




Counterfactual Analysis With Artificial Controls: Inference, High Dimensions, and Nonstationarity

Ricardo Masini & Marcelo C. Medeiros


To cite this article: Ricardo Masini & Marcelo C. Medeiros (2021) Counterfactual Analysis With Artificial Controls: Inference, High Dimensions, and Nonstationarity, Journal of the American Statistical Association, 116:536, 1773-1788, DOI: [10.1080/01621459.2021.1964978](https://doi.org/10.1080/01621459.2021.1964978)

To link to this article: <https://doi.org/10.1080/01621459.2021.1964978>

 View supplementary material [↗](#)

 Published online: 20 Sep 2021.

 Submit your article to this journal [↗](#)

 Article views: 1066

 View related articles [↗](#)

 View Crossmark data [↗](#)

 Citing articles: 7 View citing articles [↗](#)



Counterfactual Analysis With Artificial Controls: Inference, High Dimensions, and Nonstationarity

Ricardo Masini^{*a} and Marcelo C. Medeiros^b

^aSao Paulo School of Economics, Getulio Vargas Foundation, Sao Paulo, Brazil; ^bDepartment of Economics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil

ABSTRACT

Recently, there has been growing interest in developing statistical tools to conduct counterfactual analysis with aggregate data when a single “treated” unit suffers an intervention, such as a policy change, and there is no obvious control group. Usually, the proposed methods are based on the construction of an artificial counterfactual from a pool of “untreated” peers, organized in a panel data structure. In this article, we consider a general framework for counterfactual analysis for high-dimensional, nonstationary data with either deterministic and/or stochastic trends, which nests well-established methods, such as the synthetic control. We propose a resampling procedure to test intervention effects that does not rely on postintervention asymptotics and that can be used even if there is only a single observation after the intervention. A simulation study is provided as well as an empirical application. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received January 2020
Accepted August 2021

KEYWORDS

Cointegration; Comparative studies; panel data; Intervention; Policy evaluation; Resampling; Synthetic control

1. Introduction

Since the proposal of synthetic control (SC) method by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010), measuring treatment (intervention) effects on a single treated unit based on counterfactuals constructed from artificial controls has become a popular practice. Usually, artificial (synthetic) controls are built from a panel of untreated peers observed over time, before and after the intervention.

This article has two major contributions. First, we investigate the consequences of estimating counterfactuals when the data are nonstationary, with deterministic and/or stochastic trends and when the dimensionality of the counterfactual model grows with the sample size. We propose a modification of Tibshirani's (1996) least absolute and selection operator (LASSO), which has been proven to be consistent for the parameters of interest under weak sparsity of the model. Our estimator is a special case of the adaptive LASSO of Zou (2006). Our results have implications for cointegration analysis in high dimensions. Second, we develop inferential procedures based on partial resampling that can be applied in situations where the number of observations after the intervention is small when compared to the number of time periods before it. Our testing procedure can be used even when there is a single observation after the intervention. Moreover, the test can be extended to the stationary case with virtually no modifications. The statistical framework considered here nests the SC method and many of its variants as well as the panel factor (PF) method of Hsiao, Ching, and Wan (2012) and the artificial counterfactual (ArCo) of Carvalho, Masini, and Medeiros (2018).

We believe our results are of general importance for the following reasons. First, several applications of the SC method are for trending data. With nonstationary data, the usual inferential procedures to evaluate the effects of the intervention can be misleading. Second, although it is not usual for applications involving counterfactual estimation to be truly high dimensional, compared to the number of variables in the model, the number of pre-intervention observations is frequently small. Therefore, deriving the statistical properties of counterfactual estimators under high dimensions and nonstationarity at the same time is of considerable importance. Finally, recent methods consider that the number of postintervention observations grows with the sample size. In this scenario, the tests have very little power when effects diminish in the aftermath of the intervention or when effects concern the variance of the variable of interest. More worrisome is that with a long postintervention period, there could be a larger probability of contamination effects; that is, the peers may be affected by the intervention. Our inferential procedure fits nicely when the time period after the intervention is very small.

1.1. Overview

The method is divided into steps. Suppose we are interested in estimating the effects on a variable Y_t of an intervention that occurred at time $t = T_0 + 1$. We estimate a counterfactual based on a number of covariates, $\mathbf{X}_t \in \mathbb{R}^p$, constructed from a number of peers that are assumed to be unaffected by the intervention.

We allow the dimension of \mathbf{X}_t to grow with the sample size T , that is, $p \equiv p_T$. The procedure is thus summarized as:

1. Based on the sample $\{Y_t, \mathbf{X}'_t\}_{t=1}^{T_0}$ estimate $Y_t = \mathbf{X}'_t \boldsymbol{\theta}_0 + V_t$, where V_t is an error term that will be specified later. To cope with high-dimensionality and nonstationarity, estimate the model by the modification of the LASSO method proposed in this article.
2. For $t = T_0 + 1, \dots, T$, estimate the intervention effects by $\hat{\delta}_t = Y_t - \mathbf{X}'_t \hat{\boldsymbol{\theta}}_{T_0}$, where $\hat{\boldsymbol{\theta}}_{T_0}$ is the estimated coefficient in the first step.
3. Test for $\mathcal{H}_0 : \mathbf{g}(\delta_{T_0+1}, \dots, \delta_T) = \mathbf{0}$ by using the partial resampling procedure that will be described later. $\mathbf{g}(\cdot)$ is a vector-valued continuous function.

In this article, we show consistency of $\hat{\boldsymbol{\theta}}_{T_0}$ to $\boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ will be defined both under stationarity and nonstationarity. We show consistency of the estimated average intervention effect, $\hat{\Delta} = \frac{1}{T-T_0} \sum_{t=T_0+1}^T \hat{\delta}_t$. Finally, we propose a statistic to test for the general null hypothesis defined above.

1.2. Comparison With the Literature

Recent articles have discussed the effects of nonstationarity on counterfactual estimation in low dimensions. Bai, Li, and Ouyang (2014) show consistency of the Hsiao, Ching, and Wan's (2012) panel approach when the data are integrated of order one. Masini and Medeiros (2019) provided the asymptotic distribution of the counterfactual estimation under nonstationarity in low dimensions and develop the necessary results to conduct inference using the methods proposed here. Ferman and Pinto (2016) studied the SC estimator in cases with explosive common factors and imperfect pre-intervention fit. Finally, Li (2020) analyzed the properties of counterfactual estimators under both trend-stationary and unit-root cases. We complement the analysis in the previous articles by simultaneously providing a full and general treatment of counterfactual estimation with both nonstationary and high-dimensional data.

High dimensionality has been considered in settings less general than the ones considered here. For example, Bléhaut et al. (2020) considered the case of independent and identically distributed data, Li and Bell (2017) studied the case where the data are stationary, and Carvalho, Masini, and Medeiros (2018) derived results under a setup where the data are either stationary or have bounded deterministic trends, that is, deterministic functions of t/T . As we combine nonstationarity with high-dimensions, our approach generalizes the above cited articles.

Several articles have proposed methods to conduct inference for counterfactual and treatment effect estimation. Many of them derive the results from an asymptotic argument over the postintervention sample and under a less general framework than the one considered in this article. For example, the high-dimensional results in Carvalho, Masini, and Medeiros (2018) were derived under either stationarity or bounded deterministic trends. Chernozhukov, Wuthrich, and Zhu (2020) proposed a generalization of the previous article with a new inference method to test hypotheses on average treatment effects under high dimensionality and potential nonstationarity. Different from our assumptions, they impose that exactly the same (stochastic) trend is shared among all variables in the model.

Other articles tackle the problem of inference with a small number of observations after the intervention. Chernozhukov, Wuthrich, and Zhu (2018) proposed a general conformal inference method to test hypotheses on the counterfactuals. Different from the authors, we consider the case where the number of regressors grows at a faster rate than the sample size. Cattaneo, Feng, and Titiunik (2019) constructed prediction intervals in the canonical SC framework and provide conditions under which these intervals offer finite-sample probability guarantees in low dimensions and stationary data. Brodersen et al. (2015) considered a Bayesian structural time-series model to estimate the counterfactuals and advocated posterior inference to measure the effects of the intervention. In the low-dimensional case, Ferman and Pinto (2016) and Li (2020) discussed inference in the SC framework based on Andrews's (2003) end-of-sample tests. Shaikh and Toulis (2019) considered randomization tests with staggered adoption of treatment in the SC framework with low dimensional and stationary data. See also Amjad, Shah, and Shen (2018), Arkhangelsky et al. (2019), and Ben-Michael, Feller, and Rothstein (2019). Another nice extension of the SC method is Abadie and L'Hour (2019).

This article is also related to the literature on unit roots and cointegration in high dimensions. To our knowledge, this is one of the first works to derive the properties of LASSO estimators for cointegrating regressions in the case where the number of regressors is larger than the sample size. For fixed dimension, Liao and Phillips (2015), Lee, Shi, and Gao (2018), and Kock (2016) derived the limiting distribution of LASSO-type estimators under several setups with nonstationary variables. Liang and Schienle (2019) proposed a shrinkage methodology for simultaneous model selection and estimation of vector error correction models when the dimension is large and can increase with sample size. Another related article is Onatski and Wang (2018), where the authors derived the distribution of cointegration test statistics in a high-dimensional. The previous two articles consider the setting when the dimension of the model grows at slower rate than the sample size. Recently, Wijler and Smeekes (2020) considered the estimation of error correction models in high dimensions. However, their framework is quite different from ours.

1.3. Summary of the Article

The rest of the article is organized as follows. We present the setup and assumptions in Section 2 and derive the theoretical results in Section 3. In Section 3.2, we describe the inferential procedure considered in this article. A guide to practical implementation of the methods is presented in Section 4. We present the results of a simulation experiment in Section 5 and discuss the empirical application in Section 6. Section 7 concludes the article. Finally, we present additional material in the appendix. The supplementary material provides additional results and all the proofs.

2. Setup and Assumptions

2.1. Notation

All random variables (real-valued scalars, vectors and matrices) are defined in a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We

denote random variables by an upper case letter, X , for instance, and its realization by a lower case letter, $X(\omega) = x$. The expected value operator is with respect to the \mathbb{P} law such that $\mathbb{E}(X) := \int_{\Omega} X(\omega) d\mathbb{P}(\omega)$. Matrices and vectors are written in bold letters \mathbf{X} . Sets are denoted by calligraphic upper case such as in \mathcal{X} ; in that case, $|\mathcal{X}|$ denotes the cardinality of the set \mathcal{X} .

We reserve the symbol $\|\cdot\|$ without subscript for a generic (semi)norm. We use $\|\cdot\|_q$ and $\|\cdot\|_{\mathcal{L}^q}$ to denote, respectively, the ℓ^q and \mathcal{L}^q norms for $q \in [1, \infty]$. Such that for a d -dimensional (possibly random) vector $\mathbf{X} = (X_1, \dots, X_d)'$, we have $\|\mathbf{X}\|_q := (\sum_{i=1}^d |X_i|^q)^{1/q}$ for $q \in [1, \infty)$ and $\|\mathbf{X}\|_{\infty} := \max_{i \leq d} |X_i|$; and, for a scalar random variable X , $\|X\|_{\mathcal{L}^q} = (\mathbb{E}|X|^q)^{1/q}$ for $q \in [1, \infty)$ and $\|X\|_{\mathcal{L}^{\infty}}$ is the essential supremum of X . If \mathbf{X} is a $(m \times n)$ (random) matrix, then $\|\mathbf{X}\|_{\max} := \max_{i \leq m, j \leq n} |X_{ij}|$. We also use the $\|\mathbf{X}\|_0 := |\{i : X_i \neq 0\}|$ to denote the ℓ^0 “norm” of a vector \mathbf{X} . Moreover, for a d -dimensional square matrix \mathbf{M} , we use $\|\mathbf{X}\|_{\mathbf{M}}^2$ to denote the quadratic form $\mathbf{X}'\mathbf{M}\mathbf{X}$. For any vector \mathbf{X} , we use $\text{diag}(\mathbf{X})$ to denote the diagonal matrix whose diagonal consists of the elements of \mathbf{X} . $\mathbb{1}(A)$ represents an indicator function on the event A , that is, $\mathbb{1}(A) = 1$ if A is true or $\mathbb{1}(A) = 0$, otherwise.

Finally, unless stated otherwise, all the asymptotics are taken as $T_0 \rightarrow \infty$, and the $o(1)$ and $o_P(1)$ terms are with respect to the limit as $T_0 \rightarrow \infty$. We denote convergence in probability and in distribution by “ \xrightarrow{P} ” and “ \Rightarrow ,” respectively. See the supplemental material for a full list of symbols used in the article and presented in the appendix.

2.2. Basic Setup

Suppose we have n units (countries, states, municipalities, firms, etc.) indexed by $i = 1, \dots, n$. For every time period $t = 1, \dots, T$, we observe a realization of a real-valued random vector $\mathbf{Z}_t := (Z_{1t}, \dots, Z_{nt})'$. We consider a scalar variable for each unit for the sake of simplicity. The results in the article can be easily extended to the multivariate case. Furthermore, we assume that an intervention took place at $T_0 + 1$, where $1 < T_0 < T$. Let $D_t \in \{0, 1\}$ be a binary variable flagging the periods where the intervention (treatment) was in place. Therefore, following the potential outcome notation, we can express Z_{it} as

$$Z_{it} = D_t Z_{it}^{(1)} + (1 - D_t) Z_{it}^{(0)},$$

where $Z_{it}^{(1)}$ denotes the potential outcome when the unit i is exposed to the intervention and $Z_{it}^{(0)}$ is the potential outcome of unit i when it is not exposed to the intervention.

