Journal of Econometrics 207 (2018) 352–380

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

ArCo: An artificial counterfactual approach for high-dimensional panel time-series data

Carlos Carvalho [a,b,1], Ricardo Masini [c], Marcelo C. Medeiros [b,*,2]

[a] Central Bank of Brazil, Brazil
[b] Department of Economics, Pontifical Catholic University of Rio de Janeiro, Brazil
[c] São Paulo School of Economics, Getulio Vargas Foundation, Brazil

ARTICLE INFO

Article history:
Received 15 March 2017
Received in revised form 14 May 2018
Accepted 12 July 2018
Available online 15 September 2018

JEL classification:
C22
C23
C32
C33

Keywords:
Counterfactual analysis
Comparative studies
Multivariate treatment effects
Synthetic control
Policy evaluation
LASSO
Structural break
Factor models

ABSTRACT

We consider a new, flexible and easy-to-implement method to estimate the causal effects of an intervention on a single treated unit when a control group is not available and which nests previous proposals in the literature. It is a two-step methodology where in the first stage, a counterfactual is estimated based on a large-dimensional set of variables from a pool of untreated units by means of shrinkage methods, such as the *least absolute shrinkage and selection operator* (LASSO). In the second stage, we estimate the average intervention effect on a vector of variables, which is consistent and asymptotically normal. Our results are valid uniformly over a wide class of probability laws. We show that these results hold even when the exact date of the intervention is unknown. Tests for multiple interventions and for contamination effects are derived. By a simple transformation of the variables, it is possible to test for multivariate intervention effects on several moments of the variables of interest. Existing methods in the literature usually test for intervention effects on a single variable and assume that the time of the intervention is known. In addition, high-dimensionality is frequently ignored and inference is either conducted under a set of more stringent hypotheses and/or by permutation tests. A Monte Carlo experiment evaluates the properties of the method in finite samples and compares it with other alternatives. As an application, we evaluate the effects on inflation, GDP growth, retail sales and credit of an anti tax-evasion program.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

We propose a methodology to evaluate the impact of interventions which nests previous proposals of the literature. Our approach is especially useful in situations where there is a single treated unit and no available "controls", is easy to implement in practice and is robust to the presence of confounding effects, such as a global shock.[3] 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 methodology shares roots with the panel factor (PF) model of Hsiao et al. (2012) and the Synthetic Control (SC) method pioneered by Abadie and Gardeazabal (2003) and Abadie et al. (2010). Nevertheless, our proposal differs from prior methods in several dimensions as will become clear in the next paragraphs.

[*] Corresponding author.
E-mail addresses: cvianac@econ.puc-rio.br (C. Carvalho), ricardo.masini@fgv.br (R. Masini), mcm@econ.puc-rio.br (M.C. Medeiros).

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Central Bank of Brazil.

[2] The research of Marcelo C. Medeiros is partially funded by CNPq and FAPERJ.

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

https://doi.org/10.1016/j.jeconom.2018.07.005
0304-4076/© 2018 Elsevier B.V. All rights reserved.

Causality is a topic of major interest in economics. Causal statements with respect to a given treatment usually rely on the construction of counterfactuals based on the outcomes from a similar group of individuals not affected by the treatment. Notwithstanding, definitive cause-and-effect statements are often 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 (Angrist and Imbens, 1994; Heckman and Vytlacil, 2005; Belloni et al., 2014, 2017).

On the other hand, the econometric tools for cases where there is a single treated unit and no controls, which is usually the case with aggregate data, have evolved at a slower pace and much of the work has focused on simulating counterfactuals from structural models. Recently, some authors have proposed new techniques that are able, under some assumptions, to estimate counterfactuals with aggregate data (Hsiao et al., 2012; Pesaran and Smith, 2012).

1.1. Contributions of the paper

This paper fits into the literature of counterfactual analysis when a control group is not available and only one element is subjected to the treatment. We propose a two-step approach called the **artificial counterfactual (ArCo)** method to estimate the average multivariate treatment (intervention) effects on the treated unit. In contrast to the cross-section literature, the average is taken over the post-intervention period and not over the treated units. In the first step, we estimate a multivariate model based on a high-dimensional panel of time-series data from a pool of untreated peers without any stringent assumptions about the data generating process (DGP). Then, we compute the counterfactual by extrapolating the model with data after the intervention. High-dimensionality is relevant when the number of parameters to be estimated is large compared to the sample size. This can occur either when the number of peers and/or the number of variables for each peer is large or when the sample size is small. We use the *least absolute shrinkage and selection operator* (LASSO) proposed by Tibshirani (1996) to estimate the parameters. Nonlinearities can be handled by including some transformations of the explanatory variables, such as polynomials or splines, in the model. Furthermore, we propose a test of no intervention effects with a standard limiting distribution that is uniformly valid in a wide class of DGPs, without imposing any strong restrictions on the model parameters, as is usually the case when the LASSO is the estimation method, or modifying the estimator, as in Belloni et al. (2017). We also show that it is not necessary to consider two-step extensions of the LASSO, such as the adaptive LASSO of Zou (2006), to handle highly collinear regressors. Contrary to other methods, the ArCo methodology is able to simultaneously test for effects on different variables and on multiple moments of a set of variables, such as the mean and the variance. In addition, we accommodate situations where the exact time of the intervention is unknown, which is important in the case of anticipation effects. We also propose an $\mathcal{L}_p$ statistic inspired by the literature on structural breaks and show that the asymptotic properties of the method remain unchanged. Finally, we extend the methodology to the cases of multiple interventions and to contamination effects among units.

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

A Monte Carlo experiment is conducted to evaluate the small-sample performance of the methodology in comparison to well-established alternatives, namely, the before-and-after (BA) estimator, the differences-in-differences (DiD) estimator, the panel factor model of Gobillon and Magnac (2016) (PF-GM) and the SC method. We show that the bias of the ArCo method is negligible and much smaller than some of the alternatives. Simulations show that the variance and the mean squared error of the ArCo estimator are smaller than those of its competitors and that the test for the null hypothesis of no intervention effect has good size and power properties.

As an illustration, we evaluate the impacts on inflation and other macroeconomic variables of an anti tax-evasion program implemented in Brazil. The mechanism works by giving 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. Under the assumptions that (i) a certain degree of tax evasion was occurring before the intervention, (ii) the seller has some degree of market power and (iii) the penalty for tax evasion is sufficient to alter seller behavior, one is expected to see an upward movement in prices due to an increase in marginal cost. Compared to the counterfactual, the program caused a price increase of 10.72% over 23 months. This is an important result as most of the studies in the literature focus only of the effects of such policies on reducing tax evasion and neglect the effects on inflation. To highlight the multivariate nature of the ArCo methodology we also test for joint effects on GDP, retail sales and credit. We find no effect of the program on these variables.

1.2. Connections to the literature

Hsiao et al. (2012) considered a two-step method where in the first step, the counterfactual for a single treated variable is constructed as a linear combination of a low-dimensional set of covariates from pre-selected elements from a pool of peers. The model is estimated by ordinary least squares using pre-intervention data. Their theoretical results are 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 performed by information criteria. Recently, several extensions have been proposed. Ouyang and Peng (2015) relaxed the linear conditional expectation assumption. Du and Zhang (2015) and Li and Bell (2017) improved the selection mechanism for the donor pool.

The ArCo method generalizes the above papers in important directions. First, we do not restrict the analysis to a single treated variable. We can measure the impact of interventions on several variables of the treated unit simultaneously. We also allow for tests on several moments of a set of variables. For example, consider the case where the interest is on measuring the effects of a new policy on the first two moments of inflation or on inflation and output growth simultaneously. A test for joint effects is not possible with the previously proposed methods and the econometrician can only test the effects on each series separately. Second, we demonstrate that our methodology can be applied even when the intervention time is unknown. Third, we develop tests for multiple interventions and contamination effects. Finally, all previous results are derived in a high-dimensional framework where the first step estimation is carried out by LASSO, allowing for a large number of covariates/peers to be included and do not require any pre-estimation selection, which can bias the estimates. Shrinkage estimation is appealing when the sample size is small compared to the number of parameters to be estimated. Although Li and Bell (2017) have advocated the use of LASSO as a selection mechanism, the authors do not provide any theoretical results. We not only derive the asymptotic properties of the estimators but also show that all our convergence results are uniform on a wide class of probability laws under mild conditions. All our theoretical results are derived under no stringent assumptions about the DGP. We do not need to estimate the true conditional expectation as we consider the estimation of a linear projection on a set of conditioning variables. This is a positive feature of the ArCo methodology as models are usually misspecified.

Compared to DiD estimators, the advantages of the ArCo method are threefold. 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 for situations where the $n-1$ untreated units differ substantially from the treated unit and cannot form a control group, even after conditioning on a set of observables. Finally, the ArCo estimator is consistent even without the parallel trends hypothesis.[4]

More recently, Gobillon and Magnac (2016) generalize DiD estimators by estimating a correctly specified linear panel model with strictly exogenous regressors and interactive fixed effects represented as a number of common factors with heterogeneous loadings. Their theoretical results rely on double asymptotics when both T and n go to infinity. The authors allow the common confounding factors to have nonlinear deterministic trends, which is a generalization of the linear parallel trend hypothesis assumed when DiD estimation is considered.

The ArCo method differs from Gobillon and Magnac (2016) in many ways. First, as mentioned previously, we do not assume the model to be correctly specified, and we do not need to estimate the common factors. Consistent estimation of factors requires both the time-series and cross-section dimensions diverge to infinity and can be severely biased in small samples. The ArCo methodology requires only the time-series dimensions to diverge. Furthermore, we do not require the regressors to be strictly exogenous, which is an unrealistic assumption in most applications with aggregate (time-series) data. We also provide consistency results under heterogeneous nonlinear trends, but there is no need to estimate them (either explicitly or via common factors). Finally, as in the DiD case, we do not require the number of treated units to grow or to have a reliable control group (after conditioning on covariates).

Although, both the ArCo and the SC methods construct a counterfactual as a function of observed variables from a pool of peers, the two approaches have important differences. First, the SC method relies on a convex combination of peers to construct the counterfactual, which as pointed out by Ferman and Pinto (2016), biases the estimator. This is clearly evidenced in our simulation experiment. The ArCo solution is a general, possibly nonlinear, function. Even in the case of linearity, the method does not impose restrictions on the parameters. For example, the restriction that the weights in the SC method are all positive appear to be too strong. The SC method also requires an unrealistic identification assumption about the (perfect-)fit of the model in the pre-intervention period. 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 are removed, and the weights are determined in a pure cross-sectional setting. In some applications of the SC method, the number of observations used to estimate the weights is much smaller than the number of parameters to be determined. For example, in Abadie and Gardeazabal (2003), the authors have 13 observations to estimate 16 parameters.[5] In addition, the SC method was designed to evaluate the effects of the intervention on a single variable: the method has to be applied several times to evaluate the effects on a vector of variables. The ArCo methodology can be directly applied to a vector of variables of interest. In addition, there is no formal inferential procedure for hypothesis testing in the SC method, whereas in the ArCo methodology, a simple, uniformly valid and standard test can be applied. Finally, as discussed in Ferman et al. (2016), the SC method does not provide any guidance on how to select the variables that determine the optimal weights.[6]

With respect to the methodology of Pesaran and Smith (2012), the major difference is that the authors construct the counterfactual based on variables that belong to the treated unit and 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 this could be a reasonable hypothesis in some specific cases, in a general framework, this assumption is clearly restrictive.

Angrist et al. (2018) propose a semiparametric method to evaluate the effects of monetary policy. The authors rely on only information about the treated unit, and no donor pool is available. As before, this is a major difference from our approach.

[4] The first difference can be attenuated in light of the recent results of Conley and Taber (2011) and Ferman and Pinto (2015), who proposed inferential procedures when the number of treated groups is small.

[5] In these cases, the estimation is only possible due to the imposed restrictions, which can be seen as a sort of shrinkage. A similar issue appears in Abadie et al. (2010, 2015).

[6] Doudchenko and Imbens (2016) advocate the use of shrinkage methods to estimate the pre-intervention model, but no theory is provided.

Furthermore, their methodology seems to be particularly appealing to monetary economics but is difficult to apply in other settings without major modifications.

It is important to compare the ArCo methodology with the work of Belloni et al. (2014) and Belloni et al. (2017). Both papers consider the estimation of intervention effects in large dimensions. The first one consider a pure cross-sectional setting where the intervention is correlated with 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, instead of relying on instrumental variables, a set of peers is used to construct an artificial counterfactual, and the intervention is assumed to be exogenous with respect to this set of peers. Notwithstanding, the intervention may not be (and probably is not) independent of the variables belonging to the treated unit. This key assumption enables us to construct honest confidence bands by using the LASSO to estimate the conditional model in the first step. In the second paper the authors proposed a general and flexible extension of the DiD 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 when an artificial (synthetic) counterfactual must be computed to evaluate the intervention effects. Finally, it is not clear how to apply their methods to aggregate (macro) data where time-series dynamics must be considered.

Finally, it is worth comparing the ArCo method with the structural change literature. The intervention considered here can be viewed as a structural break in the DGP of the variables of interest and a possible test for its effects is to check for parameter instability. However, the difficulty of this approach is to control for confounding effects as there is no control group available, not even a synthetic one, and this test is equivalent to a "before-and-after" comparison.

1.3. Applications

There is a number of studies that require the estimation of intervention effects with no group of controls. The ArCo method can be applied to the same types of applications as SC or PF methods, which have been widely used (Athey and Imbens, 2016).

