Generalized Synthetic Control for TestOps at ABI: Models, Algorithms, and Infrastructure

Luis Costa[1], Vivek F. Farias[1], Patricio Foncea[1], Jingyuan (Donna) Gan[1], Ayush Garg[2], Ivo Rosa Montenegro[2], Kumarjit Pathak[2], Tianyi Peng[1], and Dusan Popovic[2]

[1]Massachusetts Institute of Technology
[2]Anheuser-Busch InBev

Abstract

We describe a novel optimization-based approach – Generalized Synthetic Control (GSC) – to learning from experiments conducted in the world of physical retail. GSC solves a long-standing problem of learning from physical retail experiments when treatment effects are small, the environment is highly noisy and non-stationary, and interference and adherence problems are commonplace. The use of GSC has been shown to yield an approximately 100x increase in power relative to typical inferential methods and forms the basis of a new large-scale testing platform: 'TestOps'. TestOps was developed and has been broadly implemented as part of a collaboration between Anheuser Busch Inbev (ABI) and an MIT team of operations researchers and data engineers. TestOps currently runs **physical experiments** impacting approximately **135M USD** in revenue **every month** and routinely identifies innovations that result in a **1-2%** increase in sales volume. The vast majority of these innovations would have remained unidentified absent our novel approach to inference: prior to our implementation, statistically significant conclusions could be drawn on only $\sim 6\%$ of all experiments; a fraction that has now increased by over an order of magnitude.

1 Introduction

Consider a retailer seeking to decide whether or not to implement a new innovative idea. Such innovations (or 'interventions' in the parlance of experiment design) can range from new promotions, to new assortment strategies, to new algorithms for logistics and supply chain problems.

In theory, the impact of innovative interventions can be learned via *randomized experiments*. For example, to learn the incremental effect of a new type of promotion on sales volumes, a retailer might apply this new promotion to a random subset of 'units' – individual stores for a brick-and-mortar retailer – over a fixed period of time. Comparing the growth in sales volumes in this 'test' group from the period prior to the intervention, with the growth in sales in a separate randomly selected 'control' group of stores, would then yield an estimate of the incremental impact of the new promotion.

In practice, the standard experiment alluded to above (a randomized A/B test with Differences-in-Differences estimation) is difficult to conduct in physical retail. The reasons include:

1. First, **scale and cost**. Unlike an online A/B test where one might simply funnel some fraction of traffic to a test arm, these tests typically involve physical stores. It is not unusual for a test to impact thousands of stores and the associated implementation requires significant cost and coordination.
2. Second, there usually are issues with **interference, adherence, and biased selection**: control stores are accidentally treated, test stores do not receive an intervention over the appropriate period, or else, multiple interventions are inadvertently implemented at the same store. Further due to issues of convenience, incentives, or otherwise, the selection of test and control stores is often non-random (and thus potentially, systematically biased).
3. Finally, the observed data is **noisy and non-stationary**, and the treatment effects are typically small resulting in low SNR. Even tests with *thousands* of stores that have been carefully matched have **very low power**.

Taken together, we see that **running such experiments is expensive but there is a very high chance that any given experiment is inconclusive**. In fact, this is exactly the dilemma faced by Anheuser-Busch InBev (ABI), the world largest beer producer[1]. The collaboration between MIT and ABI we report on here is an attempt to solve this issue: given the large investment ABI has made in running physical retail experiments, how do we improve our ability to draw statistically significant conclusions from their outcomes?

Our Work. In this work, we developed a new, rigorous, optimization-based approach to inference in the context of experiments conducted in physical retail. An equivalent way of casting the experimental challenges we have alluded to above is this: *the control units in a given experiment do not in fact form a valid control group.* As such, the crux of the inference challenge may be viewed as follows: *we must find a* ***synthetic*** *control, which can be broadly thought of as a certain optimal convex combination of the pre-assigned controls that serves as a more appropriate control to the test arm.* By doing so, the variance of estimating treatment effects due to the underlying problems of non-stationarity and non-random selection can be largely eliminated so that statistically significant conclusions can be drawn on a much larger fraction of experiments.

The analytics challenge then lies in understanding how to compute such a synthetic control in a manner that (a) is robust to arbitrary treatment assignment and noise (indeed, the treatment assignment can depend on the historical noise and the noise in observations is highly correlated both temporally and cross-sectionally), (b) robust to corruption so that it can accommodate having to ignore the corruption of some of the control data and (c) statistically optimal so that power is maximized. The typical approaches of computing synthetic controls (e.g., constrained linear regression) fail in the presence of these challenges. In recent work, the MIT team showed how this problem can be viewed as one of learning in panels with arbitrary intervention patterns and noise

[1] ABI is the world's largest beer producer. Their diverse portfolio of well over 500 beer brands includes global brands Budweiser, Corona, and Stella Artois; multi-country brands Beck's, Hoegaarden, Leffe, and Michelob ULTRA; and local champions such as Aguila, Antarctica, Bud Light, Brahma, Cass, Castle Lite, Cristal, Harbin, Modelo Especial, Quilmes, Victoria, Sedrin, and Skol. ABI leverages the collective strengths of approximately 169,000 employees based in nearly 50 countries worldwide. For 2021, ABI reported revenue was 54.3 billion USD.

[Farias et al., 2021]. That theoretical work showed that the problem above can be viewed as a certain large-scale regularized *non-convex* optimization problem but that despite being non-convex an *efficient* algorithm could be constructed to solve the problem. The algorithm itself relies on an alternating minimization approach that seeks to learn a factor model that describes the observed data across testing units (stores) and time. This is a breakthrough that was presented as one of 55 oral presentations at the NeurIPS 2021 conference out of over 10,000 papers. Our work is the first large-scale implementation that generalizes this breakthrough and renders it practical.

Using the state-of-the-art algorithmic framework described, we co-developed the Global Test and Learn Platform (also called *TestOps*) at ABI, a digital product that aims to provide measurement, tracking and inference in physical retail experiments. This platform allows users to configure experiments, simulate scenarios, measure experiment outcomes, and track value generation across multiple initiatives (i.e. interventions) through valid inference. Through a standardized test intake process, the platform allows users to create tickets for tracking experiments to be deployed on the front lines. TestOps is set up to be a self-service tool. The platform logs all the relevant information pertaining to a given experiment and allows for automated scheduling of test scenarios. Figure 1 presents a screenshot of the software tool.

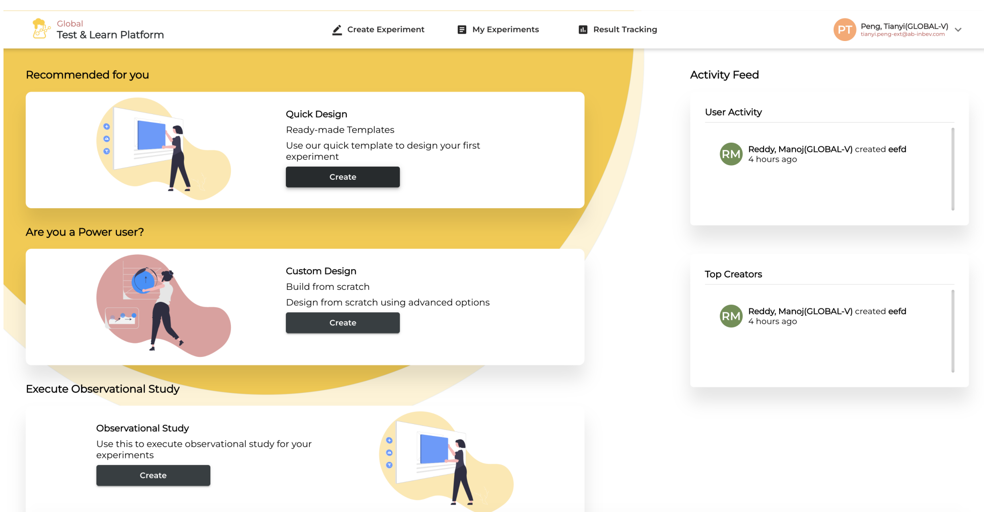


Figure 1: A screenshot of the TestOps software tool.

