\boldsymbol{x}_t;\boldsymbol{\theta}_0) \equiv \boldsymbol{\alpha}_0 + \widetilde{\mathcal{M}}(\boldsymbol{x}_t;\boldsymbol{\beta}_0)$. Since the estimation of $\boldsymbol{\beta}_0$ is asymptotically unaffected by a step intervention, under the conditions of Proposition 1, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. Consequently, $\widehat{\boldsymbol{\alpha}}(\lambda) \xrightarrow{p} \boldsymbol{\alpha} + \frac{\lambda-\lambda_0}{\lambda}\boldsymbol{\Delta}$, $\forall\lambda \in (0,1)$. □

Proof of Theorem 3.

Proof. Note that: (i) The limiting function $J_{p,0}(\lambda) \equiv \phi(\lambda)\|\boldsymbol{\Delta}\|_p$ is uniquely maximized at $\lambda = \lambda_0$ under the assumption that $\boldsymbol{\Delta}_T \neq 0$, (ii) The parametric space Λ is compact; (iii) $J_{0,p}(\cdot)$ is a continuous function as consequence of the continuity of $\phi(\cdot)$, (iv) $J_{p,T}(\lambda)$ converges uniformly in probability to $J_{p,0}(\lambda)$ (shown below). Therefore, from Theorem 2.1 of Newey and McFadden (1994) we have that $\widehat{\lambda}_{0,p} \xrightarrow{p} \lambda_0$.

In Theorem 2 we show that $\boldsymbol{S}_T$ converges in distribution to $\boldsymbol{S}_T$. Hence, $\boldsymbol{S}_T$ is uniformly tight (in particular with respect to λ). Therefore, $\frac{1}{\sqrt{T}}\boldsymbol{S}_T(\lambda)$ is $o_p(1)$ uniformly in λ. Or equivalently, $\widehat{\boldsymbol{\Delta}}_T(\lambda) \xrightarrow{p} \boldsymbol{\Delta}_T(\lambda)$, uniformly in $\lambda \in \Lambda$.

Now consider any real valued function $f(\cdot)$ that is continuous on a compact set $K \subset \mathbb{R}^k$. In that case $f(\cdot)$ is uniformly continuous on K as every continuous function on a compact domain. By definition then, for a given $\epsilon > 0$, there is a $\delta > 0$ such that for every $(\boldsymbol{x},\boldsymbol{y}) \in K^2$, $\{|f(\boldsymbol{x}) - f(\boldsymbol{y})| > \epsilon\} \Rightarrow \{\|\boldsymbol{x}-\boldsymbol{y}\| > \delta\}$. Therefore, $\mathbb{P}(|\|\boldsymbol{x}\|_p - \|\boldsymbol{y}\|_p| > \epsilon) \leq \mathbb{P}(\|\boldsymbol{x}-\boldsymbol{y}\| > \delta) + \mathbb{P}(K^c)$.

Finally, note that $\|\cdot\|_p$ is a a continuous function on $\mathbb{R}^q$ so given any $\epsilon > 0$, we can take a arbitrary large compact $K_\epsilon \subset \mathbb{R}^q$ such that $P(K^c) \leq \epsilon$. The result then follows since the first term above converges uniformly to zero in probability. □

Proof of Proposition 5.

Proof. Follows directly from Theorem 1 applied to each unit of $\mathcal{I}$ individually combined with the Cramèr-Wold device device. □

References

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505.

——— (2014): "Politics and the Synthetic Control Method," *American Journal of Political Science*, In press.

Abadie, A., and J. Gardeazabal (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132.

An, S., and F. Schorfheide (2007): "Bayesian Analysis of DSGE Models," *Econometric Reviews*, 26, 113–172.

Andrews, D. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Andrews, D., and J. Monahan (1992): "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator," *Econometrica*, 60, 953–966.

Angrist, J., and G. Imbens (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 61, 467–476.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–472.

Angrist, J., Ó. Jordá, and G. Kuersteiner (2013): "Semiparametric estimates of monetary policy effects: String theory revisited," Working Paper 2013-24, Federal Reserve Bank of San Francisco.

Bai, C.-E., Q. Li, and M. Ouyang (2014): "Property taxes and home prices: A tale of two cities," *Journal of Econometrics*, 180, 1–15.

Bai, J. (1997): "Estimating Multiple Breaks One at a Time," *Econometric Theory*, 13, 315–352.

——— (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

Belasen, A., and S. Polachek (2008): "How Hurricanes Affect Wages and Employment in Local Labor Markets," *The American Economic Review: Papers and Proceedings*, 98, 49–53.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2016): "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in Z-Estimation Framework," Working Paper 1512.07619, arXiv.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2016): "Program Evaluation with High-Dimensional Data," *Econometrica*, In press.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on treatment Effects After Selection Amongst High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

Billmeier, A., and T. Nannicini (2013): "Assessing Economic Liberalization Episodes: A Synthetic Control Approach," *The Review of Economics and Statistics*, 95, 983–1001.

Bülhmann, P., and S. van der Geer (2011): *Statistics for High Dimensional Data*. Springer.

Brockmann, H., P. Genschel, and L. Seelkopf (2016): "Happy taxation: increasing tax compliance through positive rewards?," *Journal of Public Policy*, FirstView, 1–26.

CARUSO, G., AND S. MILLER (2015): "Long run effects and intergenerational transmission of natural disasters: A case study on the 1970 Ancash Earthquake," *Journal of Development Economics*, 117, 134–150.

CAVALLO, E., S. GALIANI, I. NOY, AND J. PANTANO (2013): "Catastrophic Natural Disasters and Economic Growth," *The Review of Economics and Statistics*, 95, 1549–1561.

CHEN, H., Q. HAN, AND Y. LI (2013): "Does Index Futures Trading Reduce Volatility in the Chinese Stock Market? A Panel Data Evaluation Approach," *Journal of Futures Markets*, 33, 1167–1190.

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B, pp. 5549—-5632. Elsevier Science.

CONLEY, T., AND C. TABER (2011): "Inference with Difference in Differences with a Small Number of Policy Changes," *Review of Economics and Statistics*, 93, 113–125.

DEES, S., F. D. MAURO, M. PESARAN, AND L. SMITH (2007): "Exploring the International Linkages of the Euro area: A Gobal VAR analysis," *Journal of Applied Econometrics*, 22, 1–38.

DOUKHAN, P., AND S. LOUHICHI (1999): "A new weak dependence condition and applications to moment inequalities," *Stochastic Processes and their Applications*, 84, 313–342.

DU, Z., H. YIN, AND L. ZHANG (2013): "The macroeconomic effects of the 35-h workweek regulation in France," *The B.E. Journal of Macroeconomics*, 13, 881–901.

DU, Z., AND L. ZHANG (2015): "Home-purchase restriction, property tax and housing price in China: A counterfactual analysis," *Journal of Econometrics*, 188, 558–568.

FATAS, E., D. NOSENZO, M. SEFTON, AND D. ZIZZO (2015): "A Self-Funding Reward Mechanism for Tax Compliance," Working Paper 2650265, SSRN.

FERMAN, B., AND C. PINTO (2015): "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," Working paper, São Paulo School of Economics - FGV.

FERNÁNDEZ-VILLAVERDE, J., J. RUBIO-RAMÍREZ, T. SARGENT, AND M. WATSON (2007): "ABCs (and Ds) of Understanding VARs," *American Economic Review*, 97, 1021–1026.

FUJIKI, H., AND C. HSIAO (2015): "Disentangling the effects of multiple treatments - Measuring the net economic impact of the 1995 great Hanshin-Awaji earthquake," *Journal of Econometrics*, 186, 66–73.

GOBILLON, L., AND T. MAGNAC (2016): "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," *Review of Economics and Statistics*, forthcoming.

GRIER, K., AND N. MAYNARD (2013): "The economic consequences of Hugo Chavez: A Synthetic control analysis," *Journal of Economic Behavior and Organization*, 95, 1549–1561.

HAAN, W. D., AND A. LEVIN (1996): "Inferences from Parametric and Non-Parametric Covariance Matrix Estimation Procedures," .

HANSEN, B. (2000): "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575–604.

