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

Ricardo Masini

Sao Paulo School of Economics, Getulio Vargas Foundation

E-mail: ricardo.masini@fgv.br

Marcelo C. Medeiros

Department of Economics

Pontifical Catholic University of Rio de Janeiro

E-mail: mcm@econ.puc-rio.br

March 9, 2019

Abstract

Recently, there has been growing interest in developing econometric 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 paper, we consider a general framework for counterfactual analysis for high dimensional, non-stationary data with either deterministic and/or stochastic trends, which nests well-established methods, such as the synthetic control. Furthermore, we propose a resampling procedure to test intervention effects that does not rely on post-intervention 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 where the effects of price changes on the sales of a product are measured.

JEL Codes: C22, C23, C32, C33.

Keywords: counterfactual analysis, comparative studies, panel data, synthetic control, policy evaluation, intervention, cointegration, spurious regression, nonstationarity, inference, resampling.

Acknowledgements: The work of Marcelo C. Medeiros is partly funded by CNPq and Faperj. The authors gratefully acknowledge insightful comments from Jianqing Fan, Marcelo Fernandes, Guido Imbens, Anders B. Kock, Eduardo F. Mendes, Normam Swanson, Michael Wolf, and participants during seminars at the Department of Operations Research and Financial Engineering (ORFE) at Princeton University and the Department of Economics at Rutgers University. We are also thankful to the comments received during the 2018 Latin American Meeting of the Econometric Society, Guayaquil, Ecuador and the Barcelona GSE summer forum.

1 Introduction

Since the seminal synthetic control (SC) paper by Abadie and Gardeazabal (2003), measuring treatment (intervention) effects on a single treated unit based on counterfactuals constructed from artificial controls has become a popular practice in econometrics. Usually, these artificial (synthetic) controls are built from a panel of untreated peers observed over time, before and after the intervention; see, for example, Doudchenko and Imbens (2016) and Athey and Imbens (2017) for recent discussions.

However, in the original SC setting, the econometric estimation was traditionally viewed as a cross-sectional problem, and the time-series nature of the data is often ignored.

1.1 Main Takeaways

This paper has two major contributions. First, we investigate the consequences of estimating counterfactuals when the data are non-stationary, displaying either deterministic and/or stochastic trends in a high-dimensional setting, where the dimensionality of the model grows with the sample size. We propose a simple modification of Tibshirani's (1996) least absolute and selection operator (LASSO), which is proved to deliver consistent estimates of the parameters of interest. Our results also 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 the intervention. Our testing procedure can be used even when there is a single observation after the intervention. Moreover, the proposed test can be applied with no modification to the original stationary counterfactual setup. The econometric framework considered here nests the SC method originally proposed by Abadie and Gardeazabal (2003) and further studied in Abadie, Diamond, and Hainmueller (2010), Doudchenko and Imbens (2016) and Ferman and Pinto (2016), as well as the panel factor (PF) method of Hsiao, Ching, and Wan (2012), and the artificial counterfactual (ArCo) put forward by Carvalho, Masini, and Medeiros (2018).

We believe our results are of general importance for the following reasons. First, several applications of the SC method and its many variants are for trending data. A key example is the original SC application of Abadie and Gardeazabal (2003) where the variable of interest was the levels of the Basque Country's GDP. With nonstationary data, the usual inferential procedures to evaluate the effects of the intervention (treatment) can be extremely misleading; see, for example, the discussion in Masini and Medeiros (2019). Second, although it is not usual for applications involving counterfactual estimation to be truly high dimensional, the number of pre-intervention observations is frequently rather small compared to the number of variables used to estimate the artificial control and the use shrinkage estimation methods have been frequently advocated. Therefore, deriving the statistical properties of counterfactual estimators under high-dimensions and nonstationarity at the same time is of considerable importance.

Finally, inference in the original SC framework of Abadie and Gardeazabal (2003) and

Abadie, Diamond, and Hainmueller (2010) is carried out by permutation tests, which tend to frequently over-reject the null hypothesis; see Ferman and Pinto (2016) for a recent discussion. Furthermore, it is not clear that their approach will be valid in a non-stationary setting. On the other hand, recent extensions consider that the number of post-intervention observations grows as $T \to \infty$. In this scenario, the tests have very little power when effects fade away in the aftermath of the intervention or when effects concern solely the variance of the variable of interest. More worrisome, with a long post-intervention period, there could be a larger probability of contamination effects; that is, the peers used to construct the counterfactual may be affected by the intervention, which in turn, invalidates the main identification assumption supporting such methods. Our inferential procedure fits nicely when the time period after the intervention is very small and can be used in both stationary and non-stationary settings.

We conduct a vast simulation study to evaluate the finite-sample properties of the estimators and inferential procedures discussed in the paper. We show that the proposed methods works reasonable well even in very small samples. Furthermore, as an empirical illustration, we estimate the impact of price changes on product sales by using a novel dataset from a major retail chain in Brazil with more than 1,000 stores in the country. We show how the methods discussed in the paper can be used to estimate demand price elasticities, which can be further used to determine optimal prices for a wide class of products.

1.2 Overview

The econometric method considered in this paper is divided in 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, $\boldsymbol{X}_t \in \mathbb{R}^p$, constructed from a number of peers that are assumed to be unaffected by the intervention. We allow the dimension of $\boldsymbol{X}_t$ to grow with the sample size T, i.e. $p \equiv p_T$. The procedure is thus summarized by the following steps:

1. Based on the sample $\{Y_t, \boldsymbol{X}_t'\}_{t=1}^{T_0}$ estimate a regression

   $$Y_t = \boldsymbol{X}_t' \boldsymbol{\beta}_0 + V_t,$$

   where V_t is an error term that will be specified later. To cope with potential high-dimensionality and potential nonstationarity, the above regression should be estimated by a modification of the original LASSO method, as proposed in this paper.

2. For $t = T_0 + 1, \dots, T$, estimate the intervention effects by

   $$\widehat{\delta}_t = Y_t - \boldsymbol{X}_t' \widehat{\boldsymbol{\beta}}_{T_0},$$

   where $\widehat{\boldsymbol{\beta}}_{T_0}$ is the estimated coefficient in the first step.

3. Test for null hypothesis in the form

$$\mathscr{H}_0 : \boldsymbol{g}(\delta_{T_0+1}, \dots, \delta_T) = \mathbf{0}$$

   by using the partial resampling procedure that will be described later in the paper. $\boldsymbol{g}(\cdot)$ is a vector valued continuous function.

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

1.3 Comparison with the Literature

Recently, extensions of the SC method, which explore the time dimension, have been proposed in the literature. Nonetheless, most of the theoretical results have been derived in a more restrictive setting than the one considered in this paper. Hsiao, Ching, and Wan (2012) was probably one of the first papers to apply time-series models to the SC framework. However, neither nonstationarity nor high-dimensionality were formally discussed.[1] Li and Bell (2017) and Carvalho, Masini, and Medeiros (2018) considered counterfactual estimation when data are high-dimensional. However, the former did not take nonstationarity into account and the later impose a sort of (trend-)stationarity condition. More specifically, Carvalho, Masini, and Medeiros (2018) derived the theory for counterfactual estimation based on LASSO under either stationarity or bounded deterministic trends, that is, deterministic functions of t/T. Extending the work of Carvalho, Masini, and Medeiros (2018), Chernozhukov, Wuthrich, and Zhu (2018b) proposed new inference methods to test hypothesis on average treatment effects when both the number of pre-intervention and post-intervention are large. Similar to the previous papers, nonstationary data are ruled out. Li (2017) also considered estimators similar to Carvalho, Masini, and Medeiros (2018) in a low dimensional setting. As before, trend-stationarity is imposed.

Three recent papers discussed the effects of nonstationarity on counterfactual estimation. The first one is Bai, Li, and Ouyang (2014). The authors show consistency of the Hsiao, Ching, and Wan's (2012) panel approach when the data are integrated of first order. Masini and Medeiros (2019) provide the asymptotic distribution of the counterfactual estimation under nonstationarity and develop as well the necessary results to conduct inference. Finally, Ferman and Pinto (2016) studied the SC estimator in cases with explosive common factors and imperfect pre-intervention fit. However, all these three papers consider only the low-dimensional case. Therefore, we are not aware of any other paper that simultaneously consider counterfactual estimation with both nonstationarity and high-dimensional data. High-dimensionality has been considered in settings much less general than the ones considered in this paper; see, for example,

[1] Hsiao, Ching, and Wan (2012) conjectured that if the data are cointegrated, their results would still hold. Masini and Medeiros (2019) showed that this conjecture turns out to be imprecise.

Li and Bell (2017), Carvalho, Masini, and Medeiros (2018), and Abadie and L'Hour (2019).

A considerable limitation of those methods is that inference, whenever available, is conducted on the average effects of the intervention, usually derived from an asymptotic argument over the post-intervention sample or by some sort of permutation test as in Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010). While the former could be justified when the post-intervention sample is relatively large, it is inappropriate in most applications of interest. In addition, as stated before, Ferman and Pinto (2016) showed that permutation tests tend to frequently over-reject the null hypothesis. Furthermore, there are no results concerning the construction of confidence intervals for the estimated counterfactuals for each point in time after the intervention. One of the few exceptions is Brodersen, Galluser, Koehler, Remy, and Scott (2015), where a Bayesian structural time-series model is used to estimate the counterfactuals and posterior inference is advocated to measure the effects of the intervention. However, the paper is silent about under which hypothesis such approach is valid.

We should also compare our results with Chernozhukov, Wuthrich, and Zhu (2018a), where the authors propose a very general conformal inference method to test hypothesis on the counterfactuals when the number of observations after the intervention is small. In their setup nonstationarity is precluded.[2] Furthermore, although the authors considered a high-dimensional setting, they do not consider the case where the number of regressors grows at a faster rate than the sample size.

Ferman and Pinto (2016) and Li (2017) discussed inference in the SC framework when the post-intervention sample is small based Andrews's (2003) end-of-sample tests. However, high-dimensionality and unit-roots have not been considered in their papers.

This paper is also related to the literature of unit-roots and cointegration in high-dimensions. To our knowledge this is the first work to derive the properties of LASSO estimators for cointegrating regressions in the case where the number of regressors is potentially much larger than the number of observations. Lee, Shi, and Gao (2018) derived the limiting distribution of LASSO-type estimators under several setups involving nonstationary variables. However, in the author's framework the number of regressors is fixed. This is also the case of Liao and Phillips (2015) and Kock (2016). Recently, Liang and Schienle (2019) proposed a shrinkage methodology for simultaneous model selection and estimation of vector error correction models (VECM) when the dimension is large and can increase with sample size. However, the number of variables is allowed to grow at a smaller rate than the sample size.[3] Furthermore, the determinist terms in their model is more restrictive than ours. Another related paper is Onatski and Wang (2018), where the authors derive the distribution of cointegration test statistics in a high-dimensional setting. As before, they consider only the case where the dimension of the model grows at slower rate than the sample size.

[2] See Assumption 5 and Theorem 3 as well as Lemma 10 in Chernozhukov, Wuthrich, and Zhu (2018a). Assumption 5, which is required by Theorem 3, explicitly states that the data is stationary and β-mixing. Lemma 10 requires that $\mathbb{P}\left(\max_{1\leq t\leq T}\|\boldsymbol{X}_t\|_\infty \leq \text{finite constant}\right) = 1$. This last assumption is clearly violated with unit-root processes.

[3] See, for example, the assumptions in Corollary 2.1 in Liang and Schienle (2019).

1.4 Summary of the Paper

The rest of the paper is organized as follows. We present the setup and assumptions in Section 2 and derive the theoretical results concerning the counterfactual estimation and inference in Section 3. In Section 3.3 we describe the inferential procedure considered in this paper. We present the results of a simulation experiment in Section 4 and discuss the empirical application in Section 5. Section 6 concludes the paper. Finally, we present all proofs in the Appendix.

2 Setup and Assumptions

2.1 Notation

All random variables (real-valued scalars, vectors and matrices) are defined in a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We denote random variables by an upper case letter, X for instance, and its realization by a lower case letter, $X = x$. Matrices and vectors are written in bold letters. The expected value operator is with respect to the $\mathbb{P}$ law such that $\mathbb{E}(X) := \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega)$.

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 $\boldsymbol{X} = (X_1, \ldots, X_d)'$, we have $\|\boldsymbol{X}\|_q := (\sum_{i=1}^d |X_i|^q)^{1/q}$ and, for a scalar random variable X, $\|X\|_{\mathcal{L},q} = (\mathbb{E}|X|^q)^{1/q}$. $\|\boldsymbol{X}\|_\infty := \mathsf{sup}_{i\leq d}|X_i|$. If $\boldsymbol{X}$ is a $(m \times n)$ (random) matrix then $\|\boldsymbol{X}\|_\infty := \mathsf{sup}_{i\leq m, j\leq n}|X_{i,j}|$. We also use the $\|\boldsymbol{X}\|_0 := \#\{i : X_i \neq 0\}$ to denote the ℓ^0 "norm". Moreover, for a d-dimensional square matrix $\boldsymbol{M}$, we use $\|\boldsymbol{X}\|_{\boldsymbol{M}}$ to denote the quadratic form $\boldsymbol{X}'\boldsymbol{M}\boldsymbol{X}$. For any vector $\boldsymbol{X}$, we use $\mathsf{diag}(\boldsymbol{X})$ to denote the diagonal matrix whose diagonal is the elements of $\boldsymbol{X}$. $\mathbb{1}(A)$ represents an indicator function on the event A, i.e,

$$\mathbb{1}(A) = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

Finally, unless stated otherwise, all the asymptotics are taken as $T_0 \to \infty$, and the $o(1)$ and $o_P(1)$ terms are with respect to the limit as $T_0 \to \infty$. We denote convergence in probability and in distribution by "$\xrightarrow{p}$" and "$\Rightarrow$", respectively. A full list of symbols used in the paper is presented in the Appendix.

2.2 Basic Setup

Suppose we have n units (countries, states, municipalities, firms, etc.) indexed by $i = 1, \ldots, n$. For every time period $t = 1, \ldots, T$, we observe a realization of a real valued random vector $\boldsymbol{Z}_t := (Z_{1t}, \ldots, Z_{nt})'$.[4] Furthermore, we assume that an intervention took place at T_0+1, where $1 < T_0 < T$. Let $\mathcal{D}_t \in \{0, 1\}$ be a binary variable flagging the periods where the intervention (treatment) was in place. Therefore, following the potential outcome notation, we can express

[4] We consider a scalar variable for each unit for the sake of simplicity, and the results in the paper can be easily extended to the multivariate case.

Z_{it} as

$$Z_{it} = \mathcal{D}_t Z_{it}^{(1)} + (1 - \mathcal{D}_t) Z_{it}^{(0)},$$

where $Z_{it}^{(1)}$ denotes the potential outcome when the unit i is exposed to the intervention and $Z_{1t}^{(0)}$ 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. The null hypothesis to be tested is:

$$\mathscr{H}_0 : \delta_t := Z_{1t}^{(1)} - Z_{1t}^{(0)} = 0, \quad \forall t > T_0. \tag{2.1}$$

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 $t = T_0 + 1$ onwards. For this reason, we henceforth call it the *counterfactual* – i.e., what would Z_{1t} have been like had there been no intervention (potential outcome).

To construct the counterfactual, let $\boldsymbol{Z}_{0t}^{(0)} := \left[Z_{2t}^{(0)}, \dots, Z_{nt}^{(0)}\right]'$ be the collection of all control variables (all other variables except the ones belonging to unit 1).[5] Panel-based methods, such as the PF and ArCo methodologies, as well as the SC extensions discussed in Doudchenko and Imbens (2016), construct an artificial counterfactual by considering the following model in the absence of an intervention:

$$Z_{1t}^{(0)} = \mathcal{M}\left(\boldsymbol{Z}_{0t}^{(0)}; \boldsymbol{\theta}\right) + V_t, \quad t = 1, \dots, T, \tag{2.2}$$

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

The main idea is to estimate (2.2) using just the pre-intervention sample, $t = 1, \dots, T_0$, since in this case, $\boldsymbol{Z}_{0t}^{(0)} := \boldsymbol{Z}_{0t} = (Z_{2t}, \dots, Z_{nt})'$. Consequently, the estimated counterfactual for the post-intervention period, $t = T_0 + 1, \dots, T$, becomes $\widehat{Z}_{1t}^{(0)} := \mathcal{M}(\boldsymbol{Z}_{0t}; \widehat{\boldsymbol{\theta}}_{T_0})$. Under some sort of stationarity assumption on $\boldsymbol{Z}_{0t}$ 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), show that $\widehat{\delta}_t := Z_{1t} - \widehat{Z}_{1t}^{(0)}$ is an unbiased estimator for δ_t as the pre-intervention sample size grows to infinity and

$$\widehat{\Delta}_T = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \widehat{\delta}_t, \tag{2.3}$$

is $\sqrt{T}$-consistent for $\Delta_T := \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \delta_t$ and is asymptotically normal.

Consider the following assumption.

[5] We could also have included lags of the variables and/or exogenous regressors into $\boldsymbol{Z}_{0t}$, but again, to keep the argument simple, we have considered only contemporaneous variables; see Carvalho, Masini, and Medeiros (2018) for more general specifications.

Assumption 1. *$Z_t^{(0)}$ is independent of $\mathcal{D}_s$ for all $1 \leq s, t \leq T$.*

To recover the effects of the intervention, Assumption 1 is key. Roughly speaking, it suffices that $\mathbb{E}(\boldsymbol{Z}_{0t}|\mathcal{D}_s = 1) = \mathbb{E}(\boldsymbol{Z}_{0t}|\mathcal{D}_s = 0)$.[6]

The main purpose of the paper is to extend those results to include both deterministic and stochastic trends in the DGP of $\boldsymbol{Z}_t^{(0)}$. This presents some new challenges, to begin with, the population parameter $\boldsymbol{\theta}_0$ can no longer be identified as the linear projection parameters of $Z_{1t}^{(0)}$ onto a constant and $\boldsymbol{Z}_{0t}^{(0)}$ due to the non-stationarity of the regressors. Before we present the general setup in Section 2.4, next section describes a simple, yet instructive, example of the proposed methodology.

2.3 Factor Model Example

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

$$Z_{it}^{(0)} = c_i + \mu_i F_t + U_{it}^Z, \tag{2.4}$$

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

$$F_t = f_t^F + F_{t-1} + U_t^F, \quad t \geq 1 \tag{2.5}$$

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

$$F_t = f_t^F + U_t^F, \tag{2.6}$$

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

For now consider that $(U_{1t}^Z, \ldots, U_{nt}^Z, U_t^F)$ is a zero-mean, independent and identically distributed Gaussian random vector which trivially fulfills both conditions described later in Assumption 3 in Section 3.

