Trajectory Balancing: A General Reweighting Approach to Causal Inference with Time-Series Cross-Sectional Data

Chad Hazlett∗       Yiqing Xu†

July 15, 2018

Abstract
We introduce trajectory balancing, a general reweighting approach to causal inference with time-series cross-sectional (TSCS) data. We focus on settings in which one or more units is exposed to treatment at a given time, while a set of control units remain untreated throughout a time window of interest. First, we show that many commonly used TSCS methods imply an assumption that a unit’s non-treatment potential outcomes in the post-treatment period are linear in that unit’s pre-treatment outcomes as well as time-invariant covariates. Under this assumption, we introduce the mean balancing method that reweights the control units such that the averages of the pre-treatment outcomes and covariates are approximately equal between the treatment and (reweighted) control groups. Second, we relax the linearity assumption and propose the kernel balancing method that seeks an approximate balance on a kernel-based feature expansion of the pre-treatment outcomes and covariates. The resulting approach inherits the property of handling time-vary confounders as in synthetic control and latent factor models, but has the advantages of: (1) improving feasibility and stability with reduced user discretion compared to existing approaches; (2) accommodating both short and long pre-treatment time periods with many or few treated units; and (3) achieving balance on the high-order “trajectory” of pre-treatment outcomes rather than their simple average at each time period. We illustrate this method with simulations and two empirical examples.

Keywords: causal inference, time-series cross-sectional data, panel data, difference-in-differences, synthetic control, interactive fixed effects, kernel balancing, reweighting

∗Departments of Statistics and Political Science, University of California, Los Angeles. Email: chazlett@ucla.edu.
†Department of Political Science, University of California, San Diego. Email: yiqingxu@ucsd.edu.
1. Introduction

Causal inference with observational data is an everyday challenge for many applied researchers. A wide variety of tools attempt to make causal claims using time-series cross-sectional (TSCS) data or panel data, in which over-time variations in the outcomes and treatment status in at least a subset of the units can assist in making credible inferences.

One popular set of tools for estimation in this context are the difference-in-differences (DID) and two-way fixed effects models. While time-invariant confounders are “differenced out” under these approaches, they share a common assumption that there exists no unobserved time-varying confounders influencing both the outcome and the treatment. This assumption, often referred to as “parallel trends” in the DID tradition, requires that average change in the non-treatment outcome is the same among the controls and treated, making their paths parallel. This assumption is not directly testable, and remains a core challenges of DID and fixed effects approaches.

In addition, Imai and Kim (2013) show that fixed effects models carry additional modeling assumptions, including (1) the treatment effect is homogeneous; (2) the treatment only affects the contemporaneous outcome and (3) past outcomes do not affect future treatment. To alleviate these concerns, in this paper, we focus on a generalized DID setting in which all units under consideration begin as untreated and a subset of units receive a treatment that begins at a given time. In this particular setting, because only two histories of treatment status are under consideration, we can allow the treatment to have a long-lasting effect on the outcome, as long as we make the direct comparison between potential outcomes under the two treatment histories.

Two approaches attempt to deal with the seemingly intractable problem of unobserved time-varying confounders in this setting, both by taking advantage of information in the pre-treatment outcomes. The synthetic control method (SCM) is a weighting-based approach that finds weights on control units that form a “synthetic control” unit whose pre-treatment history closely matches that of a single treated unit (Abadie and Gardeazabal 2003, Abadie, Diamond and Hainmueller

We further note that these models may have hierarchical/multilevel elements, random or mixed effects, or data structures that account for auto-regression of many forms, but these modeling assumptions do not typically weaken the core identifying assumptions of no unobserved time-varying confounders.
One interpretation of the SCM is as follows: assume that there exist one or several omitted time varying confounders and some time-fixed linear combination of these confounders affect the outcomes of both the treated and control units. Because these confounders appear in varying degrees in the control units, weighting the control units to make their averaged pre-treatment trend match that of the treated effectively replicates the combination of confounders that must be influencing the treated unit as well. Hence, the time-varying confounders are differenced out.

A closely related approach explicitly assumes a latent factor model (LFM), a data generating process (DGP) for which the SCM would be approximately unbiased under reasonable assumptions. While we examine the details below, in brief this approach presumes the scenario described above based on an interactive fixed effects (IFE) model: a set of time-varying influences (i.e., latent factors) exist, and each unit takes some (fixed) linear combination of them (Bai 2009). As described in Xu (2017), predicting treated counterfactuals using a LFM provides an alternative estimation procedure to that of synthetic control with greater flexibility. Both the SCM and LFM are attractive in that they allow omitted time-varying confounders of a particular form, but also have several important limitations we seek to address.