Our contribution, compared to the state-of-the-art experimental platforms available commercially today (such as the **MasterCard APT, Optimizely or Adobe** platforms), is thus three-fold:

1. *(Much) Higher Power*: Our efficient non-convex panel data solver yields an effectively 100x increase in power relative to the standard DID approach inherent in commercial experimentation platforms. Indeed, because of the fact that the control data and portions of the tests data are routinely corrupted, there are no obvious inference methods outside of DID available on a commercial platform that would apply prior to this implementation.
2. *Far Fewer Assumptions*: Relative to existing techniques such as DID, our approach dispenses with the need of the so-called parallel trends assumption which is very difficult to justify in the

physical retail environment. Our approach also dispenses with the need to make assumptions on independence in observed noise and can flexibly accommodate complex temporal and cross-sectional correlations as well as endogenous treatment assignments (existing sophisticated alternatives to DID do not allow for this). Both these features give a great deal of confidence in inferential results that would otherwise be questioned given the high-stakes nature of a rollout.

3. *Robust Optimization*: Our differentiation is facilitated by viewing the inherent estimation task through the lens of robust optimization. This is a first on a commercially deployed platform, and we believe this is a lens that is particularly useful and interpretable in business environments such as ABI.

Implementation and Impact. The TestOps platform has been live in Mexico for almost a year. In recent months, the platform has run on average **∼50** large scale experiments **a month**[2] in Mexico (the region it has been rolled out initially). A typical experiment impacts approximately **tens of thousands** of test stores with interventions that range from store-specific assortment recommendations to personalized promotions. Typically, the size of the control group in these experiments is about thousands of stores. Standard DID inference would not be sufficiently powerful in this setting; **over 90% of experiments would not yield significant results at the 5% level**. The primary reason for this is (a) the treatment effects themselves are in the low single digits and (b) due to issues of interference and adherence, and the nature of the available data, observed sales are incredibly volatile. The platform equipped with our new inferential techniques increases power dramatically, resulting in a roughly **10x increase in the number of experiments that yield significant results** despite all of the aforementioned challenges: as one documented example, in October and November last year, out of 99 experiments run, the platform was able to draw statistically significant conclusions on 65. Had we instead used the legacy DID approach all but 6 experiments would have been inconclusive. It is not unusual using TestOps to identify an intervention with a treatment effect that increases volume by **1-2%**. The **average monthly volume impacted by these experiments is 1,400 kHL associated with average monthly revenues of $135M**, so that these treatment effects are very meaningful. In light of this success, the platform is currently being rolled out across ABIs 'Middle America' zone (all of Central America, from Mexico to Colombia) and subsequently, will be rolled out *globally* in 2023.

The change brought by the new inference method can be best summarized by a quote from Felipe Aragao, the Global Vice President of ABI who is leading the overall development of data-driven analytics at ABI:

> *It is hard to overstate the impact of this new inference algorithm; absent the algorithm, we would simply be unable to conduct meaningful inference, and as a consequence our ability to learn from tests on the platform would be severely constrained. Instead, with this core innovation, we are learning high value interventions in a setting where experimental learning is extraordinarily challenging.*

[2] All experiments are implemented at the physical-store level.

1.1 Related Literature

Our work lies in the broad field of causal inference in quasi-experimental design. Differences-in-Differences has been used as early as the 1850's by John Snow in the context of medical experimentation and epidemiology. Today it is the workhorse in experimentation to evaluate the effects of marketing interventions, public policy, drug experimentation, and education interventions (Bertrand et al. 2004, Angrist and Pischke 2009, Lechner et al. 2011, Li et al. 2012). In spite of this, DID can perform quite poorly in practice because the violation of so-called 'parallel-trend' assumption (e.g. we will see this in the implementation at ABI momentarily): an over-simplified assumption for the counterfactuals that require treated and control groups follow the same trend across time if no intervention is occurred (Dimick and Ryan 2014, Wing et al. 2018, Kahn-Lang and Lang 2020).

In recent years, the approaches that allow a more sophisticated counterfactual model (e.g., a low-rank model that captures the non-stationarity) start to attract significant attentions. Among them the 'synthetic control' approach is the most prominent one, which can be broadly thought of as a certain optimal convex combination of the pre-assigned controls that serves as a better control to the test arm. The synthetic control literature pioneered by Abadie and Gardeazabal (2003), Abadie et al. (2010) has grown to encompass sophisticated learning and inferential methods; see Abadie (2019) for a review. Doudchenko and Imbens (2016), Li and Bell (2017), Ben-Michael et al. (2021) consider a variety of regularized regression techniques to learn the linear combination of untreated units that yields a synthetic control. Amjad et al. (2018, 2019), Agarwal et al. (2021) consider instead the use of principal component regression techniques. Arkhangelsky et al. (2019) proposes alternative approaches to imputing counterfactuals by averaging across both untreated units (rows) and time (untreated columns). Li (2020), Chernozhukov et al. (2021) address inferential questions that arise in synthetic control, with the latter providing a permutation test that is generally applicable. Further generalizations that utilize the low-rank factor model to handle non-block general treatment patterns include the panel regression method (Bai 2009, Moon and Weidner 2017, 2018) as well as the matrix completion method (Athey et al. 2021, Xiong and Pelger 2019, Xu 2017, Bai and Ng 2019).

All of the sophisticated alternatives to DID alluded to in the previous paragraph **require that the assignment of a treatment to a given unit be exogenous**. *This completely rules out the ABI setting, where as we have discussed, the selection of test and control stores is non-random and systematically biased for a number of reasons. In addition to this issue, the treatment 'patterns' one observes do not fall neatly into the 'block' patterns required by the approaches described above.* A recent breakthrough by Farias et al. (2021) proposes an estimator that can be shown to be mini-max optimal for general endogenous treatment patterns and noises, thus generalizing existing synthetic control broadly and overcoming the above ABI-centric challenges. Our estimator, in effect, provides a surprising, novel and *efficient* solution to a hard non-convex optimization problem, by developing new ideas relevant to the theory of matrix completion.

2 Model and Problem

The problem at hand can be formulated as the following. Suppose there are N stores, where each store is indexed by i. Suppose the total period of time that is of interest can be divided into T time epochs, indexed by t (each epoch corresponds to a day or week in the ABI implementation). Let $M^* \in \mathbb{R}^{n\times T}$ be a fixed, unknown matrix where, for the sake of concreteness, say M^*_{it} represents the expected sales of the store i during the time epoch t. Let $E \in \mathbb{R}^{n\times T}$ be a zero-mean random matrix, where E_{it} is the noise in observed sales for store i at time t. Taken together, we refer to $M^*_{it} + E_{it}$ as the 'counterfactual' sales which encode the hypothetical sales if no intervention were applied.

Now consider a single intervention whose impact on sales we wish to measure; our generalized model will allow multiple simultaneous interventions. In the ABI context, such an intervention might correspond to, say, an assortment optimization innovation, a promotion strategy, or even a new approach to logistics and supply chain tasks. Let a known 'intervention pattern' matrix $Z \in \{0,1\}^{n\times T}$ encode when the intervention is applied, i.e., $Z_{it} = 1$ indicates that the intervention is used at the store i at time t; otherwise $Z_{it} = 0$.