We are ultimately concerned with testing the hypothesis on the potential effects of the intervention in the unit of interest. Without loss of generality, we set unit 1 to be the one of interest. Our framework accommodates the case of more than one treated unit with minor changes as long as the number of treated units is held fixed as the number of periods (and potentially the number of untreated units) grows. Otherwise, the result would have to reflect the limit ratio of treated to untreated units as well. The null hypothesis to be tested is:

$$\mathcal{H}_0 : \delta_t := Z_{1t}^{(1)} - Z_{1t}^{(0)} = 0, \quad \forall t > T_0. \tag{1}$$

It is evident that for each unit $i = 1, \dots, n$ and at each period $t = 1, \dots, T$, we observe either $Z_{it}^{(0)}$ or $Z_{it}^{(1)}$. In particular, $Z_{1t}^{(0)}$

is not observed from $t = T_0 + 1$ onward. For this reason, we henceforth call it the *counterfactual*—that is what would Z_{1t} have been like had there been no intervention (potential outcome).

To construct the counterfactual, let $\mathbf{Z}_{0t}^{(0)} := (Z_{2t}^{(0)}, \dots, Z_{nt}^{(0)})'$ be the collection of control variables (all other variables except the those belonging to unit 1). We could have also included lags of the variables and/or exogenous regressors into \mathbf{Z}_{0t} , but to keep the argument simple, we have considered only contemporaneous variables; see Carvalho, Masini, and Medeiros (2018) for more general specifications. Panel-based methods, such as the PF and ArCo methodologies, as well as the SC extensions discussed in Doudchenko and Imbens (2016), construct an ArCo by considering the following model in the absence of an intervention:

$$\mathbf{Z}_{1t}^{(0)} = \mathbf{M}(\mathbf{Z}_{0t}^{(0)}; \boldsymbol{\theta}_0) + V_t, \quad t = 1, \dots, T, \tag{2}$$

where $\mathbf{M} : \mathcal{Z} \times \Theta \rightarrow \mathbb{R}$, $\mathcal{Z} \subseteq \mathbb{R}^{n-1}$, is a known measurable mapping up to a vector of parameters indexed by $\boldsymbol{\theta}_0 \in \Theta$ and Θ is a parameter space. A linear specification (including a constant) for the model $\mathbf{M}(\mathbf{Z}_{0t}; \boldsymbol{\theta}_0)$ is the most common choice among counterfactual models for the pre-intervention period.

The main idea is to estimate (2) using just the pre-intervention sample, $t = 1, \dots, T_0$, since in this case, $\mathbf{Z}_{0t}^{(0)} = \mathbf{Z}_{0t} := (Z_{2t}, \dots, Z_{nt})'$ under Assumption 1 below. Consequently, the estimated counterfactual for the postintervention period, $t = T_0 + 1, \dots, T$, becomes $\widehat{\mathbf{Z}}_{1t}^{(0)} := \mathbf{M}(\mathbf{Z}_{0t}; \widehat{\boldsymbol{\theta}}_{T_0})$. Under some sort of stationarity assumption on \mathbf{Z}_{0t} and, more importantly, under the assumption that the control units are not affected by the intervention, Hsiao, Ching, and Wan (2012) and Carvalho, Masini, and Medeiros (2018) showed that $\widehat{\delta}_t := Z_{1t} - \widehat{\mathbf{Z}}_{1t}^{(0)}$ is an unbiased estimator for δ_t as the pre-intervention sample size grows to infinity and

$$\widehat{\Delta}_T = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \widehat{\delta}_t, \tag{3}$$

is \sqrt{T} -consistent for $\Delta_T := \frac{1}{T - T_0} \sum_{t=T_0+1}^T \delta_t$ and is asymptotically normal as both T and T_0 grow to infinity and keep the (sample) ratio T_0/T unaltered.

Consider the following assumption.

Assumption 1. \mathbf{Z}_{0t} is independent of D_s for all $1 \leq s, t \leq T$.

To recover the effects of the intervention, Assumption 1 is key. However, for an unbiased estimate, it would be enough to impose $\mathbb{E}(\mathbf{Z}_{0t} | D_s = 1) = \mathbb{E}(\mathbf{Z}_{0t} | D_s = 0)$. For a thorough discussion on Assumption 1, including the potential bias resulting from its failure in the stationary setup, refer to Carvalho, Masini, and Medeiros (2018).

The main purpose of this article is to extend the above results in the presence of deterministic and/or stochastic trends in the data-generating process (DGP) of $\mathbf{Z}_t^{(0)}$. This leads to new challenges. For instance, due to the nonstationarity of the regressors, $\boldsymbol{\theta}_0$ can no longer be identified as the linear projection parameter of $Z_{1t}^{(0)}$ onto a constant and $\mathbf{Z}_{0t}^{(0)}$.

2.3. Nonstationarity

We model the units in the absence of the intervention as a nonstationary (vector) process $\{Z_t^{(0)} := (Z_{1t}, \dots, Z_{nt})'\}_{t \geq 1}$.

Assumption 2. (DGP)

(1) Consider that the process $\{Z_{it}^{(0)} : 1 \leq i \leq n, t \geq 1\}$ is either generated by

(a) Stochastic Trend:

$$Z_{it}^{(0)} = Z_{it-1}^{(0)} + f_{it} + U_{it}, \quad t \geq 1, \quad \text{with } Z_{i0}^{(0)} = O_P(1), \text{ or} \quad (4)$$

(b) Deterministic Trend:

$$Z_{it}^{(0)} = f_{it} + U_{it}, \quad t \geq 1. \quad (5)$$

In both cases, $\{f_{it}\}_{t \geq 1}$ is a deterministic sequence, and $\{U_t := (U_{1t}, \dots, U_{nt})'\}_{t \geq 1}$ takes values in $\mathcal{U} \subset \mathbb{R}^n$ and is a zero-mean weakly dependent stochastic process fulfilling one of the two conditions described in [Assumption 3](#) in [Appendix A](#).

(2) Furthermore, assume that there is at least one linear combination of the elements of $Z_t^{(0)}$, with a nonzero coefficient for the first element ($i = 1$), that results in a process integrated of order 0, $I(0)$. For a formal definition of integrated process we refer to [Definition 1](#) in [Appendix A](#).

The deterministic sequence $\{f_{it}\}_{t \geq 1}$ in [Assumption 2](#) is considered idiosyncratic, that is, unit-specific. However, in most applications, we expect to have a common (up to a constant) trend such that $f_{it} = \mu_i' f_t$ where μ_i and f_t are multidimensional. The DGP (4) may involve an $I(1)$ (integrated of order 1) process depending upon the choice of the sequence f_{it} . If we take $f_{it} = \mu_i \in \mathbb{R}$, we have a unit-root process with drift μ_i . Thus, a constant f_{it} generates a linear (deterministic) trend plus a pure unit-root process.

Example 1 (Nonstationary factor model). Consider model (4) and assume there exists a nonstationary factor driving the dynamics of the units, that is, $F_t = \mu^F + F_{t-1} + U_t^F$. Thus, the nonstationary factor model $Z_{it}^{(0)} = c_i + \mu_i F_t + U_{it}^Z$, where U_{it}^Z is a weakly dependent process, is equivalent to (4) with $U_{it} = \mu_i U_t^F + U_{it}^Z - U_{it-1}^Z$ and $f_{it} = \mu_i \mu^F$. Furthermore, if $\mu_i = 1$, for all $i = 1, 2, \dots, n$, we have the nonstationary model considered in Chernozhukov, Wuthrich, and Zhu (2020). This example can be easily extended to the case where there are multiple factors.

Example 2 (Unit roots and cointegration). Consider the triangular cointegration model:

$$Z_{it}^{(0)} = \theta' Z_{0t} + V_t \quad \text{and} \quad \Delta Z_{0t} = U_{0t},$$

where V_t and U_{0t} are weakly stationary stochastic processes. This representation is equivalent to (4) with $f_{it} = 0$ and $U_{1t} = \theta' U_{0t} + V_t - V_{t-1}$.

Example 3 (Deterministic trends). Models with heterogeneous deterministic trends can be easily handled by setting $f_{it} = \mu_i f_t$, where f_t is a general deterministic function of t .

Importantly, failure to comply with [Assumption 2\(2\)](#) results in what is known as a spurious regression. We acknowledge that the name “spurious” might be misleading since, in some cases, it might be possible to construct a nonlinear function of the units that results in an $I(0)$ process. Therefore, the DGP is considered spurious only in the sense that all linear combinations of the units are not an $I(0)$ process.

2.4. The Target Model and High Dimensionality

To simplify the notation, we rename the variable of interest as $Y_t := Z_{1t}^{(0)}$ and denote the final regressors as a p -dimensional vector X_t , where $X_t := (1, Z_{0t}')'$. Note that in the current setup $p = n$, but we choose to use another notation to explicitly allow for the case where there are more variables observed for each one of the units. As mentioned before, the results in this article are easily generalized to the multivariate case. We can now properly define the target model together with its “true parameters”.

Ideally (in the mean squared error sense), we would like $M(x) := \mathbb{E}(Y_t | X_t = x)$. However in the presence of trends, we would be most likely to have the model $M = M_t$ time dependent. In fact, even a common approximation of the conditional expectation model by a linear projection of Y_t onto the space spanned by the columns of X_t would result in time-varying parameters again due to the nonstationary setup.

Let $r \in \{0, 1, \dots, n-1\}$ be the number of independent linear relations among the n units that results in an $I(0)$ process. By [Assumption 2](#), we have that $r \geq 1$ and at least one of those relations includes unit 1 such that its coefficient can be normalized to one. For the DGP (4), r also represents the number of cointegration relations as per Engle and Granger (1987). For the DGP (5), if $f_{it} = \mu_i f_t$, we have $r = n - 1$ because for any vector $\theta \in \mathbb{R}^{n-1}$ such that $(1, \theta') \mu = 0$, the trend f_t is canceled; therefore, $(1, \theta') Z_t^{(0)} \sim I(0)$.

Let $\tilde{\Gamma}$ be an $(n \times r)$ matrix containing the r independent linear relations resulting in an $I(0)$ process as described in the previous paragraph. Without loss of generality, since $\tilde{\Gamma}$ is rank r by definition, we can normalize it such that

$$J_t := \tilde{\Gamma}' Z_t^{(0)} \sim I(0), \quad \tilde{\Gamma} := (I_r : -\Gamma)'. \quad (6)$$

Furthermore, let J_{1t} be the first component of the vector J_t and $J_{0t} = 1$ if $r = 1$ and $J_{0t} = (1, J_{2t}, \dots, J_{rt})'$ for $r > 1$. Since $J_t \sim I(0)$, we can then define the limit of the average linear projection of J_{1t} onto J_{0t} as

$$\pi_{(r \times 1)} := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T [\mathbb{E}(J_{0t} J_{0t}')]^{-1} [\mathbb{E}(J_{0t} J_{1t})] \quad (7)$$

We can now define the pseudo-true parameters as

$$\theta_0 := \theta_0(r) := \begin{cases} (\pi, \Gamma)' & r = 1 \\ [\pi', (1, -\pi_0') \Gamma]' & 2 \leq r \leq n - 1, \end{cases} \quad (8)$$

where $\Gamma(n - r \times r)$ is defined by (6), π is defined by (7) and $\pi_0 := (\pi_2, \dots, \pi_r)'$.

Hence, the “pseudo-true” model in the absence of an intervention becomes

$$Y_t = X_t' \theta_0 + V_t, \quad 1 \leq t \leq T, \quad (9)$$

where the p -dimensional vector θ_0 is defined in Equation (8) depending on the DGP appearing in Assumption 2 and the number of independent linear $I(0)$ relations.

Remark 1. Whenever the cointegration subspace is of dimension $r > 1$, we cannot single out the cointegration vector even after normalization. However, θ_0 is uniquely defined by (8) as long as π defined by (7) is well defined. In fact, the vector $\beta_0 := (1, -\theta_0')$ is the vector that minimizes the variance of $\beta'(Y_t, X_t)'$ among all the vectors β defined in the cointegration subspace. Precisely for this reason, β_0 was chosen to be the “pseudo-true” parameter of interest, as it results in a counterfactual estimator with minimum variance among all other counterfactuals that could be constructed via linear combinations of the columns of $\tilde{\Gamma}$.

Example 4 (Nonstationary factor model (Example 1) revisited). Consider the model described in Example 1 and set $r = 1$ with $\mu_j \neq 0$, for $j = 1, 2$, and $\mu_j = 0$, for $j > 2$. Therefore, $\tilde{\Gamma} = (1, -\frac{\mu_1}{\mu_2}, 0, \dots, 0)'$ and the “pseudo-true” parameter vector is given by $\theta_0 = (c_1 - \frac{\mu_1}{\mu_2}c_2, \frac{\mu_1}{\mu_2}, 0, \dots, 0)$.

We consider the case where the number of regressors X_t in (9) can be much larger than the number of observations, such that p is a function of the sample size. Our motivation to move to a high-dimensional setup is to accommodate two cases: when the number of units is much larger than the number of observations available ($n \gg T$) or the number of units is small but T is also small, such that $n \approx T$ or $n > T$. High dimensionality is also important in the case where more than one variable is observed for each unit.

3. Estimation and Theoretical Results

The challenge to consistently estimate the parameters of model (9) arises because the model combines both high dimensionality and nonstationarity. Any of these two features taken separately are well studied in the literature. For estimation of (9) in a high-dimension and stationary time-series framework, see Kock and Callot (2015), Medeiros and Mendes (2016), or Carvalho, Masini, and Medeiros (2018), for example. For estimation in the low-dimensional and nonstationary case, see, Phillips (1986,1987) or Masini and Medeiros (2019). We propose to estimate the parameters of (9) via a weighted least absolute shrinkage and selection operator (WLASSO), that is, $\hat{\theta} := \hat{\theta}_{T_0}(\lambda, \mathbf{w})$ is a minimizer of $\theta \mapsto Q(\theta, \lambda, \mathbf{w})$ defined as

$$Q(\theta, \lambda, \mathbf{w}) := \frac{1}{T_0} \sum_{t=1}^{T_0} (Y_t - X_t'\theta)^2 + \lambda \sum_{i=1}^p w_i |\theta_i|, \quad (10)$$

where $\lambda \geq 0$ is the common penalty term and $\mathbf{w} := (w_1, \dots, w_p)'$ is a vector of almost surely nonnegative weights specific for each parameter. In a low-dimension setup, namely, $p \ll T$, we might choose $\lambda = 0$. Although, (10) resembles the adaptive LASSO estimator, the choice of the weights w_i , $i = 1, \dots, p$ will be very different. The weights should be determined according to the nature of the trend in each series. Table 1 shows weights for a typical process encountered in

Table 1. Weight selection.