Measuring the impacts of regional policies is a potential application. Hsiao et al. (2012) measured the impact of the economic and political integration of Hong Kong with mainland China on Hong Kong's economy, and Abadie et al. (2015) estimated the spillover of the 1990 German reunification in West Germany. Pesaran et al. (2007) study the effects of the launching of the Euro. Gobillon and Magnac (2016) considered the impact on unemployment of a new policy implemented in France in the 1990s. The effects of trade agreements were discussed in Billmeier and Nannicini (2013) and Jordan et al. (2014). The rise of a new government is also a relevant "intervention" to study. Grier and Maynard (2013) considered the economic impacts of the Chavez era.

Other applications are new regulations on housing prices, as in Bai et al. (2014) and Du and Zhang (2015), new labor laws, as considered in Du et al. (2013), and the 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 et al. (2018). Estimating the economic consequences of natural disasters, as in Belasen and Polachek (2008), Cavallo et al. (2013), Fujiki and Hsiao (2015), and Caruso and Miller (2015), is also a promising area of research.

The effects of market regulation and the introduction of new financial instruments on the risk and returns of stock markets have been considered in Chen et al. (2013) and Xie and Mo (2013). Testing the intervention effects on multiple moments of the data is a special interest in finance, where the goal could be to determine the effects of different corporate governance policies on the returns and risk of firms (Johnson et al., 2000).

1.4. Plan of the paper

In Section 2, we present the ArCo method and discuss the conditional model used in the first step. In Section 3, we derive the asymptotic properties of the ArCo estimator. Sub-Section 3.3 addresses the test for the null hypothesis of no causal effect. Extensions for trending regressors, unknown intervention time, multiple interventions and possible contamination effects are described in Section 4. In Section 5, we discuss potential sources of bias in the method. A Monte Carlo study is conducted in Section 6, and Section 7 considers an empirical exercise. Section 8 concludes the paper. Tables, figures and proofs are provided in the Appendix. A supplementary material provides additional results. Equations, tables and figures numbered with an "S" refer to this supplement.

2. The artificial counterfactual estimator

Suppose we have n units (countries, states, 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 a set of variables $\boldsymbol{z}_{it} = (z_{it}^1, \ldots, z_{it}^{q_i})' \in \mathbb{R}^{q_i}$, $q_i \geq 1$. Furthermore, assume that an intervention occurred in unit $i = 1$ and only unit 1 at time $T_0 = \lfloor \lambda_0 T \rfloor$, where $\lambda_0 \in (0, 1)$ and $\lfloor \cdot \rfloor$ is the floor function.

Let $\mathcal{D}_t$ be a binary variable indicating 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 or equals 0 otherwise, $\boldsymbol{z}_{1t}^{(1)}$ denotes the outcome when 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 concerned with testing hypotheses on the effects of the intervention on unit 1 for $t \geq T_0$. In particular, we are interested in measuring the effects on a transformation of $\boldsymbol{z}_{1t}$ defined as $\boldsymbol{y}_t \equiv \boldsymbol{h}(\boldsymbol{z}_{1t})$, where $\boldsymbol{h} : \mathbb{R}^{q_1} \mapsto \mathbb{R}^q$ is a measurable function of $\boldsymbol{z}_{1t}$. The choice of transformation $\boldsymbol{h}(\cdot)$ depends on which moments of the data the econometrician is interested in testing for effects of the intervention. In other words, the goal is to test for a break in a set of unconditional moments of the data and to check whether this break is solely due to the intervention or has other (global) causes. Typical choices for $\boldsymbol{h}(\cdot)$ are presented below.

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

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

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

$\operatorname{vech}(\boldsymbol{A})$ is the half vectorization of a symmetric matrix $\boldsymbol{A}$ (a column vector obtained by vectorizing only the lower triangular part of $\boldsymbol{A}$).

Example 3. We can also conduct joint tests by combining the different choices of $\boldsymbol{h}$. For example, to simultaneously test for a change in the mean and variance in the univariate case, we can set $\boldsymbol{h}(a) = (a, a^2)'$.

Set $\boldsymbol{y}_t = \mathcal{D}_t \boldsymbol{y}_t^{(1)} + (1 - \mathcal{D}_t)\boldsymbol{y}_t^{(0)}$. As before, $\boldsymbol{y}_t^{(1)}$ denotes the outcome when unit 1 is exposed to the intervention, and $\boldsymbol{y}_t^{(0)}$ is the potential outcome of unit 1 when there is no intervention. The exact dimensions of $\boldsymbol{y}_t$ depend on the chosen $\boldsymbol{h}(\cdot)$. However, regardless of the choice of $\boldsymbol{h}(\cdot)$, we consider, without loss of generality, that $\boldsymbol{y}_t \in \mathcal{Y} \subset \mathbb{R}^q$, $q > 0$ and that we have a sample $\{\boldsymbol{y}_t\}_{t=1}^{T}$, where the first $T_0 - 1$ observations are before the intervention and the $T - T_0 + 1$ remaining observations are after the intervention.

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 $\{\boldsymbol{\delta}_t\}_{t=T_0}^{T}$ is a deterministic sequence. Due to the flexibility of the mapping $\boldsymbol{h}(\cdot)$, interventions modeled as (1) are quite general and include 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 similar to the traditional *average treatment effect on the treated* (ATET) vastly discussed in the literature.[7] 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 is a special case of uniform treatments by setting $\boldsymbol{\delta}_t = \boldsymbol{\delta}$, $\forall t \geq T_0$.

Clearly, we do not observe $\boldsymbol{y}_t^{(0)}$ after $T_0 - 1$. We call $\boldsymbol{y}_t^{(0)}$ the *counterfactual*, i.e., what $\boldsymbol{y}_t$ would have been like had there been no intervention (potential outcome). To construct the counterfactual, let $\boldsymbol{z}_{0t} = (\boldsymbol{z}_{2t}', \ldots, \boldsymbol{z}_{nt}')'$ and $\boldsymbol{Z}_{0t} = \left(\boldsymbol{z}_{0t}', \ldots, \boldsymbol{z}_{0t-p}'\right)'$ be the collection of all the untreated units' observables up to an arbitrary lag $p \geq 0$. The exact dimensions of $\boldsymbol{Z}_{0t}$ depend upon the number of peers $(n-1)$, the number of variables per peer, q_i, $i = 2, \ldots, n$, 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 approximating model for $\boldsymbol{y}_t$ in the absence of the intervention

$$\boldsymbol{y}_t^{(0)} = \mathcal{M}(\boldsymbol{Z}_{0t}, \boldsymbol{\theta}_0) + \boldsymbol{v}_t,\ t = 1, \ldots, T, \tag{3}$$

where $\mathcal{M} : \mathcal{Z}_0 \times \boldsymbol{\Theta} \to \mathcal{Y}$ is a measurable mapping for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a finite dimensional parametric space and we assume $\mathbb{E}(\boldsymbol{v}_t) = \boldsymbol{0}$.[8] We defer the discussion about the functional form of $\mathcal{M}(\cdot, \cdot)$ and the precise definition of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ until Section 3.

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. As $\boldsymbol{y}_t^{(0)}$ is observed for $t < T_0$, we can thus define:

Definition 1. The artificial counterfactual (ArCo) estimator is

$$\widehat{\boldsymbol{\Delta}}_T = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \widehat{\boldsymbol{\delta}}_t, \tag{4}$$

where $\widehat{\boldsymbol{\delta}}_t \equiv \boldsymbol{y}_t - \mathcal{M}(\boldsymbol{Z}_{0t}, \widehat{\boldsymbol{\theta}}_{T_1})$, for $t = T_0, \ldots, T$ and $\widehat{\boldsymbol{\theta}}_{T_1}$ is a consistent estimator for $\boldsymbol{\theta}_0$ using only the first T_1 observations of the data (pre-intervention).

[7] However, as noted in the Introduction, the average is taken over time periods and not over cross-section elements.

[8] Which can be ensured by either including a constant in the model $\mathcal{M}$ or by centering the variables in a linear specification. Please note that Eq. (3) does not necessarily represent the true data generating process for $\boldsymbol{y}_t^{(0)}$ but rather an approximating model.

Therefore, the ArCo is a two-stage estimator where in the first stage, we choose and estimate the parameters in the model $\mathcal{M}$ using the pre-intervention sample, and in the second, we compute $\widehat{\boldsymbol{\Delta}}_T$ defined by (4). At this point, the following remarks are in order.

Remark 1. The ArCo estimator in (4) 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 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 remain valid when λ_0 is unknown.

There are two major advantages of applying the ArCo estimator instead of computing a simple difference in the mean of $\boldsymbol{y}_t$ before and after the intervention. The first is an efficiency argument. Note that the "before-and-after" estimator defined as $\widehat{\boldsymbol{\Delta}}_T^{BA} \equiv \frac{1}{T-T_0+1}\sum_{t=T_0}^{T}\boldsymbol{y}_t - \frac{1}{T_0-1}\sum_{t=1}^{T_0-1}\boldsymbol{y}_t$ is a particular case of our estimator when there are "bad peers" as they are uncorrelated with 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 variance of the ArCo estimator. The second argument in favor of the ArCo method is related to its capability to isolate the intervention from aggregate shocks. When attempting to measure the effect of an intervention, we are usually in a scenario that other aggregate shocks occurred at the same time. The ability to disentangle these effects is vital to provide a meaningful estimation of the intervention effect.

To recover the effects of the intervention by the ArCo we need the following key assumption.

Assumption 1. $\boldsymbol{z}_{0t}$ is independent of $\mathcal{D}_s$ for all t, s.

The assumption above is sufficient for the peers to be unaffected by the intervention. Assumption 1 has also been assumed in the case of SC and PF methods.

3. Asymptotic properties and inference

3.1. Choice of the pre-intervention model

The first stage of the ArCo method requires the choice of the model $\mathcal{M}$ which should capture most of the information from the available peers. Once the choice is made, the model must be estimated using the pre-intervention sample. We do not assume that the selected model is the true model and we consider it only as an approximation to the conditional mean $\boldsymbol{m}(\boldsymbol{Z}_{0t}) \equiv \mathbb{E}(\boldsymbol{y}_t^{(0)}|\boldsymbol{Z}_{0t})$.

Motivated by the fact that the dimensions of $\boldsymbol{Z}_{0t}$ can grow quite fast (by either including more peers, more covariates, or simply considering more lags), we propose a fully parametric specification to approximate $\boldsymbol{m}(\cdot)$ as opposed to attempting 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. Specifically, $\boldsymbol{h}_x$ could be a dictionary of functions, such as polynomials, splines, interactions, dummies or any another family of elementary transformations of $\boldsymbol{Z}_{0t}$, in the spirit of sieve estimation (Chen, 2007).

Therefore, $\mathcal{M}(\boldsymbol{Z}_{0t}, \boldsymbol{\theta}_0) = (\boldsymbol{\theta}_{0,1}'\boldsymbol{x}_{1,t}, \ldots, \boldsymbol{\theta}_{0,q}'\boldsymbol{x}_{q,t})'$ in (3) where both $\boldsymbol{x}_{j,t}$ and $\boldsymbol{\theta}_{0,j}$ are d_j-dimensional vectors for $j = 1, \ldots, q$. We allow d_j to be a function of T. Hence, $\boldsymbol{x}_{j,t}$ and $\boldsymbol{\theta}_{0,j}$ depend on T, but the subscript T is omitted. Set $\boldsymbol{r}_t \equiv \boldsymbol{m}(\boldsymbol{Z}_{0t}) - \mathcal{M}(\boldsymbol{Z}_{0t}, \boldsymbol{\theta}_0)$ as the approximation error and $\boldsymbol{\varepsilon}_t \equiv \boldsymbol{y}_t^{(0)} - \boldsymbol{m}(\boldsymbol{Z}_{0t})$ as the projection error. We can write the model as in (3), with $\boldsymbol{\nu}_t = \boldsymbol{r}_t + \boldsymbol{\varepsilon}_t$. Hence,

$$y_{jt}^{(0)} = \boldsymbol{\theta}_{0,j}'\boldsymbol{x}_{j,t} + \nu_{jt}, \quad j = 1, \ldots, q, \tag{5}$$

where $\boldsymbol{\theta}_{0,j}$ are the best (in terms of MSE) 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 the dependence of the joint distribution on the sample size T explicit, 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 $T \to \infty$.[9]

[9] We denote convergence in probability and in distribution by "$\xrightarrow{p}$" and "$\xrightarrow{d}$", respectively.

3.2. Assumptions and asymptotic theory in high dimensions

The dimensions d_j of $\boldsymbol{x}_{j,t}$ can be 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_j, and consequently $\boldsymbol{\theta}_{0,j}, j = 1\ldots, q$, to be a function of the sample size, such that $d_j \equiv d_{j,T}$ and $\boldsymbol{\theta}_{0,j} \equiv \boldsymbol{\theta}_{0,j,T}$. 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 (5), equation by equation, by the LASSO, allowing the dimension of $d_j > T$ to grow faster than the sample size. We drop the subscript j from now on to focus on a generic equation. Therefore, we estimate $\boldsymbol{\theta}_0$ via

$$\widehat{\boldsymbol{\theta}} = \arg\min\left\{\frac{1}{T_1}\sum_{t<T_0}(y_t - \boldsymbol{x}_t'\boldsymbol{\theta})^2 + \varsigma\|\boldsymbol{\theta}\|_1\right\}, \tag{6}$$

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 let $S_0 = \{i : \theta_{0,i} \neq 0\}$ with cardinality s_0. We consider the following set of assumptions.[10]

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

Assumption 4 (*Regularity*).

(a) The regularization parameter $\varsigma = \kappa\frac{d^{1/\gamma}}{\sqrt{T}}$, for some constant $\kappa > 0$.
(b) $s_0\frac{d^{2/\gamma}}{\sqrt{T}} = o(1)$.