In this paper, we introduce trajectory balancing as a general solution to causal inference with observational data in a generalized DID setting. First, we show that a surprising variety of existing modeling approaches imply an assumption that the non-treatment potential outcomes in the post-treatment period are linear in the pre-treatment outcomes and covariates. We refer to this as the Linearity in Prior Outcomes (LPO) assumption. As a first contribution, the LPO assumption naturally suggests a simpler and more general estimation procedure: obtain equal means for the treatment and control groups, on the period-wise pre-treatment outcomes, and optionally on auxiliary covariates. Because of linearity, if the treated and control groups have equal means on these features, then they have equal means in any functions linear in these features, without having to estimate the corresponding coefficients. Therefore, under the proposed group of trajectory balancing approaches, we thus first consider such a mean balancing procedure. It chooses weights for the control units such that the mean-balance constraints are satisfied while the entropy or maximum
empirical likelihood of weights are maximized. An estimate of the Average Treatment Effects on the Treated (ATT) is obtained by taking the difference between the average of treated outcome and the weighed average of control outcome in the post-treatment period. This procedure is in the spirit of the SCM as it seeks balance on pre-treatment outcomes and covariates between the treated and controls. It is different in that it first seeks exact balance on feature when it is feasible; when exact balance is infeasible, it seeks balance on the first $P$ principal components of the features, where $P$ is chosen automatically. This approach sidesteps several additional challenges of the estimation procedures of SCM and LFMs. First, unlike the SCM, it can accommodate multiple treated units in a single run with little increase in computational time or risk of non-convergence. Second, it does not require a large number of pre-treatment periods, which is necessary for both the SCM and LFMs to work properly.

Our second contribution is to overcome the limitations inherent in the LPO assumption. With longer pre-treatment periods, the LPO assumption may be adequate: there are more opportunities for important time-varying “factors” to be visible in the pre-treatment outcomes, and more balance constraints to ensure balance on these factors. However, with fewer pre-treatment periods, there are fewer balance constraints to solve, and solutions may exist that obtain good mean balance on each pre-treatment period, but fail to obtain balance on factors that influence potential outcomes. Mean balancing alone does not ensure that higher order features of the pre-treatment trajectory—such as “volatility,” “variance” or “curviness”—are balanced between treated and control groups. Relatedly, mean balance can be obtained without any guarantee that the distribution of pre-treatment trajectories looks similar in the treated and control groups, as it ensures only the means look similar periodwise. To solve these problems, we take a higher-dimensional feature expansion of the pre-treatment trajectory as well as covariates, and seek approximate balance on this feature expansion.

This corresponds to a weakening of the LPO assumption, requiring linearity of the non-treatment potential outcome in the higher-dimensional feature set rather than the original data. A natural choice for such an expansion is a kernel-based procedure, that effectively considers the similarity of each unit’s trajectory to each other unit’s trajectory in a multivariate space, and takes this vector
of similarities as the features to be balanced. The *kernel balancing* procedure has the desirable property of ensuring that any function of the pre-treatment outcomes in a large space of smooth functions will have equal means in the treated and control groups (Hazlett 2018). Again, due to feasibility concerns, the procedure seeks balance on the first $P$ eigenvectors of the kernel based features. We describe the bound on the worst-case bias due to obtaining approximate balance, and this bound is minimized through the optimization procedure.

The trajectory balancing methods we propose inherit the same useful properties as the SCM and LFMs in coping with time-varying confounding through explicitly using the pre-treatment outcome data, but offers additional advantages by (1) improving feasibility and stability with reduced user discretion compared to existing approaches; (2) accommodating both short and long pre-treatment time periods, with many or few treated units; and (3) achieving balance on the high-order “trajectory” of pre-treatment outcomes rather than their simple averages at each time period. We illustrate this method with simulations and two empirical examples. We provide access to these tools in the `tjbal` package for R.

In what follows, Section 2 provides the framework and setup; Section 3 proposes the trajectory balancing method, details, and extensions. Section 4 provides a simulated example to fix ideas, and Section 5 provides two empirical applications (with two additional applications in the Online Appendix). We conclude with Section 6.

### 2. Setup

In this section, we set up the analytical framework and describe a family of existing models for causal inference with TSCS data. Though apparently diverse and flexible, these methods all require that for each unit $i$, its non-treatment potential outcomes in the post-treatment period are linear transformations of $i$'s non-treatment potential outcomes in the pre-treatment period. We define the assumption more precisely below and refer to it as the *Linearity in Prior Outcomes* (LPO) assumption. We find that many existing approaches fall in this category by either relying on models that directly assume LPO, or involving procedures that produce provably unbiased results only when this assumption holds. For simplicity, we begin by describing these methods in a context
without covariates, i.e. only pre-treatment outcomes are used for estimation purposes. We will later generalize our methods with the inclusion of covariates.