HECKMAN, J., AND E. VYTLACIL (2005): "Structural equations, treatment effects and econometric policy evaluation," *Econometrica*, 73, 669–738.

HSIAO, C., H. S. CHING, AND S. K. WAN (2012): "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China," *Journal of Applied Econometrics*, 27, 705–740.

Johnson, S., P. Boone, A. Breach, and E. Friedmand (2000): "Corporate Governance in the Asian Financial Crisis," *Journal of Financial Economics*, 58, 141–186.

Jordan, S., A. Vivian, and M. Wohar (2014): "Sticky prices or economically-linked economies: the case of forecasting the Chinese stock market," *Journal of International Money and Finance*, 41, 95–109.

Leeb, H., and B. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

——— (2008): "Sparse estimators and the oracle property, or the return of Hodge's estimator," *Journal of Econometrics*, 142, 201–211.

——— (2009): "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding.," *Journal of Multivariate Analysis*, 100, 1065–2082.

McLeish, D. (1974): "Dependent Central Limit Theorems and Invariance Principles," *Annals of Probability*, 2, 620–628.

Newey, W., and K. West (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703–708.

Niemi, H. (1979): "On the construction of Wold decomposition for multivariate stationary processes," *Journal of Multivariate Analysis*, 9, 545–559.

Ouyang, M., and Y. Peng (2015): "The treatment-effect estimation: A case study of the 2008 economic stimulus package of China," *Journal of Econometrics*, 188, 545–557.

Pesaran, M., T. Schuermann, and S. Weiner (2004): "Modeling Regional Interdependencies Using a Global Error-Correcting Macroeconometric Model," *Journal of Business and Economic Statistics*, 22, 129–162.

Pesaran, M., L. Smith, and R. Smith (2007): "What if the UK or Sweden had joinded the Euro in 1999? An Empirical Evaluation using a Global VAR," *International Journal of Finance and Economics*, 12, 55–87.

Pesaran, M., and R. Smith (2012): "Counterfactual Analysis in Macroeconometrics: An Empirical Investigation into the Effects of Quantitative Easing," Discussion Paper 6618, IZA.

Pötscher, B., and I. Prucha (1997): *Dynamic Nonlinear Econometric Models: Asymptotic Theory.* Springer.

Rio, E. (1994): "A new weak dependence condition and applications to moment inequalities," *Comptes rendus Acad. Sci. Paris, Série I*, 318, 355–360.

Slemrod, J. (2010): "Cheating Ourselves: The Economics of Tax Evasion," *Journal of Economic Perspectives*, 21, 25–48.

Souza, F. (2014): "Tax Evasion and Inflation," Master's dissertation, Department of Economics, Pontifical Catholic University of Rio de Janeiro, http://www.econ.puc-rio.br/biblioteca.php/trabalhos/show/1413.

Tibshirani, R. (1996): "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Wan, J. (2010): "The Incentive to Declare Taxes and Tax Revenue: The Lottery Receipt Experiment in China," *Review of Development Economics*, 14, 611–624.

Xie, S., and T. Mo (2013): "Index Futures Trading and Stock Market Volatility in China: A Difference-in-Difference Approach," *Journal of Futures Markets*, 34, 282–297.

Appendix B: Figures and Tables

Table 1. Critical Vales for Unknown Intervention Time Inference: $\mathbb{P}(\|\boldsymbol{S}\|_p > c) = 1 - \alpha$

		Confidence Level					
	$\Lambda = [\underline{\lambda}, \bar{\lambda}]$	$\alpha = 0.2$	0.15	0.1	0.05	0.0025	0.001
$p = 1$	$[0.5, 0.95]$	2.5679	2.7824	3.0732	3.5457	3.9844	4.5346
	$[0.1, 0.9]$	2.4332	2.6569	2.9550	3.4530	3.9218	4.4805
	$[0.15, 0.85]$	2.3786	2.6164	2.9375	3.4482	3.9138	4.4728
	$[0.2, 0.8]$	2.3366	2.5833	2.9167	3.4399	3.9115	4.4655
$p = 2$	$[0.5, 0.95]$	3.0633	3.2814	3.5706	4.0228	4.4378	4.9674
	$[0.1, 0.9]$	2.8230	3.0441	3.3340	3.8138	4.2602	4.7792
	$[0.15, 0.85]$	2.7052	2.9400	3.2448	3.7391	4.1859	4.7235
	$[0.2, 0.8]$	2.6169	2.8579	3.1795	3.6787	4.1466	4.7159
$p = \infty$	$[0.5, 0.95]$	8.6192	9.1867	9.9400	11.1562	12.2190	13.5604
	$[0.1, 0.9]$	6.4807	6.8974	7.4353	8.2781	9.0400	10.0020
	$[0.15, 0.85]$	5.6000	5.9506	6.4041	7.1014	7.7328	8.5187
	$[0.2, 0.8]$	5.0630	5.3815	5.7957	6.4303	7.0047	7.7473

NB: All critical values were obtained as the quantile of the empirical distribution using 100,000 draws from a multivariate normal distribution with covariance $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}$ via a grid of 500 points between $\underline{\lambda}$ and $\bar{\lambda}$ inclusive.

Table 2. Rejection Rates under the Null (Test Size)

	Bias	**Var**[a]	$\widehat{s}_0$	$\alpha = 0.1$	**0.05**	**0.01**
			Innovation Distribution [b]			
Normal	0.0006	1.1304	5.4076	0.1057	0.0555	0.0128
$\chi^2(1)$	-0.0014	1.1004	5.9287	0.1227	0.0652	0.0154
t-stud(3)	0.0035	1.1026	5.6437	0.1077	0.0543	0.0103
Mixed-Normal	0.0069	1.1267	5.5457	0.1134	0.0607	0.0136
			Sample Size			
$T = 100$	0.0006	1.1304	5.4076	0.1057	0.0555	0.0128
75	-0.0030	1.1449	6.3992	0.1075	0.0546	0.0124
50	0.0021	1.1747	6.1219	0.1092	0.0626	0.0155
25	-0.0050	0.8324	3.2463	0.1330	0.0763	0.0226
			Number of Total Covariates			
$d = 100$	0.0006	1.1304	5.4076	0.1057	0.0555	0.0128
200	-0.0016	1.1655	5.7314	0.1102	0.0565	0.0135
500	-0.0043	1.2112	5.6625	0.1119	0.0556	0.0114
1000	0.0012	1.2477	5.5275	0.1054	0.0566	0.0115
			Number of Relevant (non-zero) Covariates			
$s_0 = 0$	0.0038	1.0981	0.6105	0.1059	0.0550	0.0136
5	0.0006	1.1304	5.4076	0.1057	0.0555	0.0128
10	0.0003	1.0373	9.5813	0.1103	0.0581	0.0120
100	0.0003	-	20.1624	0.1114	0.0574	0.0145
			Determinist Trend $(t/T)^{\varphi}$			
$\varphi = 0$	0.0006	1.1304	5.4076	0.1057	0.0555	0.0128
0.5	0.0142	1.1245	5.6285	0.1101	0.0598	0.0199
1	0.0183	1.1313	5.5030	0.1188	0.0613	0.0168
2	0.0221	1.1398	5.4259	0.1273	0.0675	0.0261
			Serial Correlation[c]			
$\rho = 0.2$	-0.0001	1.4109	5.5246	0.1160	0.0640	0.0158
0.4	0.0002	1.6909	5.9276	0.1223	0.0678	0.0184
0.6	0.0031	1.8895	6.9012	0.1440	0.0871	0.0283
0.8	0.0033	1.9977	7.9464	0.1546	0.0927	0.0329

Baseline DGP: (6) with $T = 100$, iid normally distributed innovations; $T_0 = 50$; $n = 100$ units; $d = n = 100$ covariates (including the constant); $s_0 = 5$, $q = 1$; $10,000$ Monte-Carlo simulations per case. The penalization parameter is chosen via Bayesian Information Criteria (BIC). We set the maximum number of included variables to be $T^{0.8}$ in the glmnet package in R.

[a] Relative to the variance of the oracle/OLS estimator in the fist stage knowing the relevant regressors.

[b] All distributions are standardized (zero mean and unit variance); Mixed normal equal to 2 Normal distributions with probability $(0.3, 0.7)$, mean $(-10, 10)$ and variance $(2, 1)$.