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

Assumption 3 controls for the heterogeneity and the dependence structure of the process that generates the sample. Specifically, 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. Second-order stationarity would be sufficient for asymptotic normality, but we require fourth-order stationarity to consistently estimate the covariance matrix of the estimator. Assumption 3(b) uniformly bounds some higher moments, 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 many relevant variables we have in (5). Assumption 4 impose regularity conditions on the growth rate of the penalty parameter and the number of parameters, respectively. They are smaller than the analogous results in the literature for the Gaussian case with fixed design.[11]

We can now define $\mathcal{P}$ as the class of probability laws that satisfies Assumptions 2, 3 and 4(b). Here is our main result.

Theorem 1 (*Main*). *Consider the estimator in (4) with the model given by* $\mathcal{M}(\boldsymbol{Z}_{0t}, \boldsymbol{\theta}_0) = (\boldsymbol{\theta}_{0,1}'\boldsymbol{x}_{1,t}, \ldots, \boldsymbol{\theta}_{0,q}'\boldsymbol{x}_{q,t})'$ *as in (5) whose parameters are estimated by (6) using only the pre-intervention sample* ($t < T_0$), *i.e.,*

$$\widehat{\Delta}_T = \frac{1}{T-T_0+1}\sum_{t=T_0}^{T}\boldsymbol{y}_t - (\widehat{\boldsymbol{\theta}}_{1,T_1}'\boldsymbol{x}_{1,t}, \ldots, \widehat{\boldsymbol{\theta}}_{q,T_1}'\boldsymbol{x}_{q,t})',$$

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

[11] Under those conditions, 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.

where $\widehat{\boldsymbol{\theta}}_{j,T_1}$ for $j = 1, \ldots, q$ is a minimizer of

$$\boldsymbol{\theta} \mapsto \frac{1}{T_1}\sum_{t<T_0}(y_{jt} - \boldsymbol{x}_{jt}'\boldsymbol{\theta})^2 + \varsigma\|\boldsymbol{\theta}\|_1.$$

Then, under Assumptions 1–4,

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

where the event $\{\boldsymbol{a} \le \boldsymbol{b}\} \equiv \{a_i \le b_i, \forall i\}$, $\Phi(\cdot)$ is the cumulative distribution function of a zero-mean normal random vector with identity variance–covariance matrix, $\boldsymbol{\Omega}_T \equiv \frac{\boldsymbol{\Gamma}_{T_1}}{T_1/T} + \frac{\boldsymbol{\Gamma}_{T_2}}{T_2/T}$, $\boldsymbol{\Gamma}_{T_1} = \mathbb{E}_P\left[\frac{1}{T_1}(\sum_{t\le T_1}\boldsymbol{\nu}_t)(\sum_{t\le T_1}\boldsymbol{\nu}_t')\right]$, and $\boldsymbol{\Gamma}_{T_2} = \mathbb{E}_P\left[\frac{1}{T_2}(\sum_{t\ge T_0}\boldsymbol{\nu}_t)(\sum_{t\ge T_0}\boldsymbol{\nu}_t')\right]$.

The results above are uniform with respect to the class of probability laws $\mathcal{P}$, which we believe to be sufficiently large to be of interest. Note that we do *not* require any strong separation of the parameters away from zero, which is usually accomplished in the literature by imposing a $\theta_{\min}$ that is uniformly bounded away from zero. Uniform convergence above is possible, in our case, as a consequence of the LASSO estimation error in the first step being negligible. Writing $\boldsymbol{x}_t = (1, \tilde{\boldsymbol{x}}_t)'$ and $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0')'$, careful review of the proof of Theorem 1 in Appendix A reveals that the LASSO estimation effect appears through the term

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

We show that, under the Assumptions of Theorem 1, the first term is $O_P(d^{1/\gamma}s_0/\sqrt{T})$ and the second $O_P(d^{1/\gamma})$. Therefore, the term is $O_P(d^{2/\gamma}s_0/\sqrt{T}) = o_P(1)$ uniformly in $P \in \mathcal{P}$. Informally, the potential 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 setup, Belloni et al. (2014) consider the case where the treatment is correlated with the set of regressors. Consequently, they propose estimation via a moment condition with the so-called *orthogonality property* to achieve uniform convergence. Further, Belloni et al. (2016) generalize this idea to conduct uniform inference in a broad class of Z-estimators.

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

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

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

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. Therefore, our estimator variance for any given λ_0 lies between the two polar cases: when there is a perfect artificial counterfactual and when the peers contribute no information. Thus, the peers' contribution to reducing the ArCo estimator asymptotic variance can be represented by an R^2-type statistic measuring the "ratio" between the explained long-run variance $\boldsymbol{\Omega}$ and the total long-run variance $\widetilde{\boldsymbol{\Omega}}$.

3.3. Hypothesis testing under the 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 on 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, we consider that unit 1 is the only potentially treated unit and that the moment of 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 the errors are uncorrelated, and given the consistency of $\widehat{\boldsymbol{\theta}}$, we can simply estimate the quantities of interest by the average of the squared residuals in the pre-intervention model. On the other hand, the results in Newey and West (1987) and Andrews (1991) can be used for serially correlated errors. In this case,

$$\widehat{\boldsymbol{\Gamma}}_{T_i} = \widehat{\boldsymbol{V}}_{0_i} + \sum_{k=1}^{T_i-1}\phi(k/S_T)\left(\widehat{\boldsymbol{V}}_{k_i} + \widehat{\boldsymbol{V}}_{k_i}'\right), \quad i = \{1, 2\}, \tag{7}$$

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

In practice, we need to specify the bandwidth parameter S_T and the weight function ϕ. The latter is usually a kernel function centered at zero. A common choice is a Bartlett kernel, where the weights are given simply by $\phi(k) = 1 - \frac{k}{M+1}$, where M is a positive constant. Theorem 2 of Newey and West (1987) and Proposition 1 of Andrews (1991) give general conditions under which the estimator is consistent in the low-dimensional setup. Moreover, Andrews (1991) discusses the choice of kernels and presents a sizeable list of options. To state our result, we borrow the definition of a class $\mathcal{K}$ of allowable kernels $\mathcal{K} = \{\phi(\cdot) : \mathbb{R} \to [-1,1] | \phi(0) = 1, \phi(x) = \phi(-x), \forall x \in \mathbb{R}, \int \phi^2(u)du < \infty, \phi$ is continuous at 0 and at all but finite many points in $\mathbb{R}\}$. It includes the most commonly used kernels in the literature, such as truncated, Bartlett, Parzen, Tukey–Hanning and quadratic spectral.

Theorem 2. *Under Assumptions 1–4, consider further for the estimator defined by (7)*

(a) *$\phi(\cdot)$ is in the class $\mathcal{K}$ (defined above) and $\int |\phi(u)|du < \infty$.*
(b) *$S_T = o\left(\frac{\sqrt{T}}{s_0 d^{1/\gamma}}\right)$ and $S_T \to \infty$.*

Then, $\widehat{\boldsymbol{\Gamma}}_{T_i} - \boldsymbol{\Gamma}_{T_i} = o_p(1)$ uniformly in $P \in \mathcal{P}$ for $i = \{1, 2\}$.

Therefore, if we replace $\boldsymbol{\Omega}_T$ with $\widehat{\boldsymbol{\Omega}}_T \equiv \frac{\widehat{\boldsymbol{\Gamma}}_{T_1}}{T_1/T} + \frac{\widehat{\boldsymbol{\Gamma}}_{T_2}}{T_2/T}$, we have a uniform consistent estimator, which allows us to construct honest (uniform) asymptotic confidence intervals and perform hypothesis testing as follows.

Proposition 1 (*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 as those 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 $j = 1, \dots, q$, where $\widehat{\omega}_j = \sqrt{[\widehat{\boldsymbol{\Omega}}]_{jj}}$ 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 2 (*Uniform Hypothesis Test*)**.** *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 a given $\epsilon > 0$, there exists a T_ϵ such that for all $T > T_\epsilon$:*

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

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

4. Extensions

We consider extensions of the previously developed framework. Section 4.1 presents two results for the case of trend-stationary units. In Section 4.2, we propose a procedure to account for the problem of an unknown intervention time and develop a consistent estimator for the most likely intervention time. The case of multiple intervention points is considered in Section 4.3. Finally, Section 4.4 investigates the presence of a treated unit among the controls, which is particularly useful for testing for spillover effects.

4.1. Deterministic trends

For clarity, we consider the case where $q_1 = \cdots = q_n = 1$ and $h(a) = a$, such that $y_t = z_{1t}^{(0)}$. Furthermore, we let the units have a (not necessarily common) deterministic trend. In particular, we assume that in the absence of the intervention, $z_{it}^{(0)} = s_{it} + \zeta_i(t/T)$, $i = 1, \dots, n$, where $\zeta_i(\cdot)$ is an integrable function on [0,1], as in Bai (2009). Note that since the deterministic term is normalized by T, it does not dominate the stochastic component s_{it} asymptotically. Let $\boldsymbol{x}_t = (1, \boldsymbol{z}_{0t}')'$ and $\boldsymbol{z}_{0t} = (z_{2t}, \dots, z_{nt})'$. In this setup, the pre-intervention model becomes a single-equation version of (5). However, due to the deterministic trend, $(y_t, \boldsymbol{x}_t')'$ is non-stationary, which is ruled out by Assumption 3. Hence, we consider a less restrictive assumption as follows.

Assumption 5 (*Trending Regressors*)**.** In model (5):

(a) $\boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{\zeta}(t/T)$, where $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_d)'$ and $\zeta_i(t/T)$ is an integrable function on [0,1] for $i = 1, \dots, d$, such that $\sup_{i \le d, d \ge 1} \zeta_i(t/T) < \infty$.
(b) $\{\boldsymbol{w}_t \equiv (\nu_t, \boldsymbol{s}_t')'\}$ fulfills Assumption 3.

Generally, Assumption 5 requires the regressors to be trend-stationary, i.e., stationary except for the deterministic trend. Therefore, we have the following result:

Theorem 3. *Under Assumptions* 1, 2, 4 and 5

(a) $\widehat{\Delta}_T - \Delta_T = o_P(1)$.
(b) $\sqrt{T}\Omega_T^{-1/2}(\widehat{\Delta}_T - \Delta_T + b_T) \xrightarrow{d} N(0,1)$, *where* $N(0,1)$ *denotes the standard normal distribution,* $b_T = \boldsymbol{r}_T(\boldsymbol{\zeta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$,

$$\boldsymbol{r}_T(\boldsymbol{\zeta}) \equiv \left[\frac{1}{T_2}\sum_{t \geq T_0}\boldsymbol{\zeta}(t/T) - \frac{1}{T_1}\sum_{t \leq T_1}\boldsymbol{\zeta}(t/T)\right],$$

and $\widehat{\boldsymbol{\beta}}$ *is the LASSO estimator for the vector of coefficients* $\boldsymbol{\beta}_0$ *of the non-constant regressors.*

As a consequence, the consistency is preserved in the case of trending regressors since $b_T = o_P(1)$ under the assumptions above. However, the bias term b_T appears in the asymptotic normality expression since, in general, $\sqrt{T}b_T$ does not vanish in probability.

An exact expression for the limiting distribution of the estimator in the high-dimensional setting with trending regressors does not appear to be straightforward. In contrast to the stationary case, the influence of the pre-intervention estimation does not vanish asymptotically. Hence, the limiting distribution is likely to be influenced by the limiting distribution of the LASSO estimator.

To gain some intuition about the limiting distribution of our estimator under this scenario, we consider the low-dimensional case where $d < T_1$ is fixed, and we estimate the pre-intervention model by ordinary least squares (OLS) to obtain the following result:

Theorem 4. *For a fixed number of regressors* $d < T_1$, *if the parameters in model* (5) *are estimated via OLS; then, under Assumptions* 1, 3 and 5, *and provided that* $\boldsymbol{\Sigma} \equiv \lim_{T\to\infty}\frac{1}{T_1}\sum_{t=1}^{T_1}\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ *is non-singular,*

$$\sqrt{T}\Upsilon_T^{-1/2}(\widehat{\Delta}_T - \Delta_T) \xrightarrow{d} N(0,1),$$

where $\Upsilon_T \equiv \frac{\Lambda_{T_1}}{T_1/T} + \frac{\Gamma_{T_2}}{T_2/T}$, Γ_{T_2} *is defined in Theorem* 1, $\Lambda_{T_1} = \boldsymbol{a}'\boldsymbol{M}_1\boldsymbol{a}$, $\boldsymbol{a} = [1, \boldsymbol{r}_T(\boldsymbol{\zeta})'\boldsymbol{M}_2^{-1}]'$, $\boldsymbol{M}_1 = \mathbb{E}_P\left[\frac{1}{T_1}(\sum_{t\leq T_1}\check{\boldsymbol{z}}_{0t}\nu_t)(\sum_{t\leq T_1}\check{\boldsymbol{z}}_{0t}'\nu_t)\right]$, $\boldsymbol{M}_2 = \mathbb{E}_P\left[\frac{1}{T_1}\sum_{t\leq T_1}\widetilde{\boldsymbol{z}}_{0t}\widetilde{\boldsymbol{z}}_{0t}'\right]$, $\boldsymbol{r}_T(\boldsymbol{\zeta})$ *is defined in Theorem* 3, $\widetilde{\boldsymbol{z}}_{0t} = \boldsymbol{z}_{0t} - \mathbb{E}_P\left[\frac{1}{T_1}\sum_{t\leq T_1}\boldsymbol{z}_{0t}\right]$, *and* $\check{\boldsymbol{z}}_{0t} = (1, \widetilde{\boldsymbol{z}}_{0t}')'$. *Recall that* $\boldsymbol{x}_t = (1, \boldsymbol{z}_{0t}')'$ *and* $\boldsymbol{z}_{0t} = (z_{2n}, \ldots, z_{nt})'$.