The factor model above results in a common trend (at least for those units with non-zero loadings, $\mu_i \neq 0$) and a correlation among the stochastic components of the vector $\boldsymbol{Z}_t^{(0)}$ due to the presence of U_t^F. We define a linear regression model, which we call pseudo-true model, as

$$Y_t = \boldsymbol{\beta}_0' \boldsymbol{X}_t + V_t,$$

where $Y_t := Z_{1t}^{(0)}$ and $\boldsymbol{X}_t := \left[1, \boldsymbol{Z}_{0t}^{(0)\prime}\right]'$. Suppose there are $1 < r + 1 \leq n$ units with non-zero loadings ($\mu_i \neq 0$) including unit 1. Without loss of generality, make those the first $r + 1$ units. In that case, we have r independent linear relations among those units resulting in a stationary

[6] 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). Admittedly, Assumption 1 is stronger than necessary, for an unbiased estimate for instance would be enough to impose $\mathbb{E}[\mathcal{M}(\boldsymbol{Z}_{0t}^{(0)}; \boldsymbol{\theta})|\mathcal{D}_s] = \mathbb{E}[\mathcal{M}(\boldsymbol{Z}_{0t}; \boldsymbol{\theta})]$.

process since we can cancel the trends (either deterministic and/or stochastic) by setting $\widetilde{\mathbf{\Gamma}}'\boldsymbol{Z}_t^{(0)}$, where

$$\widetilde{\mathbf{\Gamma}}' = \begin{pmatrix} 1 & -\frac{\mu_1}{\mu_2} & 0 & 0 & \\ \vdots & 0 & \ddots & 0 & \mathbf{0}_{r\times(n-r-1)} \\ 1 & 0 & 0 & -\frac{\mu_1}{\mu_{r+1}} & \end{pmatrix},$$

and $\mathbf{0}_{r\times(n-r-1)}$ is a $r \times (n-r-1)$ matrix of zero elements.

After normalizing to obtain the representation $\widetilde{\mathbf{\Gamma}}' = (I_r : -\mathbf{\Gamma}')$, we are left with:

$$\mathbf{\Gamma}' = \begin{pmatrix} \widetilde{\mu}_1 & \\ \vdots & \mathbf{0}_{r\times(n-r-1)} \\ \widetilde{\mu}_r & \end{pmatrix},$$

where $\widetilde{\mu}_i := \frac{\mu_i}{\mu_{r+1}}$ for $i \in \{1,\dots,r\}$. Then, $\boldsymbol{J}_t = \widetilde{\mathbf{\Gamma}}'\boldsymbol{Z}_t^{(0)}$ is stationary with a typical element given by

$$J_{i,t} = c_i - \widetilde{\mu}_i c_{r+1} + U_{it}^Z - \widetilde{\mu}_i U_{r+1,t}^Z = \widetilde{c}_i + \widetilde{U}_{it},$$

where $\widetilde{c}_i := c_i - \widetilde{\mu}_i c_{r+1}$ and $\widetilde{U}_{it} := U_{it}^Z - \widetilde{\mu}_i U_{r+1,t}^Z$.

When $r = 1$, the pseudo-true vector of parameters becomes:

$$\boldsymbol{\beta}_0 = \left(c_1 - \frac{\mu_1}{\mu_2}c_2, \frac{\mu_1}{\mu_2}, 0, \dots, 0\right)',$$

and the covariance structure of the vector $\left(U_t^F, U_{1t}^Z, \dots, U_{nt}^Z\right)'$ plays no role in determining the coefficients of the pseudo-true model, since there is only one possible linear combination that results in a $I(0)$ process. On the other hand, when $r \geq 2$, we have

$$\boldsymbol{\beta}_0 = (\widetilde{c}_1 - \boldsymbol{\zeta}'\widetilde{\boldsymbol{c}}_0, \boldsymbol{\zeta}', \widetilde{\mu}_1 - \boldsymbol{\zeta}'\widetilde{\boldsymbol{\mu}}_0, 0, \dots, 0)',$$

where $\widetilde{\boldsymbol{c}}_0 := (\widetilde{c}_2, \dots, \widetilde{c}_r)'$, $\widetilde{\boldsymbol{\mu}}_0 := (\widetilde{\mu}_2, \dots, \widetilde{\mu}_r)'$ and $\boldsymbol{\zeta}$ denote the linear projection of $\widetilde{U}_{1t}$ onto $\left(\widetilde{U}_{2t}, \dots, \widetilde{U}_{rt}\right)'$. Now it becomes evident that the covariance structure of $\left(U_{1t}^Z, \dots, U_{r+1,t}^Z\right)'$ affects the coefficients of the pseudo-true model through $\boldsymbol{\zeta}$. Finally, the error term for the factor model is given by

$$V_t = U_{1t}^Z - \sum_{i=2}^{r+1} \beta_{0,i} U_{it}^Z.$$

In a high-dimensional setup ($n \gg T$), with $\boldsymbol{\beta}_0$ properly defined, we could estimate the parameters using any regularized regression method. In particular let $\widehat{\boldsymbol{\beta}}$ be a minimizer of the LASSO objective function

$$\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \boldsymbol{X}_t'\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\lambda \geq 0$ is the penalty term. If we assume that the vector of loadings is sparse in the sense that the number of loadings different from zero, $r+1$, grows slower than T_0 even when n grows at a faster rate then T, we have the following corollary from our first main result (Theorem 1 in Section 3).

Corollary 1. *Under Assumptions 1 and for any $c > 0$, set the penalty parameter $\lambda = 4cn^{2/q}/\sqrt{T_0}$. Then, if $r(\log n)^2/\sqrt{T_0} = o(1)$, we have, as $T_0 \to \infty$:*

$$\widehat{\delta}_t - \delta_t - V_t = O_P[(\log n)^2 r/\sqrt{T_0}] = o_P(1); \quad T_0 < t \leq T.$$

If further $T_2 \to \infty$:

$$\widehat{\Delta}_T - \Delta_T = O_P\left[\frac{\log(n)r}{\sqrt{T_0}} \vee \frac{1}{\sqrt{T_2}}\right] = o_P(1).$$

The first part of the corollary gives us an asymptotically ($T_0 \to \infty$) unbiased estimator for the intervention effect for every period in the post-intervention sample. The second part establishes a consistent estimator for the average (across the post-intervention period) effect of the intervention. The latter relies on the asymptotics for the post-intervention period ($T_2 \to \infty$) as well, which is might not be credible in most practical applications.

For this reason, we propose a inference method applying a resampling scheme that is effective even when we there is a single period in the post-intervention. Let $\widehat{\boldsymbol{\delta}} := (\widehat{\delta}_{T_0+1}, \ldots \widehat{\delta}_T)'$. Consider the construction of blocks of size T_2 of consecutive observations from the pre-intervention sample. There are $T_0 - T_2 - 1$ such blocks which we denote by

$$\widehat{\boldsymbol{\delta}}_j := \left(\widehat{V}_j, \ldots, \widehat{V}_{j+T_2-1}\right) \quad j = 1, \ldots, T_0 - T_2 + 1,$$

where $\widehat{V}_t := Y_t - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}_t$ for $1 \leq t \leq T_0$ is residual of the pre-intervention estimation. We then estimate the distribution $\mathcal{Q}_T(\boldsymbol{x}) := \mathbb{P}(\widehat{\boldsymbol{\delta}} \leq \boldsymbol{x})$ by the empirical distribution function of $(\widehat{\boldsymbol{\delta}}_j)$, i.e:

$$\widehat{\mathcal{Q}}_T(\boldsymbol{x}) := \frac{1}{T_0 - T_2 + 1} \sum_{j=1}^{T_0-T_2+1} \mathbb{1}\left(\widehat{\boldsymbol{\delta}}_j \leq \boldsymbol{x}\right),$$

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

This is a simple, yet effective, way (as shown by our simulation results in Section 4) to conduct a point-by-point inference in the post-intervention period and recover the joint distribution of $\widehat{\boldsymbol{\delta}}$ or, in fact, any continuous function thereof. As a corollary from our second main result (Theorem 2 in Section 3) we have

Corollary 2. *Under the same conditions of Corollary 1, but with $r\log(n)\log(nT_0)/\sqrt{T_0} = o(1)$, we have under the null hypothesis* (2.1)*:*

$$\sup_{\boldsymbol{x} \in \mathbb{R}^{T_2}} |\widehat{\mathcal{Q}}_T(\boldsymbol{x}) - \mathcal{Q}_T(\boldsymbol{x})| = o_P(1),$$

for fixed T_2 as $T_0 \to \infty$.

2.4 General Setup

2.4.1 Non-stationarity

We model the units in the absence of the intervention as a non-stationary (vector) process $\{\boldsymbol{Z}_t^{(0)} := (Z_{1t}, \ldots, Z_{nt})'\}_{t \geq 1}$.

Assumption 2. *(DGP) Consider that the process $\{Z_{it}^{(0)} : 1 \leq i \leq n, t \geq 1\}$ is either generated by*

(a) ***Stochastic Trend:***

$$Z_{it}^{(0)} = Z_{it-1}^{(0)} + f_{it} + U_{it}, \quad t \geq 1, \tag{2.7}$$

with $Z_{i0}^{(0)} = O_P(1)$, or

(b) ***Deterministic Trend:***

$$Z_{it}^{(0)} = f_{it} + U_{it}, \quad t \geq 1. \tag{2.8}$$

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

We believe the DGPs in Assumption 2 cover a wide range of relevant situations for the empirical researcher to model non-stationary behavior. In particular, the factor model (2.4) can be recovered as a particular case of DGP (2.7) by setting

$$f_{it} = \mu_i f_t^F \quad \text{and} \quad U_{it} = \mu_i U_t^F + U_{it}^Z - U_{it-1}^Z,$$

when we have the factor evolving as per (2.5); or DGP (2.8) by setting

$$f_{it} = c_i + \mu_i f_t^F \quad \text{and} \quad U_{it} = \mu_i U_t^F + U_{it}^Z,$$

where the factor follows (2.6).

Even though the factor model fits perfectly in our framework and, in fact, motivated much of the idea behind the DGPs in Assumption 2, we have decided not to impose it. In fact, we have chosen not to impose any structure that we did not judged to be necessary to derive the results below.

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

(a) There exists a real $q > 2$ such that $\mathsf{sup}\,\{\mathbb{E}|U_{it}|^{q+\epsilon} : 1 \leq i \leq n, t \in \mathbb{N}\} < \infty$ for some $\epsilon > 0$;

(b) There exist reals $c_1, c_2, c_3 > 0$ such that $\mathsf{sup}\,\{\mathbb{P}(|U_{it}| > u) : 1 \leq i \leq n, t \in \mathbb{N}\} \leq c_1 \, \mathsf{exp}(-c_2 u^{c_3})$ for all $u > 0$.

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

Assumption 3 deals with the trade-off between moment conditions and serial dependency. In particular, it requires exponential decay of the strong mixing coefficient to ensure that the q-th moment of the sum of the zero-mean strong mixing variables is of order $T^{q/2}$. More importantly, 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.

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 sub-exponential growth of units in case of exponential decay of the tails. This includes sub-Gaussian ($c_3 \geq 2$), sub-exponential ($c_3 \geq 1$) and many other families of distributions of interest. Finally, we bound from below the smallest eigenvalue to ensure that $\mathbb{E}(\boldsymbol{Z}_t\boldsymbol{Z}_t')$ properly scaled is full rank and, therefore, avoid multicollinearity among the regressors.

For now, the deterministic sequence $\{f_{it}\}_{t\geq 1}$ appearing in both DGPs described in Assumption 2 is considered idiosyncratic, i.e., unit-specific. However, in most applications, we expect to have a common (up to a constant) trend such that $f_{it} = \boldsymbol{\mu}_i'\boldsymbol{f}_t$ where $\boldsymbol{\mu}_i$ and $\boldsymbol{f}_t$ are multidimensional.

The DGP described by (2.7) in Assumption 2 may involve an $I(1)$ (integrated of order 1) processes 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. To better understand the link between the sequence f_{it} and the trend it generates, it is worth considering the continuous version of f_{it} given by $f_i(t)$, such that

$$a_{it} := \sum_{s=1}^{t} f_{is} = O\left[\int f_i(t)\mathrm{d}t\right], \quad \text{for integrable } f_i(t) : \mathbb{R} \to \mathbb{R}^+. \tag{2.9}$$

Therefore, if $f_i(t) = O(t^c)$, with $c \in \mathbb{R}$, we have $a_{it} = O(t^{c+1})$ for $c \neq 0$. For the special case where $c = -1$, we have $a_{it} = O(\log t)$. More generally, model (2.7) covers a wide class of trend patterns depending upon the choice of the sequence $\{f_{it}\}$, including (we drop the subscript i in what follows):

No trend: $tf_t \to 0$, which implies $a_t \to 0$ as $t \to \infty$.

Sublinear: $f_t \to 0$ but $tf_t \to \infty$, which implies $a_t/t \to 0$ as $t \to \infty$.

Linear: $f_t \to c > 0$, which implies $a_t \to ct$ as $t \to \infty$.

Sub-exponential: $f_t \to \infty$ but $f_t/\exp ct \to 0$ for any $c > 0$, which implies $a_t/\exp ct \to 0$ as $t \to \infty$.

Exponential: $f_t \to c_1 \exp c_2 t$, which implies $a_t \to c_1/c_2 \exp c_2 t$ as $t \to \infty$ for some $c_1, c_2 > 0$.

Super-exponential: $f_t/(c_1 \exp c_2 t) \to \infty$, which implies $a_t/c_1 \exp c_2 t \to \infty$ as $t \to \infty$ for any $c_1, c_2 > 0$.

Clearly, both DGPs defined in Assumption 2 can be casted in the following format:

$$Z_{it}^{(0)} = d_{it} + \eta_{it}, \quad 1 \leq i \leq n,\ t \geq 1, \tag{2.10}$$

where d_{it} is a deterministic trend (which absorbs any constant), and η_{it} is the trend-free (not necessarily stationary) stochastic component. (2.8) becomes (2.10) by setting $d_{it} = c_i + f_{it}$ and $\eta_{it} = U_{it}$ and for (2.7) by backward recursion, we conclude that $d_{it} = a_{it} := \sum_{s=1}^{t} f_{it}$ and $\eta_{it} = Z_{i0}^{(0)} + \sum_{s=1}^{t} U_{is}$.

It is important to understand under which conditions the stochastic part of (2.10) is asymptotically dominated by the deterministic one, in the sense that $Z_{it}^{(0)}/d_{it} \to 1$, almost surely or in probability. For (2.8), this is always the case as long as $f_{it} \to \infty$, which implies $|d_{it}| \to \infty$. For (2.7), since the variance of η_{it} increases as $t \to \infty$, it is no longer enough to have $a_{it} \to \infty$. In fact, since we have $\eta_{it} = O_P(\sqrt{t})$, we must have a_{it} of an order higher that $\sqrt{t}$, which is ensured, for instance, by taking $f_{it} = t^c$ with $c > -1/2$. As an illustration, take 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 $Z_{it}/d_{it} = 1 + \sum_{s=1}^{t} U_s/\mu_i t \to 1$, almost surely or in probability, depending on the law of large numbers, which is available for the process $\{\boldsymbol{U}_t\}$. We formalize those facts in the following proposition.

Proposition 1. *Consider the DGPs in Assumption 2, assuming that $\{\boldsymbol{U}_t\}$ fulfills Assumption 3 for $1 \leq i \leq n$. Therefore, as $t \to \infty$,*

*(a) (**Growth Condition**) $Z_{it}^{(0)}/d_{it} \to 1$ in probability under DGP (2.7) if $\sqrt{t}/d_{it} = o(1)$; or $Z_{it}^{(0)}/d_{it} \to 1$ almost surely under DGP (2.8) if $f_{it} \to \infty$*

*(b) (**No-Growth**) $Z_{it}^{(0)} = O_P(\sqrt{t})$ under DGP (2.7) if $d_{it}/\sqrt{t} = O(1)$; or $Z_{it}^{(0)} = O_P(1)$ under DGP (2.8) if $f_{it} = O(1)$.*

Moreover, for (2.8) if $d_{it} = o(\sqrt{t})$, then $t^{-1/2} Z_{it}^{(0)}$ converges in distribution to a Gaussian random variable.

From Proposition 1 above, the DGP will satisfy or not the growth condition depending on the growth rate of d_{it}. In particular, for (2.7), the grown condition depends on whether $\sqrt{t}/d_{it} \to 0$ or not. For (2.8), the growth condition does not happen if $f_{it} \to c < \infty$. Therefore, to estimate (2.11) in the high-dimensional set-up (Section 2.4.2), we need to impose a separation between those two regimes as the number of units increases with the sample size. To that end, we consider the following assumption.

Assumption 4. *Let $h := \#\mathcal{H}$, where $\mathcal{H} \subseteq \{1, \ldots, n\}$ be the index set of units $\{Z_{it}^{(0)}, 1 \leq 1 \leq n\}$ 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 (2.7)* *in Assumption 2, consider*

$$\frac{h^{1/q}\sqrt{T_0}}{d_{\mathcal{H}}(T_0)} = o(1) \quad \textit{under Assumption 3(a),} \quad \textit{and}$$
$$\frac{\sqrt{T_0}\log h}{d_{\mathcal{H}}(T_0)} = o(1) \quad \textit{under Assumption 3(b).}$$

For (2.8) in Assumption 2 consider

$$\frac{h^{1/q}}{d_{\mathcal{H}}(T_0)} = o(1) \quad \textit{under Assumption 3(a),} \quad \textit{and}$$
$$\frac{\log h}{d_{\mathcal{H}}(T_0)} = o(1) \quad \textit{under Assumption 3(b).}$$

2.4.2 High Dimensionality and Sparsity

To simplify the notation, we rename the variable of interest as $Y_t := Z_{1t}^{(0)}$. Moreover, we consider the situation where the number of regressors $\boldsymbol{Z}_{0t}$ in (2.2) can be much larger than the number of observations.

Our motivation to move to a high-dimensional setup is to be able to accommodate two cases of interest: (i) when the setup is intrinsically high-dimensional, i.e, the number of units is in fact much larger than the number of observations available ($n \gg T$) or (ii) the number of units is small relative to the number of periods available ($n < T$), but the target model is well approximated by a linear combination using some transformation of the regressions through a set of basic functions, such that the number of effective regressors becomes much larger than the number of observations. Regardless of the case, we denote the final regressors including a constant as $\boldsymbol{X}_t := (1, X_{1,t}, \ldots, X_{p-1,t})'$, and throughout, we adopt the possibility of $p \gg n$. Hence, the "pseudo-true" model in the absence of an intervention becomes

$$Y_t = \boldsymbol{X}_t'\boldsymbol{\beta}_0 + V_t, \qquad 1 \leq t \leq T, \tag{2.11}$$

where the p-dimensional vector $\boldsymbol{\beta}_0$ is defined in (2.15) depending on the DGP of Assumption 2 and the number of independent linear $I(0)$ relations.

When $p \gg T$, even if $\boldsymbol{\beta}_0$ is properly identified, there is no hope to consistently estimate it without some additional assumptions. Therefore, we consider the linear model to be sparse in the sense that only a few of parameters are actually different than zero, i.e., $s_0 := \|\boldsymbol{\beta}_0\|_0 < T$; see Remark 1. Consequently, we have a linear high-dimensional sparse model (LHDSM), which will be estimated via a weighted least absolute shrinkage and selection operator (WLASSO), i.e, $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}_{T_0}(\lambda, \boldsymbol{w})$ is a minimizer of $\boldsymbol{\beta} \mapsto Q(\boldsymbol{\beta}, \lambda, \boldsymbol{w})$ defined as

$$Q(\boldsymbol{\beta}, \lambda, \boldsymbol{w}) := \frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \boldsymbol{X}_t'\boldsymbol{\beta})^2 + \lambda\sum_{i=1}^{p} w_i|\beta_i|, \tag{2.12}$$

where $\lambda \geq 0$ is the common penalty term and $\boldsymbol{w} := (w_1, \ldots, w_p)'$ is a vector of almost surely non-negative weights specific for each parameter.[7]

Remark 1. *For estimation purposes, it will be critical for the model* (2.11) *to be sparse. However, since we are not interested in conducting inference on any element of the parameter vector* $\boldsymbol{\beta}_0$*, what matters to us is a consistent forecast of the counterfactual unit with or without model selection consistency.*

2.4.3 The Target model

In the this section, we properly define the target model together with its "true parameters" appearing in (2.11). First, however, we need to properly define a $I(0)$ process as in Davidson (2009).