Description	DGP	Growth condition	Weight (w)
I(0) process	$X_{it} = f_{it} + U_{it}$ and $f_{it} = O(1)$	No	1
Linear trend	$X_{it} = a_0 + a_1t + U_{it}$	Yes	$ X_{i,T_0} $
Polynomial trend	$X_{it} = a_0 + a_1t + \dots + a_kt^k + U_{it}$	Yes	$ X_{i,T_0} $
I(1), no drift	$X_{it} = X_{it-1} + U_{it}$ and $X_{i0} = O_p(1)$	No	$\sqrt{T_0}$
I(1) with drift	$X_{it} = a_0 + X_{it-1} + U_{it}$ and $X_{i0} = O_p(1)$	Yes	$ X_{i,T_0} $

NOTE: The table shows the choice of weights w_i , $i = 1, \dots, p$ in the modified LASSO estimator of Equation (10) for typical DGPs that are common in empirical applications. The column *Growth Condition* indicates whether or not the growth condition holds.

practical applications. More details about the choice of w_i , $i = 1, \dots, p$ can be found in Section 4 and in Appendix A.

Hereafter, we outline the steps toward the proof of our main result (Theorem 1). The details such as technical assumptions and propositions can be found in Appendix A. First, due to the presence of trending regressors, not all the components of X_t are of the same order (in probability). Therefore, it is convenient to consider a reparameterization of the objective function (10). We stress that this reparameterization is only convenient in order to analyze the estimator properties and by no means should it be carried out in the application of the proposed methodology in practice.

Clearly, both DGPs defined in Assumption 2 can be written as follows:

$$Z_{it}^{(0)} = d_{it} + \eta_{it}, \quad 1 \leq i \leq n, t \geq 1, \quad (11)$$

where d_{it} is a deterministic trend and η_{it} is the stochastic component (not necessarily stationary). Equation (5) becomes (11) by setting $d_{it} = c_i + f_{it}$ and $\eta_{it} = U_{it}$. Similarly, for Equation (4), we conclude, by backward recursion, that $d_{it} = a_{it} := \sum_{s=1}^t f_{is}$ and $\eta_{it} = Z_{i0}^{(0)} + \sum_{s=1}^t U_{is}$.

Consider the following linear transformation applied to Equation (10).

$$\mathbf{y} := L\theta, \quad \mathbf{W}_t := L^{-1}X_t \quad L := \text{diag}[(\ell_1, \dots, \ell_p)'], \quad (12)$$

where $\ell_1 = 1$ and for $2 \leq i \leq p$, ℓ_i will depend on the nonstationary nature of the data and may be different for each one of the regressors. For instance, we set $\ell_i = d_{i,T_0}$ if the growth condition (Proposition 1(b) in Appendix A) is satisfied; otherwise $\ell_i = \sqrt{T_0}$ if the regressor follows DPG (4), or 1 if the regressor follows DGP (5) in Assumption 2. The reparameterized objective function then becomes

$$H(\mathbf{y}) := H(\mathbf{y}, \lambda, \mathbf{v}) := \frac{1}{T_0} \sum_{t=1}^{T_0} (Y_t - \mathbf{W}_t'\mathbf{y})^2 + \lambda \sum_{i=1}^p v_i |\gamma_i|, \quad (13)$$

where $\mathbf{v} := (v_1, \dots, v_p)'$ and $v_i := w_i/\ell_i$ for $1 \leq i \leq p$.

The importance of such reparameterization is that the new regressors \mathbf{W}_t are free of diverging trends, which makes the problem tractable. Moreover, a minimizer $\hat{\mathbf{y}}$ of $\mathbf{y} \mapsto H(\mathbf{y})$ is related to a minimizer $\hat{\theta}$ of $\theta \mapsto Q(\theta)$ through $\hat{\mathbf{y}} := L\hat{\theta}$, and the reparametrized target parameters become $\mathbf{y}_0 := L\theta_0$.

The oracle inequality (A.6) in Appendix A is the basis for our result. When it is applied to $\mathbf{y} = \mathbf{y}_0$, it yields an upper bound for both the prediction error $\|\hat{\mathbf{y}} - \mathbf{y}_0\|_{\Sigma}$ and the ℓ_1 -estimation

error $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1$. If we set that the number of nonzero “pseudo-true” coefficients (s_0) to grow slower than T_0 , even if $p \gg T_0$, then we can achieve consistency of both the prediction error and the ℓ_1 -estimation error under some conditions.

3.1. Weak Sparsity

The assumption of sparsity ($s_0 \ll p$) of $\boldsymbol{\gamma}_0$ is just one way of imposing a low-dimensional structure in a high-dimensional problem. Otherwise, under $p \gg T_0$, a consistent procedure is not possible. An alternative is through the concept of weak sparsity. Here, we follow Negahban et al. (2012) and define for $b \in [0, 1]$ the b -“radius” of $\boldsymbol{\gamma}_0 \in \mathbb{R}^p$ as

$$R_b := \sum_{j=1}^p |\gamma_{0,j}|^b. \tag{14}$$

The idea behind a weak sparsity is to impose a restriction on R_b for some $b \in (0, 1]$. When $b = 0$, we have $R_0 = |\mathcal{S}_0| := s_0$, which is called strong or exact sparsity. The concept of weak sparsity applied to (A.6) results in Proposition 3 in Appendix A. The latter combined with the probabilistic bounds in Lemmas 1 and 2 fully characterize the asymptotic behavior of $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ and consequently of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ from which we obtain our main result.

Theorem 1. Under Assumptions 1–6 (Assumptions 3–6 are stated in the Appendix A.2), as $T_0 \rightarrow \infty$:

1. $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 = O_p \left[\left(\frac{\psi(p)}{\sqrt{T_0}} \right)^{1-b} \frac{R_b}{\lambda_1} \right] = o_p(1)$
2. $\widehat{\delta}_t - \delta_t - V_t = O_p \left[\frac{\psi(p)^{2-b} R_b}{T_0^{(1-b)/2} \lambda_1} \right] = o_p(1)$ for all $T_0 < t \leq T$.

If further $T_1 := T - T_0 \rightarrow \infty$:

1. $\widehat{\Delta}_T - \Delta_T = O_p \left[\frac{\psi(p)^{2-b} R_b}{T_0^{(1-b)/2} \lambda_1} \vee \frac{1}{\sqrt{T_1}} \right] = o_p(1)$

where $\psi(x) = x^{2/q}$ under Assumption 3(a) and $\psi(x) = \log(x)$ under Assumption 3(b), R_b is given by (14) and λ_1 is defined in Assumption 6(c).

The results (a) and (b) of Theorem 1 follow under the condition of what we call *Partial Asymptotics*, that is, an asymptotic approach only for the pre-intervention period, where the number of postintervention periods $T_1 := T - T_0$ is kept fixed, while $T_0 \rightarrow \infty$. This approach is tailored to accommodate situations where T_0 is much larger than T_1 , which justifies the sampling error from the estimation of $\boldsymbol{\theta}_0$ by $\widehat{\boldsymbol{\theta}}$ to be of smaller order than V_t . In contrast, for part (c) of Theorem 1, we used the *Full Asymptotics* approach to establish the asymptotic properties by considering that the whole sample is increasing, while the proportion between the pre-intervention and the postintervention sample size is constant. In that case, $T \rightarrow \infty$.

Part (a) of Theorem 1 states the ℓ_1 -consistency for the parameter estimation, which in turn enables us to derive in part (b) an asymptotic (as $T_0 \rightarrow \infty$) mean-unbiased estimator for the treatment effect δ_t for every period in the postintervention sample. Finally, part (c) gives us a consistent estimator (as both T_0 and T_1 diverge to infinity) for the average intervention effect across the postintervention period.

All rates of convergences appearing in Theorem 1 depend upon the interplay of three main components: (i) the number of regressors/units via $\psi(p)$, (ii) the degree of (weak) sparsity of the “true” parameters through R_b and (iii) the (restricted) strong convexity of the objective function captured by the smallest (restricted) eigenvalue of $\boldsymbol{\Sigma} := \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbf{W}_t \mathbf{W}_t'$, which is assumed to be lower bounded by λ_1 . The term $\psi(p)$ is standard in the literature on high dimensionality. It is a consequence of trying to control for the sampling error for polynomial or subexponential tails.

3.2. Inference

The inference procedure presented in this section is based on the sequence of estimators $\{\widehat{\delta}_t\}_{t>T_0}$ obtained in Section 3. More specifically, we consider any continuous mapping $\boldsymbol{\phi} : \mathbb{R}^{T_1} \rightarrow \mathbb{R}^b$ whose argument is the T_1 -dimensional vector $(\widehat{\delta}_{T_0+1} - \delta_{T_0+1}, \dots, \widehat{\delta}_T - \delta_T)'$. Thus, we are ultimately interested in the distribution of $\widehat{\boldsymbol{\phi}} := \boldsymbol{\phi}(\widehat{\delta}_{T_0+1} - \delta_{T_0+1}, \dots, \widehat{\delta}_T - \delta_T)$ under the null (1) where $\delta_t = 0$ for all $t > T_0$.

We consider a situation where the pre-intervention period is substantially larger than the postintervention period, $T_0 \gg T_1$. We may even want to handle the case where $T_1 = 1$. The results in this section are based on part (b) of Theorem 1. We have that under the asymptotics on the pre-intervention period ($T_0 \rightarrow \infty$), $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \xrightarrow{p} 0$ where $\boldsymbol{\phi}_0 := \boldsymbol{\phi}(V_{T_0+1}, \dots, V_T)$. Consider the construction of $\widehat{\boldsymbol{\phi}}$ using only blocks of size T_1 of consecutive observations from the pre-intervention sample. There are $T_0 - T_1 + 1$ such blocks denoted by

$$\widehat{\boldsymbol{\phi}}_j := \boldsymbol{\phi}(\widehat{V}_j, \dots, \widehat{V}_{j+T_1-1}) \quad j = 1, \dots, T_0 - T_1 + 1,$$

where $\widehat{V}_t := Y_t - \widehat{\boldsymbol{\theta}}_{T_0}' \mathbf{X}_t$ with the subscript T_0 in $\widehat{\boldsymbol{\theta}}$ indicates that the estimator is calculated using the entire pre-intervention sample.

For each j , we have that $\widehat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j \xrightarrow{p} 0$ where $\boldsymbol{\phi}_j := \boldsymbol{\phi}(V_j, \dots, V_{j+T_1-1})$. Under a strict stationarity assumption on V_t , we have that $\boldsymbol{\phi}_j$ is equal in distribution to $\boldsymbol{\phi}_0$ for all j . Hence, we propose to estimate the distribution $Q_T(\mathbf{x}) := \mathbb{P}(\widehat{\boldsymbol{\phi}} \leq \mathbf{x})$ by

$$\widehat{Q}_T(\mathbf{x}) := \frac{1}{T_0 - T_1 + 1} \sum_{j=1}^{T_0 - T_1 + 1} \mathbb{1}(\widehat{\boldsymbol{\phi}}_j \leq \mathbf{x}),$$

where, for a pair of vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we say that $\mathbf{a} \leq \mathbf{b} \iff a_i \leq b_i, \forall i$.

Theorem 2. For any continuous $\boldsymbol{\phi} : \mathbb{R}^{T_1} \rightarrow \mathbb{R}^b$, let $\widehat{\boldsymbol{\phi}} := \boldsymbol{\phi}(\widehat{\delta}_{T_0+1} - \delta_{T_0+1}, \dots, \widehat{\delta}_T - \delta_T)$ and $\boldsymbol{\phi}_0 := \boldsymbol{\phi}(V_{T_0+1}, \dots, V_T)$. Consider the conditions of Theorem 1 but with Assumption 3(a) fulfilled with $q > 4$. Assume that $\{V_t\}$ is strictly stationary. Then, for fixed T_1 as $T_0 \rightarrow \infty$,

1. $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \xrightarrow{p} 0$
2. $\widehat{Q}_T(\mathbf{x}) - Q_T(\mathbf{x}) \xrightarrow{p} 0$ for all $\mathbf{x} \in \mathcal{C}_0 := \{\text{continuity point of } Q_0(\mathbf{x}) := \mathbb{P}(\boldsymbol{\phi}_0 \leq \mathbf{x})\}$
3. If $Q_0(\mathbf{x})$ is continuous, the result (b) holds uniformly in $\mathbf{x} \in \mathbb{R}^b$.

4. If ϕ is real-valued, then $Q_T[\widehat{Q}_T^{-1}(\tau)] \rightarrow \tau$ for all $\tau \in (0, 1)$ such that $Q_0^{-1}(\tau) \in C_0$, where f^{-1} denotes left inverse of f .

By the appropriate choice of $\phi(\cdot)$, **Theorem 2** provides a simple way to conduct inference. We could be interested in testing the intervention effects on all postintervention periods individually by setting $\phi(\widehat{\delta}_{T_0+1}, \dots, \widehat{\delta}_T) = (\widehat{\delta}_{T_0+1}, \dots, \widehat{\delta}_T)'$, or on the average intervention effect across the postintervention periods $\phi(\widehat{\delta}_{T_0+1}, \dots, \widehat{\delta}_T) = \frac{1}{T_1} \sum_{t=T_0+1}^T \widehat{\delta}_t$.

A reasonable choice for testing the null (1) would be $\phi(\widehat{\delta}_{T_0+1}, \dots, \widehat{\delta}_T) = \frac{1}{T_1} \sum_{t=T_0+1}^T \widehat{\delta}_t^2$, or, more generally, $\phi(\widehat{\delta}_{T_0+1}, \dots, \widehat{\delta}_T) = \frac{1}{T_1} \sum_{t=T_0+1}^T g(\widehat{\delta}_t)$, for some positive function $g(\cdot)$, such as $|\cdot|$. Regardless of the choice, **Theorem 2** ensures a correct asymptotic test size or a correct asymptotic coverage probability.