From Theorem 4, we can recover the low-dimensional version of Theorem 1 since we can set $\boldsymbol{r}_T(\boldsymbol{\zeta}) = \boldsymbol{0}$ in the absence of a deterministic trend. In this case, $\Lambda_{T1} = \Gamma_{T1}$ and, consequently, $\Upsilon_T = \Omega_T$. Moreover, we conjecture that the same result can be obtained in the high-dimensional setup if we use the adaptive LASSO of Zou (2006) instead of the LASSO estimator for the pre-intervention model. Provided conditions to ensure consistent model selection, the non-zero coefficient would be estimated in the same way as the OLS asymptotically due to its oracle property.

Furthermore, the variance Υ_T that appears in Theorem 4 can be consistently estimated by $\widehat{\Upsilon}_T \equiv \frac{\widehat{\Lambda}_{T_1}}{T_1/T} + \frac{\widehat{\Gamma}_{T_2}}{T_2/T}$, where $\widehat{\Gamma}_{T_2}$ is defined by (7), $\widehat{\Lambda}_{T_1} = \boldsymbol{d}'\boldsymbol{M}_3^{-1}\boldsymbol{M}_4\boldsymbol{M}_3^{-1}\boldsymbol{d}$ with $\boldsymbol{d} = \frac{1}{T_2}\sum_{t\geq T_0}\boldsymbol{x}_t$, $\boldsymbol{M}_3 = \frac{1}{T_1}\sum_{t<T_0}\boldsymbol{x}_t\boldsymbol{x}_t'$, $\boldsymbol{M}_4 = \sum_{|k|<T_1}\phi(k/S_T)\boldsymbol{D}_k$ with $\boldsymbol{D}_k = \frac{1}{T_1}\sum_{t=1+k}^{T_1}\boldsymbol{x}_t\boldsymbol{x}_{t-k}'\widehat{\nu}_t\widehat{\nu}_{t-k}$ for $k \geq 0$ and $\boldsymbol{D}_k = \boldsymbol{D}_{-k}'$ for $k < 0$, and $\widehat{\nu}_t = y_t - \widehat{\boldsymbol{\theta}}'\boldsymbol{x}_t$ for $t = 1, \ldots, T_1$. The result holds as long as the kernel $\phi(\cdot)$ belongs to the class $\mathcal{K}$ defined in condition (a) of Theorem 2 and the bandwidth parameter is chosen such that $S_T/\sqrt{T} = o(1)$.

4.2. Unknown intervention timing

There are reasons why the intervention timing might not be known with certainty, for example, anticipation effects related to rational expectations regarding an announced change in future policy or a simple delay in the response of the variable of interest. Regardless of the cause of uncertainty in the timing of the intervention, we propose a way to apply the methodology 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 $(0,1)$:

$$\widehat{\boldsymbol{\Delta}}_T(\lambda) = \frac{1}{T - T_\lambda + 1}\sum_{t \geq T_\lambda}\widehat{\boldsymbol{\delta}}_{t,T}(\lambda), \quad \forall \lambda \in \Lambda, \tag{8}$$

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. Moreover, consider a λ-dependent version of our average treatment effect, given by

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

Table 1
Critical values for unknown intervention time inference: $\mathbb{P}(\|\boldsymbol{S}\|_p > c) = 1 - \alpha$.

	$\Lambda = [\underline{\lambda}, \bar{\lambda}]$	Confidence Level					
		$\alpha = 0.2$	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 $\boldsymbol{\Sigma}_{\Lambda}$ via a grid of 500 points between $\underline{\lambda}$ and $\bar{\lambda}$ inclusive.

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

Assumption 6. $\{(\boldsymbol{y}_t', \boldsymbol{x}_t')'\}$ is a strictly stationary process.

Assumption 6 is clearly stronger than necessary. For instance, it would be sufficient to have $\{\boldsymbol{v}_t\}$ as a weakly stationary process. However, to avoid assumptions that are model-dependent (via the choice of $\mathcal{M}$) we state Assumption 6 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.

To analyze the properties of the estimator (8), it is convenient to define the stochastic process $\{\boldsymbol{S}_T\}$ indexed by $\lambda \in \Lambda$ such that for each $\lambda \in \Lambda$, we have $\boldsymbol{S}_T(\lambda) \equiv \sqrt{T}\boldsymbol{\Gamma}_T^{-1/2}[\widehat{\boldsymbol{\Delta}}_T(\lambda) - \boldsymbol{\Delta}_T(\lambda)]$. Note that unlike the notation used in Lemma 1, we do not include the factors T_1/T and T_2/T inside the asymptotic variance term, and since all the results are under stationarity (Assumption 6), we replace $\boldsymbol{\Gamma}_{T_1}$ and $\boldsymbol{\Gamma}_{T_2}$ with 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 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)'$ follows directly from Theorem 1 combined with Assumption 6 and the Cramèr–Wold device. Furthermore, the next theorem shows that $\boldsymbol{S}_T$ converges uniformly in $\lambda \in \Lambda$.

Theorem 5. *Under the conditions of Lemma 1 and Assumption 6:*

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

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* $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$.

The second part of Theorem 5 gives us a direct approach to conduct inference in the case of an unknown intervention time. We can replace $\boldsymbol{\Gamma}_T$ with a consistent estimator $\widehat{\boldsymbol{\Gamma}}_T$ (for instance, the one discussed 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 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 4.3, where treatment occurs at multiple unknown intervention times.

We assume a constant intervention, such as

Assumption 7. $\boldsymbol{\delta}_t = \boldsymbol{\Delta}$, for $t = T_0, \ldots, T$, where $\boldsymbol{\Delta} \in \mathbb{R}^q$ is non-random.

Remark 2. Recall that Assumption 7 is not overly restrictive due to the flexibility provided by the transformation $h(.)$. The mean of $\boldsymbol{y}_t$ can represent the variance, covariances or any other moment of interest of the original $\boldsymbol{z}_{1t}$ variable.

Remark 3. Assumption 7 implies an instantaneous treatment effect (step function) at $t = T_0$. In most cases, however, we encounter a continuous intervention effect, possibly reaching a distinguishable new steady-state value. We can accommodate these cases by trimming this transitory part of the sample, provided we have enough data, and then apply the methodology to the trimmed sample where Assumption 7 holds.

Proposition 3. *Under the conditions of* Lemma 1 *and* Assumptions 6 and 7, $\widehat{\boldsymbol{\Delta}}_T(\lambda) \xrightarrow{p} \phi(\lambda)\boldsymbol{\Delta}$, *where*

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

Since both $\frac{1-\lambda_0}{1-\lambda}$ and $\frac{\lambda_0}{\lambda}$ are bounded between 0 and 1, we have that $\|\text{plim}\, \widehat{\boldsymbol{\Delta}}_T(\lambda)\|_p \leq \|\boldsymbol{\Delta}\|_p$ 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 $\text{plim}\, \widehat{\boldsymbol{\Delta}}_T(\lambda) = \boldsymbol{\Delta}$ if and only if $\lambda = \lambda_0$. This result suggests a natural estimator for λ_0:

$$\widehat{\lambda}_{0,p} = \arg\max_{\lambda \in \Lambda} J_{T,p}(\lambda) \quad \text{and} \quad J_{T,p}(\lambda) \equiv \|\widehat{\boldsymbol{\Delta}}_T(\lambda)\|_p. \tag{9}$$

Theorem 6. *Let* $p \in [1, \infty]$. *Under the conditions of* Lemma 1 *and* Assumptions 6 and 7, *for* $\boldsymbol{\Delta} \neq \mathbf{0}$, $\widehat{\lambda}_{0,p} = \lambda_0 + o_p(1)$. *If* $\boldsymbol{\Delta} = \mathbf{0}$, $\widehat{\lambda}_{0,p}$ *converges in probability to any* $\lambda \in \Lambda$ *with equal probability.*

4.3. Multiple intervention points

We can readily extend our analysis to the case of more than one intervention affecting the unit of interest as long as Assumption 7 is valid for each intervention. Suppose 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 intervention point $s = \{1, \dots, S\}$, we can define the time of each intervention by $T_s \equiv \lfloor \lambda_s T \rfloor$ and construct our estimator in the same way as for the single intervention case. To simplify the notation, we define the set of all periods after intervention s but before intervention $s+1$ as $\tau_s = \{T_s, T_s+1, \dots, T_{s+1}-1\}$ and define $\#\{A\}$ as the number of elements in the set A. Then, we have S estimators given by

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

where $\widehat{\boldsymbol{\theta}}_{s,T}$ is the LASSO estimator for the sample indexed 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 estimate is obtained by choosing the same model for all intervention periods.

Under the same set of assumptions as for the single intervention case plus Assumption 7, the sequence of estimators $\{\widehat{\boldsymbol{\Delta}}_T^s\}_{s=1}^S$ are consistent for their respective intervention effects $\{\boldsymbol{\Delta}^s\}_{s=1}^S$ and are also asymptotically normal. However, we need to make a minor adjustment to the asymptotic covariance matrix to reflect the intervention timing:

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

Since under Assumption 7 all the interventions are constant, the asymptotic variance $\boldsymbol{\Gamma}$ is the same across all intervention points. Therefore, we can apply the inference for each breaking point as described for the single intervention case.

On the other hand, if the intervention points are unknown, we first need to 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 3 that there exists an interval of size $\epsilon > 0$ around every intervention point such that

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

Nonetheless, in contrast to the single intervention scenario, in the case of multiple intervention points, we first need to estimate the number and respective locations 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 intervention. We can then compute $\widehat{\lambda}_1$ as in (9) and test the null using $\widehat{\boldsymbol{\Delta}}_T^0(\widehat{\lambda}_1)$. In the case that we can reject the null, we split the sample at $\widehat{\lambda}_1$ and repeat the procedure in each of the two subsamples. Each time we reject the null, we split the sample in $\widehat{\lambda}_s$ and proceed sequentially until we can no longer reject the null in any subsample.

The sequential procedure described above was advocated by Bai and Perron (1998). It is 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 final one. When there are multiple intervention points with the same magnitude, the method converges 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:

$$\mathcal{H}_0^{(s)} : \boldsymbol{\Delta} = \mathbf{0} \quad \text{for all } \lambda \in \left[\lambda_j, \lambda_{j+1}\right)_{j=0}^{s},$$

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

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.

4.4. Testing for the unknown treated unit/untreated peers

All the analyses conducted so far rely on the knowledge of which unit is the treated unit and, more importantly, on the assumption that the remaining units are in fact untreated during the sample period (Assumption 1). However, cases may occur where we are either unsure of or would like to test for those conditions. Given any *finite* subset $\mathcal{I}$ of available units, we would like to test the following hypothesis:

$$\mathcal{H}_0^n : \boldsymbol{\Delta}_T^{(i)} = \mathbf{0} \quad \forall i \in \mathcal{I} \subseteq \{1, \ldots, n\}$$

$$\mathcal{H}_1^n : \boldsymbol{\Delta}_T^{(i)} \neq \mathbf{0} \quad \text{for some } i \in \mathcal{I}.$$

Nothing prevents 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 (4) for $i = 1, \ldots, n_{\mathcal{I}}$, where $n_{\mathcal{I}} < \infty$ is the cardinality of the set $\mathcal{I}$. We can then stack all of them in a vector as $\widehat{\boldsymbol{\Pi}}_T(\mathcal{I}) \equiv \left(\widehat{\boldsymbol{\Delta}}_T^{(1)'} \ldots \widehat{\boldsymbol{\Delta}}_T^{(n_{\mathcal{I}})'}\right)'$ as an average estimator for the true average intervention effect vector $\boldsymbol{\Pi}_T(\mathcal{I}) \equiv \left(\boldsymbol{\Delta}_T^{(1)'} \ldots \boldsymbol{\Delta}_T^{((\mathcal{I}))'}\right)'$, where $\boldsymbol{\Delta}_T^{(i)}$ is defined for each unit. Hence,

Proposition 4. *Under the conditions of Lemma 1, for any* finite *subset* $\mathcal{I} \subseteq \{1, \ldots, n\}$

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

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

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

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

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

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

We can obtain a consistent estimator for $\boldsymbol{\Sigma}_{\mathcal{I}}$ by repeating the procedure described in Section 3.3 for each pair $(ij) \in \mathcal{I}^2$ to obtain $\widehat{\boldsymbol{\Omega}}^{ij}$ and finally construct the matrix $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}}$. Then, for a desired significance level, we can use W_T^{π} to test $\mathcal{H}_0^n$. The test becomes even more useful after the (likely) treated unit is removed and the test is repeated with the remaining units (peers). In the case where we fail to reject the null, we can interpret the result as direct evidence in favor 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. Contamination and other issues

In this section we investigate the consequences when Assumption 1 fails. We consider without loss of generality a simple DGP. Each unit $i = 1, \ldots, n$ under no intervention is represented by $z_{it}^{(0)} = l_i f_t + \eta_{it}$, where η_{it} is a zero-mean independent and identically distributed (iid) idiosyncratic shock with variance $\sigma_{\eta_i}^2$. Furthermore, $\mathbb{E}(\eta_{it}\eta_{jt}) = 0$, for all $i \neq j$. Additionally, the common factor vector f_t is an iid random variable with zero mean and variance σ_f^2.