2.1. Notation, Assumptions, and Estimation Strategy

Suppose we have a time-series cross-sectional (TSCS) dataset of the generalized DID type, with $N_{tr}$ treated units and $N_{co}$ units; hence, the total number of units is $N = N_{tr} + N_{co}$. Denote a group indicator $G_i = 1$ if $i$ belongs to the treatment group and $G_i = 0$ if $i$ belongs to the control group.

All units are observed for $T$ periods from time 1 to time $T$. There may be a single unit, or multiple treated units all exposed to treatment for the first time at period $T_0 + 1$. Control units are never exposed to the treatment throughout the observed time period. Let $Y_{it}$ be the outcome of interest of unit $i$ at time $t$, and $D_{it}$ be an indicator of treatment status (i.e., $D_{it} = 1$ when $G_i = 1$ and $t > T_0$ and $D_{it} = 0$ otherwise). Denote $\{Y_{it}^1, Y_{it}^0\}$ the potential outcomes for unit $i$ at time $t$ when $D_{it} = 1$ or $D_{it} = 0$, respectively. Define $\tau_{it} = Y_{it}^1 - Y_{it}^0$ the individual contemporaneous treatment effect for unit $i$ at time $t$.

**Estimand.** The primary causal quantity of interest is the Average Treatment Effect on the Treated (ATT) at time $t$, $T_0 < t \leq T$, i.e.,

$$ATT_t = E[\tau_{it}|G_i = 1], \quad T_0 < t \leq T.$$

**Identification assumptions.** In order to identify the $ATT_t$, first, we assume the standard conditional ignorability assumption, but where the conditioning is done on the pre-treatment realizations of the non-treatment potential outcomes.

**Assumption 1 (Conditional Ignorability)**

$$Y_{it}^0 \perp\!
\perp G_i|Y_{i,pre}^0, \quad \forall t > T_0$$

in which $Y_{i,pre}^0 = (Y_{i1}^0, Y_{i2}^0, \cdots, Y_{iT_0}^0)$, a $(1 \times T_0)$ vector of pre-treatment outcomes. We note that only the non-treatment potential outcome $Y_{it}^0$, and not the treatment potential outcome $Y_{it}^1$ is involved
for purposes of estimating the ATT. This assumption is analogous to the selection-on-observable assumption in the causal inference literature (e.g. Rosenbaum and Rubin 1983) except that it is the pre-treatment outcomes are being conditioned on. Assumption 1 implies that once we condition on the pre-treatment outcome history, the treatment is “as-if” randomly assigned between the treated and control units. In other words, among units with the same pre-treatment histories, who gets the treatment is independent of (non-treatment) potential outcomes in the post-treatment periods.

Because the entire distribution of \( Y_{it}^0 \) is unchanged by knowledge of treatment status conditional on the pre-treatment history under Assumption 1, so is the expectation of \( Y_{it}^0 \), i.e.

\[
E[Y_{it}^0|Y_{i,pre}] = E[Y_{it}^0|Y_{i,pre}, G_i].
\]

In fact, it is only this “mean independence” assumption we require for the results below, but we employ the more commonly invoked Assumption 1 for consistency with the literature. It is convenient to rewrite this in terms of a generic functional form for the conditional expectation,

\[
E[Y_{it}^0|Y_{i,pre}] = f(Y_{i,pre}; \theta_t) = f(Y_{i,pre}; \theta_t; G_i), \quad T_0 < t \leq T,
\]

in which \( f_\theta(\cdot) \) is a flexible function of pre-treatment outcomes indexed by parameters \( \theta_t \). Later, we will see that the different procedures we propose correspond to different choices about both \( f(\cdot) \) and what is being conditioned upon. The starting point that we emphasize is that the conditional expectation of \( Y_{it}^0 \) is linear in that unit’s pre-treatment outcomes. We call this the Linearity in Prior Outcomes (LPO) assumption:

**Assumption 2 (Linearity in Pre-Treatment Outcomes)**

\[
E[Y_{it}^0|Y_{i,pre}] = (1 Y_{i,pre}^0)\theta_t, \quad T_0 < t \leq T,
\]

in which \( \theta_t \) is a \((T_0 + 1)\) vector of coefficients. We index \( \theta \) by \( t \) to indicate that each post-treatment time period may have a separate \( \theta \). However, we emphasize that \( \theta_t \) is fixed across units \( i \). Note