As we might be interested in a joint confidence set for the vector $\delta := (\delta_{T_0+1}, \dots, \delta_T)'$, we can take $\widehat{\phi} = \widehat{\delta} - \delta$, where $\widehat{\delta} := (\widehat{\delta}_{T_0+1}, \dots, \widehat{\delta}_T)'$. Unless $T_1 = 1$, there several ways to construct a confidence set for a given significance level. For instance, a $(1 - \tau)$ confidence cube that takes into account the potential autocorrelation among the δ_t 's is given by $\mathcal{A}_T := \times_{t=T_0+1}^T [\widehat{\delta}_t - \widehat{Q}_T^{-1}(1 - \tau/2); \widehat{\delta}_t + \widehat{Q}_T^{-1}(\tau/2)]$, where $\widehat{Q}_T^{-1}(\tau) = \inf\{x \in \mathbb{R} : \widehat{Q}_T(x) \geq \tau\}$ and \mathbf{t} is a vector of T_1 ones. As a direct corollary of **Theorem 2** assuming that Q_0 is continuous for any $\tau \in (0, 1)$, we have $\mathbb{P}(\delta \in \mathcal{A}_T) \rightarrow 1 - \tau$, as $T_0 \rightarrow \infty$.

Any test procedure based on an univariate test statistic $\widehat{\phi}$ can have its p -value evaluated by $1 - \widehat{Q}_T(\widehat{\phi})$ for a one-tailed test or $1 - \widehat{Q}_T(-|\widehat{\phi}|) + \widehat{Q}_T(|\widehat{\phi}|)$ for a double-tailed test. Technically speaking, $\widehat{\phi}$ is not a statistic since it depends on the value of the unknown $\{\delta_t\}_{t>T_0}$. However, under the null of interest (1), we have $\delta_t = 0$.

4. A Guide to Practice

To apply the method described in this article in practice, some choices must be made.

Choice of the penalty parameter: λ can be chosen via the Bayesian information criterion (BIC) where the degrees of freedom are determined by the number of nonzero estimates of θ_0 . We set the maximum penalty level to be $\|\frac{1}{T_0} \sum_{t=1}^{T_0} Y_t X_t\|_\infty$ with an exponential path down to a minimum value as, for example, $\lambda_{\min} = 0.001$ along L equally spaced intervals in the `glmnet` package in R. In our simulations and empirical example, we set $L = 100$.

Regressor weights: The weights w_i , $i = 1, \dots, p$ are chosen according to the trending behavior of each regressor. At the level of generality considered in the article, namely, DGPs with all sorts of deterministic and/or stochastic idiosyncratic trend combinations, it seems difficult to derive a rule to choose weights that would consistently estimate the parameters in all cases without relying on any previous knowledge of the DGP. First, as we do not penalize the intercept $w_1 = 0$. For the remaining regressors, the following strategy can be adopted:

1. If the nature of the nonstationarity is the same for all series, that is, $\forall i \in \{1, 2, \dots, n\}$, $Z_{it}^{(0)}$ follows either (4) or (5) with $f_{it}\mu_i = \boldsymbol{\mu}'\mathbf{f}_t$, where \mathbf{f}_t is a vector of deterministic components, all ω_i can be set to one. This is a similar case

to the one considered in Chernozhukov, Wuthrich, and Zhu (2020), where $\omega_i := \omega$.

- If all the units are at most of the order (in probability) of the unit of interest (unit one), again we can set the weights to a unit. The nature of the trend in the unit of interest can be determined by classical unit-root tests.
- Although the theoretical results in the article are general enough to cover a wide variety of trends, in most empirical applications, we expect to find four possible cases: I(0) series, I(1) with or without drift, and a trend-stationary series with a linear trend. These cases can be identified by usual unit-root tests applied to each series. However, we are fully aware that multiple pretesting will be an issue, especially when the number of series to be tested is large. While we do not provide a theory to accommodate these issues in our setup, we believe, based on the simulation results presented in **Section 5**, that the effects of pretesting are minor. Furthermore, in practice, we recommend that the practitioner run robustness tests by running the methodology with different choices of weights.
- To avoid spurious results, the practitioner can pretest for cointegration. For a high-dimension cointegration test refer to Onatski and Wang (2018) or Liang and Schienle (2019). We believe that **Theorem 1** coupled with **Assumption 5** can guide the practitioner to decide which weight to pick in any particular empirical application.

One remaining question is whether to estimate the model in levels, as advocated in this article, or in first differences, claiming the theory in Carvalho, Masini, and Medeiros (2018). The latter is certainly an option. However, that would destroy the potential cointegration (long-run) relation, yielding a less precise estimate of the counterfactual.

5. Simulations

The goal of this section is to conduct a Monte Carlo simulation to corroborate the asymptotic results in the article as well as to evaluate the finite-sample performance of the inferential approach advocated in the previous section.

Suppose that the units in the absence of intervention are modeled via a single factor F_t such that for each unit $i \in \{1, \dots, n\}$ and every $t \in \{1, \dots, T\}$, we have

$$Z_{it}^{(0)} = c_i + \mu_i F_t + U_{it}^Z, \tag{15}$$

where $c_i \in \mathbb{R}$, U_{it}^Z is an idiosyncratic shock and $\mu_i \in \mathbb{R}$ is the factor loading for unit i . We impose that the factor follows either a unit-root process with a (possibly nonlinear) drift

$$F_t = f_t^F + F_{t-1} + U_t^F, \quad t \geq 1 \tag{16}$$

for some initial condition $F_0 = O_p(1)$ or a trend-stationary process

$$F_t = f_t^F + U_t^F, \tag{17}$$

where $\{f_t^F\}_{t=1}^\infty$ is a deterministic sequence.

The factor model above results in a common trend (at least for those units with nonzero loadings, $\mu_i \neq 0$) and a correlation among the stochastic components of the vector $Z_t^{(0)}$.

We simulate three baseline models. The number of Monte Carlo replications is 10,000. The first simulated model consists

of Equations (15) and (17) with independent and identically normally distributed innovations, $n = 200$, and $s_0 = 5$. The factor loadings are determined as follows: $\mu_i = 1$ for $i = 1, \dots, s_0 + 1$ and $\mu_i = 0$ for $i > s_0 + 1$. Hence, the dataset consists of $s_0 + 1$ trend-stationary variables and $n - s_0 - 1$ I(0) processes. The second baseline DGP differs from the first one by considering equations (15) and (16). Therefore, the first $s_0 + 1$ variables have unit roots, and the remaining are covariance-stationary. Finally, in the third simulated specification, we mix the trends. The first $\lfloor s_0/2 \rfloor + 1$ variables are trend-stationary, followed by $s_0 - \lfloor s_0/2 \rfloor - 1$ ones with unit roots, and the remaining are covariance-stationary. In all cases, we simulated $T = 100$ observations and we set $T_1 = 3$. The test statistic considered is $\phi(x) = \|x\|_2$. We consider several alternatives to the baseline DGPs by changing the error distributions, the total number of observations (T), the number of posttreatment observations (T_1), the number of units (n), the sparsity (s_0), the shape of the deterministic component (f_t^F), and the degree of autocorrelation in the errors (ρ).

For each replication, the counterfactual model is estimated by setting $\omega_i = 1, i = 1, \dots, p$ for the first two specifications described above. Note that in these cases, we are not imposing the correct weights on all variables. This will be important to evaluate the potential harm of misspecifying the weights for some of the variables. For the third case considered, we pretest for unit roots in order to determine the weights. The procedure is as follows. For each series, we run an augmented Dickey-Fuller test for the null of the unit root against the alternative of a covariance-stationary process. If the null is rejected, we set the weight to a unit. Alternatively, in case of a nonrejection of the null, we test if the mean of the first-difference of the series is zero or not. If it is zero, we set the weight to $\sqrt{T_0}$. Otherwise, the weight is set to $|X_{i,T_0}|$; see Table 1 for details. In all cases, the penalty parameter is chosen by the BIC as described in Section 4.

Tables 2 and 3 report size results for model (15)–(17) and (15)–(16), respectively. The mixed-trend case is reported in the supplementary material. The tables show, for different settings, rejection rates under the null hypothesis of no intervention effect

Table 2. Rejection rates under the null (empirical size): deterministic trends.

	LASSO			Oracle			True		
	0.01	0.5	0.1	0.01	0.05	0.1	0.01	0.05	0.1
	Innovation Distribution								
Normal	0.0205	0.0637	0.1169	0.0297	0.0755	0.1275	0.0207	0.0583	0.1079
$\chi^2(1)$	0.0198	0.0602	0.1078	0.0231	0.0703	0.1277	0.0198	0.0591	0.1076
t-stud(3)	0.0187	0.0632	0.1144	0.0275	0.0781	0.1299	0.0208	0.0602	0.1086
Mixed Normal	0.0205	0.0603	0.1105	0.0300	0.0775	0.1339	0.0186	0.0572	0.1049
	Sample Size								
$T = 50$	0.0270	0.0768	0.1320	0.0494	0.1144	0.1740	0.0262	0.0694	0.1210
100	0.0205	0.0637	0.1169	0.0297	0.0755	0.1275	0.0207	0.0583	0.1079
150	0.0194	0.0632	0.1094	0.0220	0.0644	0.1212	0.0152	0.0536	0.1050
200	0.0182	0.0578	0.1042	0.0202	0.0592	0.1116	0.0164	0.0526	0.1018
500	0.0138	0.0530	0.1016	0.0140	0.0544	0.1004	0.0104	0.0514	0.1006
	Number of Total Units								
$n = 200$	0.0205	0.0637	0.1169	0.0297	0.0755	0.1275	0.0207	0.0583	0.1079
300	0.0236	0.0671	0.1175	0.0281	0.0743	0.1281	0.0198	0.0579	0.1053
500	0.0268	0.0748	0.1206	0.0289	0.0780	0.1327	0.0224	0.0626	0.1099
1000	0.0325	0.0778	0.1304	0.0273	0.0755	0.1298	0.0193	0.0554	0.1089
	Number of Relevant (nonzero) Covariates								
$s_0 = 2$	0.0201	0.0634	0.1152	0.0210	0.0653	0.1195	0.0174	0.0573	0.1036
5	0.0205	0.0637	0.1169	0.0297	0.0755	0.1275	0.0207	0.0583	0.1079
50	0.0223	0.0661	0.1153	0.2480	0.3547	0.4290	0.0196	0.0606	0.1079
97	0.0217	0.0626	0.1088	1.0000	1.0000	1.0000	0.0233	0.0607	0.1091
	Deterministic Component								
$f_t^F = \sqrt{t}$	0.0280	0.0809	0.1367	0.0255	0.0745	0.1299	0.0195	0.0572	0.1068
t	0.0205	0.0637	0.1169	0.0297	0.0755	0.1275	0.0207	0.0583	0.1079
$t^{3/2}$	0.0317	0.0823	0.1407	0.0314	0.0855	0.1413	0.0224	0.0630	0.1112
t^2	0.0253	0.0685	0.1177	0.0263	0.0742	0.1280	0.0178	0.0508	0.1005
	Serial Correlation								
$\rho = 0$	0.0205	0.0637	0.1169	0.0297	0.0755	0.1275	0.0207	0.0583	0.1079
0.5	0.0216	0.0607	0.1134	0.0278	0.0749	0.1281	0.0199	0.0574	0.1037
0.7	0.0246	0.0720	0.1245	0.0308	0.0812	0.1384	0.0191	0.0590	0.1046
0.9	0.0342	0.0889	0.1404	0.0486	0.1111	0.1745	0.0220	0.0635	0.1111
	Postintervention Periods								
$T_1 = 1$	0.0166	0.0583	0.1061	0.0151	0.0572	0.1099	0.0121	0.0562	0.1027
2	0.0198	0.0631	0.1109	0.0273	0.0685	0.1185	0.0125	0.0566	0.1033
3	0.0205	0.0637	0.1169	0.0297	0.0755	0.1275	0.0207	0.0583	0.1079
4	0.0301	0.0717	0.1247	0.0370	0.0896	0.1467	0.0256	0.0670	0.1151
5	0.0286	0.0686	0.1184	0.0448	0.0933	0.1537	0.0279	0.0650	0.1127

Baseline DGP: (15) and (17) with $T = 100$, independent and identically normally distributed innovations, $n = 200, s_0 = 5, T_1 = 3$ and 10,000 Monte Carlo simulations. The test statistic considered is $\phi(x) = \|x\|_2$. All distributions are standardized (zero mean and unit variance). Mixed normal is equal to 2 Normal distributions with probability (0.3, 0.7), mean $(-10, 10)$ and variance $(2, 1)$. The AR(1) structure with coefficient ρ is applied to the common factor innovation U_{1t}^F and the first unit idiosyncratic innovation U_{1t}^Z . The penalization parameter λ is chosen via the Bayesian information criterion (BIC). We set the maximum penalty level to be $\|\frac{1}{T_0} \sum_{t=1}^{T_0} Y_t X_t\|_\infty$ with an exponential path down to $\lambda_{\min} = 0.001$ along 100 equally spaced intervals in the `glmnet` package. Oracle means OLS estimation in the pre-intervention period with known active regressors S_0 (perfect model selection). True means no estimation in the pre-intervention period. True parameter θ_0 was used.

Table 3. Rejection rates under the null (empirical size): stochastic trends.