We observe the sales matrix $O \in \mathbb{R}^{n\times T}$ where O_{it} corresponds to actually observed sales at store i during period t.[3] When there is no intervention ($Z_{it} = 0$), we have $O_{it} = M^*_{it} + E_{it}$; when the intervention is applied ($Z_{it} = 1$), we will write $O_{it} = M^*_{it} + E_{it} + \mathcal{T}_{it}$ where $\mathcal{T}_{it}$ encodes the impact of the intervention on sales, i.e., $\mathcal{T}_{it}$ are unknown, heterogeneous treatment effects. In summary, the sales O_{it} can be written as

$$O_{it} = M^*_{it} + E_{it} + \mathcal{T}_{it} Z_{it}.$$

Goal: We seek to estimate the impact of our intervention (i.e. promotion, assortment change, etc.). Formally, given O and Z, the precise quantity we aim to measure, is the *average treatment effect on the treated* entries (ATT):

$$\tau^* := \sum_{ij} \mathcal{T}_{ij} Z_{ij} / \sum_{ij} Z_{ij}.$$

When the intervention is a promotion, this τ^* will correspond to the average incremental (or decremental) impact on sales caused by the promotion on the treated entries.[4] This quantity is of central interest and can be used to filter promising new business strategies. Furthermore, it serves as a basis for other more fine-grained treatment effects (e.g., sub-group treatment effects can be obtained using the same model by restricting the estimation on a subset of stores).

[3] Observable covariates on each unit may also be available; we suppress this aspect here for clarity.

[4] One may also consider the average treatment effects of all entries, $\sum_{ij} \mathcal{T}_{ij}/nT$, which obviously requires further assumptions to make the estimation of treatment effects for untreated entries possible. We leave this problem for future exploration.

2.1 Fundamental Challenges in Estimating ATT

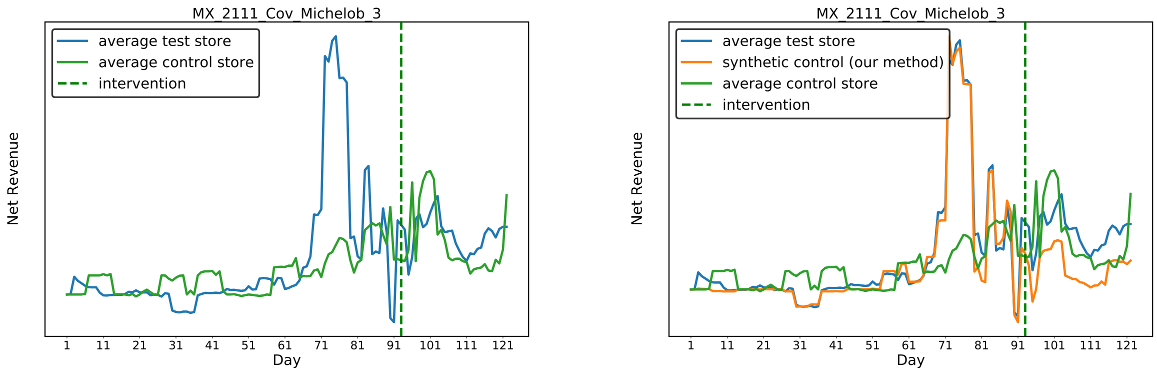
A crucial challenge in reliably estimating the ATT, is that the **pattern of interventions, Z, is typically endogenous**. This is especially the case in non-company owned brick-and-mortar retail environments, where the implementation of an intervention cannot be centrally governed. This is the case with ABI, where the network of stores are not owned by the company so that the implementation of a given intervention cannot be determined unilaterally, and is instead a function of the extent to which a sales associate is willing to execute on the experiment design, and the store-owners openness to participation. As a concrete example which we will expand on later, a typical intervention for ABI constitutes a sizable alteration to the pricing, assortment, or even physical layout of a store. This requires proper execution by at least four agents (the store managers, ABI sales representatives, ABI business teams, and ABI data analytics teams) each with their own unique incentives, so that it is no surprise that 'randomized' designs are difficult to implement. Further challenges can arise: stores may be accidentally treated, test stores inadvertently may not receive a treatment over the appropriate period, or worse, various forms of *purposeful* non-compliance might occur, e.g. a manager assigned to implement an experimental promotion might stop the promotion if they perceive sales slowing down. In summary, any reasonable estimation procedure **must allow for Z to be endogenous**: i.e. Z is fundamentally a function of M^* and E, as opposed to independent of those quantities. Parenthetically, we also must note that the noise E exhibits **complex correlations** across stores and over time and that **variance is amplified** due to the fact that the sales data available is effectively aggregated at coarse time-granularities.

2.1.1 DID: A Legacy Solution and Its Limitations

To address the challenges alluded above, one potential solution is to use the workhorse approach of Difference-in-Difference (DID), which is indeed the legacy solution used at ABI prior to this project. DID is the traditional workhorse in experimentation due to its simplicity and it ability to handle the challenges alluded to above vis-a-vis the endogeneity of Z (Ishimaru 2021, Kropko and Kubinec 2018, Imai and Kim 2021). Essentially DID makes the following structural assumption for the expected sales matrix M^*:

$$M^*_{it} = a_i + b_t \tag{1}$$

where a_i are store-fixed effects and b_t are time-fixed effects. Its worth pausing here to understand the incredibly strong assumption implicit in the above model: DID assumes that any systematic variation over time is in fact *common* across stores (the so-called 'parallel-trends' assumption), and further that outside of this common variation, any remaining variation in store sales is zero-mean and idiosyncratic. It should be clear that these are tough assumptions to make in a practical brick-and-mortar retail context. Given the simplistic counterfactual model above, DID proceeds to



(a) (b)

Figure 2: Example experiment. The plots display beer sales pre and post intervention.

estimate the ATT τ^* by fitting the observed sales O in a least-square fashion[5]:

$$(\hat{\tau}_{\mathrm{DID}}, \hat{M}) := \underset{\substack{\tau, M \\ M_{it}=a_i+b_t}}{\operatorname{argmin}} \sum_{it} (O_{it} - M_{it} - \tau Z_{it})^2 \tag{2}$$

DID effectively fails to work in the ABI setting. Ultimately, the overly simplistic counterfactual model implicit in DID, requires that we assume any variation in time that is not common across stores is idiosyncratic. This in turn results in large residuals, and consequently, wide confidence intervals. In a setting where the treatment effects we seek to measure are small, these wide confidence intervals do not permit effective inference. And this is while putting aside the even more serious issue that a failure of the underlying assumptions implies biased estimates. To see this concretely, consider Fig. 2a the experiment described by the figure is an ABI experiment run over approximately 10,000 test stores, and over 1,000 control stores. Despite the relatively large sample, the DID counterfactual model (a shift of the green line) is evidently too simplistic to capture counterfactual behavior resulting in large residuals: in particular, the standard deviation of the DID results is at least $1/\sqrt{1000}$ even in the ideal A/B testing scenario (typically much worse) but the treatment effect we aim to estimate is in fact in the order 1%, which is far smaller. As such, DID does not allow us to conclude anything from this very large scale experiment.[6]

As such, the core problem at ABI we seek to tackle is seeing a return on the investment of running non-trivial experiments where the effects one seeks to measure are small, and the environment highly non-stationary. If the experiments we ran in this environment were inconclusive, this return is non-existent. The use of DID would equate to over 90% of all experiments run being inconclusive. In the remaining sections we describe a new approach to estimation that overcomes the challenges above, and that despite the challenges of low SNR and small treatment effects, increased by 10x the

[5] This is in fact a linear regression and can be solved in an explicit form.

[6] On the other hand, a 'synthetic control' constructed by our method (can be viewed as a weighted combination of other controls) serves as a much better counterfactual estimation (the orange line in Fig. 2b) for test stores than DID, hence being able to provide more accurate estimation of treatment effects. See Section 4 for more details.

number of experiments for which we were able to draw statistically significant conclusions.