---

2To extend our results to the average treatment effect among the controls (ATC) or among all units (ATE) would require assumptions on \( Y_{it}^1 \) as well. However when only the ATT is required, the treatment potential outcomes observed directly, and the non-treatment potential outcomes of the treated are imputed from those of control units, requiring assumptions on the ignorability of treatment only for the non-treatment potential outcomes.
that the leading term in the vector \((1 \cdot Y_{i,pre}^0)\) and the corresponding first element of \(\theta\) allows an intercept of time \(t\). Thus, \(Y_{it}^0\) is affine in \(Y_{i,pre}^0\), but for ease of exposition we use the term linear.

### 2.2. Examples

A surprising number of approaches either explicitly posit a model of the DGP that implies LPO, or otherwise involve an estimation procedure that results in provably unbiased ATT estimates only when LPO holds. Below we provide several examples: (1) the difference-in-differences (DID) method and, relatedly, two-way fixed effect models: (2) a general structured time-series cross-sectional model; and (3) a latent factor model (LFM), which also provides a justification for the SCM.

**2.2.1. Examples**

**1. DID and two-way fixed effects and models.** We start with the DID model specified in Equation (2), in which \(D_{it} = 1\) for a subset of the units and 0 otherwise; \(\alpha_i\) and \(\xi_t\) are unit and time fixed effects and \(\varepsilon_{it}\) represents idiosyncratic shocks. In a two-period DID model, the treatment effect \(\tau_{it}\) is assumed to be heterogeneous both across units and over time. A closely related approach is the two-way fixed effects model specified in Equation (3). Note that a crucial difference between the two models is that the two-way fixed effects model assumes the treatment effect \(\tau\) to be constant across unit and over time.

\[
\begin{align*}
(DID) \quad Y_{it} &= \tau_{it} D_{it} + \alpha_i + \xi_t + \varepsilon_{it}, \quad t = 1, 2. \\
(Two-way) \quad Y_{it} &= \tau D_{it} + \alpha_i + \xi_t + \varepsilon_{it}, \quad t = 1, \ldots, T.
\end{align*}
\]

In either case, \(Y_{it}^0 = \alpha_i + \xi_t + \varepsilon_{it}\). In Section A.1 in the Appendix, we show that the non-treated potential outcome in the post-treatment period \(Y_{it,t>T_0}^0\) can be expressed as an average of pre-treatment outcomes plus some time-specific intercept shift and a zero-mean noise.

**2.2.2. Auto-regressive model with local trends.** Next, we investigate a structural time-series cross-sectional model that covers both traditional auto-regressive models and those with more complicated local, unit-specific trends, which is considered in Brodersen et al. (2015):
\[ Y_{it} = \tau_{it} D_{it} + \mu_{it} + \varepsilon_{it} \]

in which \( \mu_{it} = \mu_{i,t-1} + \xi_{it} + \eta_{it}^{\mu} \)

and \( \xi_{it} = \rho (\xi_{i,t-1} - \kappa) + \kappa + \eta_{it}^{\xi} \);

where the error terms \( \varepsilon_{it}, \eta_{it}^{\mu}, \) and \( \eta_{it}^{\xi} \) are independently distributed and have mean zero for all \( i \) and \( t \); \( \mu_{it} \) is a local, auto-regressive trend with a changing slope \( \xi_{xi} \); the slope of the time trend is a variation of AR(1) with a long-term slope \( \kappa \) when \( |\rho| < 1 \). When \( \rho = 0 \), \( Y_{it}^0 \) follows a random walk process. When both \( \rho \) and \( \kappa \) are equal to 0, the model is reduced to a AR(1) model. Similar to the DID setup, we allow the treatment effect \( \tau_{it} \) to be heterogeneous across unit and time. In this arrangement, too, the dependency of current non-treatment potential outcomes on past ones results in the LPO assumption. For details of the proof, see Section A.1 in the Appendix.