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

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

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

Notice from the definition above that 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 $\upsilon_T^2 \sim cT$ for $0 < c < \infty$, which in turn is necessary to ensure that $\mathbb{E}[B(s) - B(r)]^2 = s - r$ for $0 \leq s \leq r \leq 1$.

Ideally (in the mean squared error sense), we would like $\mathcal{M}(\boldsymbol{x}) := \mathbb{E}(Y_t | \boldsymbol{X}_t = \boldsymbol{x})$. However in the presence of trends, we would be most likely to have the model $\mathcal{M} = \mathcal{M}_t$ 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 $\boldsymbol{X}_t$ would result in time-varying parameters again due to the non-stationary setup.

Let $r \in \{0, 1, \ldots, n-1\}$ be the number of independent linear relations among the n units that results in a $I(0)$ process. We suppose that if $r > 0$, at least one of those relations includes unit 1 such that its coefficient can be normalize to a unit, otherwise we set $r = 0$. For the DGP (2.7), r also represents the number of cointegration relations as per Engle and Granger (1987). For the DGP (2.8), if $f_{it} = \mu_i f_t$, we have $r = n - 1$ because for any vector $\boldsymbol{\beta} \in \mathbb{R}^{n-1}$ such that $(1, \boldsymbol{\beta}')\boldsymbol{\mu} = 0$, the trend f_t is canceled; therefore, $(1, \boldsymbol{\beta}')\boldsymbol{Z}_t^{(0)} \sim I(0)$.

Assumption 5. *There is at least one linear combination among the units, with a non-zero coefficient for the unit 1, that results in a I(0) process* $(r > 0)$ *and* $\|\boldsymbol{\beta}_0\|_1 \leq c < \infty$.

Failure to comply with the Assumption above results in what is known in the literature 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 $\boldsymbol{Z}_{0t}$ that results in

[7] When $p < T$ we might choose $\lambda = 0$.

an $I(0)$ process. Therefore, the DGP is considered spurious only in the sense that all linear combinations of $\boldsymbol{Z}_{0t}$ are not an $I(0)$ process.

Under Assumption 5, let $\widetilde{\boldsymbol{\Gamma}}$ be a $(n \times r)$ matrix containing the r independent linear relations resulting in $I(0)$ process as described in the previous paragraph. Without loss of generality since $\widetilde{\boldsymbol{\Gamma}}$ is rank r by definition, we can normalize it such that

$$\underset{(r\times 1)}{\boldsymbol{J}_t} := \widetilde{\boldsymbol{\Gamma}}'\boldsymbol{Z}_t^{(0)} \sim I(0), \quad \widetilde{\boldsymbol{\Gamma}} := (I_r : -\boldsymbol{\Gamma}')'. \tag{2.13}$$

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

$$\underset{(r\times 1)}{\boldsymbol{\pi}} := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}[\mathbb{E}(\boldsymbol{J}_{0t}\boldsymbol{J}_{0t}')]^{-1}[\mathbb{E}(\boldsymbol{J}_{0t}J_{1t})] \tag{2.14}$$

We can now define the pseudo-true parameters of model (2.11) depending on the number of $I(0)$ relations among the units. If $r = 0$, by definition, there is no value for $\boldsymbol{\beta}_0 \in \mathbb{R}^n$ such that V_t in (2.11) is $I(0)$. For the remaining cases,

$$\boldsymbol{\beta}_0 := \boldsymbol{\beta}_0(r) := \begin{cases} (\boldsymbol{\pi}, \boldsymbol{\Gamma}')' & r = 1 \\ [\boldsymbol{\pi}', (1, -\boldsymbol{\pi}_0')\boldsymbol{\Gamma}']' & 2 \le r \le n-1, \end{cases} \tag{2.15}$$

where $\boldsymbol{\Gamma}(n - r \times r)$ is defined by (2.13), $\boldsymbol{\pi}$ is defined by (2.14) and $\boldsymbol{\pi}_0 := (\pi_2, \ldots, \pi_r)'$.

3 Theoretical Results

3.1 The Oracle Inequalities

Hereafter, we outline the steps towards the proof of Proposition 2 (the details are in the Appendix A), with the basis for both our main result (Theorems 1 and 2). First, due to the presence of trends in the regressors, not all the components of $\boldsymbol{X}_t$ are of the same order (in probability). Therefore, it is convenient to consider a reparametrization of the objective function (2.12) using the following linear transformation to partially cancel those trends:

$$\boldsymbol{\gamma} := \boldsymbol{L}\boldsymbol{\beta}, \quad \boldsymbol{W}_t := \boldsymbol{L}^{-1}\boldsymbol{X}_t \quad \boldsymbol{L} := \mathsf{diag}\,[(\ell_1, \ldots, \ell_p)'], \tag{3.1}$$

where $\ell_1 = 1$ and for $2 \le i \le p$ we set $\ell_i = d_{iT_0}$ if the growth condition (Proposition 1(b)) is satisfied; otherwise $\ell_i = \sqrt{T_0}$ if DPG(2.7) or 1 if DGP (2.8) in Assumption 2. The reparametrized objective function then becomes

$$H(\boldsymbol{\gamma}) := H(\boldsymbol{\gamma}, \lambda, \boldsymbol{w}) := \frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \boldsymbol{W}_t'\boldsymbol{\gamma})^2 + \lambda\sum_{i=1}^{p}\nu_i|\gamma_i|, \tag{3.2}$$

where $\nu := (\nu_1, \ldots, \nu_p)'$ and $\nu_i := w_i/\ell_i$ for $1 \leq i \leq p$.

The importance of such a reparametrization is that now the new regressors $\boldsymbol{W}_t$ are free of diverging trends, which makes the problem tractable.[8] Moreover, a minimizer $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma} \mapsto H(\boldsymbol{\gamma})$ is related to a minimizer $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} \mapsto Q(\boldsymbol{\beta})$ through $\widehat{\boldsymbol{\gamma}} := L\widehat{\boldsymbol{\beta}}$, and the reparametrized target parameters become $\boldsymbol{\gamma}_0 := L\boldsymbol{\beta}_0$. By definition, $H(\widehat{\boldsymbol{\gamma}}) \leq H(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma}$. In particular, for $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, using the fact that $Y_t = \boldsymbol{\gamma}_0'\boldsymbol{W}_t + V_t$ in the transformed variables and letting $\boldsymbol{\Sigma} := \frac{1}{T_0}\sum_{t=1}^{T_0} \boldsymbol{W}_t\boldsymbol{W}_t'$, we have the following basic inequality:

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_{\boldsymbol{\Sigma}}^2 + \lambda \sum_{i=1}^{p} \nu_i |\widehat{\gamma}_i| \leq 2(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)' \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{W}_t V_t + \lambda \sum_{i=1}^{p} \nu_i |\gamma_{0i}|. \tag{3.3}$$

Let $\mathcal{S} \subseteq \{1\ldots,p\}$ denote an index set such that for any p-dimensional vector $\boldsymbol{v}$, $\boldsymbol{v}_{\mathcal{S}}$ is the vector containing only the elements of the vector $\boldsymbol{v}$ indexed by $\mathcal{S}$; thus, $\#\boldsymbol{v}_{\mathcal{S}} = \#\mathcal{S}$ and $\mathcal{S}^c := \mathcal{S} \setminus \{1, \ldots, p\}$ its complement. We can bound from above the first term after the inequality in (3.3) using Hölder's inequality by $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \|\frac{2}{T}\sum_{t=1}^{T} \boldsymbol{W}_t V_t\|_\infty$, and we use the triangle inequality to rewrite (3.3) as

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_{\boldsymbol{\Sigma}}^2 + \lambda \sum_{i\in\mathcal{S}^c} \nu_i |\widehat{\gamma}_i| - \left\| \frac{2}{T}\sum_{t=1}^{T} \boldsymbol{W}_t V_t \right\|_\infty \|\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^c}\|_1 \leq& \\ \left\| \frac{2}{T}\sum_{t=1}^{T} \boldsymbol{W}_t V_t \right\|_\infty \|\widehat{\boldsymbol{\gamma}}_{\mathcal{S}} - \boldsymbol{\gamma}_{0,\mathcal{S}}\|_1 + \lambda \sum_{i\in S} \nu_i |\widehat{\gamma}_i - \gamma_{0i}|.\end{aligned}$$

Now consider events defined in (3.7)–(3.10) to conclude that on $\Omega_0 \cap \Omega_2$:

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_{\boldsymbol{\Sigma}}^2 + [\lambda(1-\lambda_2) - \lambda_0]\|\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^c}\|_1 \leq [\lambda_0 + \lambda(1+\lambda_2)]\|\widehat{\boldsymbol{\gamma}}_{\mathcal{S}} - \boldsymbol{\gamma}_{0,\mathcal{S}}\|_1. \tag{3.4}$$

Consequently, provided that $\lambda > 0$ satisfies condition (3.11) for some $\xi > 0$, we have that the estimation error $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ belongs to the cone $\mathscr{C}(\xi, \mathcal{S})$ given by

$$\mathscr{C}(\xi, \mathcal{S}) := \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}_{\mathcal{S}^c}\|_1 \leq \xi \|\boldsymbol{x}_{\mathcal{S}}\|_1\}. \tag{3.5}$$

As is common in the high-dimensional literature, we need a certain compatibility between the norms $\|\cdot\|_1$ and $\|\cdot\|_{\boldsymbol{\Sigma}}$. In particular, we adopt the general invertibility factor introduced by Huang and Zhang (2012) specialized to the case of a quadratic loss function.

Definition 2. *For any norm $\|\cdot\|$ and $(p \times p)$ (possibly stochastic) matrix $\boldsymbol{M}$, the general invertibility factor (GIF) over the cone* (3.5) *is given by*

$$\chi(\xi, \mathcal{S}, \|\cdot\|, \boldsymbol{M}) = \mathsf{inf}\left\{ \frac{\|\boldsymbol{x}\|_{\boldsymbol{M}}^2}{\|\boldsymbol{x}_{\mathcal{S}}\|_1 \|\boldsymbol{x}\|} : \boldsymbol{x} \in \mathscr{C}(\xi, \mathcal{S}) \right\}. \tag{3.6}$$

Moreover, we say that $\boldsymbol{M}$ satisfies the GIF condition if $\chi(\xi, \mathcal{S}, \|\cdot\|, \boldsymbol{M}) > 0$.

[8] In fact, the trend is bounded between zero and one by definition.

If we specialize to the case when $\|\boldsymbol{x}\| = \|\boldsymbol{x}_{\mathcal{S}}\|_1/\#\mathcal{S}$, the GIF becomes the $\mathscr{C}(\xi, \mathcal{S})$-restricted infimum of $\frac{\|\boldsymbol{x}\|_M^2 \#\mathcal{S}}{\|\boldsymbol{x}_{\mathcal{S}}\|_1^2}$, which is precisely the square of the compatibility constant defined in van de Geer and Bühlmann (2009). Moreover, we extend the original definition to accommodate a possibly non-deterministic $\boldsymbol{\Sigma}$. Since, as apposed to the deterministic trend case, the Σ does *not* converge to a deterministic matrix in the pure stochastic trend case. In a a low dimensional set-up Masini and Medeiros (2019), shows that Σ converges in distribution to a almost sure positive definite random matrix.

For $\mathcal{S} \supseteq \mathcal{S}_0 := \{i : \beta_{0i} \neq 0\}$ and scalars $\lambda_0, \lambda_1, \lambda_1^* > 0$ and $\lambda_2 \in (0, 1)$ we define the following auxiliary events:

$$\Omega_0 := \left\{ \left\| \frac{2}{T_0} \sum_{t=1}^{T_0} \boldsymbol{W}_t V_t \right\|_\infty \leq \lambda_0 \right\}, \tag{3.7}$$

$$\Omega_1 := \{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_\infty \leq \lambda_1\}, \tag{3.8}$$

$$\Omega_1^* := \{\chi(\xi, \mathcal{S}, \boldsymbol{\Sigma})\} \geq \lambda_1^*\}, \tag{3.9}$$

$$\Omega_2 := \left\{ \sup_{i \in \mathcal{S}} \nu_i \leq 1 + \lambda_2 \right\} \cap \left\{ \inf_{i \in \mathcal{S}^c} \nu_i \geq 1 - \lambda_2 \right\}, \tag{3.10}$$

where $\boldsymbol{\Sigma} := \frac{1}{T_0} \sum_{t=1}^{T_0} \boldsymbol{W}_t \boldsymbol{W}_t'$; $\boldsymbol{\Sigma}_0 := \mathbb{E}(\boldsymbol{\Sigma})$ and $\chi_1(\xi, \mathcal{S}, \boldsymbol{\Sigma}) := \chi(\xi, \mathcal{S}, \|\cdot\|_1/\#\mathcal{S}, \boldsymbol{\Sigma})$ with $\chi(\cdot, \cdot, \cdot, \cdot)$ defined by (3.6).

Proposition 2. *On the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$, provided that $\lambda > 0$ satisfies both*

$$\frac{\lambda_0 + \lambda(1 + \lambda_2)}{\lambda(1 - \lambda_2) - \lambda_0} \leq \xi \quad \textit{and} \quad \lambda_1 \leq \frac{\chi_1(\xi, \mathcal{S}, \boldsymbol{\Sigma}_0)}{2(1 + \xi)^2 \#\mathcal{S}} \tag{3.11}$$

for some $\xi > 0$, the following inequalities hold:

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \leq \frac{2(1 + \xi)[(1 + \lambda_2)\lambda + \lambda_0]\#\mathcal{S}}{\chi_1(\xi, \mathcal{S}, \boldsymbol{\Sigma}_0)}, \quad \textit{and}$$
$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_{\boldsymbol{\Sigma}}^2 \leq \frac{2[(1 + \lambda_2)\lambda + \lambda_0]^2 \#\mathcal{S}}{\chi_1(\xi, \mathcal{S}, \boldsymbol{\Sigma}_0)},$$

where the right hand side diverges to $+\infty$ if $\chi_1(\xi, \mathcal{S}, \boldsymbol{\Sigma}_0) = 0$.

Also, on the event $\Omega_0 \cap \Omega_1^* \cap \Omega_2$, provided that $\lambda > 0$ satisfies

$$\frac{\lambda_0 + \lambda(1 + \lambda_2)}{\lambda(1 - \lambda_2) - \lambda_0} \leq \xi \tag{3.12}$$

for some $\xi > 0$, the following inequalities hold:

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \leq \frac{(1 + \xi)[(1 + \lambda_2)\lambda + \lambda_0]\#\mathcal{S}}{\lambda_1^*}, \quad \textit{and}$$
$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_{\boldsymbol{\Sigma}}^2 \leq \frac{[(1 + \lambda_2)\lambda + \lambda_0]^2 \#\mathcal{S}}{\lambda_1^*}.$$

The set of inequalities conditional on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$ compared to those conditional on $\Omega_0 \cap \Omega_1^* \cap \Omega_2$ are similar and hold for both DGPs in Assumption 2. However, as mention before, Ω_1 cannot hold with probability approaching one in the pure integrated case since Σ does not converge in probability to Σ_0. The event Ω_1^*, however, is expected to hold with high probability as long as we pick $\lambda_1^* > 0$ small enough. In effect this is equivalent to require the smallest restricted eigenvalue of Σ to be bounded away from zero with high probability.

3.2 Estimation

Proposition 2 combined with the probabilistic bounds from the events in (3.7)–(3.10) fully characterize the asymptotic behaviour of $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ and hence, of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$.

Results (a) and (b) of Theorem 1 (below) follow under the condition of what we call *Partial Asymptotics*, i.e., an asymptotic approach only for the pre-intervention period, where the number of post intervention periods $T_2 := T - T_0$ is kept fixed, while $T_0 \to \infty$. This approach is tailored to accommodate situations where the number of pre-intervention periods T_0 is much larger than T_2, which justifies the sampling error from the estimation of $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}}$ 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 post-intervention sample size is constant. In that case, $T \to \infty$.

Theorem 1. *Under Assumptions 1-5 and, for any $c > 0$, set the penalty parameter λ of* (2.12) *by either*

*(*i*) $\lambda = 4cp^{2/q}/\sqrt{T_0}$ under Assumption 3(a); or*

*(*ii*) $\lambda = 4(c + 2\log p)/\sqrt{T_0}$ under Assumption 3(b).*

Suppose that GIF condition in Definition 2 is satisfied with $\xi = 4$. Namely, either:

*(*i*) $\chi_1(\xi, \mathcal{S}_0, \boldsymbol{\Sigma}_0) \geq \epsilon$ for some positive $\epsilon > 0$; or*

*(*ii*) For every $\epsilon > 0$ there is a $\lambda_1^* > 0$ such that $\mathbb{P}\left[\chi_1(\xi, \mathcal{S}_0, \Sigma) < \lambda_1^*\right] < \epsilon$.*

Then, provided $s_0[\psi(p)]^2/\sqrt{T_0} = o(1)$ where $s_0 := \|\boldsymbol{\beta}_0\|_0$ and, under Assumption 3(*b*), $\log p = o[(T_0^{1/4}/\log T_0)^{c_3}]$*, we have on $\boldsymbol{\Omega}_2$ as $T_0 \to \infty$:*

*(*a*) $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 = O_P[\psi(p)s_0/\sqrt{T_0}] = o_P(1)$*

*(*b*) $\widehat{\delta}_t - \delta_t - V_t = O_P[\psi(p)^2 s_0/\sqrt{T_0}] = o_P(1)$ for all $T_0 < t \leq T$*

If further $T_2 \to \infty$:

(c) $\widehat{\Delta}_T - \Delta_T = O_P\left(\frac{\psi(p)s_0}{\sqrt{T_0}} \vee \frac{1}{\sqrt{T_2}}\right) = o_P(1).$

where $\psi(x) = x^{2/q}$ under Assumption 3(a) and $\psi(x) = \log(x)$ under Assumption 3(b).

Result (b) of the theorem above give us conditionally on the event Ω_2 an asymptotic (as $T_0 \to \infty$) mean-unbiased estimator for the treatment effect δ_t for every period in the post-intervention sample. Part (c) give us a consistent estimator (as both T_0 and T_2 diverges to infinity) for the average intervention effect across the post intervention period.

In the traditional setup where all the regressors are stationary the event Ω_2 happens with probability 1 by setting $w_i = 1$ for all $1 \leq i \leq p$. Also in the case of our factor model example in 2.3, 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 to it. Since, in that case we have that $\nu_i \leq 1$ for $i \in \mathcal{S}_0$ and $\nu_i = 1$ otherwise. This fortunate result is a consequence of 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 S_0^c are of smaller or equal order in probability of the process of in variables in S_0.