	LASSO			Oracle			True		
	0.01	0.5	0.1	0.01	0.05	0.1	0.01	0.05	0.1
	Innovation Distribution								
Normal	0.0324	0.0824	0.1384	0.0319	0.0770	0.1348	0.0220	0.0611	0.1095
$\chi^2(1)$	0.0260	0.0765	0.1385	0.0244	0.0727	0.1308	0.0209	0.0598	0.1060
t-stud(3)	0.0282	0.0831	0.1444	0.0261	0.0779	0.1355	0.0194	0.0581	0.1118
Mixed Normal	0.0357	0.0912	0.1444	0.0330	0.0862	0.1426	0.0208	0.0615	0.1103
	Sample Size								
$T = 50$	0.0566	0.1155	0.1791	0.0512	0.1071	0.1663	0.0247	0.0641	0.1086
100	0.0324	0.0824	0.1384	0.0319	0.0770	0.1348	0.0220	0.0611	0.1095
150	0.0226	0.0686	0.1208	0.0216	0.0664	0.1174	0.0156	0.0526	0.0988
200	0.0193	0.0630	0.1145	0.0190	0.0617	0.1143	0.0156	0.0542	0.1022
500	0.0106	0.0546	0.1026	0.0108	0.0544	0.1010	0.0104	0.0520	0.0966
	Number of Total Units								
$n = 200$	0.0324	0.0824	0.1384	0.0319	0.0770	0.1348	0.0220	0.0611	0.1095
300	0.0391	0.0875	0.1479	0.0274	0.0748	0.1290	0.0184	0.0581	0.1039
500	0.0471	0.0953	0.1520	0.0281	0.0802	0.1358	0.0198	0.0610	0.1088
1000	0.0583	0.1085	0.1575	0.0293	0.0764	0.1300	0.0224	0.0590	0.1042
	Number of Relevant (nonzero) Covariates								
$s_0 = 2$	0.0256	0.0698	0.1272	0.0225	0.0667	0.1213	0.0188	0.0558	0.1054
5	0.0324	0.0824	0.1384	0.0319	0.0770	0.1348	0.0220	0.0611	0.1095
50	0.0497	0.1117	0.1797	0.2541	0.3636	0.4441	0.0174	0.0572	0.1058
97	0.0574	0.1251	0.1950	1.0000	1.0000	1.0000	0.0203	0.0579	0.1060
	Deterministic Component								
$f_t^F = 0$	0.0324	0.0824	0.1384	0.0319	0.0770	0.1348	0.0220	0.0611	0.1095
1	0.0314	0.0815	0.1373	0.0316	0.0815	0.1393	0.0205	0.0615	0.1122
\sqrt{t}	0.0264	0.0693	0.1191	0.0294	0.0814	0.1380	0.0215	0.0605	0.1083
t	0.0265	0.0711	0.1225	0.0292	0.0768	0.1334	0.0184	0.0560	0.1050
	Serial Correlation								
$\rho = 0$	0.0324	0.0824	0.1384	0.0319	0.0770	0.1348	0.0220	0.0611	0.1095
0.5	0.0297	0.0785	0.1313	0.0280	0.0761	0.1320	0.0178	0.0572	0.1019
0.7	0.0275	0.0773	0.1335	0.0264	0.0781	0.1342	0.0211	0.0575	0.1064
0.9	0.0299	0.0752	0.1278	0.0323	0.0823	0.1359	0.0222	0.0631	0.1107
	Postintervention Periods								
$T_1 = 1$	0.0321	0.0753	0.1273	0.0304	0.0714	0.1201	0.0295	0.0690	0.1151
2	0.0289	0.0777	0.1316	0.0271	0.0762	0.1311	0.0219	0.0759	0.1224
3	0.0324	0.0824	0.1384	0.0319	0.0770	0.1348	0.0220	0.0611	0.1095
4	0.0396	0.0930	0.1522	0.0345	0.0879	0.1430	0.0212	0.0608	0.1087
5	0.0516	0.1088	0.1695	0.0464	0.1021	0.1641	0.0293	0.0661	0.1181

NOTE: Baseline DGP: (15) and (16) with $T = 100$, independent and identically normally distributed innovations, $n = 200$, $s_0 = 5$, $T_1 = 3$ and 10,000 Monte Carlo simulations. The test statistic considered is $\phi(x) = \|x\|_2$. All distributions are standardized (zero mean and unit variance). Mixed normal is equal to 2 Normal distributions with probability (0.3, 0.7), mean $(-10, 10)$ and variance (2, 1). The AR(1) structure with coefficient ρ is applied to the common factor innovation U_{1t}^F and the first unit idiosyncratic innovation U_{1t}^Z . The penalization parameter λ is chosen via the Bayesian information criterion (BIC). We set the maximum penalty level to be $\|\frac{1}{T_0} \sum_{t=1}^{T_0} Y_t X_t\|_\infty$ with an exponential path down to $\lambda_{\min} = 0.001$ along 100 equally spaced intervals in the `glmnet` package. *Oracle* means OLS estimation in the pre-intervention period with known active regressors S_0 (perfect model selection). *True* means no estimation in the pre-intervention period. True parameter θ_0 was used.

under three different nominal size values: 0.01, 0.05 and 0.1. The rejection rates are computed for three estimation frameworks: *LASSO* means that the counterfactual is estimated by LASSO with all the n units included in the model. The penalization parameter λ is chosen as described in Section 4. *Oracle* means that the counterfactual is estimated by ordinary least squares (OLS) using only the s_0 relevant units. Finally, *True* means no estimation, that is, the counterfactual is estimated with the true values of the parameters (θ_0). All distributions are standardized (zero mean and unit variance). Mixed normal means two Normal distributions with probability (0.3, 0.7), mean $(-10, 10)$ and variance (2, 1). The autoregressive of order one, AR(1), structure with coefficient ρ is applied to the common factor innovation U_{1t}^F and the first unit idiosyncratic innovation U_{1t}^Z .

Several conclusions emerge from the tables. First, the size distortions of the LASSO are comparable to the ones from the Oracle and slightly superior to those from the true model. Note that the size distortions from the true model reflect only the estimation error of the cumulative distribution of V_t . On the other hand, the other two cases also reflect the estimation

error of θ_0 . Second, it seems that different error distributions do not affect the rejection rates. As expected, the total sample size (T) has a strong influence on the size distortions, which approached close to zero as the sample increased. The number of units (n) seems to have more influence in the case of stochastic trends, where the distortions for the case when $n = 1000$ can be nonnegligible. In addition, high residual autocorrelation, as expected, can cause more distortions. Finally, the number of observations after the intervention also seems to have an effect on the text. However, the distortions are not large. Overall, the proposed inference procedure works extremely satisfactorily, especially for the 0.1 significance level. Furthermore, either potential misspecification of the weights or pretesting for unit roots does not seem to cause any visible harm to the inferential procedure described in the article.

Table 4 presents rejection rates under the alternative for the baseline DGP case. We consider two types of intervention. The first one has only mean effects while the second causes variance effects. It is clear from that the test has nontrivial power against the alternatives.

Table 4. Rejection Rates under the alternative (empirical power).

		<i>Deterministic Trends</i>									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
		Mean intervention $\delta_t = c\sigma 1\{t > T_0\}$									
$c = 0.2$		0.10	0.12	0.14	0.16	0.17	0.19	0.20	0.22	0.23	0.25
0.4		0.23	0.27	0.32	0.35	0.37	0.40	0.43	0.46	0.47	0.48
0.6		0.48	0.51	0.56	0.60	0.63	0.65	0.67	0.69	0.70	0.71
0.8		0.76	0.79	0.82	0.86	0.88	0.89	0.91	0.91	0.92	0.93
1.0		0.94	0.95	0.97	0.97	0.97	0.98	0.98	0.98	0.99	0.99
		Variance intervention $\delta_t = c\sigma Z1\{t > T_0\}$ where $Z \sim N(0, 1)$									
$c = 0.2$		0.09	0.12	0.13	0.15	0.17	0.18	0.20	0.22	0.24	0.25
0.4		0.26	0.29	0.32	0.36	0.38	0.39	0.41	0.44	0.46	0.48
0.6		0.50	0.54	0.58	0.63	0.66	0.69	0.70	0.71	0.73	0.74
0.8		0.78	0.81	0.85	0.88	0.89	0.91	0.92	0.92	0.92	0.93
1.0		0.93	0.95	0.96	0.97	0.97	0.97	0.98	0.98	0.99	0.99
		<i>Stochastic trends</i>									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
		Mean intervention $\delta_t = c\sigma 1\{t > T_0\}$									
$c = 0.1$		0.19	0.20	0.24	0.28	0.30	0.32	0.33	0.36	0.38	0.39
0.2		0.63	0.67	0.72	0.73	0.76	0.78	0.80	0.81	0.81	0.83
0.3		0.95	0.96	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
0.4		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Variance intervention $\delta_t = c\sigma Z1\{t > T_0\}$									
$c = 0.1$		0.17	0.20	0.22	0.25	0.27	0.30	0.32	0.33	0.35	0.37
0.2		0.57	0.60	0.65	0.68	0.70	0.72	0.75	0.76	0.78	0.79
0.3		0.91	0.92	0.94	0.96	0.96	0.97	0.97	0.98	0.98	0.98
0.4		0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		<i>Mixed Trends</i>									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
		Mean intervention $\delta_t = c\sigma 1\{t > T_0\}$									
$c = 0.1$		0.20	0.22	0.22	0.30	0.31	0.33	0.35	0.36	0.38	0.39
0.2		0.61	0.67	0.75	0.78	0.78	0.79	0.81	0.82	0.82	0.82
0.3		0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
0.4		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Variance intervention $\delta_t = c\sigma Z1\{t > T_0\}$									
$c = 0.1$		0.20	0.22	0.22	0.26	0.28	0.31	0.33	0.34	0.38	0.39
0.2		0.62	0.60	0.65	0.68	0.70	0.72	0.75	0.76	0.78	0.79
0.3		0.95	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98
0.4		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

NOTE: Baseline DGP: (5) and (4) with $T = 100$, iid normally distributed innovations, $n = 200$ units, $s_0 = 5$, $T_1 = 3$ and 10,000 Monte Carlo simulations per case. Empirical rejection rate of the test statistic $\phi(x) = \|x\|_2$. The penalization parameter λ is chosen via the Bayesian Information Criteria (BIC). We set the maximum penalty level to be $\frac{1}{T_0} \sum_{t=1}^{T_0} Y_t X_t$ with an exponential path down to $\lambda_{\min} = 0.001$ along 100 equally spaced intervals in the `glmnet` package. σ^2 is the variance of unit 1 at $t = T_0$.

Results concerning parameter estimation are reported in the supplementary material.

6. Empirical Illustration

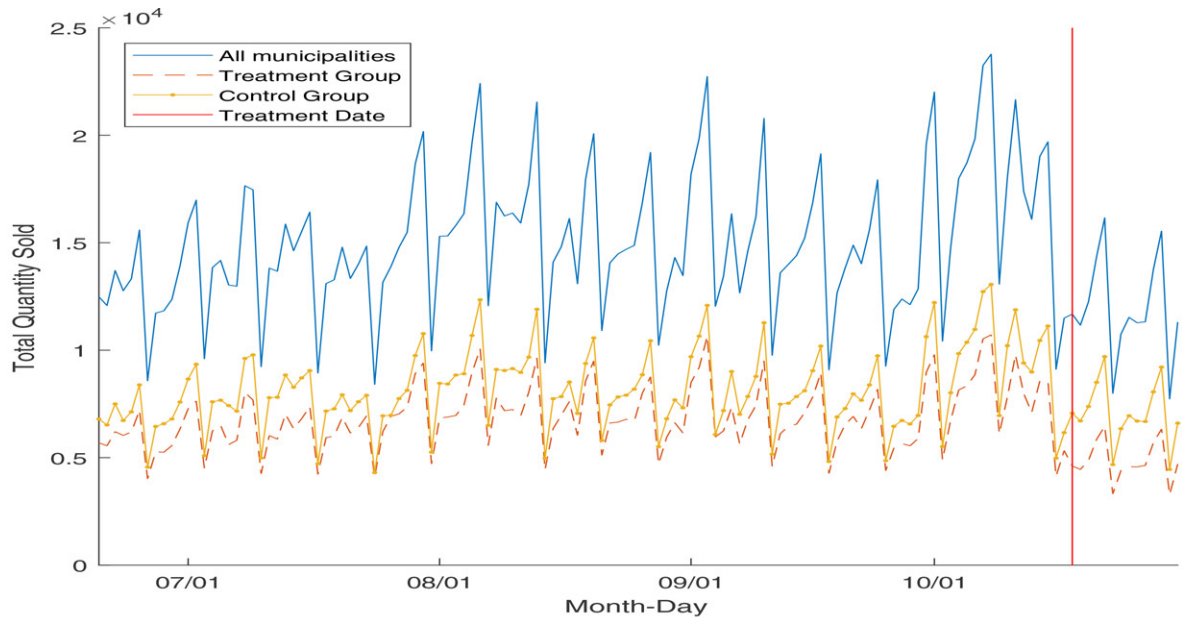
We consider an application to optimal price setting in the retail industry in Brazil. Our dataset consists of the daily prices and quantities sold of a product commercialized by one of the major retail chains in Brazil, which has approximately 1,400 stores distributed in more than 700 municipalities over the country. Due to a confidentiality agreement, we are not allowed to disclose either the name of the product or the name of the retail chain. On average, the company sells more than 29,000 units of this product per day across the country, which represents an important share of the company’s total revenue. The quantities are aggregated at the municipal level. Our sample consists of approximately 30% of the municipalities where there are stores. The number and size of stores differ across municipalities.