(3) The SCM and LFMs. Less obviously, we next consider a LFM, which also justifies the SCM. This model is also called the interactive fixed effect (IFE) models in the econometrics literature:

\[ Y_{it}^0 = f_t' \lambda_i + \xi_{it} + \varepsilon_{it}, \quad \forall i, t ; \]

in which \( \mathbb{E}[\varepsilon_{it}] = 0 \) for all \( i, t \). The bilinear form of \( f_t' \lambda_i \) gives this model two interpretations. One interpretation is that at any time \( t \), a vector of characteristics \( f_t \) exists, and each unit takes a linear combination of these characteristics, as given by a vector \( \lambda_i \). Unit \( i \) must always take combination \( \lambda_i \) of the characteristics in \( f_t \) while \( f_t \) can vary freely over time. The alternative interpretation reverses the emphasis of these two components: each unit \( i \) has some unknown time-invariant characteristics, \( \lambda_i \); though they are time invariant, their effects on the outcome can change freely over time, according to \( f_t \). However, all units must share the same set of coefficients \( f_t \) at a given time. This arrangement may seem quite flexible, but it turns out that it also requires \( Y_{it}^0 \) in the post-treatment period to be linear in unit \( i \)'s pre-treatment \( Y_{it}^0 \). Specifically, we can rewrite Equation (4) as:

\[ Y_{it}^0 = f_t' \lambda_i + \xi_{it} + \varepsilon_{it}, \quad \forall i, t ; \]

\(^3\)We omit (time-invariant) covariates for the time-being.
\[
Y_{it}^0 = \sum_{s=1}^{T_0} f_t'(\sum_{p=1}^{T_0} f_p f_p')^{-1} f_s Y_{is}^0 + \left[ \xi_t - \sum_{s=1}^{T_0} f_t'(\sum_{p=1}^{T_0} f_p f_p')^{-1} f_s \xi_s \right] \\
+ \left[ \varepsilon_{it} - \sum_{s=1}^{T_0} f_t'(\sum_{p=1}^{T_0} f_p f_p')^{-1} f_s \varepsilon_{is} \right] \quad \forall i, t > T_0 ;
\]

which reveals that \( Y_{it}^0 \) is a linear combination of the pre-treatment outcome (the first term on the right hand side (RHS)), plus some time-specific intercept shift (the second term on the RHS) and a zero-mean error term (the last term on the RHS). Again, the proof is given in Section A.1 in the Appendix.

3. Proposed Method

The LPO assumption implicated in these models suggests a more general estimation procedure that sidesteps a number of practical concerns with existing approaches. In this section, we begin describing the trajectory balancing method first with a mean balancing procedure that operates under the same LPO assumption. Then, we extend the method to a kernel based feature expansion that relaxes the LPO assumption.

3.1. Trajectory Balancing with Mean Balancing Weights

We first present a simple method for estimation when the LPO is expected to hold. We refer to this method as the mean balancing procedure, because it simply implies choosing weights on the controls such that the weighted average of the outcome is equal to the (unweighted) average of the treated units at each time-point in the pre-treatment period. In other words, mean balancing seeks to find a set of the weights \( w \) such that:

\[
\frac{1}{N_{tr}} \sum_{G_i=1} \sum_{G_i=0} Y_{i,pre}^0 = \sum_{G_i=0} w_i Y_{i,pre}^0
\]

There may be many sets of weights satisfying these constraints. Methods for choosing them generally maximize some criterion subject to the mean balance conditions. For example, maximum entropy weights maximize \( \sum_i w_i \log(w) \); and maximum empirical likelihood weights maximize
\[ \sum_i \log(w_i). \] Such weights can only be obtained when it is feasible, so an additional assumption is required:

**Assumption 3 (Feasibility of balancing weights)** There exists a set of non-negative weights \( \{w_i\}_{G_i=0} \) for the control units such that \( \sum_{G_i=0} w_i = 1 \) and pre-treatment outcomes are balanced between the treatment and reweighted control groups:

\[
\frac{1}{N_{tr}} \sum_{G_i=1} Y^0_{it} = \sum_{G_i=0} w_i Y^0_{it}, \quad t = 1, 2, \cdots, T_0.
\]

In the usual non-parametric setting for selection-on-observables, a “common support” assumption is also required, whereby at every value that the history of pre-treatment outcomes \( Y^0_{i,pre} \) could take, the probability of being a treated unit is greater than 0 and less than 1. In the current setup, this is not strictly necessary, and the feasibility assumption plays a similar role. The LPO assumption suggests that it is the means of \( Y^0_{i,pre} \) for the treated and controls that must be made equal, and doing so ensures equal expected non-treatment potential outcomes for the two groups.\(^5\) This, as shown next, eliminates bias in the ATT estimate. Although common support is not required, Assumption 3 can still fail if the treated and control group are too starkly different, for example, if the treated units have more extreme values on pre-treatment outcomes than any controls.