3 Algorithm: Generalized Synthetical Control

As seen above, the DID approach is limited by the simple structure it assumes for the mean counterfactual data, M^*, i.e., Eq. (1). Specifically, DID assumed that the mean sales pattern at any store was the sum of a store specific fixed effect and a temporal effect that was *common* to all stores. A substantially more general pattern is to allow for M^* to be described by a 'factor' model.

$$M^*_{it} = \sum_{k=1}^{r} u_{i,k} v_{t,k} \tag{3}$$

where r is the factor dimension, and $u_i \in \mathbb{R}^r, v_t \in \mathbb{R}^r$ are latent store-specific and time-specific factors respectively. Frequently referred to as interactive-fixed-effects models, models described by Eq. (3) can be made arbitrarily general through the choice of r and form the basis of modern time-series and panel data models. It is worth noting that the counterfactual model assumed by DID (i.e. $M^*_{it} = a_i + b_t$) is a special case of Eq. (3). As r increases, the 'low-rank' factor model of Eq. (3) can capture increasingly complex interactions between stores and times: e.g., capturing effects occurring at a specific type of store during specific periods of time. Such low-rank models form a core ingredient in modern machine learning with applications ranging from recommendation systems to Natural Language Processing.

Given the interactive fixed-effects model assumed above for M^*, one may naturally consider solving an optimization problem similar to the one for DID:

$$(\hat{\tau}, \hat{M}) = \underset{\tau, M, \text{rank}(M) \leq r}{\text{argmin}} \sum_{it} (O_{it} - M_{it} - \tau Z_{it})^2 \tag{4}$$

Unfortunately, this problem is non-convex and challenging to optimize and analyze. Instead, motivated by the work of a subset of the co-authors in Farias et al. (2021), we solve a certain convex surrogate for this optimization problem and apply a novel de-biasing procedure to the solution so-obtained. In Farias et al. (2021), we showed that such an estimator recovers the treatment effect *at a min-max optimal rate for a maximally large class of treatment patterns Z.* In the following subsections, we paraphrase our algorithms and results from Farias et al. (2021). We then present a sequence of experiments that serve to illustrate our approach dominates state-of-the-art alternatives.

3.1 Algorithm and Optimal Guarantees

Before describing our estimator formally, we introduce some notation: For a matrix $M \in \mathbb{R}^{n \times T}$, let $\|M\|_{\text{F}} = \sqrt{\sum_{ij} M_{ij}^2}$ be the Frobenius norm and $\|M\|_*$ be the nuclear norm of M. Further, we denote by $\langle A, B \rangle = \sum_{ij} A_{ij} B_{ij}$ be the inner product between two matrices A and B. Our estimator τ^d for the ATT τ^* is constructed in two steps:

$$(\hat{\tau}, \hat{M}) \in \underset{\tau, M}{\text{argmin}} \quad \|O - M - \tau Z\|_{\text{F}}^2 + \lambda \|M\|_*, \tag{5a}$$

$$\tau^d := \hat{\tau} - \frac{\lambda \langle Z, \hat{U}\hat{V} \rangle}{\left\| (I - \hat{U}\hat{U}^\top) Z (I - \hat{V}\hat{V}^\top) \right\|_{\mathrm{F}}^2}. \tag{5b}$$

where in the second step, (5b), we denote by $\hat{M} = \hat{U}\hat{\Sigma}\hat{V}^\top$, the SVD of $\hat{M}$. To understand this estimator, note that the first step, (5a), is a natural convex relaxation to the non-convex problem (4). The objective function's first term penalizes choices of M and τ which differ from the observed O, and the second term seeks to penalize the rank of M using the nuclear norm as a (convex) proxy. The quantity of λ encodes the relative weight of these two objectives. After the first step, having $(\hat{M}, \hat{\tau})$ as a minimizer of (5a), we could simply use $\hat{\tau}$ as our estimator for τ^*. However, we show in (Farias et al. 2021) that the second step, (5b), serves to eliminate first-order biased introduced by the regularization implicit in our convex relaxation. We see in our experiments that this de-biasing has a powerful practical impact.

Optimal guarantees: Our main results in (Farias et al. 2021), characterize the estimation error $\tau^d - \tau^*$. For simplicity, we paraphrase those results under the scenario where E consists of i.i.d. zero-mean Gaussian entries and the treatment effect is homogeneous ($\mathcal{T}_{ij} = \tau^*$). Nonetheless, Z is allowed to depend on the E:

Theorem 1. *Under regularity conditions, with high probability,*

$$|\tau^d - \tau^*| = \tilde{O}\left(\frac{\sigma}{\sqrt{\sum_{ij} Z_{ij}}} \right)$$

The $\tilde{O}$ here hides factors depending on r and other factors that depend logarithmically on the dimension (i.e. number of stores or time-periods), and polynomially on the condition number of M^*. It is worth noting that by the law of the iterated logarithm, the error rate achieved by our estimator is essentially the *optimal* rate one can achieve even when M^* is known and Z is fixed.

The unstated regularity conditions for Theorem 1 consist of two parts: The first part limits the 'collinearity' between M^* and Z by requiring the projection of Z onto the tangent space of M^* to be small. This condition is in fact essentially *necessary* for identifying τ^* using any estimator. The second part relates to to how Z is allowed to depend on the noise E: it turns out that any 'casual' dependence structure is admissible. We refer the interested reader to our paper (Farias et al. 2021)[7] for more details and more general results (e.g., allowing correlated noises).

Taken together, the regularity conditions are mild enough to allow for various endogenous treatment patterns and broadly expand on the set of patterns possible in the existing literature (Bai 2009, Abadie et al. 2010, Moon and Weidner 2015, Xu 2017, Moon and Weidner 2018, Arkhangelsky et al. 2019, Xiong and Pelger 2019, Bai and Ng 2019, Agarwal et al. 2020, Athey et al. 2021). In particular, the results here generalize the *'synthetic control'* paradigm, which addresses the special case of an exogenous Z with the support on a single row or block.

[7] See the journal version in `https://arxiv.org/abs/2106.02780`.

3.2 Experiments

Whereas the next section described the outcome of our implementation at ABI, we end this section with our experience on semi-synthetic datasets where the treatment is introduced artificially and thus ground-truth treatment-effect values are known. The results show that our estimator τ^d is more accurate than existing algorithms and its performance is robust to various treatment patterns, in particular for the treatment that is **adaptively assigned depending on the historical outcomes**.

The following four benchmarks were implemented: (i) Synthetic Difference-in-Difference (SDID) Arkhangelsky et al. (2019); (ii) Matrix-Completion with Nuclear Norm Minimization (MC-NNM) Athey et al. (2021) (iii) Robust Synthetic Control (RSC) Amjad et al. (2018) (iv) Ordinary Least Square (OLS): Selects $a, b \in \mathbb{R}^n, \tau \in \mathbb{R}$ to minimize $\|O - a1^T - 1b^T - \tau Z\|_{\mathrm{F}}^2$, where $1 \in \mathbb{R}^n$ is the vector of ones. As described, this corresponds to the canonical Difference-in-Difference (DID) method with two-way fixed effects. It is also worth noting that SDID and RSC only apply to traditional synthetic control patterns (*block* and *stagger* below).