[c] All units are simulated as AR(1) processes. The variance estimator is computed as Andrews and Monahan (1992) with an AR(1) pre-whitening followed by a standard HAC estimator with Quadratic Spectral Kernel on the residuals. Optimal bandwidth selection for AR(1) as per Andrews (1991).

TABLE 3. Rejection Rates under the Alternative (Test Power)

	$\boldsymbol{\alpha = 0.1}$	**0.075**	**0.05**	**0.025**	**0.01**
	Step Intervention[1] $\delta_t = c\,\sigma_1 1\{t \geq T_0\}$				
$c = 0.15$	0.2045	0.1695	0.1287	0.0805	0.0436
0.25	0.3783	0.3266	0.2686	0.1890	0.1108
0.35	0.5769	0.5235	0.4545	0.3465	0.2414
0.5	0.8314	0.7945	0.7440	0.6478	0.5227
0.75	0.9876	0.9831	0.9741	0.9520	0.9094
1	0.9998	0.9995	0.9992	0.9983	0.9943
	Linear Increasing $\delta_t = c\,\sigma_1 \frac{t-T_0+1}{T-T_0+1} 1\{t \geq T_0\}$				
$c = 1$	0.8318	0.7938	0.7379	0.6397	0.5121
1.25	0.9877	0.9813	0.9717	0.9459	0.8948
1.5	0.9997	0.9997	0.9990	0.9969	0.9922
	Linear Decreasing $\delta_t = c\,\sigma_1 \frac{T-t+1}{T-T_0+1} 1\{t \geq T_0\}$				
$c = 1$	0.8298	0.7956	0.7434	0.6492	0.5107
1.25	0.9868	0.9818	0.9720	0.9490	0.8985
1.5	0.9995	0.9994	0.9989	0.9968	0.9933

All simulations above as per DGP in (6) with the parameters in the baseline scenario as described in the footnote of Table 2.

[1] All interventions intensity are measured as a factor $c > 0$ of the standard deviation of unit of interest, σ_1.

Table 4. Estimators Comparison

	BA	**SC**	**DiD***	**DiD**	**GM***	**GM**	**ArCo***	**ArCo**
	No Time Trend ($\varphi = 0$) and No Serial Correlation ($\rho = 0$)							
Bias[1]	-0.001	-0.678	0.005	0.008	-0.280	-0.273	0.000	0.000
Var	3.151	50.555	17.870	51.444	0.544	0.510	1.001	1.000
MSE	3.152	86.075	17.871	51.449	6.601	6.255	1.001	1.000
	No Time Trend ($\varphi = 0$)							
Bias	-0.003	-0.596	0.000	0.000	-0.353	-0.294	-0.002	-0.002
Var	2.997	12.293	7.215	18.506	3.057	0.705	0.998	1.000
MSE	2.996	27.634	7.214	18.502	8.438	4.427	0.998	1.000
	Common Linear Time Trend ($\varphi = 1$)							
Bias	0.218	-0.579	0.034	0.033	-0.128	-0.195	0.028	0.029
Var	2.900	19.590	6.741	17.720	0.522	0.499	1.007	1.000
MSE	4.677	32.165	6.558	17.159	1.151	1.985	1.004	1.000
	Idiosyncratic Linear Time Trend ($\varphi = 1$)							
Bias	0.744	1.391	0.597	0.577	0.766	0.766	0.161	0.158
Var	0.288	0.564	0.392	1.720	1.499	1.113	0.996	1.000
MSE	2.270	7.544	1.651	2.771	3.493	3.142	0.999	1.000
	Common Quadratic Time Trend ($\varphi = 2$)							
Bias	0.288	-0.562	0.051	0.053	-0.170	-0.170	0.049	0.048
Var	2.809	18.486	6.571	17.199	0.512	0.488	1.007	1.000
MSE	5.583	28.407	6.105	15.837	1.520	1.498	1.010	1.000
	Idiosyncratic Quadratic Time Trend ($\varphi = 2$)							
Bias	0.994	-0.179	0.780	0.758	0.465	0.465	0.154	0.153
Var	1.443	0.377	3.499	8.878	0.282	0.274	0.992	1.000
MSE	14.786	0.701	10.868	14.002	3.216	3.210	0.998	1.000

$S = 10,000$ simulations from DGP (14); $T = 100$ observations; Intervention at $T_0 = 50$ only on the first variable of the first unit of intensity one standard deviation; r_f chosen such that $R^2 = 0.5$; $n = 5$ units; $q = 3$ variables per unit; innovations are iid normally distributed; $\rho = 0.5$ and $\mathsf{diag}\,(\boldsymbol{A})$ are independent draws from uniform $[-1, 1]$; All the loads (for the constant, the time trend and the stochastic factor) are independent draws from uniform distribution $[-5, 5]$, except for the common trend cases where the time trend loads are equal to unit for all variables of all units and for the cases with no time trend where they are all set to zero.

* Estimators using the $q - 1$ covariates of unit 1. Hence, unfeasible if we expect the intervention to affect all the variables in unit 1

[1] Bias measured as a ratio to the intervention intensity, defined by one standard deviation of the first variable of the first unit; Variance and MSE measured as a ratio to the ArCo Variance and MSE, respectively.

Table 5. Estimated Effects on Food Outside Home (FOH) Inflation.

Panel (a): ArCo Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	0.2500 (0.1726)	0.4441 (0.1487)	0.4870 (0.1414)	0.7973 (0.2431)	0.4478 (0.2017)	0.3796 (0.1613)	0.4046 (0.1539)	0.4422 (0.1467)
Inflation	Yes	No	No	No	Yes	Yes	Yes	No
GDP	No	Yes	No	No	Yes	Yes	Yes	No
Retail Sales	No	No	Yes	No	No	Yes	Yes	No
Credit	No	No	No	Yes	No	No	Yes	No
R-squared	0.6849	0.1240	0.3856	0.3106	0.7993	0.8948	0.8072	0
Number of regressors	10	9	10	10	19	29	39	0
Number of relevant regressors	10	3	6	9	16	15	13	0
Number of observations ($t < T_0$)	33	33	33	33	33	33	33	33
Number of observations ($t \geq T_0$)	23	23	23	23	23	23	23	23

Panel (b): Alternative Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
BA	0.4472 (0.1464)	0.4478 (0.1466)	0.4390 (0.1471)	0.4538 (0.1464)	0.4501 (0.1467)	0.4422 (0.1467)
DiD	0.2195 (0.1467)	0.2111 (0.1460)	0.2171 (0.1467)	0.2112 (0.1460)	0.2088 (0.1461)	0.2194 (0.1467)
GM	0.3699 (0.1237)	0.3785 (0.1246)	0.3759 (0.1234)	0.3759 (0.1234)	0.3607 (0.1226)	−−
GDP	Yes	No	No	Yes	Yes	No
Retail Sales	No	Yes	No	Yes	Yes	No
Credit	No	No	Yes	No	Yes	No

The upper panel in the table reports, for different choices of conditioning variables, the estimated Average Treatment Effect on the Treated (ATET) after the adoption of the program (*Nota Fiscal Paulista* – NFP). The standard errors are reported between parenthesis. Diagnostic tests do not evidence any residual autocorrelation and the standard errors are computed without any correction. The table also shows the R-squared of the first stage estimation, the number of included regressors in each case as well as the number of selected regressors by the LASSO, and the number of observations before and after the intervention. The lower panel of Table presents some alternative measures of the ATET, namely the Before-and-After (BA), the method proposed by Gobillon and Magnac (2016) (GM) and the difference-in-difference (DiD) estimators.

Table 6. Estimated Effects on Food Outside Home (FOH) Inflation: Placebo Analysis.