Under Assumptions 2 and 3, these weights can be used to estimate the ATT:

**Estimator 1** The mean balancing estimator for \( ATT \) is given by:

\[
\hat{ATT}_t = \frac{1}{N_{tr}} \sum_{G_i=1} Y^1_{it} - \sum_{G_i=0} w_i Y^0_{it}, \quad t = 1, 2, \cdots, T_0 ;
\]

where \( w_j \) are chosen s.t.

\[
\frac{1}{N_{tr}} \sum_{G_i=1} Y^0_{it} = \sum_{G_i=0} w_i Y^0_{it}, \quad t = 1, 2, \cdots, T_0 ;
\]

and \( \sum_{G_i=0} w_i = 1; \ w_i \geq 0, \ for \ all \ i \ in \ the \ controls. \)

\(^4\)We are agnostic as to what objective is maximized in seeking these weights and provide support for both maximum empirical likelihood and maximum entropy approaches.

\(^5\)In other words, because the linearity assumption takes us out of the fully non-parametric setting, common support is no longer required. Instead, because we have written bases for the non-treatment potential outcomes in the post-treatment period, we simply require feasibility of obtaining equal means on these bases.
Proposition 1 (Unbiasedness of the Mean Balancing Estimator) Under Assumptions 2 and 3, we have,

$$E[\hat{ATT}_t | Y_{1,pre}^0, Y_{2,pre}^0, \cdots, Y_{N,pre}^0] = ATT_t, \quad \forall t > T_0$$

in which $Y_{1,pre}^0, Y_{2,pre}^0, \cdots, Y_{N,pre}^0$ are pre-treatment outcome trajectories, and $ATT_t = \frac{1}{Ntr} \sum_{i=1}^{G_i} \tau it$. Proof is straightforward and can be found in the Section A.1 of the Appendix. A similar approach is described in Robbins et al. (2017) in relation to the SCM, though the approximations, kernel procedures, and other components described below have not been previously proposed in this context to our knowledge.

**Approximate balance through eigen-approximation.** Assumption 3 can fail in practice. Exact weights may be infeasible in a given sample, particular when there are many moment constraints (due to many pre-treatment time periods and/or covariates) and smaller numbers of control units to work with.\(^6\) Thus, existing procedures for achieving balance often fail in both simulations and applied examples and investigators turn to other approaches.

To address this issue, we employ an approximate balancing procedure that is more robust and reliable. Briefly, the algorithm seeks balance on the first $P$ left singular vectors, or equivalently, the principal components of $Y_{\text{pre}}^0 = (Y_{1,\text{pre}}^0, Y_{2,\text{pre}}^0, \cdots, Y_{N,\text{pre}}^0)'$, a $(N \times T_0)$ matrix.\(^7\) Moreover, under the linearity assumptions we have already made, the worst-case bias in the ATT estimate that arises from the approximation can be bounded, and we choose $P$ to minimize this potential for bias. In the next section, we will describe both the approximation approach and this bound at greater length, with specific reference to the kernelized version of the procedure introduced there.

---

\(^6\)In addition, feasibility may fail because there are significant “non-overlap” in pre-treatment outcomes or covariates between the treatment and control groups. We discuss such concerns below and describe the additional assumptions and options that investigators may consider to alleviate these challenges.

\(^7\)If the left singular vectors of $Y_{\text{pre}}^0$ are given by $U$, then the principal components are $UA$, where $A$ is the diagonal matrix containing the singular values of $Y_{\text{pre}}^0$. That is, the principle component scores are simply rescalings of the left singular vectors, so balance can be achieved on either with the same consequence. In actuality we construct the “linear kernel” matrix, $Y_{\text{pre}}^0 (Y_{\text{pre}}^0)'$, an $(N \times N)$ matrix. This simplifies generalization to the cases of non-linear kernels discussed below and allows us to use the same routine, distance measures, etc. whether a linear kernel is used or some other one. For the linear kernel, balancing on $Y_{\text{pre}}^0$ or on its linear kernel $K = Y_{\text{pre}}^0 (Y_{\text{pre}}^0)'$ is numerically equivalent. To see this, note that both have the same rank (the column rank of $Y_{\text{pre}}^0$) and the same left singular vectors, $U$. Further, because $K$ is symmetric positive definite, the left singular vectors are simply the eigenvectors. Hence, in practice we balance on the eigenvectors of $K$ and interchangeably refer to this approximation as either an “eigen-approximation” or as the use of principal components.
However, through the linear kernel, the procedure is the same under mean balancing as it is when we invoke a non-linear kernel next.