Warm-Up (block and stagger patterns). The first dataset consists of the annual tobacco consumption per capita for 38 states during 1970-2001, collected from the prominent synthetic control study (Abadie et al. 2010) (the treated unit California is removed). Similar to Athey et al. (2021), we view the collected data as M^* and introduce artificial treatments. We considered two families of patterns that are common in the economics literature: *block* and *stagger* (Athey et al. 2021). Block patterns model simultaneous adoption of the treatment, while stagger patterns model adoption at different times. In both cases, treatment continues forever once adopted. Specifically, given the parameters (m_1, m_2), a set of m_1 rows of Z are selected uniformly at random. On these rows, $Z_{ij} = 1$ if and only if $j \geq t_i$, where for block patterns, $t_i = m_2$, and for stagger patterns, t_i is selected uniformly from values greater than m_2.

To model heterogenous treatment effects, let $\mathcal{T}_{ij} = \tau^* + \delta_i$ where δ_i is i.i.d and $\delta_i \sim \mathcal{N}(0, \sigma_\delta)$ characterizes the unit-specific effect. Then the observation is $O = M^* + \mathcal{T} \circ Z$. We fix $\tau^* = \sigma_\delta = \bar{M}^*/5$ through all experiments, where $\bar{M}^*$ is the mean value of M^*. The hyperparameters for all algorithms were tuned using rank $r \sim 5$ (estimated via the spectrum of M^*).

Next, we compare the performances of the various algorithms on an ensemble of 1,000 instances with $m_1 \sim \mathrm{Uni}[1, n_1), m_2 = \mathrm{Uni}[1, n_2)$ for stagger patterns and $m_1 \sim \mathrm{Uni}[1, 5), m_2 = 18$ for block patterns (matching the year 1988, where California passed its law for tobacco control). The results are reported in the first two rows of Table 1 in terms of the average normalized error $|\tau - \tau^*|/\tau^*$.

Note that the treatment patterns here are 'home court' for the SDID and RSC synthetic control methods but our approach nonetheless outperforms these benchmarks. One potential reason is that these methods do not leverage all of the available data for learning counterfactuals: MC-NNM and SDID ignore treated observations. RSC ignores even more: it in addition does not leverage some of the *untreated* observations in M^* on treated units (i.e. observations O_{ij} for $j < t_i$ on treated units).

Adaptive Treatment Pattern. The second dataset consists of weekly sales of 167 products over 147 weeks, collected from a Kaggle competition (PredictSales 2021). In this application,

Table 1: Comparison of our estimator to benchmarks on semi-synthetic datasets (Block and Stagger correspond to Tobacco dataset; Adaptive pattern corresponds to Sales dataset). Average normalized error $|\tau - \tau^*|/\tau^*$ is reported.

Pattern	Our estimator	SDID	MC-NNM	RSC	OLS
Block	**0.15** ($\pm$0.13)	0.23 ($\pm$0.19)	0.27 ($\pm$0.24)	0.30 ($\pm$0.26)	0.38 ($\pm$0.36)
Stagger	**0.10** ($\pm$0.20)	0.16 ($\pm$0.18)	0.15 ($\pm$0.16)	0.20 ($\pm$ 0.27)	0.18 ($\pm$ 0.19)
Adaptive	**0.02** ($\pm$0.02)	-	0.13 ($\pm$0.10)	-	0.20 ($\pm$0.18)

treatment corresponds to various 'promotions' of a product (e.g. price reductions, advertisements, etc.). We introduced an artificial promotion Z, used the collected data as M^* ($\bar{M}^* \approx 12170$), and the goal was to estimate the average treatment effect given $O = M^* + \mathcal{T} \circ Z$ and Z ($\mathcal{T}$ follows the same generation process as above with $\tau^* = \sigma = \bar{M}^*/5$).

Now the challenge in these settings is that these promotions are often decided based on previous sales. Put another way, the treatment matrix Z is constructed *adaptively.* We considered a simple model for generating adaptive patterns for Z: Fix parameters (a, b). If the sale of a product reaches its lowest point among the past a weeks, then we added promotions for the following b weeks (this models a common preference for promoting low-sale products). Across our instances, (a, b) was generated according to $a \in \text{Uni}[5, 25], b \in \text{Uni}[5, 25]$. This represents a treatment pattern where it is unclear how typical synthetic control approaches (SDID, RSC) might even be applied.

The rank of M^* is estimated via the spectrum with $r \sim 35$. See Table 1 for the results averaged over 1,000 instances. The average of $|\tau - \tau^*|/\tau^*$ is $\sim 2\%$ for our algorithm, versus 13% for MC-NNM, indicating a strong improvement. This demonstrates the advantage of our algorithm for **complex adaptive treatment patterns**, which widely exist in real applications. On the other hand, the performance of matrix-completion algorithms is limited for those structured and adaptive missing-ness patterns. We overcome this limitation by leveraging the treated data[8].

4 ABI implementation

In this section, we describe the impact our algorithmic development has had on experimentation at ABI. In particular, we discussing in detail how the TestOps platform has enabled learning from experiments that previously, due to the inherent volatility in the ABI environment and issues of biased test and control selection, would have been inconclusive. To establish this, we zoom into a set of 99 experiments run in Mexico shortly after the platforms launch, and show that:

1. Our approach (we refer to as GSC, generalized synthetic control) provides a significant increase in experimental power relative to the incumbent approach used for inference (DID). Specifically, we establish – in an entirely model agnostic fashion – that the power of DID is essentially zero for treatment effects smaller than 17% (i.e. DID is only able to confidently detect very large

[8]There is a natural trade-off here: if the heterogeneity in treatment effects were on the order of the M^* (so that $\|(\mathcal{T} - \tau^*) \circ Z\| \gg \sigma_r(M^*)$, the smallest singular value of M^*) then it is unclear that the treated data would help (and it might, in fact, hurt). But for most practical applications, the treatment effects we seek to estimate are typically small relative to the nominal observed values.

treatment effects). In contrast, our approach has power exceeding 80% for treatment effects as low as 3%. It is difficult to overstate how dramatic this impact is.

2. As a result of the lower power above, DID was able to draw significant inferences on treatment effects for just 6 of the 99 experiments run shortly after launch; i.e. DID was inconclusive on over 90% of the experiments run. This is, of course, not an acceptable outcome given how costly these experiments are. In contrast, because of its dramatically higher power, GSC was able to draw significant inferences in 65 experiments: a 10x increase over what was possible with DID.

In effect, we will demonstrate a platform that allows for learning from experimentation despite the volatility and non-stationary of the ABI environment and the inherent problem of biased selection of treatment units in a setting where such selections must be made by groups with differing incentives.

4.1 Experiments

We briefly describe the set of experiments that are the focus of our comparison of relative merits: we consider a set of 99 experiments run in the Mexico geography in October and November 2021. These experiments were all setup and monitored through the TestOps platform (described in Section 4.4). Each experiment was focused on measuring the efficacy of a particular type of intervention; broadly two types of interventions were experimented with: promotion strategies (i.e., a store-level promotion with some discounts) or assortment strategies (i.e., recommendation of a particular assortment to store owners via sales representatives or online messages). These experiments were all large-scale experiments with control groups typically consisting of thousands of stores, and test groups consisting of tens of thousands of stores, with the intervention implemented over a roughly one month time frame; Table 2 provides summary statistics describing experiment information such as size, length, and impacted revenues.

Table 2: Statistics of Experiments at ABI

# Experiments (per month)	~ 50
# Treated Stores (per experiment)	(1800, 6900, 11000)*
# Control Stores (per experiment)	(130, 1700, 4000)*
Duration of Experiments	~ 30 days
Impacted Revenue (per month)	~135M USD
Sparsity	$\sim 95\%$

*(Q_1, Q_2, Q_3) refers to the the first, second, third quantile respectively.

It is worth noting that the challenges described in Section 2 are germane to these experiments: the data collected, being 'sell-in' data in *very* noisy and sparse (with a sparsity of approximately 95%) environments. Compliance was a major concern: the experiments involved at least four agents (data analytics group, business group, sales representatives, and store managers) where each had unique incentives. Finally, most experiments required a physical change at the store. All of this led to test and control group selection driven by convenience and incentives, as opposed to random selection. Put a different way, the test and control groups are endogenous (as discussed in

Section 2.1), a challenge that GSC is able to overcome.