	Placebos						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Goiás (GO)	−0.0113	0.1624	0.1606	0.1888	−0.1477	−0.1931	−0.0979
	(0.1811)	(0.1707)	(0.1557)	(0.1642)	(0.2334)	(0.2331)	(0.2032)
Pará (PA)	0.1328	0.2714	0.1933	−0.1419	0.3690	0.3690	0.2789
	(0.2021)	(0.1640)	(0.1708)	(0.2085)	(0.2407)	(0.2407)	(0.2052)
Ceará (CE)	−0.0380	0.2657	0.2223	0.2092	0.1972	0.1972	0.1358
	(0.1484)	(0.1547)	(0.1349)	(0.1368)	(0.1613)	(0.1613)	(0.2506)
Pernambuco (PE)	0.1769	0.1895	0.2698	0.5322	0.1586	0.1586	0.5021
	(0.1949)	(0.1687)	(0.1718)	(0.1741)	(0.2073)	(0.2073)	(0.2174)
Bahia (BA)	0.0125	0.0756	0.1001	0.5707	0.2800	0.2800	0.1737
	(0.2655)	(0.2228)	(0.2433)	(0.3547)	(0.3201)	(0.3201)	(0.2932)
Minas Gerais (MG)	−0.0706	0.1265	0.1417	0.3472	−0.1089	−0.1089	0.0736
	(0.1198)	(0.1007)	(0.1083)	(0.1705)	(0.1560)	(0.1560)	(0.1554)
Rio de Janeiro (RJ)	0.2245	0.2992	0.3126	0.2484	0.1723	0.1723	0.0724
	(0.1165)	(0.1278)	(0.1230)	(0.1245)	(0.1111)	(0.1111)	(0.1300)
Paraná (PR)	0.1409	0.3400	0.2238	0.1441	0.2373	0.2373	0.1732
	(0.2527)	(0.1904)	(0.1582)	(0.2658)	(0.2939)	(0.2939)	(0.2131)
Rio Grande do Sul (RS)	0.4292	0.5422	0.5315	0.4996	0.5325	0.5325	0.4450
	(0.1614)	(0.1653)	(0.1599)	(0.1580)	(0.1627)	(0.1627)	(0.2430)
Inflation	Yes	No	No	No	Yes	Yes	Yes
GDP	No	Yes	No	No	Yes	Yes	Yes
Retail Sales	No	No	Yes	No	No	Yes	Yes
Credit	No	No	No	Yes	No	No	Yes

The table presents the estimated effect of the intervention on the untreated units. Values between parenthesis are the standard error of the estimates.

Table 7. Estimated Effects on Food Outside Home (FOH) Inflation: The Case without RS.

Panel (a): ArCo Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	0.2992	0.4438	0.4913	0.5064	0.4763	0.4070	0.4046
	(0.1704)	(0.1486)	(0.1432)	(0.1480)	(0.2010)	(0.1600)	(0.1539)
Inflation	Yes	No	No	No	Yes	Yes	Yes
GDP	No	Yes	No	No	Yes	Yes	Yes
Retail Sales	No	No	Yes	No	No	Yes	Yes
Credit	No	No	No	Yes	No	No	Yes
R-squared	0.6439	0.1213	0.3928	0.1026	0.7960	0.8568	0.8072
Number of regressors	9	8	9	9	17	26	35
Number of relevant regressors	9	3	7	5	14	17	13
Number of observations ($t < T_0$)	33	33	33	33	33	33	33
Number of observations ($t \geq T_0$)	23	23	23	23	23	23	23

Panel (b): Alternative Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
DiD	0.2524	0.2407	0.2494	0.2412	0.2387	0.2520
	(0.1466)	(0.1456)	(0.1467)	(0.1556)	(0.1457)	(0.1466)
GM	0.3694	0.3788	0.3595	0.3775	0.3660	–
	(0.1234)	(0.1243)	(0.1246)	(0.1227)	(0.1228)	
GDP	Yes	No	No	Yes	Yes	No
Retail Sales	No	Yes	No	Yes	Yes	No
Credit	No	No	Yes	No	Yes	No

The upper panel in the table reports, for different choices of conditioning variables, the estimated Average Treatment Effect on the Treated (ATET) after the adoption of the program (*Nota Fiscal Paulista* – NFP). The standard errors are reported between parenthesis. Diagnostic tests do not evidence any residual autocorrelation and the standard errors are computed without any correction. The table also shows the R-squared of the first stage estimation, the number of included regressors in each case as well as the number of selected regressors by the LASSO, and the number of observations before and after the intervention. The lower panel of Table presents some alternative measures of the ATET, namely the Before-and-After (BA), the method proposed by Gobillon and Magnac (2016) (GM) and the difference-in-difference (DiD) estimators.

Figure 1. Bias Factor defined on (13) for $l_i = \sigma_{\eta_i} = 1$ for all $i = 1, \ldots, n$.

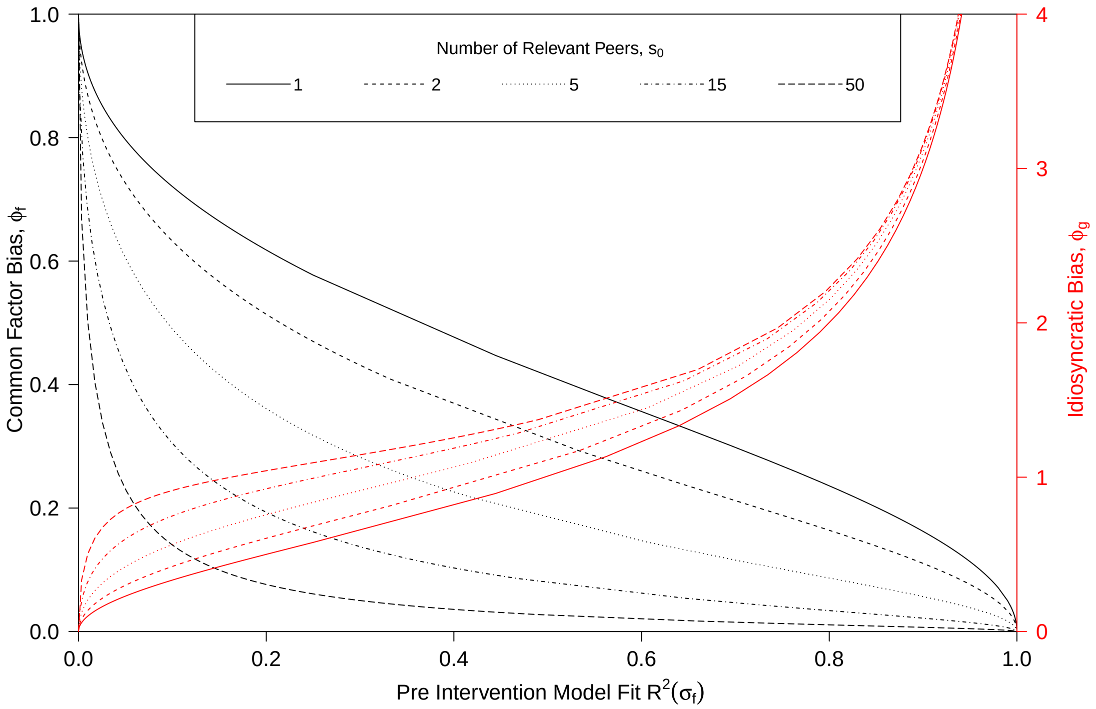


FIGURE 2. Kernel Density - Estimator Comparison with no Trend and no Serial Correlation