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

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

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

The best linear projection model is given by $\mathbb{L}(\boldsymbol{y}_t|\boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0 = \left[\boldsymbol{l}_0\boldsymbol{l}_0' + \operatorname{diag}(\boldsymbol{r}_0)\right]^{-1}(l_1\boldsymbol{l}_0)$. Furthermore, $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_0 + \nu_t$, where $\mathbb{E}(\boldsymbol{x}_t\nu_t) = \mathbf{0}$ by definition, and $\sigma_\nu^2 \equiv \mathbb{E}(\nu_t^2) = \sigma_f^2\left(l_1^2 + r_1 - \boldsymbol{\beta}_0' l_1\boldsymbol{l}_0\right)$. Therefore, $\boldsymbol{\beta}_0 \equiv \boldsymbol{\beta}_0(\boldsymbol{l}, \boldsymbol{r})$ and $\sigma_\nu^2 \equiv \sigma_\nu^2(\boldsymbol{l}, \boldsymbol{r}, \sigma_f^2)$, where $\boldsymbol{r} = (r_1, \boldsymbol{r}_0')'$ and $\boldsymbol{l} = (l_1, \ldots, l_n)'$.

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

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

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

Under Assumption 1, we have that $\mathbf{c}_t^{(0)} = c_t^f = 0$, $\forall t$; thus, $\mathbb{E}(\delta_t) = c_t^1$ and, ignoring the sampling error of estimating $\boldsymbol{\beta}_0$, the ArCo estimator is unbiased for the average of c_t^1 for the post-intervention period. On the other hand, without these assumptions, we have the following bias in the normalized statistic

$$b_t \equiv \mathbb{E}\left(\frac{\delta_t - c_t^1}{\sigma_\nu}\right) = \underbrace{\left(\frac{l_1 - \boldsymbol{l}_0'\boldsymbol{\beta}_0}{\sigma_\nu}\right)}_{\equiv \phi_f} c_t^f - \frac{\mathbf{c}_t^{0\prime}\boldsymbol{\beta}_0}{\sigma_\nu} \qquad (10)$$

The factor in the first term of the bias $\phi_f = \phi_f(\boldsymbol{l}, \boldsymbol{r}, \sigma_f^2)$ is a non-linear expression that is difficult to express in closed form. However, regardless of the choice of the factor loads $\boldsymbol{l}$ and idiosyncratic shock variances $\boldsymbol{\sigma}_\eta^2 = (\sigma_{\eta_1}^2, \dots, \sigma_{\eta_n}^2)'$, as $\sigma_f^2 \to \infty$, $r \to 0$ 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 approaches zero quite fast, as seen in the left scale of Fig. 1. Additionally, $\phi_f = \phi(s_0)$ is decreasing in the number of relevant variables s_0 for fixed R^2.

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

The second bias term in (10) can be seen as a result, for instance, of a global shock that induces breaks in peers in a non-systematic way, which makes this source of bias difficult to handle. To gain a better understanding, consider the case where the idiosyncratic shock is a fixed 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 occurs, 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 Fig. 1.

Therefore, when one expects $\mathbf{c}_t^0 \neq \mathbf{0}$, 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). 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.[12] 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 4.4.

6. Monte Carlo simulation

We conduct size and power simulations to investigate the finite sample properties of the test as well as a "horse race" to compare the ArCo estimator with potential alternatives, namely, the SC, PF, DiD and BA estimators.

6.1. Size and power simulations

The DGP is a version of the model (S.1) with the following baseline scenario: $T = 100$, $n = 100$, $q = 1$, $\lambda_0 = 0.5$ (intervention at the middle of the sample), $s_0 = 5$ relevant parameters with loading factor equal to 1 and $f = 1$. The common factor and all idiosyncratic shocks are iid and 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, a t-Student distribution with 3 degrees of freedom for fat-tails and a mixed normal distribution for bimodality.[13] As shown in first panel of Table 2, little influence on the overall size of the test is perceived.

Next, we consider $T = \{25, 50, 75, 100\}$. The size distortions are small even with only 50 observations, as shown in the second panel of Table 2.

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

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

We also investigate the influence of increasing the number of covariates. We set $d = \{100, 200, 500, 1000\}$. The third panel of Table 2 shows that the test size appears to be unaffected by the increase in model complexity. This is not surprising since consistent model selection is not an issue for the methodology. We also study a change in the number of relevant covariates (units) in the model. We consider a case where all the regressors are irrelevant, which (asymptotically) reduces 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 $1/\sqrt{j}, j = 1, \dots, 100$. In all cases, the LASSO does 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 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 becomes more pronounced as the autoregressive coefficient (ρ) approaches 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. Specifically, it underestimates $\boldsymbol{\Omega}$. We test several alternatives for $\widehat{\boldsymbol{\Omega}}_T$, including Newey and West (1987), Andrews (1991), Andrews and Monahan (1992), and Haan and Levin (1996), and we obtain the best results using the procedure proposed in Andrews and Monahan (1992). It is worth noting that the slightly oversized tests are a direct consequence of the persistence of $\{\nu_t\}$ and not necessarily of the persistence of $\{(y_t, \boldsymbol{x}_t')\}$. The problem 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 processes, the usual kernel HAC estimators are known to perform well, and the tests are not oversized.

6.2. Estimator comparison

To conduct the "horse race" among competitors for the counterfactual analysis, we consider the following DGP:

$$\boldsymbol{z}_{it}^{(0)} = \rho \boldsymbol{A}_i \boldsymbol{z}_{it-1}^{(0)} + \boldsymbol{\varepsilon}_{it}, \quad i = 1, \dots, n, ; t = 1, \dots, T, \tag{11}$$

where $\boldsymbol{\varepsilon}_{it} = \boldsymbol{\Lambda}_i \boldsymbol{f}_t + \boldsymbol{\eta}_{it}, \boldsymbol{f}_t = [1, (t/T)^{\varphi}, \upsilon_t], \boldsymbol{z}_{it} \in \mathbb{R}^q, \rho \in [0, 1), \varphi > 0, \boldsymbol{A}_i(q \times q)$ is a diagonal matrix with diagonal elements strictly between -1 and 1, $\{\upsilon_t\}$ 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 obtained by stacking all the $\boldsymbol{z}_{it}^{(0)}$, and let $\boldsymbol{\Lambda}$ be the $(nq \times 3)$ matrix after stacking all the $\boldsymbol{\Lambda}_i$. Similarly, define $\boldsymbol{\varepsilon}_t$ by stacking $\boldsymbol{\varepsilon}_{it}$ and $\boldsymbol{A}$ as 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 $\text{vec}(\boldsymbol{\Sigma}) \equiv \text{vec}[(\mathbb{V}\boldsymbol{z}_t)] = [\boldsymbol{I}_{(nq)^2} - \rho^2 \boldsymbol{A} \otimes \boldsymbol{A}]^{-1} \text{vec}(\boldsymbol{\Omega})$.

We set $y_{it}^{(1)} = y_{it}^{(0)} + \delta_t 1\{t \geq T_0 \text{ and } i = 1\}$, and 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 it is convenient, we use the following partition scheme: $\boldsymbol{z}_{it} = (y_{it}, \boldsymbol{x}_{it}')'$ and $\boldsymbol{z}_{0t} = (\boldsymbol{z}_{2t}', \dots \boldsymbol{z}_{nt}')$.

Before-and-after (BA)

The BA estimator is defined as:

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

Differences-in-differences (DiD)

The OLS estimator of the dummy coefficient in the following regression models. For the case with covariates,

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

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

Synthetic control (SC)

We use the *Synth* package.[14] 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 variable.

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

where $\widehat{y}_{1t} = \boldsymbol{w}^{*\prime}\boldsymbol{y}_{0t}$. The weight vector $\boldsymbol{w}$ must contain 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 is composed of the variables of each of the remaining $n-1$ units

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

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

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

$$\boldsymbol{V}^* = \text{arg min} \ \frac{1}{T_0-1} \sum_{t=1}^{T0-1} \left[y_{1t} - \boldsymbol{w}^*(\boldsymbol{V})'\boldsymbol{y}_{0t}\right]^2, \tag{12}$$

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 S.1–S.6. Overall, the SC and the GM are heavily biased in most of the considered cases. For the former, this might be a consequence of the instability of the algorithm to find the minimizer of (12) since the bias persists even in the absence of time trends, where any fixed linear combination of peers should give us an unbiased estimator. For the latter, it is most likely a consequence of the poor finite sample properties of the 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 appear to have comparable small bias, at least in the absence of deterministic trends, regardless of the presence of serial correlation. The ArCo seems to have better MSE performance, which is not surprising since by definition our estimator in the first stage searches for the linear combination that minimizes the MSE.

For the cases with trends, the BA estimator is severely biased since, without the information from the peers, it cannot account for the trend effect. For the common trend case, the DiD estimator have relatively small bias. Again, the ArCo estimators have comparable bias to the DiD estimators for the common trend cases but with significantly smaller variance (ranging from 6 to 16 times smaller). The clear advantage of ArCo estimation can be seen in the idiosyncratic time trend cases. Even though some small bias appears, it is clearly much smaller than that of all the other alternatives.

7. The effects of an anti tax evasion program on inflation

We apply the ArCo methodology to estimate the effects of an anti tax-evasion program in Brazil on inflation, economic growth, retail sales and credit. Although, the causes of business non-compliance and tax evasion have been extensively studied in the literature (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. In October 2007, the state government of São Paulo, Brazil, implemented an anti tax-evasion scheme called *Nota Fiscal Paulista* (NFP) program. The 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 program works as an incentive to the consumer to ask for electronic sales receipts, which give the consumer the right to participate in monthly lotteries promoted by the government. According to the rules of the program, registered consumers have also the right to receive part of the ICMS paid by the seller as tax rebate when their tax identifier 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 (Wan, 2010). The effectiveness of such programs has been discussed in Fatas et al. (2015) and Brockmann et al. (2016). In São Paulo, the 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 rapidly from 44 thousand Reais in January 2008 to an average of 70 million Reais distributed monthly by the end of the same year; see Fig. 2.

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 São Paulo, the author showed that one year after the launching of the NFP program, the accumulated

[14] R package maintained by Jens Hainmueller.

inflation on food away from home (FAH) was 5% higher in 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 and we also test for effects on other macro variables. We also consider the BA, GM, and DiD estimators.

7.1. Effects on inflation

Under the assumptions that (i) a certain degree of tax evasion was occurring before the intervention, (ii) the sellers have some degree of market power and (iii) the penalty for tax evasion is large enough to alter the seller behavior, one is expected to see an upward movement in prices due to an increase in marginal cost. We would like to investigate whether the NFP had an impact on consumer prices. 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. To highlight the potential of the ArCo methodology to test for joint effects, we also run the multivariate version of the test including GDP growth, retail sales growth and credit growth.

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 away from home (FAH) as a good indicator for price levels in those sectors. The sample then consists of monthly FAH index for 10 metropolitan areas[15] 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 (FAH), 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) criterion. The choice of the HQ instead of the BIC, for example, is driven by the fact that the latter delivers conditional models with no variables in most of the cases. The in-sample period (pre-intervention) consists of 33 months while the size of the out-of-sample period is 23.

The factors in the GM methodology are computed by principal components. The number of factors is determined as to explain 80% of the total variance in the data.

The results are depicted in Table 5. The upper panel reports, for different choices of conditioning variables, the estimated average effect 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 average effect 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 average effect is 0.4478%. The aggregate effect during the out-of-sample period is 10.72%. In the second case, column (6) of Table 5, the monthly average effect is 0.3796% and the aggregate effect is 9.04%. Two facts worth discussing. 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 average effect 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 good.

Figs. 3 and S.7 show the actual and counterfactual data, both in-sample and out-of-sample. Fig. 3 considers the case where only inflation and GDP growth are considered as conditioning variables while the plots in Figure S.7 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 average effects. 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 S.1. For this reason we re-run the analysis without Porto Alegre in the donor pool. The results are reported in Table S.2. The overall picture seems unchanged.

7.2. Effects on GDP, retail sales and credit

In order to illustrate the multivariate nature of the ArCo methodology we also test for effects of the NFP program on GDP growth, retail sales growth and credit growth. Based on the results of the previous section, we remove Porto Alegre from the sample of donors. The results are shown in Table 6. The table reports the individual effects for all the seven different set of regressors as well as the joint Wald-type test for the null of no-effects on any of the variables jointly. As can be seen from the table, the program has significant effects only on inflation.

[15] 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.

8. Conclusions and future research

We proposed a flexible method to conduct counterfactual analysis with aggregate data which is specially relevant in situations where there is a single treated unit and "controls" are not available, such as in regional policy evaluation. The ArCo methodology is easy to implement and extends and generalize previous proposals in the literature in several aspects: (1) the distribution of test for no-intervention effect is standard and asymptotically honest confidence regions for the average intervention effect can be constructed; (2) although the results rely on the number of time-series observations diverging, the LASSO estimator has good finite sample properties, even when the number of estimated parameters are much larger than the sample size; (3) we allow for nonlinear, heterogeneous confounding effects; (4) we provide a complete asymptotic theory which can be used to jointly test for intervention effects on a group of variables; (5) the methodology can be applied even if the time of the intervention is not known; (6) multiple interventions can be handled; and (7) 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 cointegration or not). A non-parametric or semiparametric estimation in the pre-intervention model can be also considered.