To extend this result to the general setup, let $w_i = w_{i,t}$ be a possibly stochastic sequence of almost surely non-negative weights. Then, the event Ω_2 happens with probability approaching 1 as long as the events $\{\limsup_t \sup_{i \in S_0} \nu_{i,t} \leq 1\}$ and $\{\liminf_t \inf_{i \in S_0^c} v_{i,t} \geq 1\}$ where $\nu_{it} := w_{it}/\ell_{it}$ also happen with probability approaching one. In particular, if we choose the vector of weights $\boldsymbol{w}$ in (2.12) according to the proposition below, the event Ω_2 occurs with probability approaching one since by the definition of ℓ_i in (3.1), $\nu_i \to 1$ for all $2 \leq i \leq p$ and we are able to state the following result.

Proposition 3. *Under the same conditions of Theorem 1, if we set $w_1 = 0$, such that the intercept is not penalized and for $2 \leq i \leq p$, set $w_i = X_{iT_0}$ under the growth condition (Proposition 1(b)); otherwise, we set $w_i = 1$ if DGP (2.7) or $\sqrt{T_0}$ if DGP (2.8) in Assumption 2, then,*

$$\mathbb{P}(\Omega_2) \to 1 \quad as \quad T_0 \to \infty.$$

In 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. Carefully analysis of the proof of Proposition 3 shows that its approach combined with Assumption 4 suffices for the result.

At this level of generality, namely DGPs with all sort of deterministic and/or stochastic idiosyncratic trends combination, 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. For instance, one could test for the order of integration and/or determinist trend in the unit of interested using classical time series tests since, as previously mention, if all the units are at most the order (in probability) of the unit of interest we can always set the weights to unit. If that is not the case, yet another approach, could be to previously test for cointegration among the variables. For a high-dimension cointegration test refer to Onatski and Wang (2018) or Liang and Schienle (2019). We believe that Theorem 1 coupled with Proposition 3 can guide the practitioner to decide which weight to pick in any particular empirical application.

3.3 Inference

The inference procedure presented in this section is based on the sequence of estimators $\{\widehat{\delta}_t\}_{t>T_0}$ obtained Section 3.2. More specifically, we consider any continuous mappings $\boldsymbol{\phi} : \mathbb{R}^{T_2} \to \mathbb{R}^b$ whose argument is the T_2-dimensional vector $(\widehat{\delta}_{T_0+1} - \delta_{T_0+1}, \ldots, \widehat{\delta}_T - \delta_T)'$. Thus, we are ultimately interested in the distribution of $\widehat{\boldsymbol{\phi}} := \boldsymbol{\phi}(\widehat{\delta}_{T_0+1} - \delta_{T_0+1}, \ldots, \widehat{\delta}_T - \delta_T)$ under the null (2.1) where $\delta_t = 0$ for all $t > T_0$.

As mentioned previously we would like to consider a situation when the pre-intervention period to be substantially longer than the post intervention period, $T_0 \gg T_2$. It could be well the case that $T_2 = 1$. The results are based on part (b) of Theorem 1. As a direct corollary, we have under the asymptotic on the pre-invention period ($T_0 \to \infty$) that $\widehat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}_0 := \boldsymbol{\phi}(V_{T_0+1}, \ldots, V_T)$ by the Continuous Mapping Theorem. Consider the construction of $\widehat{\boldsymbol{\phi}}$ using only blocks of size T_2 of consecutive observations from the pre-intervention sample. There are $T_0 - T_2 - 1$ such blocks denoted by

$$\widehat{\boldsymbol{\phi}}_j := \boldsymbol{\phi}(\widehat{V}_j, \ldots, \widehat{V}_{j+T_2-1}) \quad j = 1, \ldots, T_0 - T_2 + 1,$$

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

For fixed j, we have that $\widehat{\boldsymbol{\phi}}_j \xrightarrow{p} \boldsymbol{\phi}_j := \boldsymbol{\phi}(V_j, \ldots, V_{j+T_2-1})$. 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 $\mathcal{Q}_T(\boldsymbol{x}) := \mathbb{P}(\widehat{\boldsymbol{\phi}} \leq \boldsymbol{x})$ by

$$\widehat{\mathcal{Q}}_T(\boldsymbol{x}) := \frac{1}{T_0 - T_2 + 1} \sum_{j=1}^{T_0-T_2+1} \mathbb{1}(\widehat{\boldsymbol{\phi}}_j \leq \boldsymbol{x}),$$

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

Theorem 2. *For any continuous* $\phi : \mathbb{R}^{T_2} \to \mathbb{R}^b$, *let* $\widehat{\boldsymbol{\phi}} := \boldsymbol{\phi}(\widehat{\delta}_{T_0+1} - \delta_{T_0+1}, \ldots, \widehat{\delta}_T - \delta_T)$ *and* $\boldsymbol{\phi}_0 := \phi(V_{T_0+1}, \ldots, V_T)$. *Consider the same conditions of Theorem 1 but with Assumption 3(a) fulfilled with* $q > 4$; *assume further that* $\{V_t\}$ *is a strictly stationary process and* $s_0 = o\{\sqrt{T_0}/[\psi(p)\psi(pT_0)]\}$, *then we have for fixed* T_2 *as* $T_0 \to \infty$

(a) $\widehat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}_0$

(b) $\widehat{\mathcal{Q}}_T(\boldsymbol{x}) - \mathcal{Q}_T(x) \xrightarrow{p} 0$ *for all* $x \in \mathcal{C}_0 := \{$*continuity point of* $\mathcal{Q}_0(\boldsymbol{x}) := \mathbb{P}(\boldsymbol{\phi}_0 \leq \boldsymbol{x})\}$

(c) *If* $\mathcal{Q}_0(\boldsymbol{x})$ *is continuous, the result (b) holds uniformity in* $\boldsymbol{x} \in \mathbb{R}^b$.

(d) *If* $\boldsymbol{\phi}$ *is real-valued, then* $\mathcal{Q}_T[\widehat{\mathcal{Q}}_T^{-1}(\tau)] \to \tau$ *for all* $\tau \in (0, 1)$ *such that* $Q_0^{-1}(\tau) \in \mathcal{C}_0$, *where* $\mathcal{Q}^{-1}(\tau) := \{\inf x : \mathcal{Q}(x) \geq \tau\}$.

By the appropriate choice of $\boldsymbol{\phi}(\cdot)$, Theorem 2 provides a simple way to conduct inference. We could be interested in testing the intervention effects on all post-intervention periods individually by setting

$$\boldsymbol{\phi}(\widehat{\delta}_{T_0+1}, \ldots, \widehat{\delta}_T) = (\widehat{\delta}_{T_0+1}, \ldots, \widehat{\delta}_T)',$$

or on the average intervention effect across the post-intervention periods

$$\boldsymbol{\phi}(\widehat{\delta}_{T_0+1},\ldots,\widehat{\delta}_T)=\frac{1}{T_2}\sum_{t=T_0+1}^{T}\widehat{\delta}_t.$$

A reasonable choice for testing the null (2.1) using a univariate statistic would be

$$\boldsymbol{\phi}(\widehat{\delta}_{T_0+1},\ldots,\widehat{\delta}_T)=\frac{1}{T_2}\sum_{t=T_0+1}^{T}\widehat{\delta}_t^2,$$

or, more generally, to set

$$\boldsymbol{\phi}(\widehat{\delta}_{T_0+1},\ldots,\widehat{\delta}_T)=\frac{1}{T_2}\sum_{t=T_0+1}^{T}g(\widehat{\delta}_t),$$

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 for confidence intervals.

For instance, we might be interested in a joint confidence set for the vector $\boldsymbol{\delta}:=(\delta_{T_0+1},\ldots,\delta_T)'$; then, we might take $\widehat{\boldsymbol{\phi}}=\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}$ where $\widehat{\boldsymbol{\delta}}:=(\widehat{\delta}_{T_0+1},\ldots,\widehat{\delta}_T)'$. Notice that, unless $T_2=1$, 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

$$C_T:=\bigtimes_{t=T_0+1}^{T}[\widehat{\delta}_t-\widetilde{\mathcal{Q}}_T^{-1}(1-\tau/2);\widehat{\delta}_t-\widetilde{\mathcal{Q}}_T^{-1}(\tau/2)],$$

where $\widetilde{Q}_T^{-1}(\tau)=\mathsf{inf}\{x\in\mathbb{R}:\widehat{\mathcal{Q}}_T(x\iota)\geq\tau\}$ and ι is a vector of T_2 ones. As a direct corollary of Theorem 2 assuming that $\mathcal{Q}_0$ is continuous for any $\tau\in(0,1)$

$$\mathbb{P}\left(\boldsymbol{\delta}\in C_T\right)\to 1-\tau,\quad\text{as } T_0\to\infty.$$

Alternatively, any test procedure based on an univariate test statistic $\widehat{\phi}$ can have its p-value evaluated simply by $1-\widehat{\mathcal{Q}}_T(\widehat{\phi})$ for a one-tailed test or $1-\widehat{\mathcal{Q}}_T(-|\widehat{\phi}|)+\widehat{\mathcal{Q}}_T(|\widehat{\phi}|)$ for a double-tailed test.[9]

4 Simulations

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

[9] Technically, $\widehat{\phi}$ is not a statistic since it depends on the value of the unknown $\{\delta_t\}_{t>T_0}$. However, under the null of interest (2.1), we have $\delta_t=0$.

4.1 Inference

We simulate two baseline models. The number of Monte-Carlo replications is $10,000$. The first simulated model consists of equations (2.4) and (2.6) with independent and identically normally distributed innovations, $n = 200$, $s_0 = 5$. The second baseline DGP differs from the first one by considering equations (2.4) and (2.5). In both cases we simulated $T = 100$ observations and we set $T_2 = 3$. 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 post-treatment observations (T_2), 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). Tables 1 and 2 report size results for model (2.4)–(2.6) and (2.4)–(2.5), respectively. The tables show, for different settings, rejection rates under the null hypothesis of no intervention effect 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 via Bayesian Information Criterion (BIC). We set the maximum penalty level to be $\|\frac{1}{T_0}\sum_{t=1}^{T_0} Y_t \boldsymbol{X}_t\|_\infty$ with an exponential path down to $\lambda_{\mathsf{min}} = 0.001$ along 100 equally spaced intervals in the `glmnet` package. **Oracle** means thatthe counterfactual is estimated by 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 ($\boldsymbol{\beta}_0$). All distributions are standardized (zero mean and unit variance). Mixed normal means to two 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.

Several conclusions emerge from the tables. First the size distortions of the LASSO are comparable to the ones from the Oracle and slightly superior than the ones from the true model. Note that the size distortions from the true model reflects only the estimation error of the cumulative distribution of V_t. On the other hand, the other two cases reflect also the estimation error of the $\boldsymbol{\beta}_0$ parameter. 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 got close to zero as the sample increases. The number of units (n) seems to influence more in the case of stochastic trends, where the distortions for the case when $n = 1000$ can be non-negligible. In addition, high residual autocorrelation, as expected, can cause more distortions. Finally, the number of observations after the intervention seems also to have an effect on the text. However, the distortions are not large. Overall, the proposed inference procedure works extreme satisfactorily, specially for the 0.1 significance level.

Table 3 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 the table that the test has nontrivial power against the alternatives.

4.2 Parameter Estimation

The table reports several statistics averaged over 10,000 replications for each one of four data generating processes. More specifically, mean ℓ_1-norm is the average $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$, mean bias is the average bias $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ over the simulations, mean MSE is the average mean squared error, and mean Δ is the average intervention effect over the 10 out-of-sample periods. Note that the true value of Δ is zero. MSE Δ is the average squared error over the simulation and, finally, median Δ is the median of the estimates of Δ over the simulations. Each column in the table represents a variation of the baseline scenario, in which we set $T = 100$,$s_0 = 5$, $n = 100$ and $\rho = 0$. Model (1) is given by equations (2.4) and (2.5) where $f_t^F = 0$. Model (2) is given by equations (2.4) and (2.5) where $f_t^F = 1$. Model (3) is given by equations (2.4) and (2.6) where $f_t^F = t$. Model (4) is given by equations (2.4) and (2.6) where $f_t^F = t^2$.

As expected the ℓ_1-norm, the bias, and the MSE of the estimators decrease with the sample size, but increase as the degree of sparsity decreases (s_0 grows), as the number of covariates grows or as the autocorrelation in the errors increases. Nevertheless, the biases are negligible. Concerning the estimator of the average intervention effect (Δ), the estimators are rather precise when the trends are deterministic. On the other hand, with stochastic trends, the biases are small only with no error autocorrelation.

5 Empirical Illustrations

5.1 Heterogeneous Effects and the Price Elasticity of Demand

We illustrate the proposed inferential procedures for counterfactual analysis with an application to optimal price setting in the retail industry in Brazil. Our dataset consist of the daily prices and quantities sold of a product A, commercialized by one of the major retail chains in Brazil, which has approximately 1,000 stores distributed in more than 400 municipalities over the country.[10] 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 about 50% 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. More specifically, 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).[11] The selection of the treatment and control groups was carried out according to the 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

[10] Due to a confidentiality agreement, we are not allowed to disclosure either the name of the product or the name of the retail chain.

[11] A different experiment was running during the same period in the other half of the municipalities. Therefore, we decided to exclude these cities in order to avoid potential sources of biases.

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 valid Assumption 1. Second, although according to municipality characteristics, we keep a homogenous 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 trend. Therefore, due to the facts described before, we advocate the use of the methodology proposed in this paper and not alternatives, such as the difference-in-differences estimator.

For each day t, q_{it} represents the total quantities sold of product A in all stores of municipality i, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Our sample runs from June 20, 2016, to October 31, 2016, representing 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 reduced in Δ_p Brazilian Reais, while for the other municipalities, they were kept fixed. The first 126 municipalities are in the control group ($i = 1, \ldots, 126$), whereas the remaining 107 are in the treatment group ($i = 127, \ldots, 233$). The number of pre-treatment 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 that seems to be linear and deterministic. 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 pre-treatment sample.[12] There is a clear heterogeneity in the trend pattern that precludes the use of the traditional differences-in-differences estimator. These facts are corroborated with the results presented in Table 5. The table reports the estimated coefficients of a linear trend model for the total sold quantities in each group as well as the coefficients of the linear trend, when dummies to control for the days-of-the-week effect are included in the model. The numbers between parentheses are heteroskedastic-autocorrelation robust (HAC) standard errors. The table also presents the results of the augmented Dickey-Fuller (ADF) test for the null of unit roots against the alternative of a trend-stationary model. The null of unit-roots are strongly rejected for the control group. For the treatment group the null is rejected at a 7% level. When both groups are merged together, the rejection is at a 6% level. As it is well known that ADF tests have low power in small samples, the results provide strong evidence in favor of a trend-stationary model.

To determine the optimal price of the product, 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

[12] For each municipality, we estimate by ordinary least squares the following linear trend model: $q_{it} = \alpha_i + \beta_i t + u_t$. Panel (b) in Figure 1 displays the empirical distribution of $\widehat{\beta}$ across municipalities.

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

$$\begin{aligned} q_t &= \beta_0 + \sum_{i=1}^{126} \beta_i q_{it} + \pi_1 \mathsf{Mon}_t + \pi_2 \mathsf{Tue}_t + \pi_3 \mathsf{Wed}_t + \pi_4 \mathsf{Thu}_t + \pi_5 \mathsf{Fri}_t + \pi_6 \mathsf{Sat}_t + V_t, \\ &= \boldsymbol{X}_t' \boldsymbol{\beta} + V_t \end{aligned} \tag{5.1}$$

by the WLASSO procedure described in the paper using the 120 observations from June 20, 2016, to October 18, 2016 (pre-treatment sample). $\mathsf{Mon}_t, \ldots, \mathsf{Sat}_t$ are six dummies for the days of the week. As we include a constant in the model, we omit the dummy for Sundays. The penalty parameter of the WLASSO procedure is selected by the BIC.

2. Project the counterfactual for the treatment period as

$$\widehat{q_t} = \boldsymbol{X}_t' \widehat{\boldsymbol{\beta}}$$

and compute

$$\delta_t = q_t - \widehat{q_t}.$$

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 6. The figure shows the actual sales, the estimated counterfactual, as well as a 95% confidence interval using the partial resampling method described in Section 3.3, 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 $\phi(x) = \frac{1}{T_2}\sum_{j=1}^{T_2} x_j^2$ and $\phi(x) = \frac{1}{T_2}\sum_{j=1}^{T_2} |x_j|$. Table 6, 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 4,000 units, potentially causing a great impact in terms of revenue and profit. The table also reports the R-squared and the number of selected regressors with the WLASSO method. It is clear that the model has a good in-sample fit (R-squared=0.96).

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 (5.1) by the following model:

$$
\begin{aligned}
q_{jt} &= \beta_{k0} + \sum_{i=1}^{126} \beta_{ki} q_{it} + \pi_{k1}\mathsf{Mon}_t + \pi_{k2}\mathsf{Tue}_t + \pi_{k3}\mathsf{Wed}_t + \pi_{k4}\mathsf{Thu}_t + \pi_{k5}\mathsf{Fri}_t + \pi_{k6}\mathsf{Sat}_t + V_{jt}, \\
&= \boldsymbol{X}'_{jt}\boldsymbol{\beta}_k + V_{jt}, \quad j = 126, \ldots, 233;\ k = j - 126.
\end{aligned}
\tag{5.2}
$$

The results are displayed in Panel (b) of Table 6. 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 $\phi(x) = \frac{1}{T_2}\sum_{j=1}^{T_2} x_j^2$ and $\phi(x) = \frac{1}{T_2}\sum_{j=1}^{T_2} |x_j|$ 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.

6 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 paper allows for potentially high-dimensional and non-stationary 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 post-intervention observations), and we also developed an inferential procedure based on partial re-sampling to test the general hypothesis on the intervention effects. Our testing procedure does not rely on post-intervention asymptotics.

A Proof of the Main Results

A.1 Proof of Proposition 1

In light of representation (2.10), it is enough for the result (a) to show that η_{it}/d_{it} vanishes in the appropriate sense as $t \to \infty$. Under DGP (2.7), we have

$$\frac{\eta_{it}}{d_{it}} = \frac{Z_{i0}^{(0)}}{d_{it}} + \frac{\sum_{s=1}^{t} U_{is}}{\sqrt{t}} \frac{\sqrt{t}}{d_{it}} = o_P(1) + O_P(1)o(1) = o_P(1),$$

where $O_P(1)$ term is a consequence of Assumption 3. Under DGP (2.8), we have simply $\eta_{it}/d_{it} = U_{it}/(c_i + f_{it}) \to 0$, almost surely as $f_{it} \to \infty$.

As for result (b), we have for DGP (2.7), $Z_{it}^{(0)} = d_{it} + Z_{it}^{(0)} + \sum_{s=1}^{t} U_{it} = O(\sqrt{t}) + O_P(1) + O_P(\sqrt{t}) = O_P(\sqrt{t})$ and for DGP (2.8), $Z_{it}^{(0)} = c_i + f_{it} + U_{it} = O(1) + O(1) + O_P(1) = O_P(1)$.

Finally, under DGP (2.7), if $d_{it} = o(\sqrt{t})$, we have the result by the central limit theorem (ensured by Assumption 3) combined with Slutsky theorem, since $t^{-1/2} Z_{it}^{(0)} = o(1) + t^{-1/2} \sum_{s=1}^{t} U_{it}$.