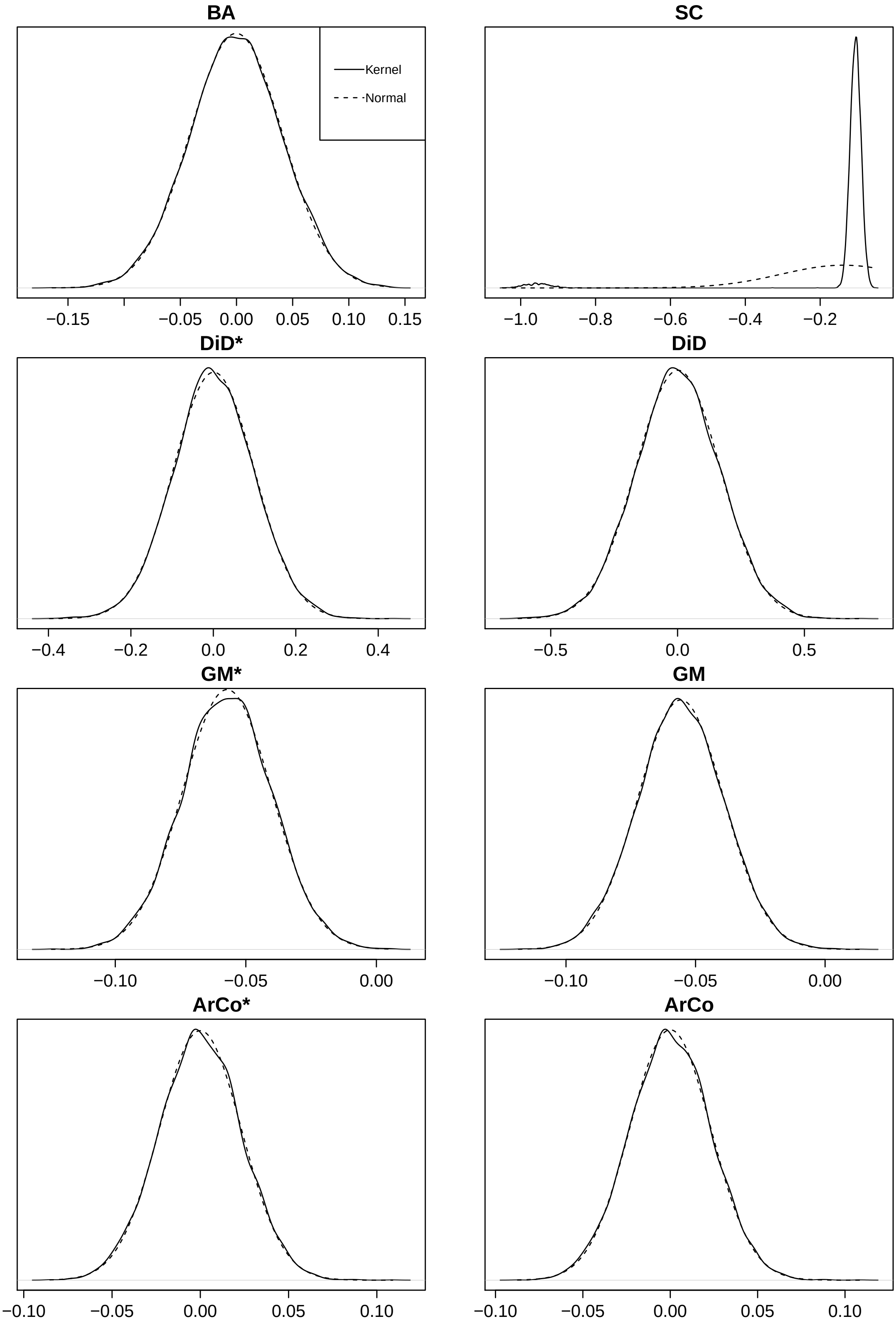


Figure 3. Kernel Density - Estimator Comparison with no Trend

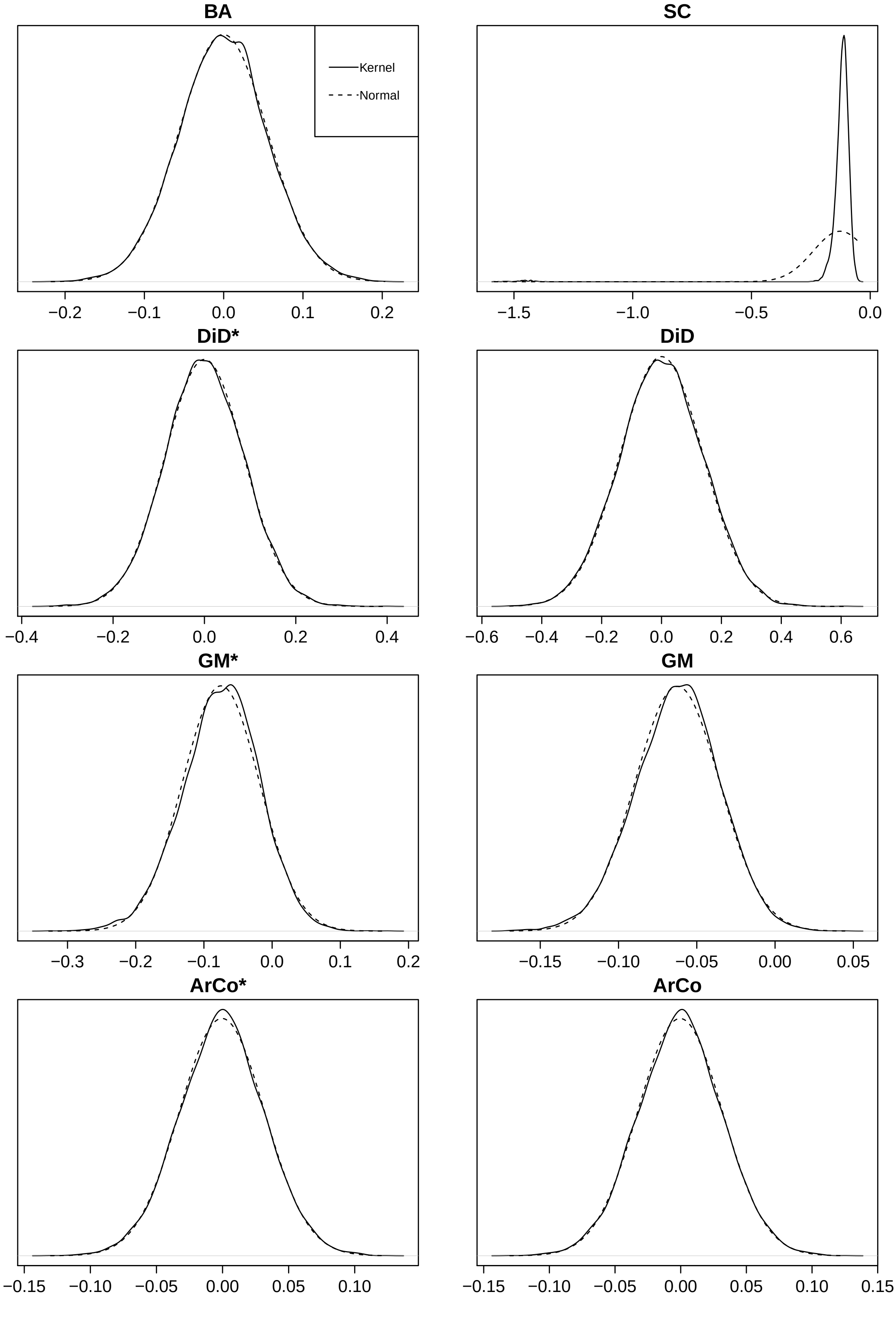


Figure 4. Kernel Density - Estimator Comparison with Common Linear Trend