To determine the optimal price of the product (in terms of profit or revenue maximization), a randomized experiment has been carried out. The price of the product was changed in 107 municipalities (treatment group), while in the other 126 municipalities, the prices were kept fixed at the original level (control group). As a different experiment was running during the same period in the remaining municipalities, we decided to exclude these cities in order to avoid potential sources of biases. The selection of the treatment and control groups was carried out according to socioeconomic and demographic characteristics of each municipality as well as to the distribution of stores in each city. Nevertheless, it is important to emphasize three facts. First, we used no information about the quantities sold of the product in each municipality, which is our output variable, in the randomization process. This way, we avoid any selection bias and can maintain the validity of **Assumption 1**. Second, although according to municipality characteristics, we keep a homogeneous balance between groups, the parallel trend hypothesis is violated, and there is strong heterogeneity with

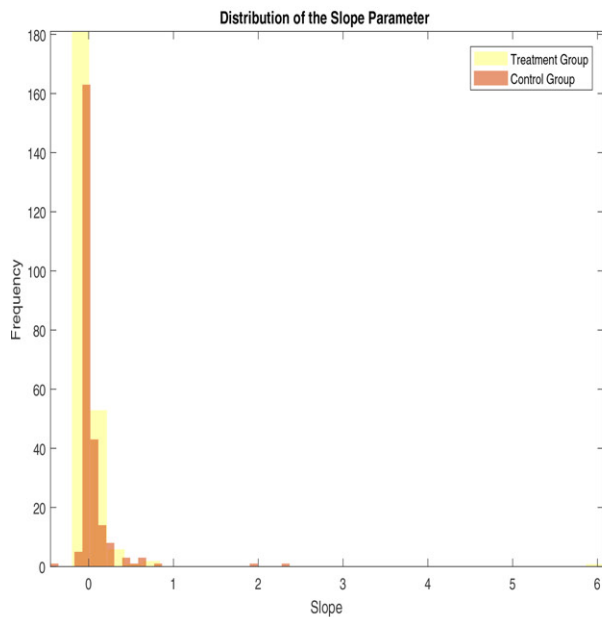
respect to the quantities sold and consumer behavior in each city, even after controlling for observables. Finally, the time series of sold quantities displays a clear heterogeneous trend. Due to the previously described facts, we advocate the use of the methodology proposed in this article.

For each day t , q_{it} represents the total quantities sold of the product in all stores of municipality i , where $i = 1, \dots, n$ and $t = 1, \dots, T$. Our sample runs from June 20, 2016, to October

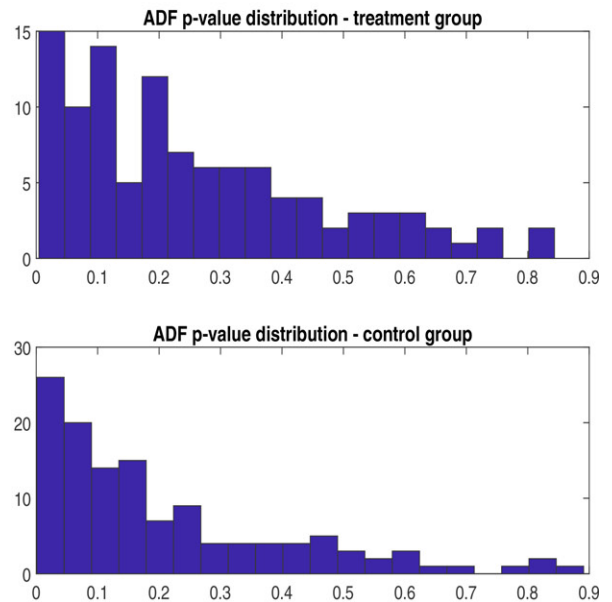
31, 2016, a total of 134 daily observations. The experiment was conducted during the period October 18–31 (14 days). During these days, the practiced prices in the municipalities belonging to the treatment group were increased in Δ_p Brazilian Reais, while for the other municipalities, they were kept fixed. The first 126 municipalities are in the control group ($i = 1, \dots, 126$), whereas the remaining 107 are in the treatment group ($i = 127, \dots, 233$). The number of pretreatment



(a) Time-series dynamics of the sold quantities



(b) Histograms of the slope parameter



(c) Histograms of the ADF test

Figure 1. Quantities sold.

NOTE: Panel (a) displays the daily evolution of total quantities sold in all municipalities and in the treatment and control groups. The sample period runs from June 20, 2016, to October 31, 2016. The experiment starts in October 18, 2016, and ends in October 31, 2016 (14 observations). The starting date of the experiment is represented by the vertical red line. Panel (b) shows the estimated slope coefficients in a pure linear trend model for the quantities sold in each municipality during the pretreatment sample. Panel (c) displays the histogram of the p -value of the augmented Dickey-Fuller test for the null of unit roots against the alternative of a trend-stationary model applied to the quantity sold in each municipality.

observations is $T_0 = 120$. Panel (a) in Figure 1 presents the time-series dynamics of the total quantity sold over all municipalities as well as in the control and treatment groups. Some facts emerge from the visual inspection of the figure. First, there is a clear trend in the data. Second, there is also a strong weekly pattern. Panel (b) in Figure 1 displays the histograms of the estimated slope parameter of a pure linear trend model for the municipalities in the control and treatment groups during the pretreatment sample. For each municipality, we estimate by OLS the following linear trend model: $q_{it} = \alpha_i + \theta_i t + u_{it}$. Panel (b) in Figure 1 displays the empirical distribution of $\hat{\theta}$ across municipalities. Panel (c) shows the histogram of the p -values of the augmented Dickey-Fuller (ADF) test for the null of unit roots against the alternative of a trend-stationary model, applied to the series of quantities sold in each municipality of the control and treatment groups. There is a clear heterogeneity in the trend pattern that precludes the use of the traditional differences-in-differences estimator as well as the methodology put forward in Carvalho, Masini, and Medeiros (2018).

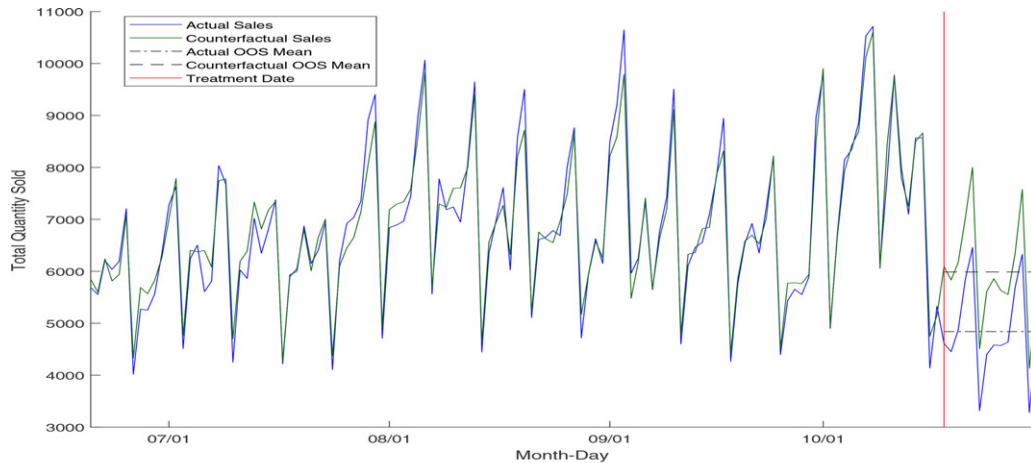
To determine the optimal price, it is necessary to obtain the effects of the price change on the quantities sold. We consider two cases. In the first case, we assume that the effects are homogeneous across municipalities, and our output variable of interest is the total quantity of the product sold in the treatment group: $q_t = \frac{1}{107} \sum_{i=126}^{233} q_{it}$.

We estimate the effect according to the following steps:

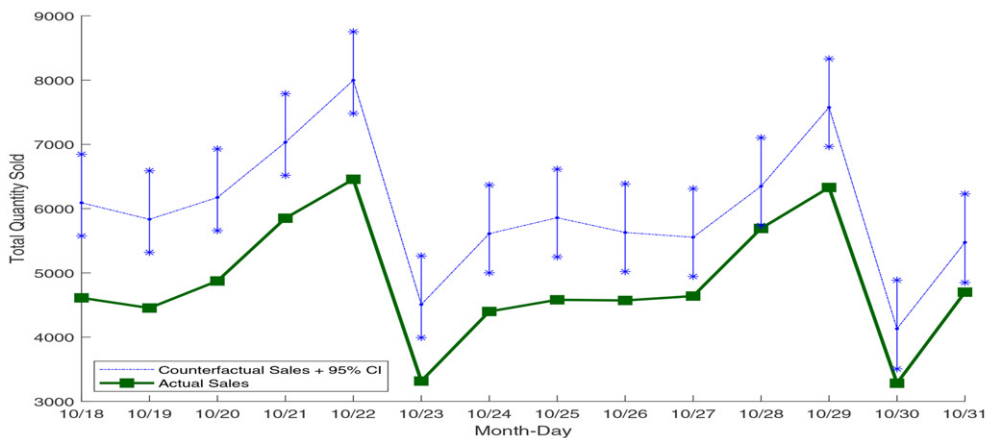
1. Estimate the parameters of the regression

$$q_t = \theta_0 + \sum_{i=1}^{126} \theta_i q_{it} + \pi_1 \text{Mon}_t + \pi_2 \text{Tue}_t + \pi_3 \text{Wed}_t + \pi_4 \text{Thu}_t + \pi_5 \text{Fri}_t + \pi_6 \text{Sat}_t + V_t = \mathbf{X}'_t \boldsymbol{\theta} + V_t$$

by the WLASSO procedure using the 120 observations from June 20, 2016, to October 18, 2016 (pretreatment sample). $\text{Mon}_t, \dots, \text{Sat}_t$ are six dummies for the days of the week. As we include a constant in the model, we omit the dummy for



(a) Actual and counterfactual sales



(b) Actual and counterfactual sales during treatment period

Figure 2. Actual and counterfactual sales.

NOTE: Panel (a) shows the aggregated actual and counterfactual sales over the pretreatment and posttreatment periods. The sample period runs from June 20, 2016, to October 31, 2016. The experiment starts in October 18, 2016, and ends in October 31, 2016 (14 observations). The starting date of the experiment is represented by the vertical red line. Panel (b) shows the aggregated actual and counterfactual sales for the posttreatment period. Confidence intervals of 95% for the counterfactual path are also displayed.

Table 5. Results.

	Panel (a): Aggregated	Panel (b): Disaggregated			
		Mean	Std. Dev.	Max.	Min.
Δ	-1,147	-12.90	52.08	5.52	-526.70
Δ / #stores	-4.33	-4.21	4.42	5.52	-23.27
p -value (square)	0	0.41	0.29	1	0
p -value (absolute)	0	0.36	0.31	1	0
Proportion (%) of rejection of the null (square)	NA	19	NA	NA	NA
Proportion (%) of rejection of the null (absolute)	NA	31	NA	NA	NA
Number of regressors	133	133	NA	NA	NA
Number of relevant regressors	26	9.46	8.06	72	0
Number of pretreatment observations	120	120	NA	NA	NA
Number of observations during treatment	14	14	NA	NA	NA

NOTE: The table reports estimation results. Panel (a) shows the average treatment effect Δ for all stores in the treatment group over the treatment period. The average effect per store is also reported (Δ / #stores), where #stores is the number of stores in the treatment group. p -value (square) and p -value (absolute) represent the p -values of the resampling-based test with $\phi(x) = \frac{1}{T_1} \sum_{j=1}^{T_1} x_j^2$ and $\phi(x) = \frac{1}{T_1} \sum_{j=1}^{T_1} |x_j|$, respectively.

Sundays. The penalty parameter of the WLASSO procedure is selected by the BIC.

- Project the counterfactual as $\hat{q}_t = X_t' \hat{\theta}$ and compute $\delta_t = q_t - \hat{q}_t$, for $t > T_0$.

We evaluate the effects on sales during each one of the 14 days following the initial price increase. The results are reported in Figure 2 and Table 5. The figure shows the actual sales, the estimated counterfactual, and a 95% confidence interval using the partial resampling method described in Section 3.2, where $\phi(x) = x$. As expected, the effects are negative and statistically significant for most of the days. We also run the resampling test for $\phi(x) = \frac{1}{T_1} \sum_{j=1}^{T_1} x_j^2$ and $\phi(x) = \frac{1}{T_1} \sum_{j=1}^{T_1} |x_j|$. Table 5, Panel (a), reports the average effect for all municipalities in the treatment group as well as the effect per store. Extrapolating the result for the entire company, the average daily effect yields a reduction in sales of more than 4000 units, potentially causing a great impact in terms of revenue and profit. The table also reports the number of selected regressors with the WLASSO method.

To measure the degree of heterogeneity of price elasticities across different municipalities, we estimate the counterfactuals for each one of the municipalities in the treatment group. We replace (18) by the following model:

$$\begin{aligned}
 q_{jt} &= \theta_{k0} + \sum_{i=1}^{126} \theta_{ki} q_{it} + \pi_{k1} \text{Mon}_t + \pi_{k2} \text{Tue}_t + \pi_{k3} \text{Wed}_t \\
 &\quad + \pi_{k4} \text{Thu}_t + \pi_{k5} \text{Fri}_t + \pi_{k6} \text{Sat}_t + V_{jt}, \\
 &= X'_{jt} \theta_k + V_{jt}, \quad j = 126, \dots, 233; \quad k = j - 126.
 \end{aligned}$$

The results are displayed in Panel (b) of Table 5. The table reports the mean, standard deviation, maximum and minimum of the average daily effects for each municipality as well as the effects normalized by the number of stores in each city in the treatment group. The table also reports the mean, standard deviation, maximum and minimum of the p -value of the resampling test conducted with $\phi(x) = \frac{1}{T_1} \sum_{j=1}^{T_1} x_j^2$ and $\phi(x) = \frac{1}{T_1} \sum_{j=1}^{T_1} |x_j|$ and the proportion of municipalities where the null of no effect has been rejected. For the squared test, in 19% of the cities, the increase in prices negatively affected the demand for the product, whereas according to the absolute test, the effects are negative and significant in 30% of the municipalities.

7. Conclusions

We discussed a flexible method to conduct counterfactual analysis with aggregate data, which is particularly relevant in situations where there is a single treated unit and “controls” are not available, such as in regional policy evaluation. The setup considered in the article allows for potentially high-dimensional and nonstationary data displaying deterministic and/or stochastic trends. We proposed a weighted version of the LASSO for parameter estimation in a high-dimensional linear regression framework, which is consistent under very general assumptions. Furthermore, we showed the consistency of the average intervention effect (over postintervention observations), and we also developed an inferential procedure based on partial resampling to test the general hypothesis on the intervention effects. Our testing procedure does not rely on postintervention asymptotics.

Appendix A: Omitted Technical Assumptions and Results

This appendix collects more technical details such as definitions, assumptions and auxiliary results that lead to our main result (Theorem 1).

A.1. Definitions

We reproduce here the definition in Davidson (2009).