A.2 Proof of Proposition 2

On the events defined by (3.7)–(3.10), we conclude from (3.4) combined with the first condition of (3.11) that the following inequalities hold:

$$\|\widehat{\gamma} - \gamma_0\|_{\Sigma}^2 \leq [\lambda_0 + \lambda(1 + \lambda_2)] \|\widehat{\gamma}_{\mathcal{S}} - \gamma_{0,\mathcal{S}}\|_1, \quad \text{and} \tag{A.1}$$

$$\|\widehat{\gamma}_{\mathcal{S}^c}\|_1 \leq \xi \|\widehat{\gamma}_{\mathcal{S}} - \gamma_{0,\mathcal{S}}\|_1 \tag{A.2}$$

Trivially, using (A.2), we can write:

$$\|\widehat{\gamma} - \gamma_0\|_1 = \|\widehat{\gamma}_{\mathcal{S}} - \gamma_{0,\mathcal{S}}\|_1 + \|\widehat{\gamma}_{\mathcal{S}^c} - \gamma_{0,\mathcal{S}^c}\|_1 \leq (1 + \xi) \|\widehat{\gamma}_{\mathcal{S}} - \gamma_{0,\mathcal{S}}\|_1. \tag{A.3}$$

Now from the definition of λ_1 in the event Ω_1, we have

$$\left| \|\widehat{\gamma} - \gamma_0\|_{\Sigma}^2 - \|\widehat{\gamma} - \gamma_0\|_{\Sigma_0}^2 \right| = |(\widehat{\gamma} - \gamma_0)'(\Sigma - \Sigma_0)(\widehat{\gamma} - \gamma_0)| \leq \lambda_1 \|\widehat{\gamma} - \gamma_0\|_1^2.$$

In addition, from the definition of the GIF condition applied to the matrix Σ_0, we have on the cone (3.5)

$$\|\widehat{\gamma}_{\mathcal{S}} - \gamma_{0,\mathcal{S}}\|_1^2 \leq \frac{\|\widehat{\gamma} - \gamma_0\|_{\Sigma_0}^2 \# \mathcal{S}}{\chi_1(\xi, \mathcal{S}, \Sigma_0)}.$$

Combine the last three displays to conclude that

$$\left| \|\widehat{\gamma} - \gamma_0\|_{\Sigma}^2 - \|\widehat{\gamma} - \gamma_0\|_{\Sigma_0}^2 \right| \leq \lambda_1 (1 + \xi)^2 \frac{\|\widehat{\gamma} - \gamma_0\|_{\Sigma_0}^2 \# \mathcal{S}}{\chi_1(\xi, \mathcal{S}, \Sigma_0)}.$$

Notice that the second condition on (3.11) implies that $2(1+\xi)^2\lambda_1\#\mathcal{S}/\chi_1(\xi,\mathcal{S},\mathbf{\Sigma}_0) \leq 1$. Then, we can re-write the previous expression as

$$\left|\frac{\|\widehat{\gamma}-\gamma_0\|_{\mathbf{\Sigma}}^2}{\|\widehat{\gamma}-\gamma_0\|_{\mathbf{\Sigma}_0}^2}-1\right| \leq \frac{2\lambda_1(1+\xi)^2\#\mathcal{S}}{2\chi_1(\xi,\mathcal{S},\mathbf{\Sigma}_0)} \leq \frac{1}{2}.$$

Once again, using the GIF condition on $\mathbf{\Sigma}_0$, the previous result and the inequality (A.1) yield

$$\|\widehat{\gamma}_{\mathcal{S}}-\gamma_{0,\mathcal{S}}\|_1 \leq \frac{\|\widehat{\gamma}-\gamma_0\|_{\mathbf{\Sigma}}^2\#\mathcal{S}}{\|\widehat{\gamma}_{\mathcal{S}}-\gamma_{0,\mathcal{S}}\|_1\chi_1(\xi,\mathcal{S},\mathbf{\Sigma}_0)}\frac{\|\widehat{\gamma}-\gamma_0\|_{\mathbf{\Sigma}_0}^2}{\|\widehat{\gamma}-\gamma_0\|_{\mathbf{\Sigma}}^2} \leq \frac{2[\lambda_0+\lambda(1+\lambda_2)]\#\mathcal{S}}{\chi_1(\xi,\mathcal{S},\mathbf{\Sigma}_0)}.$$

Finally, the last inequality combined with (A.3) yields the first result of the Proposition, and combined with (A.1), it yields the second result.

A.3 Proof of Theorem 1

We divide the proof into three steps. First, we show that under the hypotheses of the Theorem that the process $\{\boldsymbol{W}_tV_t\}_{t\geq 1}$ can be properly bounded. Then, we show that the event $\Omega_0\cap\Omega_1$ occurs with high probability. Finally, we derive the results of the Theorem.

A.3.1 Bound Control

We have $\boldsymbol{W}_t = \boldsymbol{L}^{-1}\boldsymbol{X}_t = \boldsymbol{L}^{-1}(\boldsymbol{d}_t+\boldsymbol{\eta}_t)$ where $\boldsymbol{d}_t := (d_{1t},\ldots,d_{pt})'$ and $\boldsymbol{\eta}_t := (\eta_{1t},\ldots,\eta_{pt})'$ for $t \geq 1$. Then, for the DGP (2.8) in Assumption 2, recall that $\boldsymbol{\eta}_t = \boldsymbol{U}_t$, $\boldsymbol{d}_t = c + \boldsymbol{\mu}f_t$ and $\boldsymbol{L}$ is just a deterministic diagonal matrix. Hence, the process $\{\boldsymbol{W}_t\}$ is strong mixing with the same coefficient as the process $\{\boldsymbol{U}_t\}$. Moreover the process $\{V_t\}$, as a linear combination of $\boldsymbol{U}_t$, is also strong mixing with the same mixing coefficient as the process $\{\boldsymbol{U}_t\}$. Therefore, the process $\{\boldsymbol{W}_tV_t\}$ is also strong mixing with the same mixing coefficient as the process $\{\boldsymbol{U}_t\}$ under Assumption 3. Also, by definition of the scaling matrix $\boldsymbol{L}$, all the components of the vector $\boldsymbol{L}^{-1}\boldsymbol{d}_t$ are bounded between 0 and 1. If the process $\{\boldsymbol{U}_t\}$ fulfills condition (a) of Assumption 3 so does $\{V_t\}$ because $V_t = U_{1t} - \sum_{i=2}^{n}\beta_{0,i}U_{it}$ and

$$\|V_t\|_{\mathcal{L}^q} \leq |\|U_{1t}\|_{\mathcal{L}^q} + \sum_{i=2}^{n}|\beta_{0,i}|\|U_{it}\|_{\mathcal{L}^q} = O(\|\boldsymbol{\beta}_0\|_1) = O(1).$$

Then, by Cauchy-Schwartz inequality, we have that $\{\boldsymbol{W}_tV_t\}$ fulfills the same condition with constant $q/2$ since for some $\epsilon > 0$ we have

$$\sup_{t\in\mathbb{N}}\sup_{i\leq p}\mathbb{E}|U_{it}V_t|^{q/2+\epsilon/2} \leq \left(\sup_{t\in\mathbb{N}}\sup_{i\leq p}\mathbb{E}|U_{it}|^{q+\epsilon}\sup_{t\in\mathbb{N}}\sup_{i\leq p}\mathbb{E}|V_t|^{q+\epsilon}\right)^{1/2} < \infty.$$

Furthermore, if $\{(V_t,\boldsymbol{U}_t')'\}$ also fulfills condition (b) of Assumption 3 with the triple (a_1,a_2,a_3) in the exponential bound, then the process $\{\boldsymbol{W}_tV_t\}$ complies with Assumption 3(b) with the triple $(2a_1, a_2, a_3/2)$ since for each component of the vector $\boldsymbol{U}_t V_t$ is bounded by

$$\mathbb{P}(|U_{it}V_t| > u) \leq \mathbb{P}(|U_{it}| > \sqrt{u}) + \mathbb{P}(|V_t| > \sqrt{u}) \leq 2a_1 \exp(-a_2 u^{a_3/2}).$$

Now consider DGP (2.7). We find bounds for $\| \sum_{t=1}^{T_0} W_{it}V_t\|_{\mathcal{L}^q}$ and $\| \sum_{t=1}^{T_0} W_{it}W_{jt}\|_{\mathcal{L}^q}$ uniformly in $t \leq T_0$ and $1 \leq i, j \leq p$. For the latter we have

$$\left\|\sum_{t=1}^{T_0} W_{it}W_{jt}\right\|_{\mathcal{L}^q} \leq \sum_{t=1}^{T_0} \frac{d_{it}d_{jt}}{\ell_i \ell_j} + \frac{1}{\ell_j}\left\|\sum_{t=1}^{T_0} \frac{d_{it}}{\ell_i}\eta_{jt}\right\|_{\mathcal{L}^q} + \frac{1}{\ell_i}\left\|\sum_{t=1}^{T_0} \frac{d_{jt}}{\ell_j}\eta_{it}\right\|_{\mathcal{L}^q} + \frac{1}{\ell_i \ell_j}\left\|\sum_{t=1}^{T_0} \eta_{it}\eta_{jt}\right\|_{\mathcal{L}^q}.$$

Since $d_{it}/\ell_i \in [0, 1]$ for all i by definition, the first term is $O(T_0)$. The second and third terms are $O(T_0^{3/2}/l_j)$ and $O(T_0^{3/2}/l_i)$ respectively by result (b) of Lemma 1 and the last one if $O(T_0^2/(\ell_i \ell_j))$ from result (c) of Lemma 1. From which we conclude that

$$\left\|\sum_{t=1}^{T_0} W_{it}W_{jt}\right\|_{\mathcal{L}^q} = O\left(T_0 \vee \frac{T_0^{3/2}}{\ell_i \wedge \ell_j} \vee \frac{T_0^2}{\ell_i \ell_j}\right) = O(T_0).$$

For the former, we start by the triangle inequality

$$\left\|\sum_{t=1}^{T_0} W_{it}V_t\right\|_{\mathcal{L}^q} \leq \left\|\sum_{t=1}^{T_0} \frac{d_{it}}{\ell_i}V_t\right\|_{\mathcal{L}^q} + \frac{1}{\ell_i}\left\|\sum_{t=1}^{T_0} \eta_{it}V_t\right\|_{\mathcal{L}^q}.$$

The first term is $O(\sqrt{T_0})$ by result (a) of Lemma 1. For the second term we may use result (c) and Hölder's inequality to obtain

$$\left\|\sum_{t=1}^{T_0} \eta_{it}V_t\right\|_{\mathcal{L}^q} = \left\|\sum_{t=1}^{T_0} \eta_{it}U_{1t}\right\|_{\mathcal{L}^q} + \sum_{j=2}^{n} \beta_{0,j}\left\|\sum_{t=1}^{T_0} \eta_{it}U_{jt}\right\|_{\mathcal{L}^q} = O(T_0 \vee T_0\|\boldsymbol{\beta}_0\|_1) = O(T_0).$$

Hence, second term is $O(T_0/\ell_i)$ by result (a) and therefore

$$\left\|\sum_{t=1}^{T_0} W_{it}V_t\right\|_{\mathcal{L}^q} = O(\sqrt{T_0} \vee T_0/\ell_i) = O(\sqrt{T_0}).$$

A.3.2 Probability Bounds on Ω_0 and Ω_1

In light of the results in the previous subsection we can set $\lambda_0 = \lambda/2$ with λ as stated in the theorem. For DGP (2.8), results (b) and (c) of Lemma 2 allow us to conclude that for all $c > 0$:

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

We start by showing that $\mathbb{P}(\Omega_1) \to 1$. Recall that $\mathbb{P}(\Omega_1^c) = \mathbb{P}(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_\infty > \lambda_1)$. Set $\lambda_1 = \chi_1(\xi, \mathcal{S}, \boldsymbol{\Sigma}_0)/[2(1+\xi)^2 s]$ and $x = \lambda_1\sqrt{T_0}$ in Lemma 2. Results (d) and (e) in Lemma 2 imply that

$$\mathbb{P}(\Omega_1^c) = \begin{cases} O\left[\left(\frac{p^{2/q}s}{\sqrt{T_0}}\right)^q\right] = o(1) & \text{under Assumption 3(a),} \\ O\left\{\exp\left[2\log p - \frac{\chi_1\sqrt{T_0}}{4(1+\xi)^2 s}\right]\right\} = o(1) & \text{under Assumption 3(b),} \end{cases}$$

where the $o(1)$ terms follow by assumption of the theorem since $p^{4/q}s/\sqrt{T_0} = o(1)$ and $s\log p/\sqrt{T_0} = o(1)$.

Also, from the relation $\lambda = 2\lambda_0$, we may choose $\lambda_2 > 0$ arbitrarily close to 0 such that the condition (3.11) in Proposition 2 is fulfilled with ξ arbitrarily close to 3. For instance, setting $\lambda_2 = 1/10$ yields

$$\frac{\lambda_0 + \lambda(1+\lambda_2)}{\lambda(1-\lambda_2) - \lambda_0} = \frac{1 + 2(1+\lambda_2)}{2(1-\lambda_2) - 1} = \frac{3 + 2\lambda_2}{1 - 2\lambda_2} = 4 =: \xi.$$

Provided that the GIF condition holds, i.e., $\chi_1(4, \mathcal{S}, \boldsymbol{\Sigma}_0) > 0$, we have for λ as stated in the theorem and for all $c > 0$:

$$\mathbb{P}(\Omega_0 \cap \Omega_1) \geq 1 - \begin{cases} O(c^{-q/2}) & \text{under Assumption 3(a),} \\ O[\exp(-c/2)] & \text{under Assumption 3(b).} \end{cases}$$

Similarly for the DGP (2.7) under Assumption 3(a) by setting λ as stated in the theorem yields

$$\mathbb{P}(\Omega_0^c) = \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{W}_t V_t\right\|_\infty > \frac{\lambda_0}{2}\right) = O(c^{-q/2}) \quad \text{and}$$

$$\mathbb{P}(\Omega_1^c) = \varepsilon$$

A.3.3 Final Results

Combining the previous display with the results of Proposition 2, we conclude for λ as stated in the theorem that on Ω_2 we have

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 = O_P(\lambda s_0) \quad \text{and} \quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_{\boldsymbol{\Sigma}}^2 = O_P(\lambda^2 s_0).$$

Notice that $\lambda = O[\psi(p)/\sqrt{T_0}]$ with $\psi(x) = x^{2/q}$ under Assumption 3(a) and $\psi(x) = \log x$ under Assumption 3(b). Then, result (a) of the theorem follows since $s_0 = o(\lambda^{-1})$ by assumption. For the remaining results, we use the fact that

$$\widehat{\delta}_t - \delta_t = V_t + (\widehat{\boldsymbol{\gamma}}_{T_0} - \boldsymbol{\gamma}_0)'\boldsymbol{W}_t, \quad T_0 < t \leq T.$$

For (b), we have, according to Hölder's inequality, that $|\widehat{\delta}_t - \delta_t - V_t| = |(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)'\boldsymbol{W}_t| \leq \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \|\boldsymbol{W}_t\|_\infty$. The first term is $O_P(\lambda s_0)$ from (a), and the second term is $O_P[\psi(p)]$ from Lemma 2(a). Hence, $\widehat{\delta}_t - \delta_t - V_t = O_P[\psi(p)^2 s_0/\sqrt{T_0}] = o_P(1)$ also by Assumption. For (c), we have

$$\widehat{\Delta}_T - \Delta_T := \frac{1}{T_2}\sum_{t>T_0}\widehat{\delta}_t - \delta_t = \frac{1}{T_2}\sum_{t>T_0} V_t - (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)'\frac{1}{T_2}\sum_{t>T_0}\boldsymbol{W}_t.$$

The first term is $O_P(1/\sqrt{T_2})$ under Assumption 3, and the absolute value of the second term is upper bounded by Hölder's inequality, since

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \left\|\frac{1}{T_2}\sum_{t>T_0}\boldsymbol{W}_t\right\|_\infty \leq \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \left(\left\|\frac{1}{T_2}\sum_{t>T_0}\boldsymbol{W}_t - \mathbb{E}(\boldsymbol{W}_t)\right\|_\infty + \left\|\frac{1}{T_2}\sum_{t>T_0}\mathbb{E}(\boldsymbol{W}_t)\right\|_\infty\right).$$

The first term in parentheses is $O_P[\psi(p)/\sqrt{T_2}]$ by Lemma 2(b), whereas the second is $O(1)$. Therefore, under the assumptions of the theorem, the term in parentheses is $O_P(1)$. The term outside the parentheses is $O_P[\psi(p)s_0/\sqrt{T_0}]$ by result (a). Hence, $(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)'\frac{1}{T_2}\sum_{t>T_0}\boldsymbol{W}_t = O_P[\psi(p)s_0/\sqrt{T_0}]$ and, therefore

$$\widehat{\Delta}_T - \Delta_T = O_P\left[\frac{\psi(p)s_0}{\sqrt{T_0}} \vee \frac{1}{\sqrt{T_2}}\right].$$

A.4 Proof of Proposition 3

According to the Proposition, let $\mathcal{R}$ is the index set of the stochastic (non-deterministic) w_i. From the definition of Ω_2 we conclude that

$$\Omega_2 = \left\{\sup_{i\in S}\nu_S \leq 1 + \lambda_2\right\} \cap \left\{\inf_{i\in S^c}\nu_{S^c} \geq 1 - \lambda_2\right\} \supseteq \left\{\sup_{i\in\mathcal{H}}|\nu_i - 1| \leq \lambda_2\right\}.$$

To see that it is indeed the case, recall that the intercept is always included in the model (belongs to S). Hence, $\nu_1 = 0 \leq 1 + \lambda_2$ for any $\lambda_2 \in (0,1)$. For $i > 1$, ν_i is either 1, in that case trivially $1 - \lambda_2 \leq \nu_i \leq 1 + \lambda_2$, or $\nu_i = 1 + \eta_{iT_0}/d_{iT_0}$.

We now show that $\sup_{i\in\mathcal{R}}|\eta_{it}/d_{it}| = o_P(1)$ as $t \to \infty$. For DGP (2.7), we have $\eta_{iT_0}/d_{iT_0} = \left(\frac{1}{\sqrt{T_0}}\sum_{t=1}^{T_0}U_{it}\right)\frac{\sqrt{T_0}}{d_{iT_0}}$ for $i \in \mathcal{H}$ in Assumption 2. Thus,

$$\sup_{i\in\mathcal{H}}|\eta_{iT_0}/d_{iT_0}| \leq \sup_{i\in\mathcal{H}}\left|\frac{1}{\sqrt{T_0}}\sum_{t=1}^{T_0}U_{it}\right|\frac{\sqrt{T_0}}{\inf_{i\in\mathcal{H}}|d_{iT_0}|}.$$

Let $d_{\mathcal{R}}(T_0) := \inf_{i\in\mathcal{R}}|d_{iT_0}|$. Since $\{U_t\}$ is a zero mean strong-mixing process by assumption, we can apply Lemma 2(b) to conclude that

$$\sup_{i\in\mathcal{R}}|\nu_i - 1| = \begin{cases} O_P\left[\frac{(\#\mathcal{R})^{1/q}\sqrt{T_0}}{d_{\mathcal{R}}(T_0)}\right] = o_P(1) & \text{under Assumption 3(a),}\\ O_P\left[\frac{\sqrt{T_0}\log(\#\mathcal{R})}{d_{\mathcal{R}}(T_0)}\right] = o_P(1) & \text{under Assumption 3(b).}\end{cases}$$

For DGP (2.8) in Assumption 2 we have that $\eta_{iT_0}/d_{iT_0} = U_{iT_0}/d_{iT_0}$. Then, $\sup_{i\in\mathcal{H}}|U_{iT_0}/d_{iT_0}| \leq$

$\mathsf{sup}_{i\in\mathcal{H}} |U_{iT_0}|/ \mathsf{inf}_{i\in\mathcal{H}} |d_{iT_0}|$. Applying Lemma 2(a), we have that

$$\sup_{i\in\mathcal{H}} |\nu_i - 1| = \begin{cases} O_P\left[\frac{(\#\mathcal{R})^{1/q}}{d_{\mathcal{R}}(T_0)}\right] = o_P(1) & \text{under Assumption 3(a)}, \\ O_P\left[\frac{\log(\#\mathcal{R})}{d_{\mathcal{R}}(T_0)}\right] = o_P(1) & \text{under Assumption 3(b)}, \end{cases}$$

where all the $o_P(1)$ terms follow from Assumption 4.