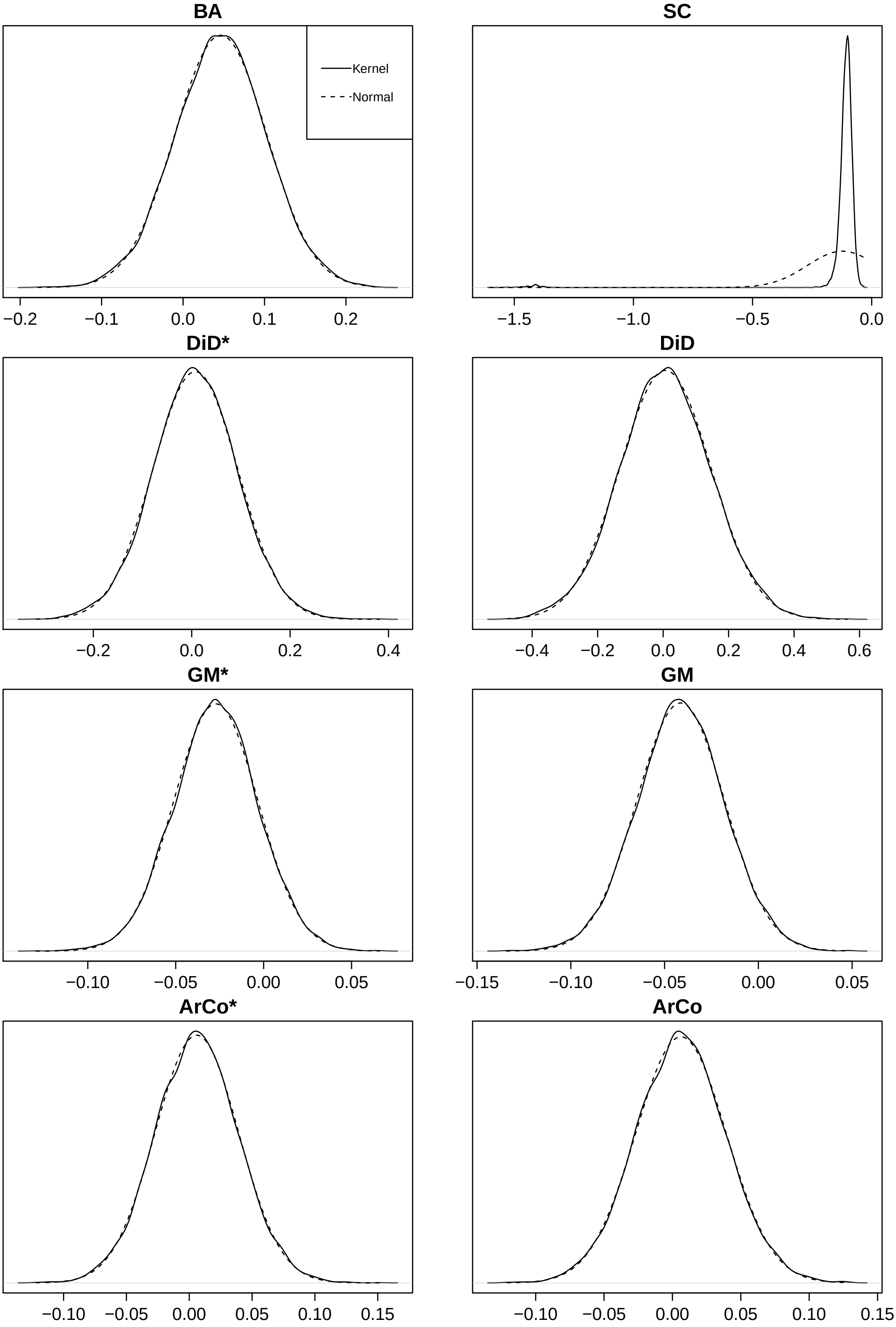


Figure 5. Kernel Density - Estimator Comparison with Idiosyncratic Linear Trend

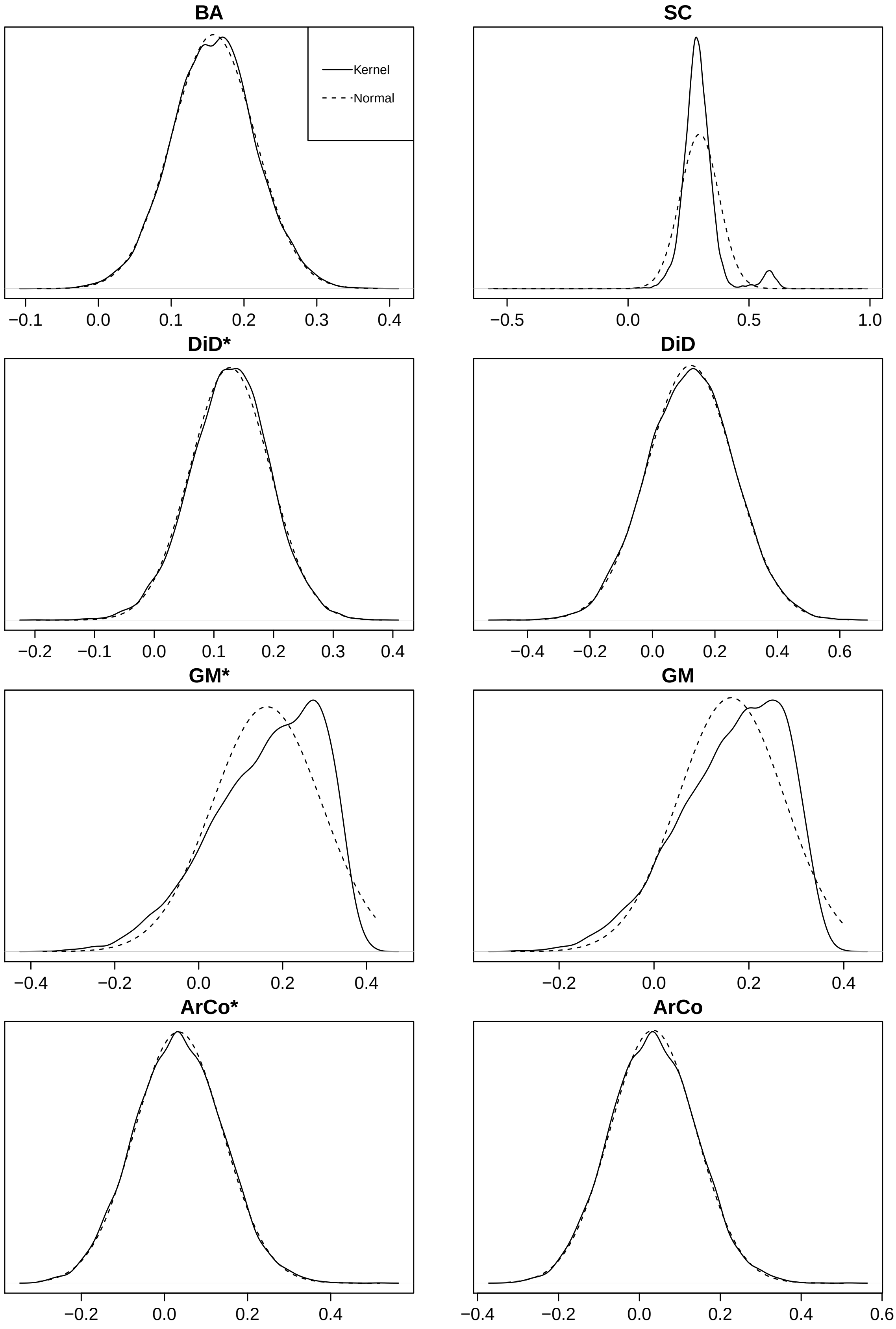


FIGURE 6. Kernel Density - Estimator Comparison with Common Quadratic Trend