4.2 Power

We begin with examining the improvement in power provided by the use of GSC relative to DID: in a nutshell, this will help establish whether, and to what extent, GSC provides an increase in inferential power. We do so, by running what is colloquially referred to as an 'A/A' test. Specifically, we consider 'test' and 'control' data where we know that in fact that neither the test nor control units were treated over the data collection period in question. As such, the true treatment effect is zero, and one cares to understand the treatment effect recovered by the inference approach under study (in our case, DID, and GSC). We consider two sets of A/A tests: First, we construct such tests using data from the pre-treatment period in each of our experiments (where we know that neither test nor control until were treated). Second, we construct tests by drawing test and control units exclusively from stores that were known to be control units during our experiments.

4.2.1 A/A Tests with Pre-Treatment Data

One approach to constructing a valid A/A test given our experimental setup is as follows. For any given experiment, we consider the test and control units corresponding to that experiment, but restrict ourselves to observations in the *pre-treatment period* where we know that neither test nor control units could have been subject to the intervention. As input to the inference procedure, we supply a fictitious treatment start time (that, of course, precedes the actual start of the experiment). In summary, the observations that we feed to any inference procedure under study are as follows:

1. The set of test and control units.
2. All observations $O_{i,t}$ for units i in the test and control groups, and all epochs $t < T_{\text{real}}$ prior to the start time of the actual intervention, T_{real}.
3. A fictitious intervention start time $T_{\text{fict}} < T_{\text{real}}$, so that $Z_{i,t} = 1$ iff $t \geq T_{\text{fict}}$ and i is in the test group.

The treatment effect we learn should of course be zero, and as such the estimates produced by any given inference procedure yield the distribution of estimates produced by that procedure when the null (i.e. the treatment effect is zero) is in fact true. The left panel of Figure 3 shows the distribution of estimated treatment effects under the DID and GSC inference procedures when we know the null to be true. We clearly see that both approaches are effectively unbiased. However, the spread in the DID estimates is significantly larger than that under GSC; in other words DID is prone to producing non-trivial treatment effect estimates when the treatment effect is in fact zero. We can leverage these distributions to calculate the power (i.e. the likelihood that we correctly reject the null when the null is not true) of either procedure, as a function of the treatment effect size of an alternate hypothesis. We want the power to be as large as possible, even for small treatment effect sizes. Here, looking at the right panel in Figure 3 reveals that the power of GSC is greater than 80% even for small treatment effects (as low as 3%); **this is dramatically higher than DID**, which has essentially zero power until we get to effect sizes that are on the order of 17%. These results show that GSC has substantially higher power than DID over a large effect size range, and

Figure 3: **A/A Tests With Pre-Treatment Data**. Left: GSC correctly produces near-zero treatment effect estimates when this is the case. Right: The power of GSC is dramatically higher than that of DID

especially at smaller effect sizes where the value of experimentation is especially large.

4.2.2 A/A Tests with Control Data

A complementary approach to constructing a valid A/A test given our experimental setup is as follows. For any given experiment, we consider just the control units, and further split these units into a fictitious test and control group. We then assume that the treatment was applied to the units in our fictitious control group (whereas, of course, this is not the case in reality). As input to the inference procedure, we supply a partition of the control units prescribing effectively a fictitious test and control group. In summary, the observations that we feed to any inference procedure under study are as follows:

1. The set of control units in the original experiment.
2. All observations $O_{i,t}$ for control units i, and the start time of the intervention in the original intervention, T_{real}.
3. A partition of the units into two groups: a fictitious 'test' groups and a fictitious 'control' group. Thus, $Z_{i,t} = 1$ iff $t > T_{\text{real}}$ and i is in the fictions test partition.

The treatment effect we learn should again, of course be zero. The left panel of Figure 3 shows the distribution of estimated treatment effects under the DID and GSC inference procedures when we know the null to be true. Like before, the spread in the DID estimates is again significantly larger than that under GSC. As before, we can leverage these distributions to calculate the power of either procedure, as a function of the treatment effect size of an alternate hypothesis. Here, looking at the right panel in Figure 4 reveals that the power of GSC is greater than 80% even for small treatment effects; **again, dramatically higher than DID**, which has essentially zero power until we get to effect sizes that are on the order of 25%.

In summary these experiment show in an entirely data-driven fashion that is independent of our modeling assumptions that GSC has substantially higher power than DID, and as such can be expected to infer treatment effects in the sorts of experiments run on the TestOps platform, where

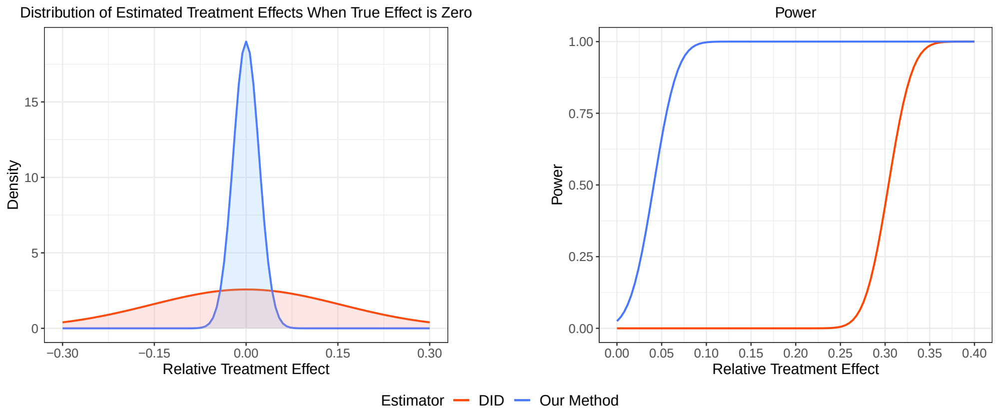


Figure 4: **A/A Tests With Control Data**. Left: GSC correctly produces near-zero treatment effect estimates when this is the case. Right: The power of GSC is dramatically higher than that of DID

treatment effects are expected to be small, and the environment is highly volatile. The next Section demonstrates that this is actually the case.

4.3 Platform Performance on Real Experiments

The previous section demonstrates that GSC has significantly higher power than DID vis-a-vis experiments run on the TestOps platform. This is likely to be of value, especially for smaller treatment effects. Here we demonstrate that this value bears out, and further illustrate the mechanism driving the improvement in inferential power.

Our main results are described in Table 3. This table focuses on 99 experiments run on the TestOps platform in October and November 2021. The average monthly volume impacted by these experiments together is 1,400 kHL corresponding to average monthly revenues of $135M, so that identifying interventions with positive treatment effects is very meaningful. As we see from the two left-most columns, the vast majority of these experiments have small treatment effects. The third column from left ('DID') shows that of these 99 experiments, DID is able to provide significant inference in just 6 experiments; in a sense this is a testament to the absence of parallel trends in the ABI environment, where the drivers of sales over time are likely to be geographically diverse. The fact that the treatment and control groups are not constructed randomly only exacerbates this issue. On the other hand we see from the right-most column that GSC is able to draw significant conclusions in 65 experiments – a remarkable 10X increase. In particular, focusing on those experiments where the treatment effect was in fact deemed to be positive (the last four grayed-out rows which together pertain to 53 experiments), GSC was able to conclude this to be the case in 36 of those experiments (relative to just 4 for DID). In very simple terms, this corresponds to 32 **innovations over two months that would have been disregarded despite having a positive impact on the business**. This is the crux of the platforms value: it makes it possible to learn from experiments where treatment effects are small and volatility in counterfactual outcomes is large; both of these are hallmarks of the physical environment ABI operates in.