### 3.2. Relaxing the LPO Assumption: Feature Expansion Using Kernel

An important limitation of mean balancing is that the number of constraints solved by the weights is limited to $T_0$. With large $T_0$ this is less troubling. Intuitively, not only does large $T_0$ imply more constraints, but factors influencing $Y^0$ at post-treatment times will have a chance to appear in the pre-treatment period so that balancing on the pre-treatment outcomes in every period ensures that these factors are themselves well balanced. That said, even then this approach does nothing to guarantee that the trajectories of individual units look similar in the treated and control groups. For example, a control group that varies wildly around a flat line could be well mean balanced to a treated group that has all “flat” trajectories by giving equal weights to each control unit. Yet, the treated and control groups would look very different on features such as variance or volatility. If features like variance or volatility later come to have a large directional impact on $Y^0$, then this imbalance can generate bias. We provide a similar example in simulations below (Figure 2).

While the LPO corresponds to the mean balance procedure, relaxing the LPO assumption corresponds to an approach of seeking balance on additional, non-linear features of the pre-treatment outcomes. Consider a non-linear feature mapping $X \mapsto \phi(X)$ from $\mathbb{R}^D \mapsto \mathbb{R}^{D'}$. We are agnostic as to what this mapping is, but generally $D' \gg D$. Below, we propose a particular mapping implied by the use of a universal kernel. The key requirement of this mapping is that it enables us to claim that each unit’s expected post-treatment $Y^0$ outcomes are approximately linear in $\phi(Y^0_{i,\text{pre}})$.

Hence, Assumption 2 is replaced by the following assumption:

**Assumption 4 (Linearity of non-treatment outcome in $\phi(Y)$)**

\[
E[Y^0_{it}|Y^0_{i,\text{pre}}] = \phi(Y^0_{i,\text{pre}})'\theta_t;
\]

in which $Y^0_{i,\text{pre}} = (Y^0_{i1}, Y^0_{i2}, \cdots, Y^0_{iT_0})$, a $(1 \times T_0)$ vector of pre-treatment outcomes.

Note that Assumption 4 is also a special case of Equation (1), implied by Assumption 1. By the same logic as above, weights that achieve mean balance on $\phi(Y^0_{i,\text{pre}})$ for the treated and control units...
group will achieve equal means on $\phi(Y_{i,pre}^0)\theta_t$ for the treated and control groups, regardless of $\theta_t$. Thus, as LPO motivated mean balance in the original data $(Y_{i,pre}^0)$, Assumption 4 motivates mean balance on the features $\phi(Y_{i,pre}^0)$.

The choice of $\phi(\cdot)$ is clearly critical to the credibility of Assumption 4. We postpone this choice to the discussion of kernels below, as kernels will give us the ability to choose high- or infinite-dimensional choices of $\phi$ with desirable properties. For analytical purposes, we begin by again requiring a feasibility assumption,

**Assumption 5 (Feasibility of $\phi$-balance)** There exist a set of non-negative weights $\{w_i\}_{G_i=0}$ for the control units such that $\sum_{G_i=0} w_i = 1$ and pre-treatment outcomes are balanced between the treatment and reweighted control groups:

$$\frac{1}{N_{tr}} \sum_{G_i=1} \phi(Y_{i,pre}^0) = \sum_{G_i=0} w_i \phi(Y_{i,1...T_0}).$$

As noted above, this feasibility assumption supplants the more usual common support assumption. As $\phi(\cdot)$ is higher dimensional, this transformation makes feasibility easier to violate, which makes the approximation approach even more necessary. If, for example, the treated units are sufficiently different from the controls that the $\phi(Y)$ describing them on average lies outside the span of the $\phi(Y)$ of the control units, then the weights will be infeasible.

We now suggest the kernel-based variant of trajectory balancing under Assumption 4. It employs the same mean balancing idea as above, but with weights that obtain mean balance on $\phi(Y)$ between the treated and control units. The estimator is unbiased for $ATT_t$ under Assumptions 4 and 5,

**Estimator 2** The kernel balancing estimator for $ATT_t$ is given by:

$$\widehat{ATT}_t^k = \frac{1}{N_{tr}} \sum_{G_i=1} Y_{it}^1 - \sum_{G_i=0} w_i Y_{jt}^0$$

where $w_i$ are chosen s.t.

$$\frac{1}{N_{tr}} \sum_{G_i=1} \phi(Y_{i,pre}^0) = \sum_{G_i=0} w_i \phi(Y_{i,pre}^0)$$

and $\sum_{G_i=0} w_i = 1; w_i > 0$, for all $i$ in the controls.
Proposition 2 (Unbiasedness of the Kernel Balancing Estimator) Under Assumptions 4 and 5, we have,
\[ E[\hat{ATT}_k|Y_{1,\text{pre}}, Y_{2,\text{pre}}, \ldots, Y_{N,\text{pre}}] = ATT_t, \quad \forall t > T_0 \]
in which \( Y_{1,\text{pre}}, Y_{2,\text{pre}}, \ldots, Y_{N,\text{pre}} \) are pre-treatment outcome trajectories, and \( ATT_t = \frac{1}{N_{tr}} \sum_{G_i=1} \tau_{it} \).