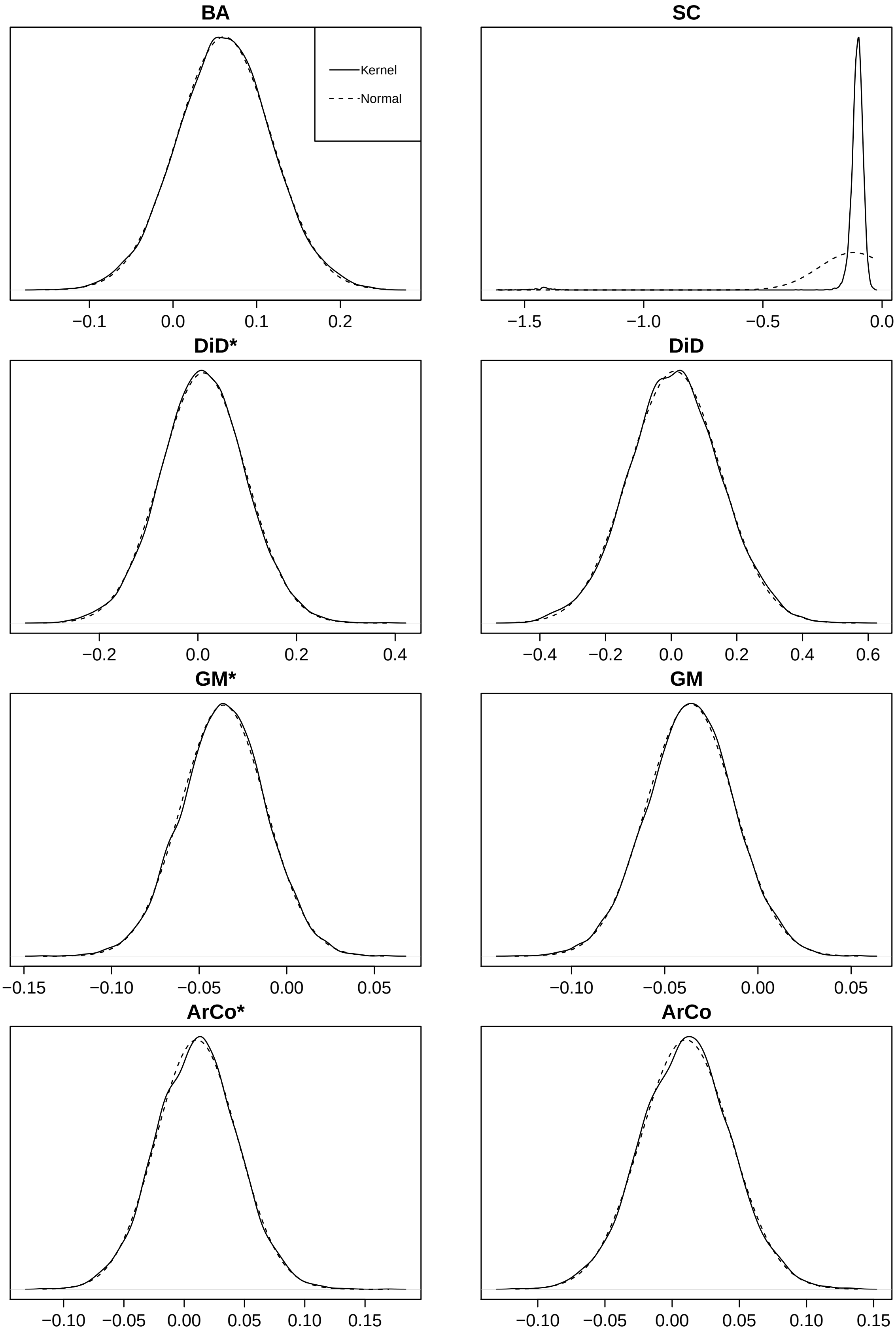


Figure 7. Kernel Density - Estimator Comparison with Idiosyncratic Quadratic Trend

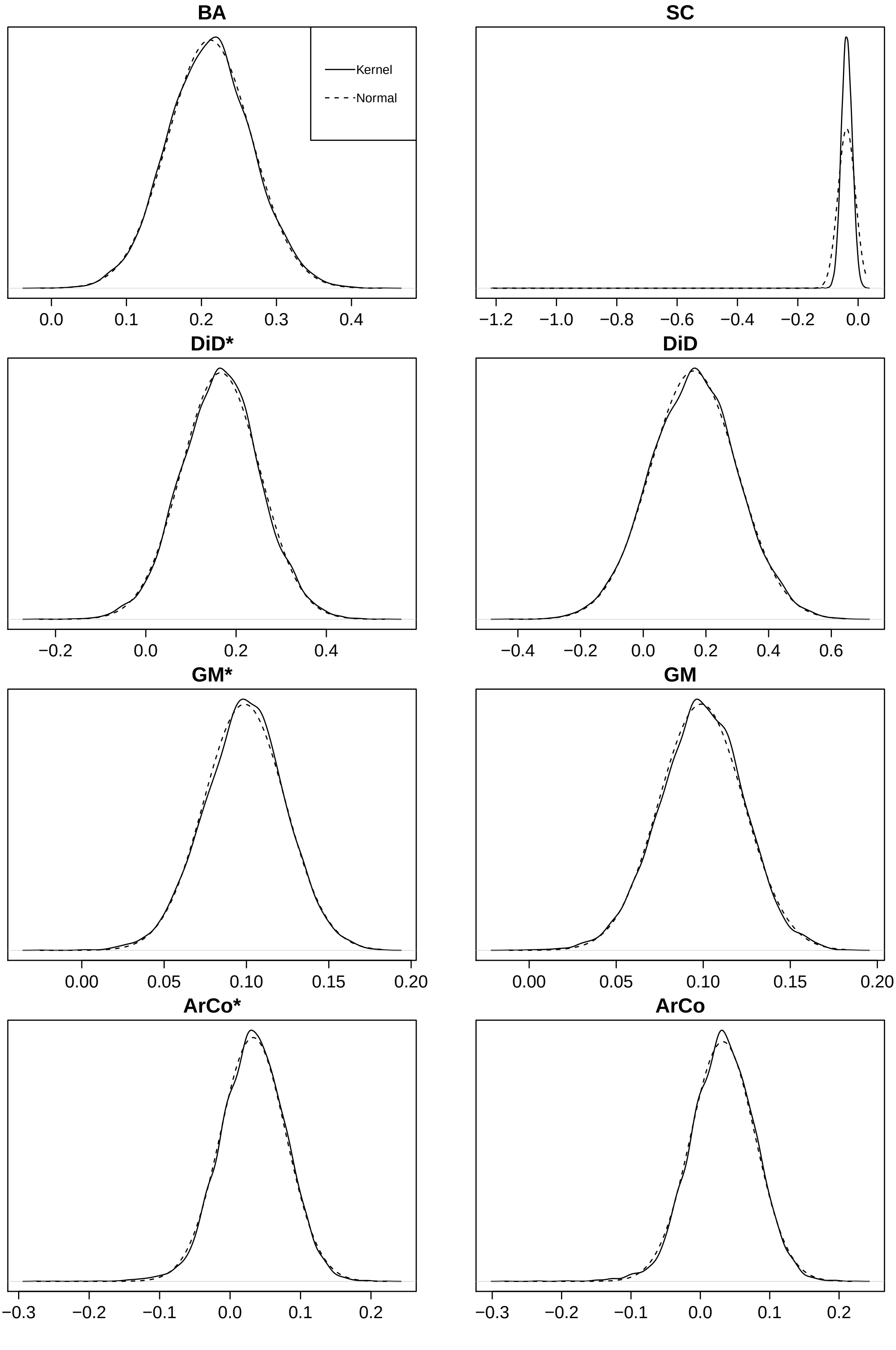
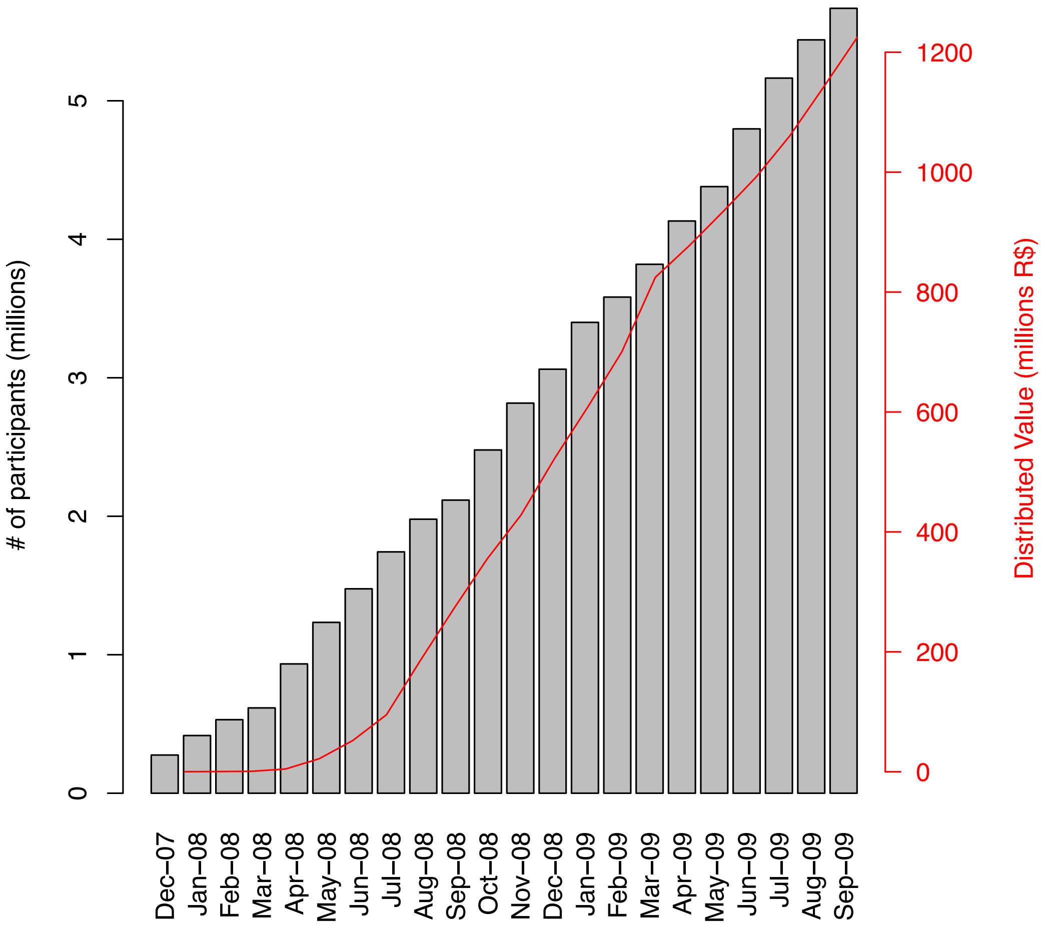
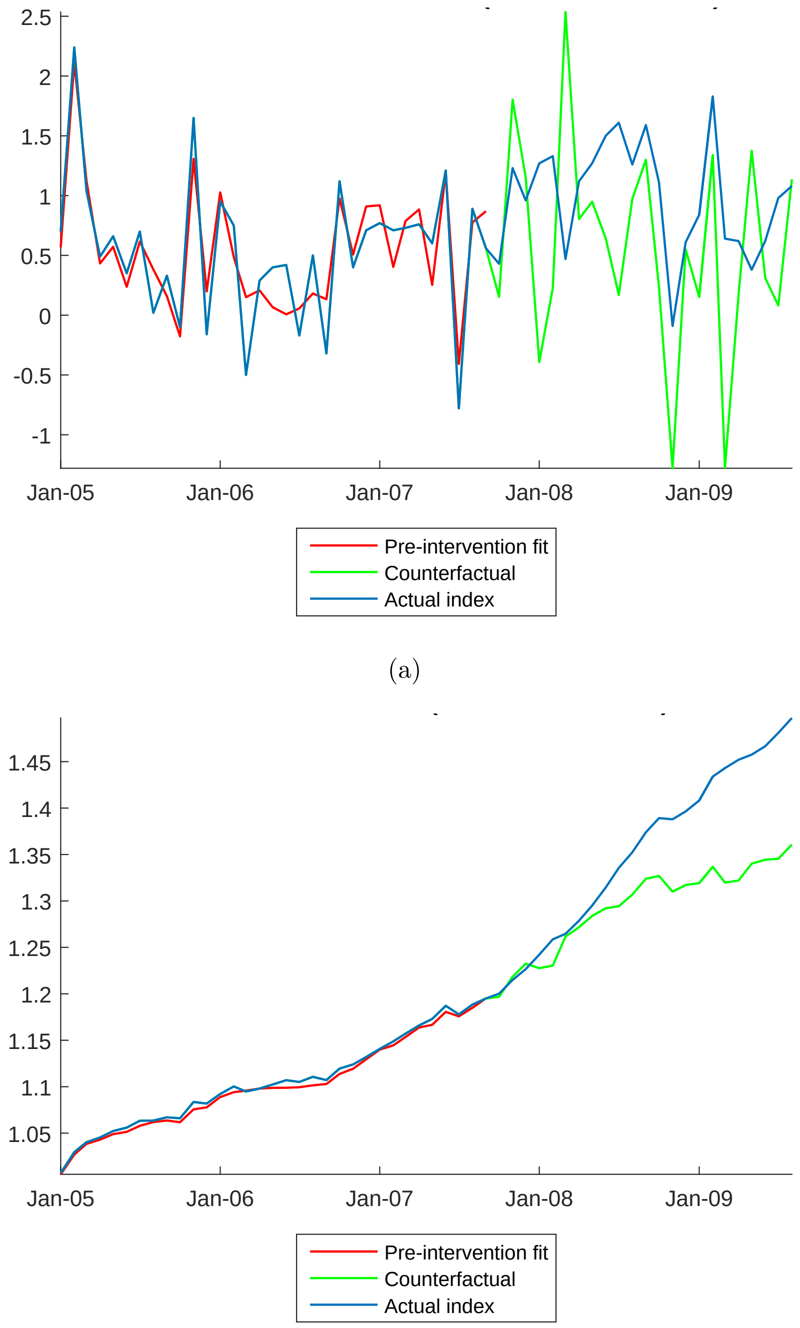