Table 3: Experiments in October and November 2021

Treatment Effect (Our Method)	#Experiments	#Conclusive Experiments (DID)	#Conclusive Experiments **(Our Method)**
$< -3\%$	11	0	11
$[-3\%, -2\%)$	2	0	2
$[-2\%, -1\%)$	10	1	7
$[-1\%, 0)$	23	1	9
$[0, 1\%)$	23	3	11
$[1\%, 2\%)$	11	0	8
$[2\%, 3\%)$	7	0	7
$> 3\%$	12	1	10
Total	99	6	**65**

The conclusiveness is corresponding to 5% significance level. The last four grayed-out rows correspond to experiments with positive estimated treatment effects.

DID vs. GSC: A Visualization Fig. 5 correspond to a set of experiments for which GSC identified the treatment effect to be positive at a significance of 5%. Each chart in the figure shows three time series: the blue series corresponds to average revenue sold (the KPI of interest) in the test group, whereas the green series corresponds to average revenue sold in the control group. The control group is not much of a control as is evident from the wide disparity between the blue and green series prior to the start of the intervention. This is despite the fact that these are averages over large groups of stores, and illustrates two points:

1. The parallel trends assumption is a poor one.
2. The test and control groups are not randomly constructed.

The orange series on the other hand attempt to mitigate these issues by effectively constructing a *carefully weighted average* of the control units – the GSC method effectively enables the construction of this weighted average. We see that the resulting weighted average of control stores – the orange series – is a far better representative of test group performance as evidenced by the fact that the blue and orange series track closely prior to the start of the intervention. It is ultimately the fact that we have constructed this higher quality control that enables the effective inference we see.

4.4 TestOps Experimentation Platform

In this final section, we provide a virtual walk-through of how a user might leverage the TestOps platform. The platform allows users to (i) configure experiments; (ii) simulate scenarios; (iii) track experiment outcomes and ultimately, (iv) measure the value generated across multiple initiatives. Our users are of course not expected to be statisticians or experiment design experts, and as such, a key goal here is to democratize rigorous decision making without requiring a sophisticated technical background. At the same time, we strived to make the process and especially inferences drawn from data as transparent as possible, as we describe next. The procedure of conducting and analyzing an experiment using TestOps involves four steps:

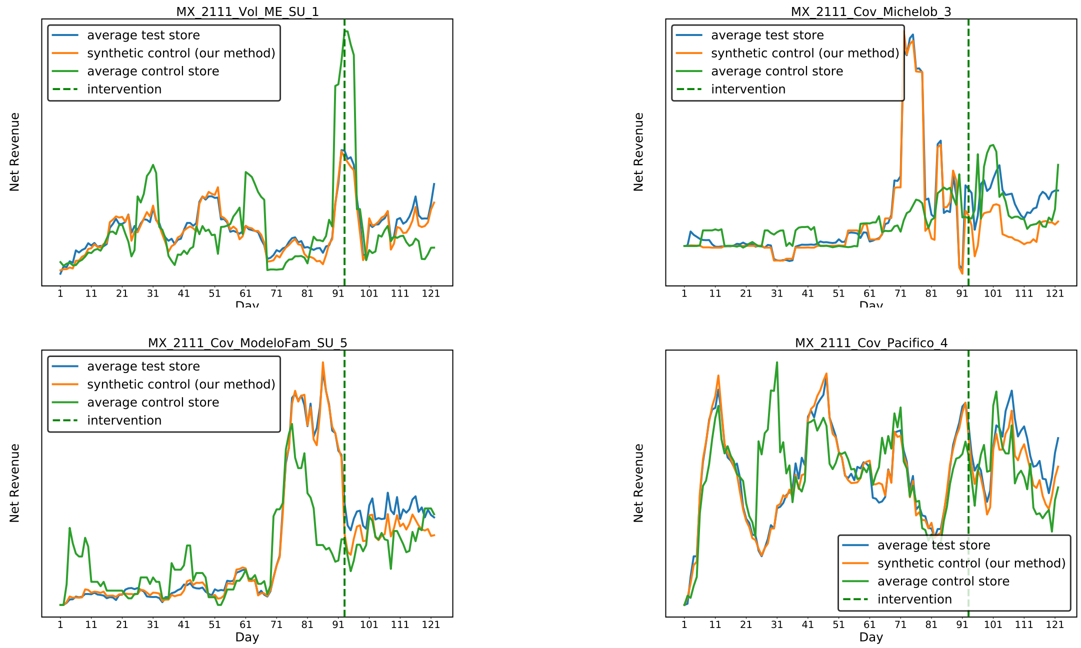


Figure 5: Illustration of experiments which are inconclusive under DID, while conclusively positive under GSC.

Test Intake (See Fig. 6). To begin, the user needs to select the market (such as Mexico, Colombia, etc..) where the experiment will run. The user can then (optionally) further filter out experimental stores based on store features such as sizes and locations in 'Select Scope and Granularity Filters'. The metrics that are of interest can also be selected in this step (such as volume or net revenues; multiple metrics may be selected simultaneously).

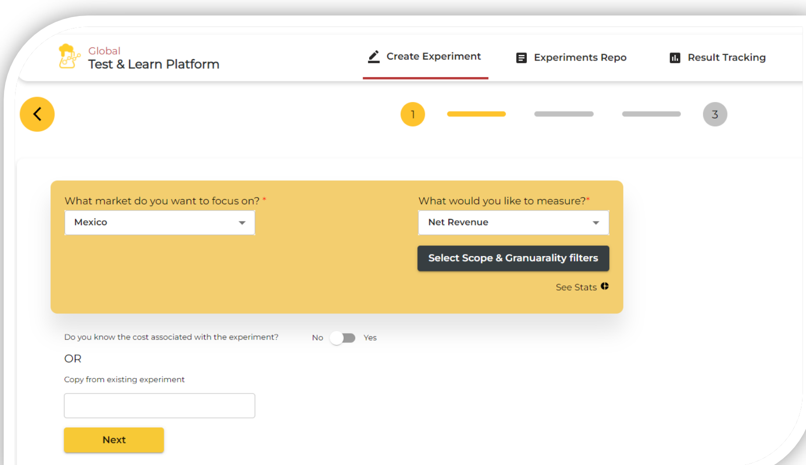


Figure 6: Test Intake at TestOps.

Experiment Design (See Fig. 7). Next, the user can start to specify more details about the experiment, such as a starting date and ending date, as well as the measurement range that one wants to get records on (so that the appropriate pre-treatment data can also be loaded). The number of control stores and the number of treatment stores can also be specified, and multiple initiatives can be scheduled (e.g., one control group with multiple treatment groups) at the same time (in 'Edit Experiment'). Here, we provide the user with guidance for how many stores to select, by automating the power analysis described in Section 4.2 so that the user can estimate the minimum number of stores required for an anticipated treatment effect. Once an experiment ticket is generated online, an operations team will reach out for further details and prepare to execute the experiment.

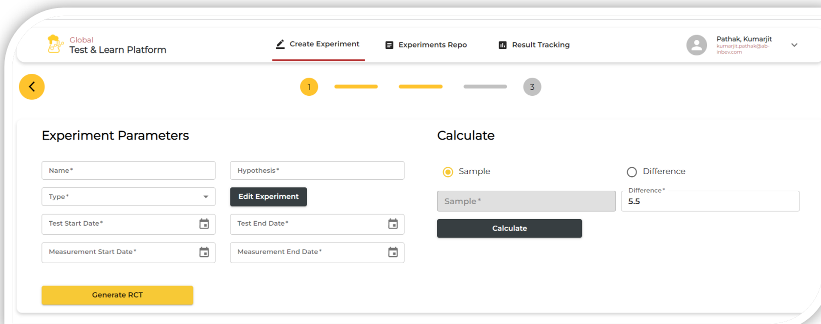


Figure 7: Experiment Design at TestOps.