A.5 Proof of Theorem 2

Part (a) follows directly from Theorem 1 (b), combined with the continuous mapping theorem. We prove (b) by showing that both $\widehat{\mathcal{Q}}_T(x) - \mathcal{Q}_0(x) = o_P(1)$ and $\mathcal{Q}_T(x) - \mathcal{Q}_0(x) = o(1)$, as $T_0 \to \infty$ for all $x \in \mathcal{C}_0$, the continuity points of $\mathcal{Q}_0(x) := \mathbb{P}(\phi_0 \leq x)$. The result then follows by the triangle inequality. For the latter, as a consequence of result (a), we have $\widehat{\boldsymbol{\phi}} \Rightarrow \boldsymbol{\phi}_0$. For the former, let $\widetilde{\mathcal{Q}}_T(x) := \frac{1}{\tau}\sum_{j=1}^{\tau} \mathbb{1}(\phi_j \leq x\}$ be the unfeasible counterpart of $\widehat{\mathcal{Q}}(x)$, where $\tau := T_0 - T_2 + 1$. We first show that $\widetilde{\mathcal{Q}}_T(x) - \mathcal{Q}_0(x)$ vanishes in probability as $T_0 \to \infty$. Due to the strict stationarity assumption, $\mathbb{E}[\widetilde{\mathcal{Q}}_T(x)] = \frac{1}{\tau}\sum_{j=1}^{\tau} \mathbb{P}(\psi_j \leq x) = \mathbb{P}(\psi_0 \leq x) =: \mathcal{Q}_0(x)$. Hence, $\widetilde{\mathcal{Q}}_T(x)$ is unbiased for $\mathcal{Q}_0(x)$. So, it is enough to show that $\mathbb{E}\left[\widetilde{\mathcal{Q}}^2_T(x)\right]$ converges to zero. Notice that the sequence $\{A_j := \mathbb{1}(\phi_j \leq x)\}_j$ is stationary. For this reason,

$$\mathbb{E}\left[\widetilde{\mathcal{Q}}^2_T(x)\right] = \frac{1}{\tau}\sum_{|k|<\tau}\left(1 - \frac{|k|}{\tau}\right)\boldsymbol{\gamma}_k, \quad \boldsymbol{\gamma}_k := \mathbb{E}(A_1 A_{1+k}).$$

In addition, $0 \leq A_j \leq 1$, so we can bound the first $T_2 - 1$ covariances by 1 and the remaining covariances using a mixing inequality due to Ibragimov (1962), as for $|k| \geq T_2$, we have $\boldsymbol{\gamma}_k \leq 4\alpha(k - T_2 + 1)$, where $\alpha(m)$ is the mixing coefficient of the process $\{V_t\}_t$. In fact, the sequence $\{A_j(\nu_j, \ldots, \nu_{j+T_2-1}\}_j$ is also strong mixing. Then,

$$\mathbb{E}\left[\widetilde{\mathcal{Q}}^2_T(x)\right] \leq \frac{2T_2 + 1}{\tau} + \frac{8}{\tau}\sum_{k=T_2}^{\tau} \alpha(k - T_2 + 1).$$

Finally, since $T_0 \to \infty$ implies $\tau \to \infty$, we have that the first term converging to zero, and the second term converges to zero due to Assumption 3, which establishes that $\widetilde{\mathcal{Q}}_T(x) - \mathcal{Q}_0(x) = o_P(1)$ for all x.

Now we write $\widehat{\mathcal{Q}}(x) = \frac{1}{\tau}\sum_{j=1}^{\tau} I[\phi_j + (\widehat{\phi}_j - \phi_j) \leq x]$ and, for any $\epsilon > 0$, we define the event $\mathscr{A}_T(\epsilon) := \{\mathsf{sup}_j \|\widehat{\phi}_j - \phi_j\|_\infty \leq \epsilon\}$. On $\mathscr{A}_T$, we have that

$$\widetilde{\mathcal{Q}}(x - \epsilon\iota) \leq \widehat{\mathcal{Q}}(x) \leq \widetilde{\mathcal{Q}}(x + \epsilon\iota),$$

where $\iota \in \mathbb{R}^b$ is a vector of 1s. If we add a further condition that $\mathscr{B}_T(\epsilon, x) := \{|\widetilde{\mathcal{Q}}(x - \epsilon\iota) - \mathcal{Q}_0(x - \epsilon\iota)| \vee |\widetilde{\mathcal{Q}}(x + \epsilon\iota) - \mathcal{Q}_0(x + \epsilon\iota)| \leq \epsilon\}$, we have

$$\mathcal{Q}_0(x - \epsilon\iota) - \epsilon \leq \widehat{\mathcal{Q}}(x) \leq \mathcal{Q}_0(x + \epsilon\iota) + \epsilon.$$

Now take $\epsilon \to 0$ to conclude that, conditional on $\mathscr{A}_T \cap \mathscr{B}_T$, we have $|\widehat{\mathcal{Q}}(x) - \mathcal{Q}_0(x)| \leq \epsilon$ for all $x \in \mathcal{C}_0$.

Therefore, it is enough to show that $\mathbb{P}(\mathscr{A}_T \cap \mathscr{B}_T) = 1$ establishes the result (b). $\mathscr{B}_T$ is a sure event as $\widetilde{\mathcal{Q}}(x) \to \mathcal{Q}_0(x)$ for all $x \in C_0$. As for $\mathscr{A}_T$, notice that for $1 \leq t \leq T_0$, we have $\widehat{V}_t - V_t = (\widehat{\boldsymbol{\gamma}}_{T_0} - \boldsymbol{\gamma}_0)'\boldsymbol{W}_t$. As a consequence, by Hölder's inequality,

$$\sup_{t \leq T_0} |\widehat{V}_t - V_t| \leq \|\widehat{\boldsymbol{\gamma}}_{T_0} - \boldsymbol{\gamma}_0\|_1 \sup_{t \leq T_0} \|\boldsymbol{W}_t\|_\infty = \|\widehat{\boldsymbol{\gamma}}_{T_0} - \boldsymbol{\gamma}_0\|_1 \sup_{t,i} |W_{it}|.$$

The first term is $O_P[s_0\psi(p)/\sqrt{T_0}]$ by Theorem 1(a), and the second term is $O_P[\psi(pT_0)]$ by Lemma 2(a). Then, under the assumptions of the theorem, we conclude that $\sup_{t \leq T_0} |\widehat{V}_t - V_t| = O_P[s_0\psi(p)\psi(pT_0)/\sqrt{T_0}] = o_P(1)$. Since $\phi(\cdot)$ is continuous, the last result implies $\sup_j \|\widehat{\phi}_j - \phi_j\|_\infty = o_P(1)$.

For (c) and (d), we use the fact that (b) is equivalent (refer to Theorem 6.3.1 of Resnick (1999)) to say that for any subsequence $\{T_j\}$, we can extract a further subsequence $\{T_{j_k}\}$ such that $\widehat{\mathcal{Q}}_{T_{j_k}}(\omega, x) \to \mathcal{Q}_0(x)$ for all $\omega \in \Omega_3$ and $x \in \mathcal{C}_0$ with $\mathbb{P}(\Omega_3) = 1$. For (c), since $\mathcal{Q}_0(x)$ is assumed continuous and for each fixed ω, $\widehat{\mathcal{Q}}_{T_{j_k}}(\omega, x)$ is a cdf, the last convergence can be made uniform by Polya's theorem, i.e., $\sup_{x \in \mathbb{R}^b} |\widehat{\mathcal{Q}}_{T_{j_k}}(\omega, x) - \mathcal{Q}_0(x)| \to 0$ for all $\omega \in \Omega_3$, where $\mathbb{P}(\Omega_3) = 1$. The result then follows by using the equivalence (in the other direction) of Theorem 6.3.1 of Resnick (1999).

For (d), we know that, for each $\omega \in \Omega_3$ and $x \in \mathcal{C}_0$, $\widehat{\mathcal{Q}}_{T_{j_k}}(\omega, x) \to \mathcal{Q}_0(x)$ is equivalent to $\widehat{\mathcal{Q}}^{-1}_{T_{j_k}}(\omega, x) \to \mathcal{Q}_0^{-1}(x)$. We refer to Lemma 21.2 of van der Vaart (2000), which implies once again by Theorem 6.3.1 of Resnick (1999) that $\widehat{\mathcal{Q}}_T^{-1}(x) \xrightarrow{p} \mathcal{Q}_0^{-1}(x)$. For the same reasoning $Q_T^{-1}(x) \to Q_0^{-1}(x)$ is equivalent to $Q_T(x) \to Q_0(x)$ for all $x \in \mathcal{C}_0$. By the triangle inequality, we have $\widehat{\mathcal{Q}}_T^{-1}(x) - \mathcal{Q}_T^{-1}(x) = o_P(1)$ for $x \in \mathcal{C}_0$, then we write

$$\mathcal{Q}_T\left[\widehat{\mathcal{Q}}_T^{-1}(\tau)\right] = \mathcal{Q}_T\left[\mathcal{Q}_0^{-1}(\tau) + \widehat{\mathcal{Q}}_T^{-1}(\tau) - \mathcal{Q}_0^{-1}(\tau)\right].$$

Then, conditional on the event $\mathscr{D}(\epsilon) := \left\{\left|\widehat{\mathcal{Q}}_T^{-1}(x) - \mathcal{Q}_0^{-1}(x)\right| \leq \epsilon\right\}$, defined for an arbitrary $\epsilon > 0$, and by the monotonicity of $\mathcal{Q}_T(\cdot)$, we have

$$\mathcal{Q}_T\left[\mathcal{Q}_0^{-1}(\tau) - \epsilon\right] \leq \mathcal{Q}_T\left[\widehat{\mathcal{Q}}_T(\tau)\right] \leq \mathcal{Q}_T\left[\mathcal{Q}_0^{-1}(\tau) + \epsilon\right].$$

Additionally, consider the event

$$\mathscr{E}(\epsilon) := \left\{\left|\mathcal{Q}_T[\mathcal{Q}_0^{-1}(\tau) - \epsilon] - \mathcal{Q}_0[\mathcal{Q}_0^{-1}(\tau) - \epsilon]\right| \vee \left|\mathcal{Q}_T[\mathcal{Q}_0^{-1}(\tau) + \epsilon] - \mathcal{Q}_0[\mathcal{Q}_0^{-1}(\tau) + \epsilon]\right| \leq \epsilon\right\}$$

to write that, conditioned on $\mathscr{D}(\epsilon) \cap \mathscr{E}(\epsilon)$, we have

$$\mathcal{Q}_0\left[\mathcal{Q}_0^{-1}(\tau) - \epsilon\right] - \epsilon \leq \mathcal{Q}_T\left[\widehat{\mathcal{Q}}_T(\tau)\right] \leq \mathcal{Q}_T\left[\mathcal{Q}_0^{-1}(\tau) + \epsilon\right] + \epsilon.$$

Take the limit as $\epsilon \to 0$ to conclude that, for fixed $\tau \in (0, 1)$, if $Q_0^{-1}(\tau) \in \mathcal{C}_0$ and on $\mathscr{D}(\epsilon) \cap \mathscr{E}(\epsilon)$, we have that $\left|\mathcal{Q}_T\left[\widehat{\mathcal{Q}}_T(\tau)\right]-\tau\right|\leq\epsilon$, as $\mathcal{Q}_0\left[\mathcal{Q}_0^{-1}(\tau)\right]=\tau$ for $x\in\mathcal{C}_0$. Finally, the conditioning event happens with probability approaching 1.

B Auxiliary Lemmas

Due to the lack of different characters, the variable denominations in this appendix are not necessarily consistent with the remainder of the article.

Lemma 1. *Let $\{X_t, t\in\mathbb{N}\}$ be a real-valued zero mean strong mixing process with mixing coefficient given by $\alpha(m)=\mathsf{exp}(-2cm)$ for some $c>0$, such that for some $q>2$, $\mathsf{sup}_{t\in\mathbb{N}}\mathbb{E}|X_t|^{q+\varepsilon}<C_q<\infty$ for some $\varepsilon>0$. Also define the partial sum $S_t:=\sum_{s=1}^{t}X_t$, then*

(a) $\|S_T\|_{\mathcal{L}^q}=O(\sqrt{T})$

(b) $\|\sum_{t=1}^{T}S_t\|_{\mathcal{L}^q}=O(T^{3/2})$

(c) $\|\sum_{t=1}^{T}S_tX_t\|_{\mathcal{L}^{q/2}}=O(T)$ *if* $q>4$

(d) $\|\sum_{t=1}^{T}S_t^2\|_{\mathcal{L}^q}=O(T^2)$

Proof. Result (a) can be found in Rio (1994); (b) follows from (a) and the triangle inequality since

$$\left\|\sum_{t=1}^{T}S_t\right\|_{\mathcal{L}^q}\leq\sum_{t=1}^{T}\|S_t\|_{\mathcal{L}^q}=\sum_{t=1}^{T}(O(\sqrt{t})=O(T^{3/2}).$$

For (c), we have that $S_t^2=(S_{t-1}+X_t)^2=S_{t-1}^2+2S_{t-1}X_t+X_t^2$. After taking summations across t and rearranging we are left with

$$\sum_{t=1}^{T}S_{t-1}X_t=\frac{1}{2}\left(S_T^2-\sum_{t=1}^{T}X_t^2\right).$$

Then, by the triangle inequality we have for $q>4$:

$$\begin{aligned}2\left\|\sum_{t=1}^{T}S_{t-1}X_t\right\|_{\mathcal{L}^{q/2}}&=\left\|S_T^2-\sum_{t=1}^{T}X_t^2\right\|_{\mathcal{L}^{q/2}}\\&=\left\|S_T^2-\sum_{t=1}^{T}(X_t^2-\mathbb{E}X_t^2)-\sum_{t=1}^{T}\mathbb{E}X_t^2\right\|_{\mathcal{L}^{q/2}}\\&\leq\|S_T^2\|_{\mathcal{L}^{q/2}}+\left\|\sum_{t=1}^{T}(X_t^2-\mathbb{E}X_t^2)\right\|_{\mathcal{L}^{q/2}}+\sum_{t=1}^{T}\mathbb{E}X_t^2.\end{aligned}$$

Since the $\mathcal{L}^q$ norm is sub-multiplicative, the first term is upper bounded by $\|S_T\|_{\mathcal{L}_{q/2}}^2$, which is $O(T)$ by (a). The second term is also $O(T)$ by (a) since $X_t^2-\mathbb{E}X_t^2$ is a zero mean strong mixing process with finite moments of order $q/2+\delta/2$. Finally the last is $O(T)$ and we conclude that $\| \sum_{t=1}^{T} S_{t-1}X_t\|_{\mathcal{L}^{q/2}} = O(T)$. The result (c) then follows from the triangle inequality because

$$\left\| \sum_{t=1}^{T} S_t X_t \right\|_{\mathcal{L}^{q/2}} \leq \left\| \sum_{t=1}^{T} S_{t-1} X_t \right\|_{\mathcal{L}^{q/2}} + \left\| \sum_{t=1}^{T} X_t^2 \right\|_{\mathcal{L}^{q/2}} = O(T).$$

Finally for (d) we have by the triangle inequality followed by (a):

$$\left\| \sum_{t=1}^{T} S_t^2 \right\|_{\mathcal{L}^q} \leq \sum_{t=1}^{T} \|S_t^2\|_{\mathcal{L}^q} = \sum_{t=1}^{T} O\left(t\right) = O(T^2).$$

□

Lemma 2. *Let $\{\boldsymbol{X}_t := (X_{1t} \dots X_{pt})', t \in \mathbb{N}\}$ be a $\mathbb{R}^p$-valued zero mean strong mixing random vector process with mixing coefficient given by $\alpha(m) = \mathsf{exp}(-2cm)$ for some $c > 0$. Also consider that following class of function*

$$\Psi := \{\psi : \mathbb{R} \to \mathbb{R} : \psi(x) = |x|^q, \psi(x) = \mathsf{exp}\, x^r, q > 2, r > 0\}.$$

Suppose that:

(i) There exists $q > 2$ such that $\mathsf{sup}_t \, \mathsf{sup}_{i \leq p} \mathbb{E}|X_{it}|^{r+\delta} < C_q < \infty$ for some $\delta > 0$ and

(ii) there exist positive constants a_1, a_2 and a_3, such that $\mathsf{sup}_t \, \mathsf{sup}_{i \leq p} \mathbb{P}(|X_{it}| > u) \leq a_1 \mathsf{exp}(-a_2 x^{a_3})$ for all $x > 0$.

Then, for every $x > 0$, we have

(a) $\mathbb{P}\left(\|\boldsymbol{X}_t\|_\infty \geq x\right) \leq C_1 p/\psi(x)$.

(b) $\mathbb{P}\left(\frac{1}{\sqrt{T}} \left\|\sum_{t=1}^{T} \boldsymbol{X}_t\right\|_\infty \geq x\right) \leq C_2 p/x^q$

(c) $\mathbb{P}\left(\frac{1}{\sqrt{T}} \left\|\sum_{t=1}^{T} \boldsymbol{X}_t\right\|_\infty \geq x\right) \leq R_{1,T}$.