Figure 8. NFP Participation (left) and Value distributed (right)

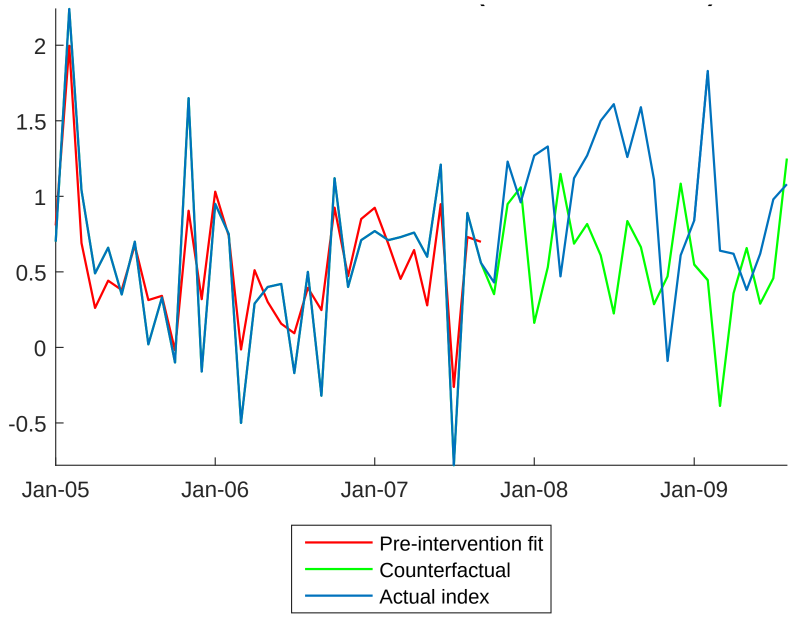




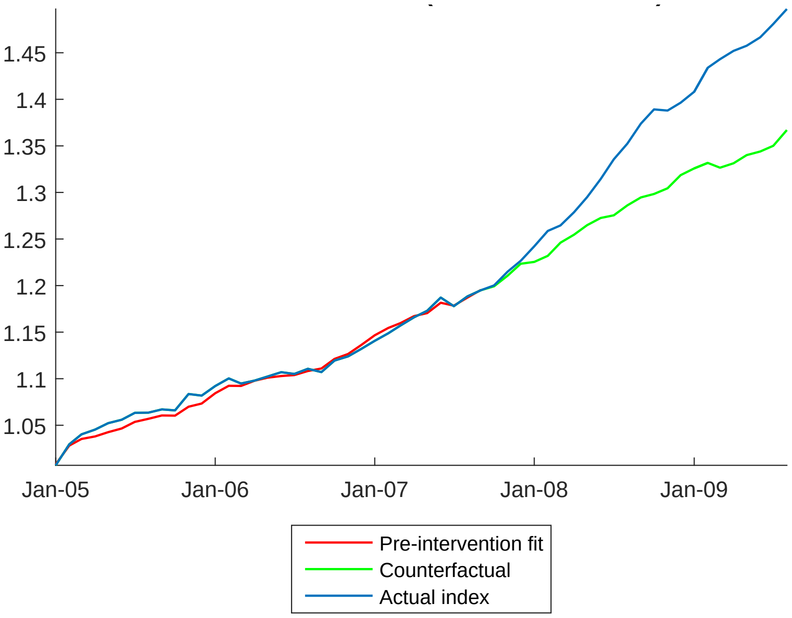
(a)

(b)

Figure 9. Actual and counterfactual data. The conditioning variables are **inflation** and **DGP growth**. Panel (a) monthly inflation. Panel (b) accumulated monthly inflation.



(a)



(b)

Figure 10. Actual and counterfactual data without RS. The conditioning variables are **inflation**, **DGP growth**, and **retail sales growth**. Panel (a) monthly inflation. Panel (b) accumulated monthly inflation.