Acknowledgments

We are extremely grateful to the co-editor (Oliver Linton), the associate editor and two anonymous referees for very constructive comments and for editorial guidance. We also wish to thank the participants at the 2014 NBER-NSF Time Series Conference, the 2014 Meeting of the Brazilian Econometric Society (SBE), the 2014 Workshop on Measuring and Modeling Financial Risk with High Frequency Data, the 2015 World Congress of the Econometric Society, the 2015 conference of the European Economic Association, the Workshop Rio-São Paulo of Econometrics, the (EC)^2 in Toulouse in 2016, and the Norges Bank conference: Big data, machine learning and the macro economy. We are also thankful to the participants at seminars at Emory University, Ohio State University, Erasmus University and at the PUC-Rio Student Workshop. We are specially 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, Leonardo Rezende, and Anders B. Kock for great comments and discussions. Finally, we are specially thankful to Eduardo Mendes for hours of rich discussions as well as Fernando Friaą de Souza for discussions on the application of the method to the "Nota Fiscal Paulista". Any mistakes are our own.

Appendix A. Proofs

Model choice and a general result

The theoretical results in the paper are based on a linear model to construct the counterfactual. However, we provide a general lemma to show that any statistical model satisfying a set of assumptions will deliver consistent and asymptotically normal estimates for the average intervention effect. For $t \geq T_0$, let $\widehat{\mathcal{M}}_{t,T_1} \equiv \mathcal{M}(\boldsymbol{Z}_{0t}, \widehat{\boldsymbol{\theta}}_{T_1})$, be an estimator of $\mathcal{M}_t \equiv \mathcal{M}(\boldsymbol{Z}_{0t}, \boldsymbol{\theta}_0)$ using only the first $T_1 = T_0 - 1$ observations to obtain $\widehat{\boldsymbol{\theta}}_{T_1}$ as an estimator of $\boldsymbol{\theta}_0$. Define $\boldsymbol{\eta}_{t,T_1} \equiv \widehat{\mathcal{M}}_{t,T_1} - \mathcal{M}_t$, $t \geq T_0$ and $\boldsymbol{\nu}_t = \boldsymbol{y}_t - \mathcal{M}_t$, then we state:

Lemma 1. *Assume that, uniformly in $P \in \mathcal{P}$ (an arbitrary class of probability laws):*

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

(b) $\frac{1}{\sqrt{T_1}}\boldsymbol{\Gamma}_{T_1}^{-1/2}\sum_{t\leq T_1}\boldsymbol{\nu}_t \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q)$, *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) $\frac{1}{\sqrt{T_2}}\boldsymbol{\Gamma}_{T_2}^{-1/2}\sum_{t\geq T_0}\boldsymbol{\nu}_t \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q)$, *where* $\boldsymbol{\Gamma}_{T_2} = \mathbb{E}_P\left[\frac{1}{T_2}(\sum_{t\geq T_0}\boldsymbol{\nu}_t)(\sum_{t\geq T_0}\boldsymbol{\nu}_t')\right]$.

Under Assumption 1 and conditions (a)–(c), uniformly in $P \in \mathcal{P}$, it holds that:

$$\sqrt{T}\boldsymbol{\Omega}_T^{-1/2}\left(\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_q\right),$$

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) ensures that the estimation error to be asymptotic negligible, ensuring the $\sqrt{T}$ rate of convergence of the estimator. Under (a) we can write:

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

Conditions (b) and (c) ensure the asymptotic normality of the terms above after appropriate normalization. Finally, given a model for the first stage of the ArCo method, one must verify that conditions (a)–(c) in the lemma hold in order to prove the converge results for $\widehat{\boldsymbol{\Delta}}_T$.

Proof. See supplementary material. □

Auxiliary lemmas

In 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 2 *(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. See supplementary material. □

Lemma 3 *(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. *See supplementary material.* □

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

Lemma 4. *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 $r > 2$, $\sup_t \mathbb{E}|u_t|^{r+\delta} < \infty$ 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 5. *On $\mathscr{A}(a) \cap \mathscr{B}(b)$, provided that $\varsigma \geq 2a$, $b \leq \frac{\psi_0^2}{32 s_0}$, and the compatibility constraint is satisfied for $\boldsymbol{\Sigma}$ with constant $\psi_0 > 0$, we have that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\widehat{\boldsymbol{\Sigma}}}^2 + \varsigma\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \leq 8\varsigma^2 \frac{s_0}{\psi_0^2}$, where $\boldsymbol{\Sigma} \equiv \mathbb{E}(\widehat{\boldsymbol{\Sigma}})$, $\widehat{\boldsymbol{\Sigma}} \equiv \frac{1}{T_1}\sum_{t=1}^{T_1} \boldsymbol{x}_t\boldsymbol{x}_t'$, and $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\widehat{\boldsymbol{\Sigma}}}^2 \equiv (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. For real constants $a, b > 0$, $\mathscr{A}(a) \cap \mathscr{B}(b)$ are defined as*

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

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

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

Proof. See supplementary material. □

Lemma 6. *Under Assumptions 2–4, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 = O_P\left(s_0 \frac{d^{1/\gamma}}{\sqrt{T}}\right)$.*

Proof. See supplementary material. □

Lemma 7. *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 of each entry $(u_{i,t})$ fulfills the conditions of Lemma 4 for some real $r > 2$ for all $i \in \{1, \ldots, d\}$. Then, $\|\boldsymbol{S}_T\|_{\max} = O_P(d^{1/r}\sqrt{T})$.*

Proof. See supplementary material. □

Proof of Theorem 1

Proof. First we verify condition (a) of Lemma 1. In the linear case we have $\boldsymbol{\eta}_{t,T_1} = [\boldsymbol{x}'_{1t}(\widehat{\boldsymbol{\theta}}_{1,T_1} - \boldsymbol{\theta}_{1,0}), \ldots, \boldsymbol{x}'_{qt}(\widehat{\boldsymbol{\theta}}_{q,T_1} - \boldsymbol{\theta}_{q,0})]'$ for $t \geq T_0$. Since all the q components of $\boldsymbol{\eta}_{t,T_1}$ have the same linear form and q is fixed (not growing with T) it is enough to show that condition (a) holds for an arbitrary index $j \in \{1, \ldots q\}$, which is omitted in what follows for clarity. 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, $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = \frac{1}{T_1}\sum_{t\leq T_1} \boldsymbol{v}_t - \frac{1}{T_1}\sum_{t\leq T_1} \widetilde{\boldsymbol{x}}_t \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$. Combining the last two expressions we can write

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

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

$$\sqrt{T}\left(\frac{1}{T_2}\sum_{t\geq T_0} \boldsymbol{\eta}_{t,T} - \frac{1}{T_1}\sum_{t\leq T_1} \boldsymbol{v}_t\right) = \left(\frac{\sqrt{T}}{T_2}\sum_{t\geq T_0} \widetilde{\boldsymbol{x}}_t - \frac{\sqrt{T}}{T_1}\sum_{t\leq T_1} \widetilde{\boldsymbol{x}}_t\right)' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right). \tag{13}$$

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 7 combined with Assumption 3(a)–(b). The second term is $O_P\left(s_0\frac{d^{1/\gamma}}{\sqrt{T}}\right)$ by Lemma 6. Hence, the last term 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 Lemma 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 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 Lemma 1 are verified and the result follows directly from Lemma 1. □

Proof of Theorem 2

Proof. The consistency proof is practically the same for both $\widehat{\boldsymbol{\Gamma}}_{T_1}$ and $\widehat{\boldsymbol{\Gamma}}_{T_2}$, so we focus on the first one and drop the subscript $i = 1$ for notation convenience. Let $\widetilde{\boldsymbol{\Gamma}}_T = \sum_{|k|<T}\phi\left(\frac{k}{S_T}\right)\widetilde{\boldsymbol{V}}_k$, where $\widetilde{\boldsymbol{V}}_k$ is the same as $\widehat{\boldsymbol{V}}_k$ but with $\widehat{\boldsymbol{v}}_t$ replaced by the (unobservable) $\boldsymbol{v}_t$. Thus $\widehat{\boldsymbol{V}}_k - \widetilde{\boldsymbol{V}}_k$ is a $q \times q$ matrix with typical element given by

$$\left(\widehat{\boldsymbol{V}}_k - \widetilde{\boldsymbol{V}}_k\right)_{ij} = \boldsymbol{\eta}'_i\left(\frac{1}{T}\sum_{t>k}^{T} \boldsymbol{p}_t^{(i,j,k)}\right) + \boldsymbol{\eta}'_j\left(\frac{1}{T}\sum_{t>k}^{T} \boldsymbol{p}_t^{(j,i,k)}\right) + \boldsymbol{\eta}'_i\left(\frac{1}{T}\sum_{t>k}^{T} \boldsymbol{M}_t^{(i,j,k)}\right)\boldsymbol{\eta}_j$$
$$+ \tfrac{T-k}{T}\left[\boldsymbol{\eta}'_i\mathbb{E}\nu_{i,t}\boldsymbol{x}_{j,t-k} + \boldsymbol{\eta}'_j\mathbb{E}\nu_{j,t-k}\boldsymbol{x}_{i,t} + \boldsymbol{\eta}'_i\mathbb{E}(\boldsymbol{x}_{i,t}\boldsymbol{x}'_{j,t-k})\boldsymbol{\eta}_j\right],$$

where to further simply notation we define

$$\boldsymbol{\eta}_i = (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{0,i}), \quad \boldsymbol{p}_t^{(i,j,k)} = \nu_{i,t}\boldsymbol{x}_{j,t-k} - \mathbb{E}(\nu_{i,t}\boldsymbol{x}_{j,t-k}), \quad \text{and} \quad \boldsymbol{M}_t^{(i,j,k)} = \boldsymbol{x}_{i,t}\boldsymbol{x}'_{j,t-k} - \mathbb{E}(\boldsymbol{x}_{i,t}\boldsymbol{x}'_{j,t-k}).$$

We now show that each of the terms in the right hand side is $o_P(1)$. First, by Assumption 3(b) and Cauchy–Schwarz inequality we have for some $\delta > 0$,

$$\mathbb{E}|p_{l,t}^{(i,j,k)}|^{\gamma+\delta/2} \leq c_\gamma \quad \text{and} \quad \mathbb{E}|M_{l,m,t}^{(i,j,k)}|^{\gamma+\delta/2} \leq c_\gamma,$$

where $1 \leq l \leq d$ and $1 \leq l, m \leq d$ index the elements of $\boldsymbol{p}_t^{(i,j,k)}$ and $\boldsymbol{M}_t^{(i,j,k)}$ respectively. Since this bound is uniform across all equations $1 \leq i, j \leq q$, for all lags $0 \leq k \leq T-1$, for all elements $1 \leq l, m \leq d_T$ and for all $T \geq 1$ we drop the superscript (i, j, k) for clarity.

Note that $\left\|\sum_{t>k}^{T}\boldsymbol{p}_t\right\|_\infty = O_P(\sqrt{T}d^{1/\gamma})$ since by the union bound followed by the Markov's inequality and Lemma 3

$$\mathbb{P}\left(\left\|\sum_{t>k}^{T}\boldsymbol{p}_t\right\|_\infty > \epsilon\sqrt{T}d^{1/\gamma}\right) \leq d\max_{1\leq l\leq d}\mathbb{P}\left(\sum_{t>k}^{T} p_{l,t} > \epsilon\sqrt{T}d^{1/\gamma}\right)$$
$$\leq \frac{d}{dT^{\gamma/2}\epsilon^\gamma}\max_{1\leq l\leq d}\mathbb{E}\left|\sum_{t>k}^{T} p_{l,t}\right|^\gamma \leq \frac{1}{T^{\gamma/2}\epsilon^\gamma}C_\gamma(T-k)^{\gamma/2} \leq \frac{C_\gamma}{\epsilon^\gamma},$$

where C_γ denote a generic constant only depending on γ of Assumption 3.

Following the same procedure we can shown that $\left\| \sum_{t>k}^{T} \boldsymbol{M}_t \right\|_{\max} = O_P(\sqrt{T} d^{2/\gamma})$ because

$$\mathbb{P}\left(\left\| \sum_{t>k}^{T} \boldsymbol{M}_t \right\|_{\max} > \epsilon \sqrt{T} d^{2/\gamma} \right) \leq \frac{d^2}{d^2 T^{\gamma/2} \epsilon^{\gamma}} \max_{1 \leq l,m \leq d} \mathbb{E} \left| \sum_{t>k}^{T} M_{l,m,t} \right|^{\gamma} \leq \frac{C_{\gamma}}{\epsilon^{\gamma}}$$

Now, the first and second terms of $(\widehat{\boldsymbol{V}}_k - \widetilde{\boldsymbol{V}}_k)_{ij}$ expression can be bounded using Holder's inequality by $\frac{1}{T} \|\boldsymbol{\eta}_i\|_1 \| \sum_{t>k}^{T} \boldsymbol{p}_t \|_{\infty}$, where $\|\boldsymbol{\eta}_i\|_1 = O_P(s_0 \frac{d^{1/\gamma}}{\sqrt{T}})$ by Lemma 6 and $\| \sum_{t>k}^{T} \boldsymbol{p}_t \|_{\infty} = O_P(\sqrt{T} d^{1/\gamma})$, hence the first and second terms are $O_P(s_0 \frac{d^{2/\gamma}}{T})$. The third term can be bounded by $\frac{1}{T} \left\| \boldsymbol{\eta}_i \right\|_1 \left\| \sum_{t>k}^{T} \boldsymbol{M}_t \right\|_{\max} \left\| \boldsymbol{\eta}_j \right\|_1$. Once again, the terms at the ends are $O_P(s_0 \frac{d^{1/\gamma}}{\sqrt{T}})$ and the term in between is $O_P(\sqrt{T} d^{2/\gamma})$, thus the third is $O_P(s_0^2 \frac{d^{4/\gamma}}{T^{3/2}})$.