Definition 1 (I(0) process). A generic scalar process $\{G_t\}$ is said to be $I(0)$, denoted $G_t \sim I(0)$, if

$$\mathcal{G}_T := \mathcal{G}_T(s) := \frac{1}{\nu_T} \sum_{t=1}^{\lfloor Ts \rfloor} [G_t - \mathbb{E}(G_t)] \Rightarrow B,$$

where $\nu_T^2 := \mathbb{E} \left\{ \sum_{t=1}^T [G_t - \mathbb{E}(G_t)]^2 \right\}$ and $B := \{B(s), s \in [0, 1]\}$ is a standard Wiener process.

From the definition above, stationarity is not (even weakly) required for a process to be $I(0)$. However, deterministic trends are not allowed, and summability of the covariance is necessary. Otherwise, if any of those conditions are violated, we could not have $\nu_T^2 \sim cT$ for $0 < c < \infty$, which is necessary to ensure that $\mathbb{E}[B(s) - B(r)]^2 = s - r$ for $0 \leq s \leq r \leq 1$.

As is common in the literature on high dimensionality, we need a certain compatibility between the norms $\|\cdot\|_1$ and $\|\cdot\|_{\Sigma}$. Many options are available in the literature, but they are usually in the form of a lower bound of the minimum of the eigenvalue of Σ . Here, we follow van de Geer and Bühlmann (2009) and Huang and Zhang (2012).

Definition 2 (Compatibility Constant). For a $(p \times p)$ (possibly stochastic) matrix M , a set $S \subseteq \{1, \dots, p\}$ and a scalar $\xi \geq 0$, the compatibility constant is given by

$$\chi(M, S, \xi) := \inf \left\{ \frac{\|x\|_M \sqrt{|S|}}{\|x_S\|_1} : x \in \mathbb{R}^p : \|x_{S^c}\|_1 \leq \xi \|x_S\|_1 \right\}. \tag{A.1}$$

Moreover, we say that (M, S, ξ) satisfies the compatibility condition if $\chi(M, S, \xi) > 0$.

Notice that the square of the compatibility constant is the minimum ℓ_1 -eigenvalue of Σ restricted to a cone in \mathbb{R}^p . Moreover, we allow for random Σ in the definition since, as opposed to the deterministic trend case, the Σ does not converge to a deterministic matrix in the pure stochastic trend case. In a low-dimensional setup Masini and Medeiros (2019), shows that Σ converges in distribution to a positive definite random matrix almost surely.

A.1.1. Growth condition

To understand the link between the sequence f_{it} appearing in Assumption 2 and the trend it generates, it is worth considering the continuous version of f_{it} given by $f_i(t)$, such that

$$a_{it} := \sum_{s=1}^t f_{is} = O \left[\int f_i(t) dt \right], \quad \text{for integrable } f_i(t) : \mathbb{R} \rightarrow \mathbb{R}^+. \tag{A.2}$$

Therefore, if $f_i(t) = O(t^c)$, with $c \in \mathbb{R}$, we have $a_{it} = O(t^{c+1})$ for $c \neq 0$. When $c = -1$, we have $a_{it} = O(\log t)$. Model (4) covers a wide class of trend patterns depending on the choice of the sequence $\{f_{it}\}$, including (we drop the subscript i in what follows):

- No trend:** $tf_t \rightarrow 0$, which implies $a_t \rightarrow 0$ as $t \rightarrow \infty$.
- Sublinear:** $f_t \rightarrow 0$ and $tf_t \rightarrow \infty$, which implies $a_t/t \rightarrow 0$ as $t \rightarrow \infty$.
- Linear:** $f_t \rightarrow c > 0$, which implies $a_t \rightarrow ct$ as $t \rightarrow \infty$.
- Subexponential:** $f_t \rightarrow \infty$ and $f_t/\exp ct \rightarrow 0$, for any $c > 0$. Thus, $a_t/\exp ct \rightarrow 0$ as $t \rightarrow \infty$.
- Exponential:** $f_t \rightarrow c_1 \exp c_2 t$, which implies $a_t \rightarrow c_1/c_2 \exp c_2 t$ as $t \rightarrow \infty$ for some $c_1, c_2 > 0$.
- Superexponential:** $f_t/(c_1 \exp c_2 t) \rightarrow \infty$. Therefore, $a_t/c_1 \exp c_2 t \rightarrow \infty$ as $t \rightarrow \infty$ for $c_1, c_2 > 0$.

It is important to understand under which conditions the stochastic part of (11) is asymptotically dominated by the deterministic one, in the sense that $Z_{it}^{(0)}/d_{it} \rightarrow 1$, almost surely or in probability. For (5), this is always the case as long as $f_{it} \rightarrow \infty$, which implies $|d_{it}| \rightarrow \infty$. For (4), since the variance of η_{it} increases as $t \rightarrow \infty$, it is no longer enough to have $a_{it} \rightarrow \infty$. In fact, since $\eta_{it} = O_p(\sqrt{t})$, a_{it} must be of an order higher than \sqrt{t} . This is ensured, for instance, by taking $f_{it} = t^c$ with $c > -1/2$. As an illustration, taking a random walk with drift as an example, $Z_{it} = \mu_i t + \sum_{s=1}^t U_{is}$. Then, $d_{it} = \mu_i t$, and we have $Z_{it}/d_{it} = 1 + \sum_{s=1}^t U_s/\mu_i t \rightarrow 1$, almost surely or in probability, depending on the law of large numbers, which is available for the process $\{U_t\}$. We formalize those facts in the following proposition, which also gives the definition of *Growth condition*, which will be used later on.

Proposition 1. Consider the DGPs in Assumption 2, assuming that $\{U_t\}$ fulfills Assumption 3 for $1 \leq i \leq n$. Therefore, as $t \rightarrow \infty$,

1. **(Growth Condition)** $Z_{it}^{(0)}/d_{it} \rightarrow 1$ in probability under DGP (4) if $\sqrt{t}/d_{it} = o(1)$; or $Z_{it}^{(0)}/d_{it} \rightarrow 1$ almost surely under DGP (5) if $f_{it} \rightarrow \infty$
2. **(No-Growth)** $Z_{it}^{(0)} = O_p(\sqrt{t})$ under DGP (4) if $d_{it}/\sqrt{t} = O(1)$; or $Z_{it}^{(0)} = O_p(1)$ under DGP (5) if $f_{it} = O(1)$.

Moreover, for (5), if $d_{it} = o(\sqrt{t})$, then $t^{-1/2}Z_{it}^{(0)}$ converges in distribution to a Gaussian random variable.

A.2. Assumptions

Assumption 3 deals with the tradeoff between moment conditions and serial dependency. The exponential decay of the strong mixing coefficient ensures that the q -th moment of the sum of the zero-mean strong mixing variables is of order $T^q/2$. The exponential decay allows us to invoke a result from Merlevède, Peligrad, and Rio (2009) and derive a Bernstein-type inequality that, combined with condition (b), results in an exponential bound for the sum of innovations.

Assumption 3 (Moments and Dependency). $\{U_t\}_{t \geq 1}$ is a zero-mean strong mixing sequence of n -dimensional random vectors with a mixing coefficient given by $\alpha(m) = \exp(-2cm)$ for some $c > 0$ fulfilling one of the conditions:

1. There exists a real $q > 2$ such that $\sup \{ \mathbb{E}|U_{it}|^{q+\epsilon} : 1 \leq i \leq n, t \in \mathbb{N} \} < \infty$, for $\epsilon > 0$;
2. There exist real numbers $c_1, c_2, c_3 > 0$ such that $\sup \{ \mathbb{P}(|U_{it}| > u) : 1 \leq i \leq n, t \in \mathbb{N} \} \leq c_1 \exp(-c_2 u^{c_3})$, for all $u > 0$.

In both cases, the smallest eigenvalue of the matrix $\mathbb{E}(U_t U_t')$ is bounded away from 0 uniformly in $t \in \mathbb{N}$.

Clearly, Assumption 3(b) implies (a) for all $q > 0$. The converse is not true even if (a) holds for all $q > 2$. We employ (a) to deal with fat tails, whereas we use (b) to handle subexponential growth of units in case of exponential decay of the tails. This includes sub-Gaussian ($c_3 \geq 2$), subexponential ($c_3 \geq 1$) and many other families of distributions of interest. Finally, we bound from below the smallest eigenvalue to ensure that $\mathbb{E}(Z_t Z_t')$ properly scaled is full rank and, therefore, avoid multicollinearity among the regressors.

From Proposition 1, we conclude that whether the DGP will satisfy the growth condition depends on the growth rate of d_{it} . For (4), the growth condition depends on whether $\sqrt{t}/d_{it} \rightarrow 0$ or not. For (5), the growth condition does not hold if $f_{it} \rightarrow c < \infty$. Therefore, to estimate (9) in high dimensions, we need to impose a separation between those two regimes as the number of units increases. Consider the following assumptions.

Assumption 4 (Strong Separation). Let $\mathcal{H} \subseteq \{1, \dots, n\}$ be the index set of units $\{Z_{it}^{(0)}, 1 \leq i \leq n\}$ that fulfill the **growth condition** of Proposition 1 and $d_{\mathcal{H}}(T_0) := \inf_{i \in \mathcal{H}} |d_{i,T_0}|$. Then, for (4) in Assumption 2, assume

$$\frac{\psi(|\mathcal{H}|\sqrt{T_0})}{d_{\mathcal{H}}(T_0)} = o(1) \quad \text{for DGP (4) in Assumption 2; or}$$

$$\frac{\psi(|\mathcal{H}|)}{d_{\mathcal{H}}(T_0)} = o(1) \quad \text{for DGP (5) in Assumption 2,}$$

where $\psi(x) = x^{1/q}$ under Assumption 3(a) and $\psi(x) = \log(x)$ under Assumption 3(b).

Assumption 5 (Hyperparameter Tuning). For some $c > 0$, set the penalty parameter λ of (10) by either

1. $\lambda = 4cp^{2/q}/\sqrt{T_0}$ under Assumption 3(a); or
2. $\lambda = 4(c + 2 \log p)/\sqrt{T_0}$ under Assumption 3(b).

Additionally, set $w_1 = 0$ such that the intercept is not penalized and, for $2 \leq i \leq p$, set

1. $w_i = |X_{iT_0}|$ under the growth condition (Proposition 1(b)) or
2. $w_i = 1$ if DGP (4) or $\sqrt{T_0}$ if DGP (5) in Assumption 2

Assumption 6 (Rates). Assume

1. $\|\theta_0\|_1 \leq c$ for some $c < \infty$
2. $\log p = o[(T_0^{1/4} / \log T_0)^{c_3}]$ under Assumption 3(b)
3. $\frac{\psi(p)^{2-b} R_b}{T_0^{(1-b)/2} \lambda_1} = o(1)$ for some $b \in [0, 1]$ where $\chi_2(\Sigma, S_b, \xi) = O_p(\lambda_1)$

where $\psi(x) = x^{1/q}$ under Assumption 3(a) and $\psi(x) = \log(x)$ under Assumption 3(b).

Assumption 6(a) is sufficient to bound the moments of V_t in terms of the moments of U_t in Lemma 1. Part (b) is necessary to apply the second part of Lemma 2 in the proof of Lemma 1. Finally, (c) states the rate conditions to ensure the consistency of the estimator. More importantly, it implies a lower (probability) bound of the compatibility constant and deserves some clarification. For the case of deterministic trends, DGP (5), under mild conditions, we have that $\|\Sigma - \mathbb{E}(\Sigma)\|_{\max}$ vanishes in probability. Therefore, we could replace the condition by a deterministic one in terms of a deterministic matrix $\mathbb{E}(\Sigma)$. Hence, we require only the existence of a constant $\lambda_1 > 0$, such that $\chi(\mathbb{E}(\Sigma), S_b, \xi) \geq \lambda_1$, where $\chi(\mathbb{E}(\Sigma), S_b, \xi)$ is given by Definition 2.

Unfortunately, that is no longer true for the stochastic trends, DGP (4), as Σ fails to converge to a deterministic matrix. The event $\{\chi^2(\Sigma, S_b, \xi) \geq \lambda_1\}$, however, is expected to hold with high probability as long as we pick $\lambda_1 > 0$ small enough or vanishing at an appropriate rate. Note such an event imposes an indirect restriction on the number of cointegration relations $r \in \{0, \dots, n - 1\}$ that could exist among the n units. It is not difficult to check that the rank of the stochastic matrix Σ is lower bounded by $\max[\min(p, T) - r, 1]$ almost surely. Hence, when p grows faster than T and r grows faster than T , we might have the rank of Σ approximately equal to 1. Therefore, it is unlikely that the event $\{\chi^2(\Sigma, S_b, \xi) \geq \lambda_1\}$ will hold with high probability regardless of the choice of $\lambda_1 \geq 0$. On the other hand, if $r = o(T)$, it is plausible to expect, as in the fixed design case, that the minimum restricted eigenvalue of Σ is bounded away from zero with high probability.