Experiment Status (See Fig. 8). In the 'Experiment Repo' page, the user can check the status of all experiments they submitted; ('Not Started', 'Submitted', 'In-Progress', 'Completed'). A search panel helps the user to quickly identify the experiment they want to check. There is also a calendar view so that the user can easily have a timeline perspective for each experiment they are running.

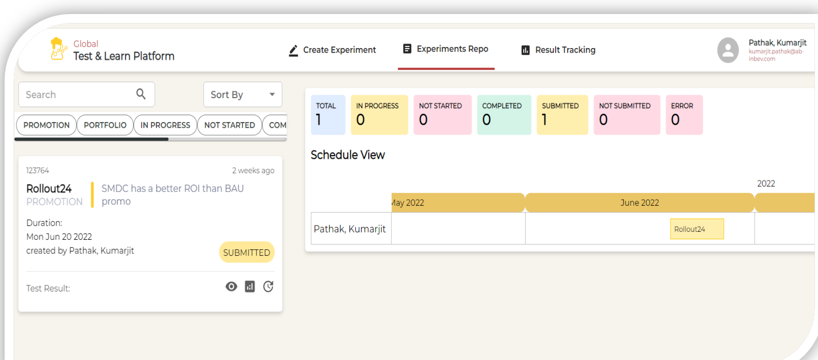


Figure 8: Experiment Status at TestOps.

Results Tracking (See Fig. 9). Finally, the user can track and analyze experimental results on the 'Results Tracking' page where real-time sales (and other metrics pertaining to the experiment) are tracked and presented. The treatment effects and p-values are automatically generated via GSC. Further, 'synthetic control'-type plots (of the type we discussed in Figure 5) are also presented to help the user transparently understand trends and the treatment effect estimate. A search panel is also provided for quickly identifying experiments (such as filtering by type, by date, etc..).

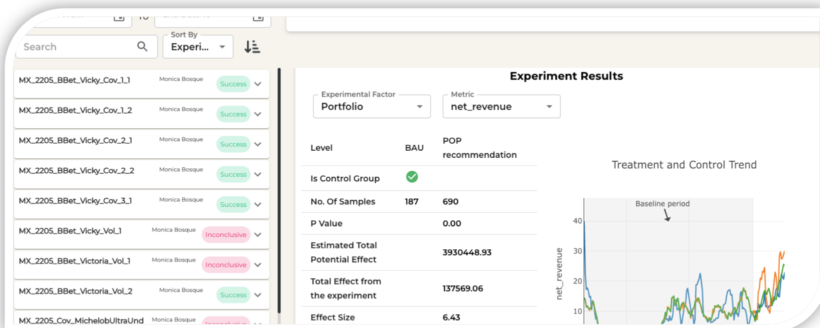


Figure 9: Experiment Tracking at TestOps.

This relatively straightforward approach to designing, running, and monitoring experiments has dramatically lowered barriers to rigorous learning from experimentation at ABI.

5 Broad Impact

Given its resounding success in Mexico, TestOps is already being scaled out to the Middle America Zone (MAZ) at ABI, which consists of all of Central America. In the next year, we anticipate a global rollout. We anticipate similar relative merits and performance improvements to what we have seen in Mexico. Further, the experimentation problems we have described are typical of physical retail, and so the ideas, models, and algorithms that we have developed in this work are likely valuable at other retailers with a large physical presence.

In addition, the mathematical/algorithmic ideas explored and extended in this work, including the surprising existence of an efficient solver for a hard non-convex problem and our implementation thereof, are of much broader significance within robust statistics, causal inference, machine learning, and optimization communities. These ideas may be applied in other contexts where factor models are appropriate and raise interesting research questions.

Beyond Retail. *Most importantly, we believe that the ability to learn efficiently, rigorously, and cost-effectively from experiments* ***fundamentally*** *determines our ability to extract economic value from* ***analytical innovations****. This work has sought to build a cutting-edge solution that facilitates such experimental learning.* We believe that the ideas here are generalizable and valuable beyond retail – such as in domains like healthcare delivery and public policy-making – where experimentation faces similar challenges.

References

Abadie A (2019) Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature* .

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association* 105(490):493–505.

Abadie A, Gardeazabal J (2003) The economic costs of conflict: A case study of the basque country. *American economic review* 93(1):113–132.

Agarwal A, Alomar A, Cosson R, Shah D, Shen D (2020) Synthetic interventions. *arXiv preprint arXiv:2006.07691* .

Agarwal A, Shah D, Shen D, Song D (2021) On robustness of principal component regression. *Journal of the American Statistical Association* (just-accepted):1–34.

Amjad M, Misra V, Shah D, Shen D (2019) mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3(2):1–27.

Amjad M, Shah D, Shen D (2018) Robust synthetic control. *The Journal of Machine Learning Research* 19(1):802–852.

Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S (2019) Synthetic difference in differences. Technical report, National Bureau of Economic Research.

Athey S, Bayati M, Doudchenko N, Imbens G, Khosravi K (2021) Matrix completion methods for causal panel data models. *Journal of the American Statistical Association* 1–41.

Bai J (2009) Panel data models with interactive fixed effects. *Econometrica* 77(4):1229–1279.

Bai J, Ng S (2019) Matrix completion, counterfactuals, and factor analysis of missing data. *arXiv preprint arXiv:1910.06677* .

Ben-Michael E, Feller A, Rothstein J (2021) The augmented synthetic control method. *Journal of the American Statistical Association* (just-accepted):1–34.

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* 119(1):249–275.

Chernozhukov V, Wüthrich K, Zhu Y (2021) An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association* (just-accepted):1–44.

Dimick JB, Ryan AM (2014) Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama* 312(22):2401–2402.

Doudchenko N, Imbens GW (2016) Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Farias V, Li A, Peng T (2021) Learning treatment effects in panels with general intervention patterns. *Advances in Neural Information Processing Systems* 34.

Imai K, Kim IS (2021) On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis* 29(3):405–415.

Ishimaru S (2021) What do we get from a two-way fixed effects estimator? implications from a general numerical equivalence. *arXiv preprint arXiv:2103.12374* .

Kahn-Lang A, Lang K (2020) The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics* 38(3):613–620.

Kropko J, Kubinec R (2018) Why the two-way fixed effects model is difficult to interpret, and what to do about it. *Kropko J, Kubinec R (2020) Interpretation and identification of within-unit and cross-sectional variation in panel data models. PLoS ONE* 15(4):e0231349.

Lechner M, et al. (2011) The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics* 4(3):165–224.

Li H, Graham DJ, Majumdar A (2012) The effects of congestion charging on road traffic casualties: A causal analysis using difference-in-difference estimation. *Accident Analysis & Prevention* 49:366–377.

Li KT (2020) Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association* 115(532):2068–2083.

Li KT, Bell DR (2017) Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics* 197(1):65–75.

Moon HR, Weidner M (2015) Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica* 83(4):1543–1579.

Moon HR, Weidner M (2017) Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory* 33(1):158–195.

Moon HR, Weidner M (2018) Nuclear norm regularized estimation of panel regression models. *arXiv preprint arXiv:1810.10987* .

PredictSales (2021) Predict future sales. `https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview`, accessed: 2021-01-15.

Wing C, Simon K, Bello-Gomez RA (2018) Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health* 39(1):453–469.

Xiong R, Pelger M (2019) Large dimensional latent factor modeling with missing observations and applications to causal inference. *arXiv preprint arXiv:1910.08273* .

Xu Y (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1):57–76.