The forth and fifth terms of $(\widehat{\boldsymbol{V}}_k - \widetilde{\boldsymbol{V}}_k)_{ij}$ expression can be bounded by $\|\boldsymbol{\eta}_i\|_1 \|\mathbb{E}(\nu_{i,t} \boldsymbol{x}_{j,t-k})\|_{\infty}$ using Holder's inequality. The last term is bounded by $\left\| \boldsymbol{\eta}_i \right\|_1 \left\| \mathbb{E}(\boldsymbol{x}_{i,t} \boldsymbol{x}_{j,t-k}') \right\|_{\max} \left\| \boldsymbol{\eta}_j \right\|_1$ Since both $\mathbb{E}(\nu_{i,t} \boldsymbol{x}_{j,t-k})$ and $\mathbb{E}(\boldsymbol{x}_{i,t} \boldsymbol{x}_{j,t-k}')$ are uniformly bounded by Assumption 3(b),[16] the forth and fifth terms are $O_P(s_0 \frac{d^{1/\gamma}}{\sqrt{T}})$ and the last term is $O_P(s_0^2 \frac{d^{2/\gamma}}{T})$.

Careful review of the bounds above show they are independent of k, hence we can state that uniformly in $k \geq 0$

$$(\widehat{\boldsymbol{V}}_k - \widetilde{\boldsymbol{V}}_k)_{ij} = O_P\left(\max\left\{ \frac{s_0 d^{2/\gamma}}{T}, \frac{s_0^2 d^{4/\gamma}}{T^{3/2}}, \frac{s_0 d^{1/\gamma}}{\sqrt{T}}, \frac{s_0^2 d^{2/\gamma}}{T} \right\} \right), \quad 1 \leq i,j \leq q$$

Under Assumption 4(b), the maximum above is dominated by the term $\frac{s_0 d^{1/\gamma}}{\sqrt{T}}$ asymptotically. Also, since the number of equation q is fixed it is equivalent to say that converge holds for an arbitrary norm $\sup_{k \geq 0} \|\widehat{\boldsymbol{V}}_k - \widetilde{\boldsymbol{V}}_k\| = O_P\left(\frac{s_0 d^{1/\gamma}}{\sqrt{T}} \right)$.

Now the argument runs parallel to the proof of Theorem 1(a) in Andrews (1991). From the triangle inequality we have $\|\widehat{\boldsymbol{\Gamma}}_T - \boldsymbol{\Gamma}_T\| \leq \|\widehat{\boldsymbol{\Gamma}}_T - \widetilde{\boldsymbol{\Gamma}}_T\| + \|\widetilde{\boldsymbol{\Gamma}}_T - \boldsymbol{\Gamma}_T\|$. For the first term, using the last result we have

$$\frac{\sqrt{T}}{s_0 d^{1/\gamma} S_T} (\widehat{\boldsymbol{\Gamma}}_T - \widetilde{\boldsymbol{\Gamma}}_T) = \frac{1}{S_T} \sum_{|k|<T} \phi(k/S) \frac{\sqrt{T}}{s_0 d^{1/\gamma}} \left(\widehat{\boldsymbol{V}}_k - \widetilde{\boldsymbol{V}}_k \right) = O_P(1),$$

where we use the uniform (in k) boundedness in probability derived above and the fact that $\frac{1}{S_T} \sum_{|k|<T} |\phi(k/S_T)| \to \int |\phi(u)| du$, which is finite by condition (a) of Theorem 2.

For the second term we have that Assumption 3(a) and (b) implies assumption A of Andrews (1991). Hence, under condition (a), $S_T \to \infty$ and $S_T/T \to 0$ (which are implied by condition (b) of Theorem 2) we have that $\|\widetilde{\boldsymbol{\Gamma}}_T - \boldsymbol{\Gamma}_T\| = o_P(1)$. Therefore, $\|\widehat{\boldsymbol{\Gamma}}_T - \boldsymbol{\Gamma}_T\| = O_P\left(\frac{s_0 d^{1/\gamma} S_T}{\sqrt{T}} \right)$.

This completes the proof by choosing the bandwidth parameters as per condition (b) of Theorem 2. Notice that we recover the low dimension (fixed s_0 and d) result of Andrews (1991) where it is required that $S_T^2/T = o(1)$. □

Proof of Propositions 1 and 2

Proof. Both follows directly from Theorem 1 combined with Lemma 3(c) □

Proof of Theorem 3

Proof. Combine expression (S.3) with (13) prior to the $\sqrt{T}$ multiplication we are left with

$$\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T = \tfrac{1}{T_2} \sum_{t \geq T_0} \boldsymbol{\nu}_t - \tfrac{1}{T_1} \sum_{t \leq T_1} \boldsymbol{\nu}_t - \left(\tfrac{1}{T_2} \sum_{t \geq T_0} \widetilde{\boldsymbol{x}}_t - \tfrac{1}{T_1} \sum_{t \leq T_1} \widetilde{\boldsymbol{x}}_t \right)' (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{14}$$

the first two terms are $O_p(1/\sqrt{T})$, the last term can be rewritten as

$$\left(\tfrac{1}{T_2} \sum_{t \geq T_0} \check{\boldsymbol{x}}_t - \tfrac{1}{T_1} \sum_{t \leq T_1} \check{\boldsymbol{x}}_t \right)' (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \left(\tfrac{1}{T_2} \sum_{t \geq T_0} \mathbb{E}\widetilde{\boldsymbol{x}}_t - \tfrac{1}{T_1} \sum_{t \leq T_1} \mathbb{E}\widetilde{\boldsymbol{x}}_t \right)' (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where $\check{\boldsymbol{x}}_t \equiv \widetilde{x}_t - \mathbb{E}\widetilde{\boldsymbol{x}}_t$.

The first term of the last display is $O_P(s_0 d^{2/\gamma}/T)$ for the same argument used in the proof of Theorem 1. The difference in second term (which is zero under stationarity) is only due to the deterministic trend $\boldsymbol{\zeta}(t/T)$ that would make $\mathbb{E}\widetilde{\boldsymbol{x}}_t$ differ

[16] $\mathbb{E}(\nu_{i,t} \boldsymbol{x}_{j,t-k}) = 0$ and for $i = j$ and $k = 0$ by assumption.

across t, then we can rewrite it as

$$b_T \equiv \left(\tfrac{1}{T_2} \sum_{t \geq T_0} \zeta(t/T) - \tfrac{1}{T_1} \sum_{t \leq T_1} \zeta(t/T) \right)' (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \equiv \boldsymbol{r}_T(\boldsymbol{\zeta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

We can then bound b_T by $\|r_T(\boldsymbol{\zeta})\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1$ using Holder's inequality. By Lemma 6 we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = O_P(s_0 d^{1/\gamma}/\sqrt{T})$ and $\|r_T(\boldsymbol{\zeta})\|_\infty$ is uniformly bounded by Assumption 5(a), which completes the proof of result (a). For (b) we can multiply (14) by $\sqrt{T}$ and use the definition of b_T to write

$$\sqrt{T}(\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T + b_T) = \tfrac{\sqrt{T}}{T_2} \sum_{t \geq T_0} \boldsymbol{\nu}_t - \tfrac{\sqrt{T}}{T_1} \sum_{t \leq T_1} \boldsymbol{\nu}_t + O_P(s_0 \tfrac{d^{2/\gamma}}{\sqrt{T}}) \tag{15}$$

where the last term is $o_P(1)$ under Assumption 4(b) and the result follows. □

Proof of Theorem 4

Set $\boldsymbol{\Upsilon}_T \equiv \frac{\boldsymbol{\Lambda}_{T_1}}{T_1/T} + \frac{\boldsymbol{\Gamma}_{T_2}}{T_2/T}$, where $\boldsymbol{\Gamma}_{T_2}$ is defined in Theorem 1, $\boldsymbol{\Lambda}_{T_1} = \boldsymbol{a}'\boldsymbol{M}_1\boldsymbol{a}$, $\boldsymbol{a} = (1, \boldsymbol{r}_T(\boldsymbol{\zeta})'\boldsymbol{M}_2^{-1})'$, $\boldsymbol{M}_1 = \mathbb{E}_P\left[\frac{1}{T_1}(\sum_{t \leq T_1} \check{\boldsymbol{z}}_{0t}\nu_t)(\sum_{t \leq T_1} \check{\boldsymbol{z}}_{0t}'\nu_t)\right]$, $\boldsymbol{M}_2 = \mathbb{E}_P\left[\frac{1}{T_1}(\sum_{t \leq T_1} \widetilde{\boldsymbol{z}}_{0t}\widetilde{\boldsymbol{z}}_{0t}')\right]$, $\boldsymbol{r}_T(\boldsymbol{\zeta})$ is defined in Theorem 3, $\widetilde{\boldsymbol{z}}_{0t} = \boldsymbol{z}_{0t} - \mathbb{E}_P\boldsymbol{z}_{0t}$, $\check{\boldsymbol{z}}_{0t} = (1, \widetilde{\boldsymbol{z}}_{0t}')'$. Recall that $\boldsymbol{x}_t = (1, \boldsymbol{z}_{0t}')'$ and $\boldsymbol{z}_{0t} = (z_{2n}, \ldots, z_{nt})'$.

Proof. We take from (15) to write

$$\sqrt{T}(\widehat{\Delta}_T - \Delta_T) = \tfrac{\sqrt{T}}{T_2} \sum_{t \geq T_0} \nu_t - \tfrac{\sqrt{T}}{T_1} \sum_{t \leq T_1} \nu_t - r_T(\boldsymbol{\zeta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_P(1)$$

and use the fact that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \widehat{\boldsymbol{M}}_2^{-1}\widehat{\boldsymbol{M}}_1$ where $\widehat{\boldsymbol{M}}_2 = \frac{1}{T_1}\sum_{t \leq T_1} \boldsymbol{z}_{0t}^* \boldsymbol{z}_{\boldsymbol{0t}}^{*\,\prime}$, $\widehat{\boldsymbol{M}}_1 = \frac{1}{T_1}\sum_{t \leq T_1} \boldsymbol{z}_{0t}^* \nu_t$ and $\boldsymbol{z}_{0t}^* = \boldsymbol{z}_{0t} - \frac{1}{T_1}\sum_{t \leq T_1} \boldsymbol{z}_{0t}$ for $t = 1, \ldots, T_1$, to rewrite as

$$\sqrt{T}(\widehat{\Delta}_T - \Delta_T) = \tfrac{\sqrt{T}}{T_2} \sum_{t \geq T_0} \nu_t - \tfrac{\sqrt{T}}{T_1} \sum_{t \leq T_1} \nu_t - \boldsymbol{r}_T(\boldsymbol{\zeta})'\widehat{\boldsymbol{M}}_2^{-1} \tfrac{\sqrt{T}}{T_1} \sum_{t \leq T_1} \boldsymbol{z}_{0t}^* \nu_t + o_P(1)$$

Let $\dot{\boldsymbol{z}}_{0t} = (1, \boldsymbol{z}_{0t}^{*\,\prime})'$ and $\widehat{\boldsymbol{a}} = (1, \boldsymbol{r}_T(\boldsymbol{\zeta})'\widehat{\boldsymbol{M}}_2^{-1})'$, then the expression above becomes

$$\sqrt{T}(\widehat{\Delta}_T - \Delta_T) = \sqrt{\tfrac{T}{T_2}} \left(\tfrac{1}{\sqrt{T_2}} \sum_{t \geq T_0} \nu_t \right) - \sqrt{\tfrac{T}{T_1}} \widehat{\boldsymbol{a}}' \left(\tfrac{1}{\sqrt{T_1}} \sum_{t \leq T_1} \dot{\boldsymbol{z}}_{0t} \nu_t \right) + o_P(1)$$

now $\widehat{\boldsymbol{M}}_2 = \boldsymbol{M}_2 + o_P(1)$ by the law of large numbers and $\frac{1}{\sqrt{T_1}}\sum_{t \leq T_1} (\check{\boldsymbol{z}}_{0t} - \dot{\boldsymbol{z}}_{0t})\nu_t = o_P(1)$ then by the continuous mapping theorem we have

$$\sqrt{T}(\widehat{\Delta}_T - \Delta_T) = \sqrt{\tfrac{T}{T_2}} \left(\tfrac{1}{\sqrt{T_2}} \sum_{t \geq T_0} \nu_t \right) - \sqrt{\tfrac{T}{T_1}} \boldsymbol{a}' \left(\tfrac{1}{\sqrt{T_1}} \sum_{t \leq T_1} \check{\boldsymbol{z}}_{0t} \nu_t \right) + o_P(1).$$

The result then follows by the Central Limit Theorem and the fact that the two summations are over non-overlapping intervals as per part (b) and (c) of Lemma 1. □

Proof of Theorem 5

Proof. From (S.3) in the Proof of Lemma 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 Lemma 1, Assumption 6 and the fact that Λ is compact.

For fix $\lambda \in \Lambda$, the pointwise convergence in distribution follows under the conditions of Lemma 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

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

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
	Deterministic 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: (S.1) 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 first 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).

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

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

Proof of Proposition 3

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 6: Let $\lambda \le \lambda_0$:

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

Table 3
Rejection rates under the alternative (Test power).

	$\alpha = 0.1$	0.075	0.05	0.025	0.01
	Step Intervention[a] $\delta_t = c\,\sigma_1 1\{t \geq T_0\}$				
$c = 0.15$	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 (S.1) with the parameters in the baseline scenario as described in the footnote of Table 2.

[a] 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[a]	DiD	GM[a]	GM	ArCo[a]	ArCo
	No Time Trend ($\varphi = 0$) and No Serial Correlation ($\rho = 0$)							
Bias[b]	−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 (11); $T = 100$ 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 $\mathrm{diag}(\boldsymbol{A})$ 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.

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

[b] 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 away from home (FAH) 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 intervention effect 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 average intervention effect, 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 GDP growth, retail sales growth and credit growth: The Case without Porto Alegre.