(d) $\mathbb{P}\left[\frac{1}{\sqrt{T}} \left\|\sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t' - \mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t')\right\|_\infty \geq x\right] \leq C_3 p^2/x^q$

(e) $\mathbb{P}\left[\frac{1}{\sqrt{T}} \left\|\sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t' - \mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t')\right\|_\infty \geq x\right] \leq R_{2,T}$

where $C_j, j = 1, 2, 3$ are constants depending on q and c. Also,

$$\begin{aligned} R_{1,T} &= p\, \mathsf{exp}\left\{2c_2 \left[\sigma + \frac{1}{4c_1^2 (\mathsf{log}\, T)^4}\right] - \frac{x}{2}\right\} \\ &\quad + \sqrt{T} p \left\{\mathbb{1}\left[\frac{x}{2} \leq \mu_1\left(\frac{M}{2}\right)\right] + \mathbb{1}\left[\frac{x}{2} > \mu_1\left(\frac{M}{2}\right)\right] a_1 \,\mathsf{exp}\left[-a_2 (M/2)^{a_3}\right]\right\} \\ R_{2,T} &= p^2\, \mathsf{exp}\left\{2c_2 \left[\kappa + \frac{1}{4c_1^2 (\mathsf{log}\, T)^4}\right] - \frac{x}{2}\right\} \\ &\quad + \sqrt{T} p^2 \left\{\mathbb{1}\left[\frac{x}{2} \leq \omega\sqrt{\mu_2\left(\sqrt{\frac{M}{2}}\right)}\right] + \mathbb{1}\left[\frac{x}{2} > \omega\sqrt{\mu_2\left(\sqrt{\frac{M}{2}}\right)}\right] 2a_1 \,\mathsf{exp}\left[-a_2 \left(\frac{M}{2}\right)^{a_3/2}\right]\right\}, \end{aligned}$$

where $M := \frac{\sqrt{T}}{2c_1(\log T)^2}$ *and, for* $k > 0$,

$$\mu_k(x) := |\mathbb{E}X_{it}^k \mathbb{1}(|X_{it}| > x)| \leq 2\frac{a_1}{a_2^{k/a_3}}\boldsymbol{\gamma}\left(\frac{k}{a_3} + 1, a_2 x^{a_3}\right), \tag{B.1}$$

where $\boldsymbol{\gamma}(s, a) := \int_a^\infty x^{s-1}\exp(-x)\mathrm{d}x$ *is the incomplete upper Gamma function. For instance, when* $k = a_3 = 1$, *(B.1) turns out to be* $2\frac{a_1}{a_2^2}(1 + a_2 x)\exp(-a_2 x)$.

If we further impose that $\log p = o(M^{a_3/2})$, *then, as* $T \to \infty$,

$$R_{1,T} \to p\exp\left(2c_2\sigma - \frac{x}{2}\right)$$
$$R_{2,T} \to p^2\exp\left(2c_2\kappa - \frac{x}{2}\right).$$

Proof. First, for any $(p_1 \times p_2)$ real-valued random matrix $\boldsymbol{Y}$ and $\psi \in \Psi$, we have by Markov's inequality that, for any $x > 0$,

$$\mathbb{P}(\|\boldsymbol{Y}\|_\infty \geq x) \leq \frac{\mathbb{E}[\psi(\|\boldsymbol{Y}\|_\infty)]}{\psi(x)} \leq \frac{p_1 p_2 \sup_{i\leq p_1; j\leq p_2}\mathbb{E}[\psi(|Y_{i,j}|)]}{\psi(x)}. \tag{B.2}$$

Part (a) then follows by setting $\boldsymbol{Y} = \boldsymbol{X}_t$ in (B.2) and apply the definition C_ψ. In case $\psi(x) = |x|^q$, for part (b) set $\boldsymbol{Y} = \frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{X}_t$ or for the part (d) set $\boldsymbol{Y} = \frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{X}_t\boldsymbol{X}_t - \mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t')$ in (B.2), and we have Lemma 6 of Carvalho, Masini, and Medeiros (2018).

For part (c), if $\psi(x) = \exp(x)$, we use a truncation argument. For now, fix $M > 0$ and let $X_{it}^{\leq} := X_{it}\mathbb{1}(|X_{it}| \leq M/2) - \mathbb{E}[X_{it}\mathbb{1}(|X_{it}| \leq M/2)]$ and $X_{it}^{>} := X_{it}\mathbb{1}(|X_{it}| > M/2) - \mathbb{E}[X_{it}\mathbb{1}(|X_{it}| > M/2)]$ for $1 \leq i \leq p$ and $t \geq 1$. Since $\boldsymbol{X}_t$ is zero mean by assumption, we have that $X_{it} = X_{it}^{\leq} + X_{it}^{>}$. Furthermore, by construction, $X_{it}^{\leq}$ is a bounded (by M) zero-mean random variable. Therefore, from Theorem 2 in Merlevède, Peligrad, and Rio (2009), there exist positive constants c_1 and c_2, depending only on c, such that for all $T \geq 2$ and $0 < q < \frac{1}{c_1 M(\log T)^2}$, the following inequality holds:

$$\log\mathbb{E}\left[\exp\left(q\sum_{t=1}^T X_{i,t}^{\leq}\right)\right] \leq \frac{c_2 q^2(T\sigma_i^2 + M^2)}{1 - c_1 M q(\log T)^2}, \quad i = 1, \ldots, p,$$

where $\sigma_i^2 := \sup_t \sum_{k\in\mathbb{Z}} |\mathbb{E}\left(X_{it}^{\leq}X_{it+k}^{\leq}|\right) < \infty$. If we set $q = \frac{1}{\sqrt{T}}$, take $M = \frac{\sqrt{T}}{2c_1(\log T)^2}$ and $\sigma^2 := \sup_{i\leq p}\sigma_i^2$, we have

$$\log\mathbb{E}\left[\exp\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T X_{i,t}^{\leq}\right)\right] \leq 2c_2\left[\sigma^2 + \frac{1}{4c_1^2(\log T)^4}\right].$$

Let $\boldsymbol{X}_t^{\leq} := (X_{1t}^{\leq}, \ldots, X_{pt}^{\leq})'$. Then, applying (B.2) with $\boldsymbol{Y} = \frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{X}_t^{\leq}$ and $\psi(x) = \exp(x)$, we have

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{X}_t^{\leq}\right\|_\infty \geq x\right) \leq p\exp\left[2c_2\left(\sigma + \frac{1}{4c_1^2(\log T)^4}\right) - x\right].$$

We now bound $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{X}_t^{>}$, where $\boldsymbol{X}_t^{>} := (X_{1t}^{>},\ldots,X_{pt}^{>})'$. First, notice that

$$\mathbb{P}\left[|X_{it}\mathbb{1}(|X_{it}|>M/2)|\geq x\right]\leq \mathbb{P}(|X_{it}|>M/2)\leq a_1\exp(-a_2(M/2)^{a_3}).$$

Also,

$$|\mathbb{E}[X_{it}\mathbb{1}(|X_{it}|>M/2)]|\leq\int_{\mathcal{X}_i}|x|\mathbb{1}(|x|>M/2)\mathrm{d}F_{it}(x)\leq 2\int_{M/2}^{\infty}xf(x)\mathrm{d}x,$$

where $F_{it}(x):=\mathbb{P}(X_{it}\leq x)$ and $f(x)=a_1a_2a_3x^{a_3-1}\exp(-a_2x^{a_3})$, i.e., $f:=\frac{\mathrm{d}F}{\mathrm{d}x}$ with $F(x):=1-a_1\exp(-a_2x^{a_3})$. The last integral cannot be solved analytically when a_3 is not a positive integer. Apart from a change in variable, it is related to the incomplete upper gamma function as defined above.

Then, by the triangle inequality, we have

$$\begin{aligned}\mathbb{P}(|X_{it}^{>}|\geq x)&=\mathbb{P}\left\{|X_{it}\mathbb{1}(|X_{it}|>M/2)-\mathbb{E}[X_{it}\mathbb{1}(|X_{it}|>M/2)]|\geq x\right\}\\&\leq\mathbb{P}\left[|X_{it}\mathbb{1}(|X_{it}|>M/2)|\geq x-\mu_1\left(\frac{M}{2}\right)\right]\\&\leq\mathbb{1}\left[x\leq\mu_1\left(\frac{M}{2}\right)\right]+\mathbb{1}\left[x>\mu_1\left(\frac{M}{2}\right)\right]\mathbb{P}(|X_{it}|>M/2)\\&\leq\mathbb{1}\left[x\leq\mu_1\left(\frac{M}{2}\right)\right]+\mathbb{1}\left[x>\mu_1(\frac{M}{2})\right]a_1\exp\left[-a_2(M/2)^{a_3}\right].\end{aligned}$$

Apply the union bound to conclude that

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{X}_t^{>}\right\|_\infty\geq x\right)&\leq\sqrt{T}p\sup_t\sup_{i\leq p}\mathbb{P}(|X_{it}^{>}|\geq x)\\&\leq\sqrt{T}p\left\{\mathbb{1}\left[x\leq\mu_1\left(\frac{M}{2}\right)\right]+\mathbb{1}\left[x>\mu_1\left(\frac{M}{2}\right)\right]a_1\exp\left[-a_2(M/2)^{a_3}\right]\right\}.\end{aligned}$$

Combining both bounds using the fact that $\{|A+B|\geq x\}\subseteq\{|A|\geq x/2\}\cup\{|B|\geq x/2\}$, we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{X}_t\right\|_\infty\geq x\right)&\leq p\exp\left\{2c_2\left[\sigma+\frac{1}{4c_1^2(\log T)^4}\right]-\frac{x}{2}\right\}\\&+\sqrt{T}p\left\{\mathbb{1}\left[\frac{x}{2}\leq\mu_1\left(\frac{M}{2}\right)\right]+\mathbb{1}\left[\frac{x}{2}>\mu_1\left(\frac{M}{2}\right)\right]a_1\exp(-a_2(M/2)^{a_3})\right\}.\end{aligned}$$

For (e) set $\psi(x)=\exp(x)$ and $\boldsymbol{Y}=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{W}_t$ where $\boldsymbol{W}_t:=\boldsymbol{X}_t\boldsymbol{X}_t'-\mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t')$ in (B.2) to obtain

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{W}_t\right\|_\infty\geq x\right)\leq\frac{p^2\sup_{1\leq i,j\leq p}\mathbb{E}\left[\exp\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}W_{i,j,t}\right)\right]}{\exp(x)}.$$

We can conduct a similar truncation argument to the proof of part (c). Let $W_{i,j,t}=W_{i,j,t}^{\leq}+W_{i,j,t}^{>}$ where $W_{i,j,t}^{\leq}:=X_{it}X_{jt}\mathbb{1}\left[(|X_{it}|\vee|X_{jt}|)\leq\sqrt{M/2}\right]-\mathbb{E}\left\{X_{it}X_{jt}\mathbb{1}\left[(|X_{it}|\vee|X_{jt}|)\leq\sqrt{M/2}\right]\right\}$ and

$W^{>}_{i,j,t} = X_{it}X_{jt}\mathbb{1}\left[(|X_{it}| \vee |X_{jt}|) > \sqrt{M/2}\right] - \mathbb{E}\left\{X_{it}X_{jt}\mathbb{1}\left[(|X_{it}| \vee |X_{jt}|) > \sqrt{M/2}\right]\right\}$; then by construction, for each $1 \leq i,j \leq p$, we have that $\{W^{\leq}_{i,j,t}\}_{t\geq 1}$ is a zero mean, bounded by M, a strong mixing sequence with the same exponential decay of $\{\boldsymbol{X}_t\}_{t\geq 1}$. For that reason,

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{W}^{\leq}_{t}\right\|_{\infty} \geq x\right) \leq p^2 \exp\left\{2c_2\left[\kappa + \frac{1}{4c_1^2(\log T)^4}\right] - x\right\},$$

where $\kappa^2 := \sup_{1\leq i,j\leq p}\sup_t \sum_{k\in\mathbb{Z}}|\mathbb{E}\left(W_{i,j,t}W_{i,j,t+k}\right)| < \infty$. For the second term, we have, by Hölder's inequality,

$$\begin{aligned}\left|\mathbb{E}\left(X_{it}X_{jt}\right)\mathbb{1}\left(|X_{it}| \vee |X_{jt}| > \sqrt{M/2}\right)\right| &\leq \mathbb{E}\left[|X_{it}X_{jt}|\mathbb{1}\left(|X_{it}| \vee |X_{jt}| > \sqrt{M/2}\right)\right] \\ &\leq \left\{\mathbb{E}\left(X_{it}^2\right)\mathbb{E}\left[X_{jt}^2\mathbb{1}\left(|X_{it}| \vee |X_{jt}| > \sqrt{M/2}\right)\right]\right\}^{1/2} \\ &\leq \left\{\mathbb{E}X_{it}^2\mathbb{E}\left[X_{jt}^2\mathbb{1}\left(|X_{jt}| > \sqrt{M/2}\right)\right]\right\}^{1/2} \\ &\leq \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2},\end{aligned}$$

where $\sup_t \sup_i \mathbb{E}\left(X_{it}^2\right) \leq \omega^2 < \infty$ and $\mu_2(\cdot)$ is defined in (B.1).

Then, by the triangle inequality,

$$\begin{aligned}\mathbb{P}(|W^{>}_{i,j,t}| \geq x) &= \mathbb{P}\left\{\left|X_{it}X_{jt}\mathbb{1}\left(|X_{it}| \vee |X_{jt}| > \sqrt{M/2}\right) - \mathbb{E}\left[X_{it}X_{jt}\mathbb{1}\left(|X_{it}| \vee |X_{jt}| > \sqrt{M/2}\right)\right]\right| \geq x\right\} \\ &\leq \mathbb{P}\left\{\left|X_{it}X_{jt}\mathbb{1}\left(|X_{it}| \vee |X_{jt}| > \sqrt{M/2}\right)\right| \geq x - \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\right\} \\ &\leq \mathbb{1}\left\{x \leq \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\right\} \\ &\quad + \mathbb{1}\left\{x > \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\right\}\mathbb{P}\left(|X_{it}| \vee |X_{jt}| > \sqrt{M/2}\right) \\ &\leq \mathbb{1}\{x \leq \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\} \\ &\quad + \mathbb{1}\left\{x > \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\right\}2a_1\exp\left[-a_2(M/2)^{a_3/2}\right].\end{aligned}$$

Once again, apply the union bound to conclude

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{W}^{>}_{t}\right\|_{\infty} \geq x\right) \leq \sqrt{T}p^2\sup_{t\leq T}\sup_{1\leq i,j\leq p}\mathbb{P}(|W^{>}_{i,j,t}| \geq x).$$

Combining both bounds using the fact that $\{|A+B| \geq x\} \subseteq \{|A| \geq x/2\} \cup \{|B| \geq x/2\}$, we have

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{W}_t\right\|_\infty \geq x\right) \leq p^2 \operatorname{exp}\left\{2c_2\left[\kappa + \frac{1}{4c_1^2(\operatorname{log} T)^4}\right] - \frac{x}{2}\right\}$$
$$+ \sqrt{T}p^2\mathbb{1}\left\{\frac{x}{2} \leq \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\right\}$$
$$+ p^2\mathbb{1}\left\{\frac{x}{2} > \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\right\} 2a_1 \operatorname{exp}\left[-a_2(M/2)^{a_3/2}\right].$$

For the second part of the Lemma, we use the upper bound for the incomplete upper gamma function given by Natalini and Palumbo (2000), which states that for $s > 1$, $b > 1$ and $a > \frac{b}{b-1}(s-1)$, we have $\boldsymbol{\gamma}(s,a) < ba^{s-1}\operatorname{exp}(-a)$. Applying this bound in (B.1) with $b = 2$, we have that for all $k > 0$ and $y > 2k/a_3$:

$$\mu_k(y) := 2\frac{a_1}{a_2^{k/a_3}}\boldsymbol{\gamma}(k/a_3 + 1, a_2 y^{a_3}) < 4a_1 y^k \operatorname{exp}(-a_2 y^{a_3}),$$

from which we conclude that $\mu_k(y) \to 0$ as $y \to \infty$.

Since $M \to \infty$ is $T \to \infty$, we have for each $x > 0$, there is a $T_x \in \mathbb{N}$ such that $x > 2\left\{\mu_1(M/2) \vee \omega\left[\mu_2\left(\sqrt{M/2}\right)\right]^{1/2}\right\}$, whenever $T > T_x$. Thus, for $T > T_x$, we have

$$R_{1,T} = p\operatorname{exp}\left\{2c_2\left[\sigma + \frac{1}{4c_1^2(\operatorname{log} T)^4}\right] - \frac{x}{2}\right\} + \sqrt{T}pa_1\operatorname{exp}\left[-a_2(M/2)^{a_3}\right]$$
$$R_{2,T} = p^2\operatorname{exp}\left\{2c_2\left[\kappa + \frac{1}{4c_1^2(\operatorname{log} T)^4}\right] - \frac{x}{2}\right\} + \sqrt{T}p^2 2a_1\operatorname{exp}\left[-a_2\left(\frac{M}{2}\right)^{a_3/2}\right].$$

Hence, as long as $\operatorname{log} p = o(M^{a_3/2})$, we have the second result of the Lemma. □

C List of Symbols

C.1 The Romans

$c, c_1, c_2, \ldots$	Generic positive constants
d	Generic Deterministic Trend
e	Exponential
f	Deterministic Trends
g	
h	Cardinality of set $\mathcal{H}$
i, j, k, t, s	Units/regressors, time index
ℓ, L	Scaling matrix and its entries
m	Lag of alpha mixing
q	Number of moments
r	Number of I(0) relations
s, s_0	Cardinality of index set
n, p	Number of units and regressors
w	Individual weights of the LASSO
A	Random element of proof of Theorem 3
B	Standard Brownian motion
F	Factor of the common factor model
G	Generic random vector of Assumption 3 and Definition 1
H	Transformed objective function
M	Generic matrix used in GIF
$I(\cdot)$	Integrated process
J	Linear combination of I(0) processes
O, o, O_P, o_P	Landou notation
Q	LASSO Objective function
R	Remainder of Lemma 2
T, T_0, T_1, T_2	Sample size and Treatment, Pre and Post
U, U^Z, U^F	Innovation
V	Regression error
$X, Y, W, Z^{(0)}, Z^{(1)}$	Units and its transformation

C.2 The Greeks

α	Mixing coefficient
$\boldsymbol{\beta}, \boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}$	Parameter, True and Estimated
$\boldsymbol{\gamma}, \boldsymbol{\gamma}_0, \widehat{\boldsymbol{\gamma}}$	Transformed parameter, True and Estimated
$\delta, \widehat{\delta}, \Delta, \widehat{\Delta}$	Treatment effect, ATE and Estimates
ϵ	Arbitrary small positive constant
ζ	Linear Projecion in the Factor Model
η	The stochastic component of the DGP
θ, Θ	Parameters of the generic model
ι	Vector of 1's
κ	Auxiliry Lemma 1 Appendix
λ, λ_0	Penalty parameter
μ	Constant of the deterministic trend
ν	Combined weight trend
ξ	Cone constant
$\boldsymbol{\pi}$	Projection of I(0) process
ρ	Simulation autocorrelation coefficient
σ	Variance of the innovation
τ	Quantiles
υ	Variance of the defining I(0) process
$\phi, \widehat{\phi}, \phi_j$	The Inference function
χ	GIF Constant
ψ, Ψ	Deterministic Trends
$\Omega, \Omega_0, \Omega_1, \ldots, \omega$	Sample space, events
$\boldsymbol{\gamma}, \widetilde{\boldsymbol{\gamma}}$	Cointegration matrix
$\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0$	Covariance matrix of WW'

C.3 Miscellaneous

$\mathbb{N}, \mathbb{Z}, \mathbb{R}$	Naturals, integers and real
$\mathscr{C}$	Cone
$\mathscr{H}$	Test hypothesis
$\mathscr{F}$	Sigma algebra
$\mathbb{P}, \mathbb{E}$	Probability and expectation operator
$\mathcal{D}$	Intervention indicator
U	Innovation
$\mathcal{M}$	Generic model
$\mathcal{G}$	Process to define I(0)
$\mathcal{H}$	Set index of growth condition
$\mathcal{S}, \mathcal{S}_0$	Set index
R	index set in the proof of Proposition 3

Table 1: **Rejection Rates under the Null (empirical size): Deterministic Trends**

Baseline DGP: (2.4) and (2.6) with $T = 100$, independent and identically normally distributed innovations, $n = 200$, $s_0 = 5$, $T_2 = 3$ 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 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 Bayesian Information Criterion (BIC). We set the maximum penalty level to be $\|\frac{1}{T_0}\sum_{t=1}^{T_0} Y_t \boldsymbol{X}_t\|_\infty$ with an exponential path down to $\lambda_{\mathsf{min}} = 0.001$ 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 β_0 was used.