A.3. Results

A.3.1. Oracle Inequalities

For $\mathcal{S} \subseteq \{1, \dots, p\}$ and scalars $\lambda_0, \lambda_1 > 0, a \geq 0, b \in [0, 1]$ and $\lambda_2 \in (0, 1)$ consider the following auxiliary events:

$$\Omega_0 := \left\{ \left\| \frac{2}{T_0} \sum_{t=1}^{T_0} W_t V_t \right\|_{\infty} \leq \lambda_0 \right\}, \tag{A.3}$$

$$\Omega_1(\mathcal{S}) := \{R_b^a \chi^2(\Sigma, \mathcal{S}, \xi) \geq \lambda_1\}, \tag{A.4}$$

$$\Omega_2(\mathcal{S}) := \left\{ \sup_{i \in \mathcal{S}} v_i \leq 1 + \lambda_2 \right\} \cap \left\{ \inf_{i \in \mathcal{S}^c} v_i \geq 1 - \lambda_2 \right\}. \tag{A.5}$$

Then, we have the following oracle inequality

Proposition 2. Then, on the event $\Omega_0 \cap \Omega_2$ and provided that $\lambda > \lambda_0$, the following inequality holds for all $\delta \in [0, 1), \mathcal{Y} \in \mathbb{R}^p$ and $\mathcal{S} \subseteq \{1, \dots, p\}$:

$$\begin{aligned} \|\widehat{\mathcal{Y}} - \mathcal{Y}_0\|_{\Sigma} + 2\delta \underline{\lambda} \|\widehat{\mathcal{Y}} - \mathcal{Y}\|_1 &\leq \|\mathcal{Y} - \mathcal{Y}_0\|_{\Sigma} + \frac{\bar{\lambda}^2 |\mathcal{S}|}{\chi^2(\Sigma, \mathcal{S}, \xi)} \\ &\quad + 4\lambda (\|\mathcal{Y}_{\mathcal{S}^c}\|_v \vee \|\mathcal{Y}_{\mathcal{S}^c}\|_1), \end{aligned} \tag{A.6}$$

where $\underline{\lambda} := \lambda(1 - \lambda_2) - \lambda_0, \bar{\lambda} := \lambda(1 + \lambda_2) + \lambda_0 + \delta \underline{\lambda}, \xi := \bar{\lambda}((1 - \delta)\underline{\lambda})^{-1}$ and $\Sigma := \frac{1}{T_0} \sum_{t=1}^{T_0} W_t W_t'$. Additionally, the right-hand side is taken to be $+\infty$ whenever $\chi(\Sigma, \mathcal{S}, \xi) = 0$.

If we set $\mathcal{S} = \mathcal{S}_0 := \{j : |\gamma_{0j}| > 0\}$, that is, the set of active regressors, then

$$\|\widehat{\mathcal{Y}} - \mathcal{Y}_0\|_{\Sigma} + 2\delta \underline{\lambda} \|\widehat{\mathcal{Y}} - \mathcal{Y}_0\|_1 \leq \frac{\bar{\lambda}^2 |\mathcal{S}_0|}{\chi^2(\Sigma, \mathcal{S}_0, \xi)}.$$

Proposition 3. Let $\mathcal{S}_b := \{j : |\gamma_j^0| > \frac{\bar{\lambda}^2}{\lambda} \mathbb{1}\{b > 0\}\}$ for $b \in [0, 1]$. Then, under the same conditions of Proposition 2, for any $b \in [0, 1]$:

$$\begin{aligned} \|\widehat{\mathcal{Y}} - \mathcal{Y}_0\|_{\Sigma} + 2\delta \underline{\lambda} \|\widehat{\mathcal{Y}} - \mathcal{Y}_0\|_1 \\ \leq \left[\frac{1}{\chi^2(\Sigma, \mathcal{S}_b, \xi)} + 4(1 + \lambda_2) \right] \bar{\lambda}^{2(1-b)} \lambda^b R_b. \end{aligned} \tag{A.7}$$

Therefore, if we set $\lambda = k\lambda_0$ for some $k > 1/(1 - \lambda_2)$, then on $\Omega_0 \cap \Omega_1 \cap \Omega_2$:

$$\|\widehat{\mathcal{Y}} - \mathcal{Y}\|_1 \leq C_1 \left[\frac{1}{\chi^2(\Sigma, \mathcal{S}_b, \xi)} + 4(1 + \lambda_2) \right] \lambda^{1-b} R_b, \tag{A.8}$$

and $\mathcal{S}_b := \{j : |\gamma_j^0| > C_2 \lambda\}$ where both $C_1, C_2 > 0$ are constants depending only on b , the constant λ_2 and the choice of $\delta \in (0, 1)$ and k given by

$$\begin{aligned} C_1 &:= \frac{[1 + \delta + (1 - \delta)(\lambda_2 + 1/k)]^{2(q-1)}}{2\delta(1 - \lambda_2 - 1/k)} \quad \text{and} \\ C_2 &:= \mathbb{1}\{b > 0\} [1 + \delta + (1 - \delta)(\lambda_2 + 1/k)]^2. \end{aligned}$$

A.3.2. Probability Bounds on the events Ω_0 and Ω_2

Lemma 1. Under assumption, set $\lambda_0 = \lambda/2$, then

$$\begin{aligned} \mathbb{P}(\Omega_0^c) &= \mathbb{P} \left(\left\| \frac{1}{T_0} \sum_{t=1}^{T_0} W_t V_t \right\|_{\infty} > \frac{\lambda_0}{2} \right) \\ &= \begin{cases} O(c^{-q/2}) & \text{under Assumption 3(a)} \\ O[\exp(-c/2)] & \text{under Assumption 3(b)}. \end{cases} \end{aligned}$$

In the setup where all the regressors are stationary, the event Ω_2 happens with probability 1 by setting $w_i = 1$ for all $1 \leq i \leq p$. In the factor model example, setting $w_i = 1$ for all units results in Ω_2 occurring surely regardless of the factor DGP considered and/or the deterministic trend associated with it. Since, in that case, we have that $v_i \leq 1$ for $i \in \mathcal{S}_0$ and $v_i = 1$ otherwise. This fortunate result is a consequence that all regressors that do not load on the factor are $I(0)$ processes. The same would be true whenever the process of the units in \mathcal{S}_0^c is of smaller or equal order in probability of the process in variables in \mathcal{S}_0 .

To extend this result to the general setup, let $w_i = w_{i,t}$ be a possibly stochastic sequence of almost surely nonnegative weights. Then, the event Ω_2 happens with probability approaching 1 as long as $\{\limsup_t \sup_{i \in \mathcal{S}} v_{i,t} \leq 1\}$ and $\{\liminf_t \inf_{i \in \mathcal{S}^c} v_{i,t} \geq 1\}$, where $v_{i,t} := w_{i,t}/\ell_{i,t}$ also happens with probability approaching one. If we choose the vector of weights w in (10) according to Assumption 5, the event Ω_2 occurs with probability approaching one since by the definition of ℓ_i in (12), $v_i \rightarrow 1$ for all $2 \leq i \leq p$. For the case when the growth condition holds for X_{it} , we would like to penalize it setting $w_{it} = d_{i,T_0}$. However, since we do not directly observe it, we are using X_{i,T_0} instead, and we are able to state the following result.

Lemma 2. Under the same conditions of Theorem 1 we have $\mathbb{P}(\Omega_2) \rightarrow 1$ as $T_0 \rightarrow \infty$.

Acknowledgments

The work of Marcelo C. Medeiros is partly funded by CNPq and CAPES. The authors gratefully acknowledge the invaluable comments and guidance of the guest coeditors, Alberto Abadie and Matias Cattaneo as well as three anonymous referees. The authors are thankful for the comments from Frank Diebold, Jianqing Fan, Marcelo Fernandes, Guido Imbens, Anders B. Kock, Sophocles Mavroeidis, Eduardo F. Mendes, Pedro Souza, Normam Swanson, Michael Wolf, and participants during seminars at Princeton University, Rutgers University, University of Pennsylvania, Warwick University, Oxford University, São Paulo School of Economics, Pontifical Catholic University of Rio de Janeiro, and University of Brasilia as well as during the 2018 Latin American Meeting of the Econometric Society, Guayaquil, Ecuador and the Barcelona GSE summer forum. A special acknowledgment goes to Étienne Wijler for insightful and technical discussions.

Supplementary Material

The supplementary material contains additional simulation results, all the proofs for the theoretical results in the paper and a list of symbols used.

References

- Abadie, A., and Gardeazabal, J. (2003), “The Economic Costs of Conflict: A Case Study of the Basque Country,” *American Economic Review*, 93, 113–132. [1773]
- Abadie, A., and L’Hour, J. (2019), “A Penalized Synthetic Control Estimator for Disaggregated Data,” Technical report, CREST. [1774]
- Abadie, A., Diamond, A., and Hainmueller, J. (2010), “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program,” *Journal of the American Statistical Association*, 105, 493–505. [1773]
- Amjad, M. D., Shah, M., and Shen, D. (2018), “Robust Synthetic Control,” *Journal of Machine Learning Research*, 19, 802–852. [1774]
- Andrews, D.W.K. (2003), “End-of-Sample Instability Tests,” *Econometrica*, 71, 1661–1694. [1774]
- Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W., and Wager, S. (2019), “Synthetic Difference in Differences,” Working Paper 1812.09970. [1774]
- Bai, C.-E., Li, Q., and Ouyang, M. (2014), “Property Taxes and Home Prices: A Tale of Two Cities,” *Journal of Econometrics*, 180, 1–15. [1774]
- Ben-Michael, E., Feller, A., and Rothstein, J. (2019), “The Augmented Synthetic Control Method,” Working Paper 1811.04170. [1774]
- Bléhaut, M., D’Haultfoeuille, X., L’Hour, J., and Tsybakov, A.B. (2020), “An Alternative to Synthetic Control for Models With Many Covariates Under Sparsity,” arxiv:2005.12225. [1774]
- Brodersen, K. H., Galluser, F., Koehler, J., Remy, N., and Scott, S. L. (2015), “Inferring Causal Impact Using Bayesian Structural Time-Series Models,” *Annals of Applied Statistics*, 9, 247–274. [1774]
- Carvalho, C. V., Masini, R., and Medeiros, M. C. (2018), “Arco: An Artificial Counterfactual Approach for High-Dimensional Panel Time-Series Data,” *Journal of Econometrics*, 207, 352–380. [1773,1774,1775,1777,1779,1784]
- Cattaneo, M. D., Feng, Y., and Titiunik, R. (2019), “Prediction Intervals for Synthetic Control Methods,” arXiv:1912.07120. [1774]
- Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2018), “An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls,” arxiv:1712.09089. [1774]
- Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2020), “Practical and Robust t-Test Based Inference for Synthetic Control and Related Methods,” arxiv:1812.10820. [1774,1776,1779]
- Davidson, J. (2009), “When is a Time Series $i(0)$?” in *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, p. 322–342. Oxford: Oxford University Press, 2009. [1785]
- Doudchenko, N., and Imbens, G. W. (2016), “Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis,” 22791, NBER. arXiv:1610.07748. [1775]
- Engle, R., and Granger, C. (1987), “Co-Integration and Error Correction: Representation, Estimation, and Testing,” *Econometrica*, 55, 251–276. [1776]
- Ferman, B., and Pinto, C. (2016), “Synthetic Controls With Imperfect Pre-Treatment Fit,” Working paper, São Paulo School of Economics - FGV. [1774]
- Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012), “A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong With Mainland China,” *Journal of Applied Econometrics*, 27, 705–740. [1773,1774,1775]
- Huang, J., and Zhang, C.-H. (2012), “Estimation and Selection Via Absolute Penalized Convex Minimization and Its Multistage Adaptive Applications,” *Journal of Machine Learning Research*, 13, 1839–1864. [1786]
- Kock, A., and Callot, L. (2015), “Oracle Inequalities for High Dimensional Vector Autoregressions,” *Journal of Econometrics*, 186, 325–344. [1777]
- Kock, A.B. (2016), “Consistent and Conservative Model Selection With the Adaptive Lasso in Stationary and Nonstationary Autoregressions,” *Econometric Theory*, 32, 243–259. [1774]
- Lee, J. H., Shi, Z., and Gao, Z. (2018), “On Lasso for Predictive Regressions,” arxiv:1810.03140, arXiv. [1774]
- Li, K.T. (2020), “Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods,” *Journal of the American Statistical Association*, 115, 2068–2083. [1774]
- Li, K. T., and Bell, D. R. (2017), “Estimation of Average Treatment Effects With Panel Data: Asymptotic Theory and Implementation,” *Journal of Econometrics*, 197, 65–75. [1774]
- Liang, C., and Schienle, M. (2019), “Determination of Vector Error Correction Models in High Dimensions,” *Journal of Econometrics*, 208, 418–441. [1774,1779]
- Liao, Z., and Phillips, P. C. B. (2015), “Automated Estimation of Vector Correction Models,” *Econometric Theory*, 31, 581–646. [1774]
- Masini, R. P., and Medeiros, M. C. (2019), “Counterfactual Analysis and Inference With Non-Stationary Data,” *Journal of Business and Economic Statistics*. [1774,1777,1786]
- Medeiros, M. C., and Mendes, E. F. (2016), “ ℓ_1 -Regularization of High-Dimensional Time-Series Models With Flexible Innovations,” *Journal of Econometrics*, 191, 255–271. [1777]
- Merlevède, F., Peligrad, M., and Rio, E. (2009), “Bernstein Inequality and Moderate Deviations Under Strong Mixing Conditions,” in *High Dimensional Probability V: The Luminy Volume*, eds. C. Houdré, V. Koltchinskii, D.M. Mason, and M. Peligrad, Vol. 5, pp. 273–292. Beachwood, OH: Institute of Mathematical Statistics. [1786]
- Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), “A Unified Framework for High-Dimensional Analysis of m -Estimators With Decomposable Regularizers,” *Statistical Science*, 27, 538–557. [1778]
- Onatski, A., and Wang, C. (2018), “Alternative Asymptotics for Cointegration Tests in Large Vars,” *Econometrica*, 86, 1465–1478. [1774,1779]
- Phillips, P. C. B. (1986), “Understanding Spurious Regressions in Econometrics,” *Journal of Econometrics*, 33, 311–340.
- (1987), “Time Series Regression With a Unit Root,” *Econometrica*, 55, 277–301.
- Shaikh, A., and Toulis, P. (2019), “Randomization Tests in Observational Studies With Staggered Adoption of Treatment,” arXiv:1912.10610. [1774]
- Tibshirani, R. (1996), “Regression Shrinkage and Selection Via the LASSO,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1773]
- van de Geer, S., and Bühlmann, P. (2009), “On the Conditions Used to Prove Oracle Results for the Lasso,” *Electronic Journal of Statistics*, 3, 1360–1392. [1786]
- Wijler, W., and Smeeke, S. (2020), “An Automated Approach Towards Sparse Single-Equation Cointegration Modelling,” *Journal of Econometrics*. [1774]
- Zou, H. (2006), “The Adaptive LASSO and Its Oracle Properties,” *Journal of the American Statistical Association*, 101, 1418–1429. [1773]