ArCo Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Inflation	0.2992 (0.1704)	0.4438 (0.1486)	0.4913 (0.1432)	0.5064 (0.1480)	0.4763 (0.2010)	0.4070 (0.1600)	0.4046 (0.1539)
GDP	−0.0020 (0.0043)	−0.0002 (0.0032)	−0.0016 (0.0034)	−0.0024 (0.0034)	−0.0028 (0.0043)	0.0006 (0.0039)	−0.0001 (0.0036)
Retail	0.0020 (0.0040)	0.0016 (0.0045)	0.0020 (0.0041)	0.0012 (0.0039)	0.0027 (0.0055)	−0.0001 (0.0049)	0.0003 (0.0056)
Credit	0.0018 (0.0027)	0.0024 (0.0026)	0.0018 (0.0027)	−0.0008 (0.0017)	0.0031 (0.0027)	0.0029 (0.0027)	0.0003 (0.0017)
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							
Inflation	0.6439	0.1213	0.3928	0.1026	0.7960	0.8568	0.8072
GDP	0.2943	0.2022	0.0782	0.1017	0.7488	0.6482	0.5968
Retail Sales	0.1152	0.4180	0.5664	0.0374	0.7106	0.7695	0.8406
Credit	0	0.1132	0	0.6014	0.2085	0.1719	0.7543
Number of relevant regressors							
	9	3	7	5	14	17	13
	8	4	1	2	17	17	15
	3	6	8	1	13	13	16
	0	1	0	6	5	4	14
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
Joint $\chi_2(4)$ (p-value)	0.0900	0.0356	0.0090	0.0073	0.0028	0.0235	0.0252

The table reports the estimated average intervention effect 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. Finally, the last row of the table reports the p-value of the χ_2 statistic for the test of the null of no-effects on any of the four macroeconomic variables considered.

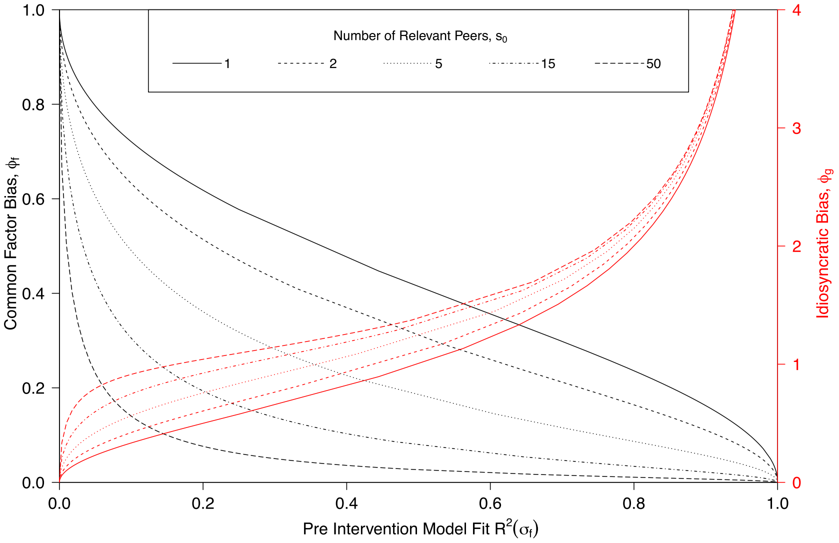


Fig. 1. Bias Factor defined on (10) for $l_i = \sigma_{\eta_i} = 1$ for all $i = 1, \ldots, n$.

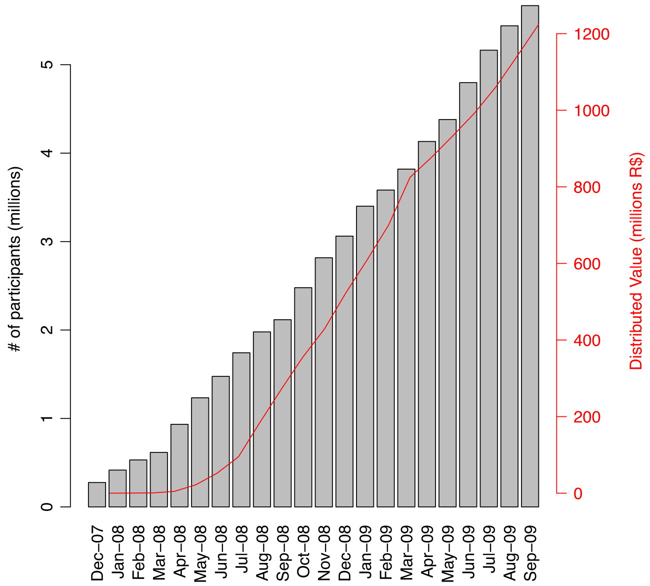
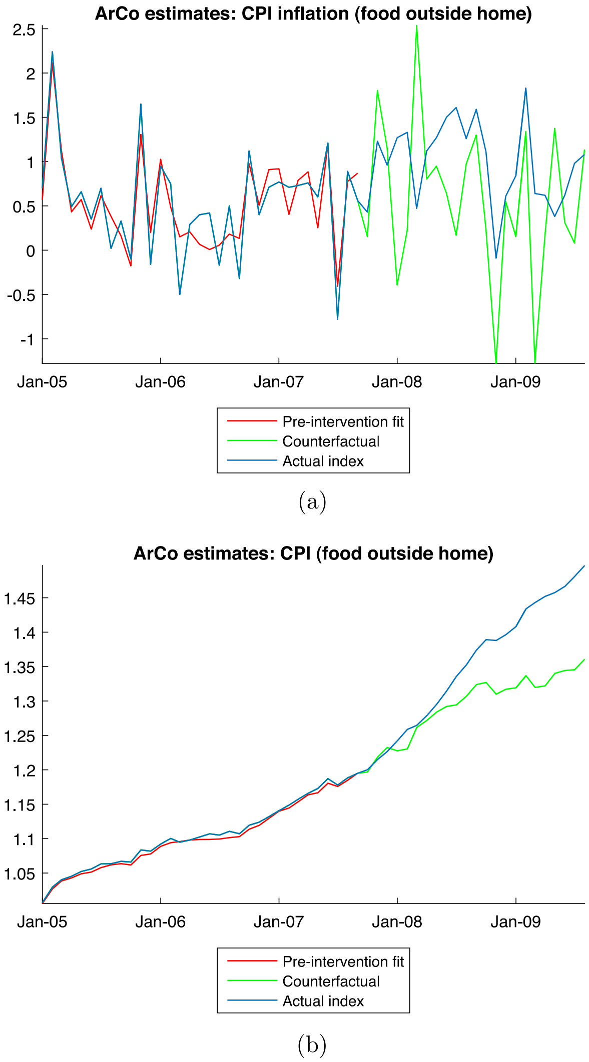


Fig. 2. NFP Participation (left) and Value distributed (right).

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

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

where the second equality follows from Assumption 7, 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



(a)

(b)

Fig. 3. Actual and counterfactual data. The conditioning variables are **inflation** and **DGP growth**. Panel (a) monthly inflation. Panel (b) accumulated monthly inflation.

$\boldsymbol{\beta}_0$ the remaining parameters. Then,

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

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 Lemma 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 6

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 5 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 continuous function on $\mathbb{R}^q$ so given any $\epsilon > 0$, we can take an 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 4

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

Appendix B. Figures and tables

See Tables 2–6 and Figs. 1–3.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2018.07.005.

References

Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. J. Amer. Statist. Assoc. 105, 493–505.

Abadie, A., Diamond, A., Hainmueller, J., 2015. Politics and the synthetic control method. Am. J. Polit. Sci. 59, 495–510.

Abadie, A., Gardeazabal, J., 2003. The economic costs of conflict: A case study of the Basque country. Amer. Econ. Rev. 93, 113–132.

Andrews, D., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

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

Angrist, J., Imbens, G., 1994. Identification and estimation of local average treatment effects. Econometrica 61, 467–476.

Angrist, J., Jordá, Ó., Kuersteiner, G., 2018. Semiparametric estimates of monetary policy effects: String theory revisited. J. Bus. Econom. Statist. 36, 371–387.

Athey, S., Imbens, G., 2016. The State of Applied Econometrics - Causality and Policy Evaluation, Working Paper arXiv:1607.00699v1.

Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77, 1229–1279.

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

Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66, 47–78.

Belasen, A., Polachek, S., 2008. How hurricanes affect wages and employment in local labor markets. Am. Econ. Rev.: Pap. Proc. 98, 49–53.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection amongst high-dimensional controls. Rev. Econom. Stud. 81, 608–650.

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

Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85, 233–298.

Billmeier, A., Nannicini, T., 2013. Assessing economic liberalization episodes: A synthetic control approach. Rev. Econ. Stat. 95, 983–1001.

Brockmann, H., Genschel, P., Seelkopf, L., 2016. Happy taxation: increasing tax compliance through positive rewards?. J. Publ. Policy, FirstView 1–26.

Bülhmann, P., van der Geer, S., 2011. Statistics for High Dimensional Data. Springer.

Caruso, G., Miller, S., 2015. Long run effects and intergenerational transmission of natural disasters: A case study on the 1970 Ancash Earthquake. J. Dev. Econ. 117, 134–150.

Cavallo, E., Galiani, S., Noy, I., Pantano, J., 2013. Catastrophic natural disasters and economic growth. Rev. Econ. Stat. 95, 1549–1561.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier Science, pp. 5549–5632.

Chen, H., Han, Q., Li, Y., 2013. Does index futures trading reduce volatility in the Chinese stock market? A panel data evaluation approach. J. Futures Mark. 33, 1167–1190.

Conley, T., Taber, C., 2011. Inference with difference in differences with a small number of policy changes. Rev. Econ. Stat. 93, 113–125.

Doudchenko, N., Imbens, G., 2016. Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis, 22791. NBER.

Doukhan, P., Louhichi, S., 1999. A new weak dependence condition and applications to moment inequalities. Stochastic Process. Appl. 84, 313–342.

Du, Z., Yin, H., Zhang, L., 2013. The macroeconomic effects of the 35-h workweek regulation in France. B.E. J. Macroecon. 13, 881–901.

Du, Z., Zhang, L., 2015. Home-purchase restriction, property tax and housing price in China: A counterfactual analysis. J. Econometrics 188, 558–568.

Fatas, E., Nosenzo, D., Sefton, M., Zizzo, D., 2015. A Self-Funding Reward Mechanism for Tax Compliance, Working Paper 2650265, SSRN.

Ferman, B., Pinto, C., 2015. Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity, Working paper, São Paulo School of Economics - FGV.

Ferman, B., Pinto, C., 2016. Revisiting the Synthetic Control Estimator, Working paper, São Paulo School of Economics - FGV.

Ferman, B., Pinto, C., Possebom, V., 2016. Cherry Picking with Synthetic Controls, Working paper, São Paulo School of Economics - FGV.

Fujiki, H., Hsiao, C., 2015. Disentangling the effects of multiple treatments - Measuring the net economic impact of the 1995 great Hanshin-Awaji earthquake. J. Econometrics 186, 66–73.

Gobillon, L., Magnac, T., 2016. Regional policy evaluation: interactive fixed effects and synthetic controls. Rev. Econ. Stat. 98, 535–551.

Grier, K., Maynard, N., 2013. The economic consequences of Hugo Chavez: A Synthetic control analysis. J. Econ. Behav. Organ. 95, 1549–1561.

Haan, W.D., Levin, A., 1996. Inferences from Parametric and Non-Parametric Covariance Matrix Estimation Procedures.

Heckman, J., Vytlacil, E., 2005. Structural equations, treatment effects and econometric policy evaluation. Econometrica 73, 669–738.

Hsiao, C., Ching, H.S., Wan, S.K., 2012. A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong Kong with Mainland China. J. Appl. Econometrics 27, 705–740.

Johnson, S., Boone, P., Friedmand, E., 2000. Corporate governance in the Asian financial crisis. J. Financ. Econ. 58, 141–186.

Jordan, S., Vivian, A., Wohar, M., 2014. Sticky prices or economically-linked economies: the case of forecasting the Chinese stock market. J. Int. Money Finance 41, 95–109.

Leeb, H., Pötscher, B., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21, 21–59.

Leeb, H., Pötscher, B., 2008. Sparse estimators and the oracle property, or the return of Hodge's estimator. J. Econometrics 142, 201–211.

Leeb, H., Pötscher, B., 2009. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. J. Multivariate Anal. 100, 1065–2082.

Li, K., Bell, D., 2017. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. J. Econometrics 197, 65–75.

McLeish, D., 1974. Dependent central limit theorems and invariance principles. Ann. Probab. 2, 620–628.

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

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

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

Pesaran, M., Smith, L., Smith, R., 2007. What if the UK or Sweden had joinded the Euro in 1999? An Empirical Evaluation using a Global VAR. Int. J. Finance Econ. 12, 55–87.

Pötscher, B., Prucha, I., 1997. Dynamic Nonlinear Econometric Models: Asymptotic Theory. Springer.

Rio, E., 1994. A new weak dependence condition and applications to moment inequalities. C. R. Acad. Sci. Paris Sér. I 318, 355–360.

Slemrod, J., 2010. Cheating ourselves: the economics of tax evasion. J. Econ. Perspect. 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. J. R. Stat. Soc. Ser. B Stat. 58, 267–288.

Wan, J., 2010. The incentive to declare taxes and tax revenue: The lottery receipt experiment in China. Rev. Dev. Econ. 14, 611–624.

Xie, S., Mo, T., 2013. Index futures trading and stock market volatility in China: A difference-in-difference approach. J. Futures Mark. 34, 282–297.

Zou, H., 2006. The adaptive LASSO and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.