	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 (non-zero) 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
					Determinist 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
					Post Intervention Periods				
$T_2 = 1$	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

Table 2: **Rejection Rates under the Null (empirical size): Stochastic Trends**

Baseline DGP: (2.4) and (2.5) with $T = 100$, independent and identically normally distributed innovations, $n = 200$, $s_0 = 5$, $T_2 = 3$ 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 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^F_{1t} and the first unit idiosyncratic innovation U^Z_{1t}. The penalization parameter λ is chosen via Bayesian Information Criterion (BIC). We set the maximum penalty level to be $\|\frac{1}{T_0}\sum_{t=1}^{T_0} Y_t \boldsymbol{X}_t\|_\infty$ with an exponential path down to $\lambda_{\mathsf{min}} = 0.001$ 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 β_0 was used.

	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 (non-zero) 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
				Post Intervention Periods					
$T_2 = 1$	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

Table 3: **Rejection Rates under the alternative (empirical power).**

Baseline DGP: (2.8) and (2.7) with $T = 100$, iid normally distributed innovations, $n = 200$ units, $s_0 = 5$, $T_2 = 3$ 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 Bayesian Information Criteria (BIC). We set the maximum penalty level to be $\|\frac{1}{T_0}\sum_{t=1}^{T_0} Y_t X_t\|_\infty$ with an expoential path down to $\lambda_{\mathsf{min}} = 0.001$ along 100 equally spaced intervals in the `glmnet` package. σ^2 is the variance of unit 1 at $t = T_0$.

Deterministic Trends

	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 $\delta_t = c\sigma Z 1\{t > T_0\}$ 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

Stochastic Trends

	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 $\delta_t = c\sigma Z 1\{t > T_0\}$									
$c = 0.1$	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

Table 4: **Monte Carlo Results: Estimation**

The table reports several statistics averaged over 10,000 replications for each one of four data generating processes. More specifically, mean ℓ_1-norm is the average $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$, mean bias is the average bias $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ over the simulations, mean MSE is the average mean squared error, and mean Δ is the average intervention effect over the 10 out-of-sample periods. Note that the true value of Δ is zero. MSE Δ is the average squared error over the simulation and, finally, median Δ is the median of the estimates of Δ over the simulations. Each column in the table represents a variation of the baseline scenario, in which we set $T = 100$,$s_0 = 5$, $n = 100$ and $\rho = 0$. Model (1) is given by equations (2.4) and (2.5) where $f_t^F = 0$. Model (2) is given by equations (2.4) and (2.5) where $f_t^F = 1$. Model (3) is given by equations (2.4) and (2.6) where $f_t^F = t$. Model (4) is given by equations (2.4) and (2.6) where $f_t^F = t^2$.

Model	**Statistic**	**Baseline**	**Sample Size**		**Sparsity**		**Regressors**		**Autocorrelation**	
			$T = 500$	$T = 1000$	$s_0 = 1$	$s_0 = 10$	$n = 50$	$n = 200$	$\rho = 0.2$	$\rho = 0.5$
(1)	mean ℓ_1-norm	1.36	0.26	0.13	0.19	3.04	0.99	1.72	1.46	1.87
	mean bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mean MSE	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	mean Δ	-0.03	-0.03	0.02	0.01	-0.04	0.01	0.01	0.03	-0.19
	MSE Δ	1.57	0.25	0.17	0.33	3.48	1.00	2.27	2.13	4.99
	median Δ	-0.03	-0.03	0.02	0.01	-0.04	0.01	0.01	0.03	-0.19
(2)	mean ℓ_1-norm	2.46	0.34	0.15	0.63	4.38	1.52	3.55	2.91	3.83
	mean bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mean MSE	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	mean Δ	0.10	-0.02	-0.01	-0.28	-0.08	-0.17	-0.30	0.08	-0.17
	MSE Δ	3.20	0.29	0.15	0.93	6.24	1.56	5.72	4.53	13.21
	median Δ	0.10	-0.02	-0.01	-0.28	-0.08	-0.17	-0.30	0.08	-0.17
(3)	mean ℓ_1-norm	3.45	0.66	0.32	1.02	5.82	1.96	4.61	3.68	3.95
	mean bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mean MSE	0.01	0.00	0.00	0.00	0.02	0.01	0.01	0.01	0.01
	mean Δ	0.01	-0.02	0.00	-0.08	0.00	0.13	0.00	-0.11	-0.08
	MSE Δ	4.81	0.39	0.23	1.73	7.41	2.25	7.74	5.87	15.51
	median Δ	0.01	-0.02	0.00	-0.08	0.00	0.13	0.00	-0.11	-0.08
(4)	mean ℓ_1-norm	1.46	0.64	0.58	0.33	2.93	1.24	1.66	1.52	1.93
	mean bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mean MSE	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
	mean Δ	-0.06	0.01	-0.01	-0.29	-0.03	-0.06	-0.07	-0.06	-0.08
	MSE Δ	0.22	0.12	0.12	0.25	0.30	0.18	0.26	0.32	0.73
	median Δ	-0.06	0.01	-0.01	-0.29	-0.03	-0.06	-0.07	-0.06	-0.08

Table 5: **Aggregated Data – Descriptive Statistics.**

The table reports the estimated coefficients of a linear trend model for the total sold quantities in each group as well as the coefficients of the linear trend, when dummies to control for the days-of-the-week effect are included in the model. The numbers between parentheses are heteroskedastic-autocorrelation robust (HAC) standard errors. The table also presents the results of the augmented Dickey-Fuller (ADF) test for the null of unit roots against the alternative of a trend-stationary model. The table also reports the p-values of Johansen's cointegration tests.

Panel (a): ADF and cointegration tests

	All	Treatment Group	Control Group
ADF (p-value)	0.06	0.07	0.00
PP (p-value)	0.00	0.00	0.00

	No coint.	At most 1
Johansen Trace (p-value)	0.00	0.00
Johansen Rank (p-value)	0.00	0.00

Panel (b): Linear trend parameters

	All	Treatment Group	Control Group	All	Treatment Group	Control Group
Intercept	$13,458.57$	$6,157.138$	$7,301.43$	$8,359.75$	$3,958.18$	$4,401.58$
	(530.68)	(238.33)	(301.35)	(614.94)	(257.12)	(371.18)
Slope	26.08	11.93	14.16	26.55	12.09	14.46
	(10.14)	(4.49)	(5.73)	(10.62)	(4.57)	(6.30)
Days-of-the Week Dummies	No	No	No	Yes	Yes	Yes

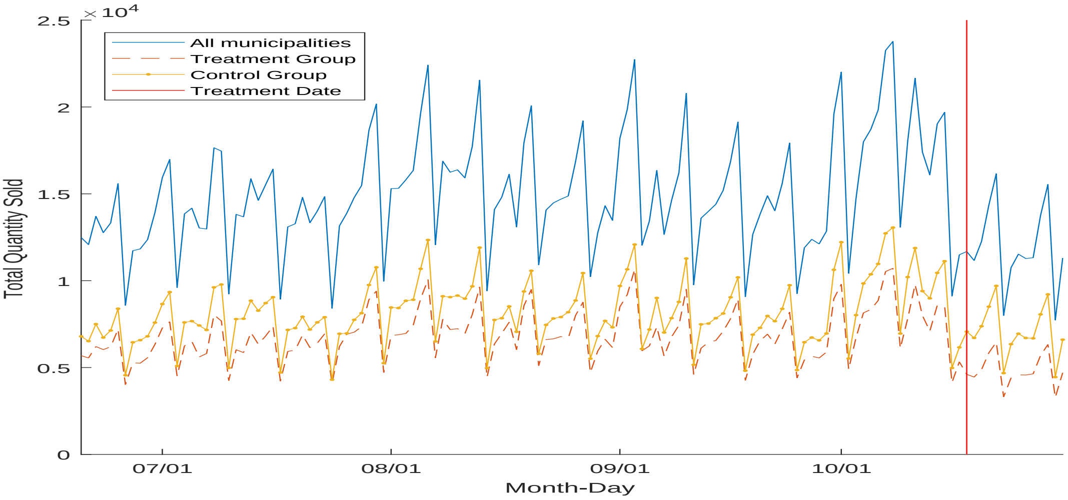
Table 6: **Results.**

The table reports estimation results. Panel (a) show 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 $\phi(x) = \frac{1}{T_2}\sum_{j=1}^{T_2} x_j^2$ and $\phi(x) = \frac{1}{T_2}\sum_{j=1}^{T_2} |x_j|$, respectively.

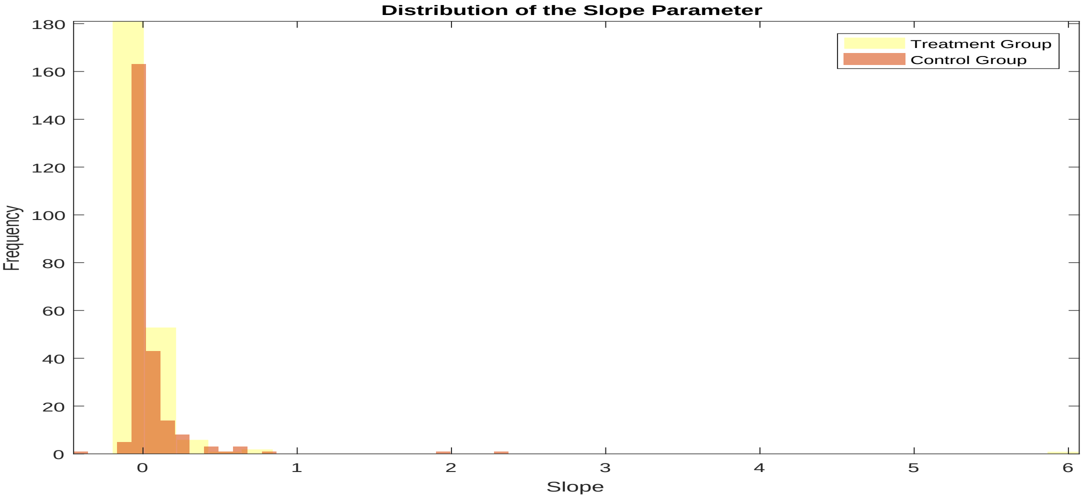
	Panel (a): **Aggregated**	Panel (b) **Disaggregated**			
		Mean	Std. Dev.	Max.	Min.
Δ	-1,147	-12.90	52.08	5.52	-526.70
Δ/#stores	-4.33	-4.21	4.42	5.52	-23.27
p-value (square)	0	0.41	0.29	1	0
p-value (absolute)	0	0.36	0.31	1	0
Proportion (%) of rejection of the null (square)	NA	19	NA	NA	NA
Proportion (%) of rejection of the null (absolute)	NA	31	NA	NA	NA
R-squared	0.96	0.44	0.25	0.95	0
Number of regressors	133	133	NA	NA	NA
Number of relevant regressors	26	9.46	8.06	72	0
Number of pre-treatment observations	120	120	NA	NA	NA
Number of observations during treatment	14	14	NA	NA	NA

Figure 1: **Quantities sold.**

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 pre-treatment sample.



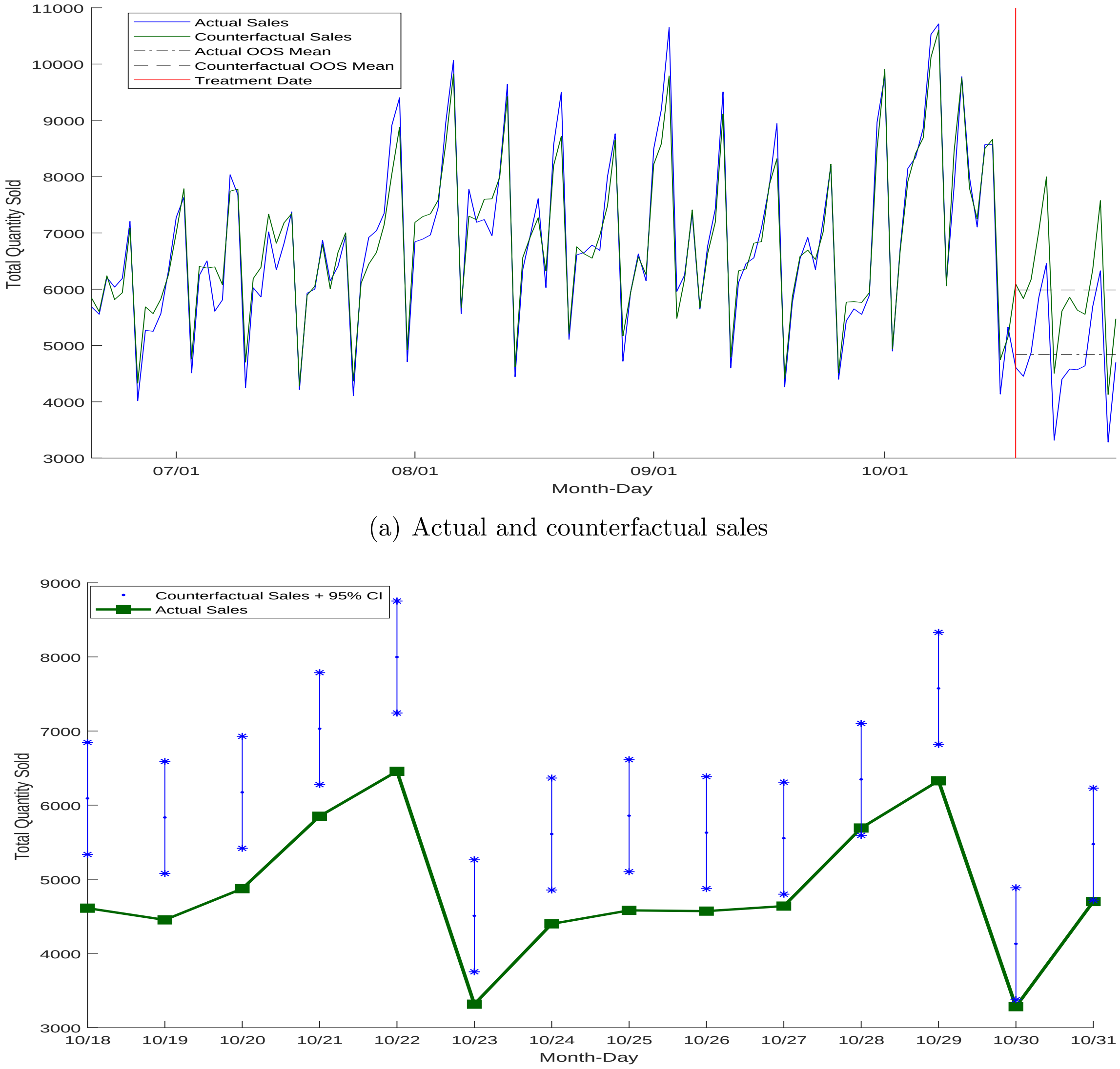
(a) Time-series dynamics of the sold quantities



(b) Histograms of the slope parameter

Figure 2: **Actual and counterfactual sales.**

Panel (a) shows the aggregated actual and counterfactual sales over the pre-treatment and post-treatment 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 post-treatment period. 95% confidence intervals for the counterfactual path is also displayed.



(a) Actual and counterfactual sales

(b) Actual and counterfactual sales during treatment period

References

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

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

Abadie, A., and J. L'Hour (2019): "A Penalized Synthetic Control Estimator for Disaggregated Data," Discussion paper, CREST.

Andrews, D. (2003): "End-of-sample instability tests," *Econometrica*, 71, 1661–1694.

Athey, S., and G. Imbens (2017): "The State of Applied Econometrics - Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32.

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

Brodersen, K., F. Galluser, J. Koehler, N. Remy, and S. Scott (2015): "Inferring Causal Impact using Bayesian Structural Time-Series Models," *Annals of Applied Statistics*, 9, 247–274.

Carvalho, C., R. Masini, and M. Medeiros (2018): "ArCo: An Artificial Counterfactual Approach for High-Dimensional Panel Time-Series Data," *Journal of Econometrics*, 207, 352–380.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2018a): "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," arxiv:1712.09089, arXiv.

——— (2018b): "Inference on average treatment effects in aggregate panel data settings," arxiv:1812.10820, arXiv.

Davidson, J. (2009): *When is a Time Series I(0)?*pp. 322–342. Oxford University Press.

Doudchenko, N., and G. Imbens (2016): "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis," 22791, NBER, available at arXiv:1610.07748.

Engle, R., and C. Granger (1987): "Co-integration and error correction: Representation, estimation, and testing," *Econometrica*, 55, 251–276.

Ferman, B., and C. Pinto (2016): "Synthetic Controls with Imperfect Pre-Treatment Fit," Working paper, São Paulo School of Economics - FGV.

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

Huang, J., and C.-H. Zhang (2012): “Estimation and Selection via Absolute Penalized Convex Minimization And Its Multistage Adaptive Applications,” *Journal of Machine Learning Research*, 13, 1839–1864.

Ibragimov, A. (1962): “Some Limit Theorems for Stationary Processes,” *Theory of Probability and its Applications*, 7, 349–382.

Kock, A. (2016): “Consistent and conservative model selection with the adaptive LASSO in stationary and nonstationary autoregressions,” *Econometric Theory*, 32, 243–259.

Lee, J., Z. Shi, and Z. Gao (2018): “On Lasso for predictive regressions,” arxiv:1810.03140, arXiv.

Li, K. (2017): “Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods,” Working paper, The Wharton School, the University of Pennsylvania.

Li, K., and D. Bell (2017): “Estimation of average treatment effects with panel data: Asymptotic theory and implementation,” *Journal of Econometrics*, 197, 65–75.

Liang, C., and M. Schienle (2019): “Determination of vector error correction models in high dimensions,” *Journal of Econometrics*, 208, 418–441.

Liao, Z., and P. Phillips (2015): “Automated estimation of vector correction models,” *Econometric Theory*, 31, 581–646.

Masini, R., and M. Medeiros (2019): “Counterfactual Analysis and Inference with Non-Stationary Data,” Discussion Paper 2894065, SSRN.

Merlevède, F., M. Peligrad, and E. Rio (2009): “Bernstein inequality and moderate deviations under strong mixing conditions,” in *High Dimensional Probability V: The Luminy Volume*, ed. by C. Houdré, V. Koltchinskii, D. Mason, and M. Peligrad, vol. Volume 5, pp. 273–292. Institute of Mathematical Statistics.

Natalini, P., and B. Palumbo (2000): “Inequalities for the incomplete Gamma function,” *Mathematical Inequalities and Applications*, 3, 69–77.

Onatski, A., and C. Wang (2018): “Alternative asymptotics for cointegration tests in large VARs,” *Econometrica*, 86, 1465–1478.

Resnick, S. (1999): *A Probability Path*. Birkhäuser Boston.

Rio, E. (1994): “Inégalités de moments pour les suites stationnaires et fortement mélangeantes.,” *Comptes rendus Acad. Sci. Paris, Série I*, 318, 355–360.

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

van de Geer, S. A., and P. Bühlmann (2009): “On the conditions used to prove oracle results for the Lasso,” *Electron. J. Statist.*, 3, 1360–1392.

van der Vaart, A. W. (2000): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.