We omit the proof as it is very similar to that of Proposition 1. While both the mean balancing and kernel balancing procedures will attempt to make the average trajectory of control units match that of treated units in the pre-treatment period, the latter is more fully in keeping with the name “trajectory balancing.” This is because the higher-order representation of the history of pre-treatment outcome given by \( \phi(Y_{\text{pre}}) \) allows us to go beyond balancing on the average trajectory, rather, it ensures that the control units looks more similar to the treated units in their trajectories receive higher weights. We demonstrate how such a step improves higher-order comparability of the chosen controls to the treated in simulations and applied examples below.

Kernel-based choice of \( \phi \). We now briefly discuss the kernel based approach that effectively chooses \( \phi(\cdot) \) and determines the weights. The basis for this method is kernel balancing (kBal, Hazlett 2018). While a much more technical discussion of kernels is possible, for present purposes, we can pose them simply as functions that assess similarity in some sense. Let \( k(Y_i, Y_j) \) be a function from \( \mathbb{R}^{T_0} \times \mathbb{R}^{T_0} \rightarrow \mathbb{R} \) that measures the similarity of \( Y_i \) and \( Y_j \). With \( N \) observations, an \((N \times N)\) kernel matrix \( K \) can be constructed such that \( K_{i,j} = k(Y_i, Y_j) \). Note that row \( i \) of \( K \), which we will designate \( K_i \) has the form \([k(Y_i, Y_1), k(Y_i, Y_2), \ldots, k(Y_i, Y_N)]\). The simplest view is one that regards \( \phi(Y_{i,\text{pre}}) \) as simply \( K_i \), or in matrix form, our feature expansion simply replaces \( Y_{\text{pre}} \) with \( K \). Thus, for each unit \( i \), we are proposing to replace the original \( T_0 \)-dimensional sequence of pre-treatment outcomes with a new, \( N \)-dimensional feature vector, \( K_i \). This \( K_i \) encodes how similar unit \( i \) is to unit 1, unit 2, and so on, making it a very rich representation of the data.

In theory, we would like to find weights that make the weighted average of \( K_i \) among the controls equal to the unweighted average of \( T_i \) among the treated,
\[
\frac{1}{N_{tr}} \sum_{G_j=1} K_j = \sum_{G_i=0} w_i K_i \text{ s.t. } \sum_{G_i=0} w_i = 1, w_i > 0, \forall i, G_i = 0
\]
though we will relax this requirement with an approximation momentarily. While many choices of
the kernel function $k(\cdot, \cdot)$ are possible, here we use the Gaussian kernel, the workhorse kernel in
machine learning,

$$k(Y_i, Y_j) = \exp(-||Y_i - Y_j||^2/h)$$

where $||Y_i - Y_j||$ is the Euclidean distance. It can be shown that all the functions that are linear
in $K_i$ are also linear in an infinite-dimensional feature expansion, one for which is $\langle \phi(Y_i), \phi(Y_j) \rangle = k(Y_i, Y_j)$, and that as $N \to \infty$, this space of functions contains all continuous functions.

**Approximate balance on $K$.** As exact balance on all $N$ dimensions of $K$ is typically infeasible,
we instead seek approximate balance. The approach is identical to that briefly described regarding
the linear kernel and mean balancing above, but we describe it at greater length here now that
the kernel matrix has been introduced. The basic idea is to achieve approximate balance while
minimizing the (worst-case) bias due to this approximation: (1) take the eigenvectors of $K$ based
on singular value decomposition (SVD), and (2) achieve balance on the first $P$ eigenvectors, leaving
those whose eigenvalues rank $P+1$ to $N$ unbalanced, where (3) the value of $P$ is chosen to minimize
the “worst-case” bias that could arise due to remaining imbalances.

The needed bound on the bias is derived as follows. Recalling Assumption 4 and replacing
$\phi(Y_i^{pre})$ with $K_i$, we have:

$$E[Y_i^0] = Kc = VAV'c = Vd$$

where $c$ is a set of coefficients on $K$ that exist due to the linearity assumption; $V$ is the matrix of
eigenvectors of $K$, $A$ is the matrix whose diagonal contains the eigenvalues of $K$, and $d = AV'c$.
Also note that the (reproducing kernel Hilbert space) norm for the chosen function is given by
c'Kc = d'A^{-1}Vd$. In conceptual terms, this norm describes how complicated or “wiggly” this
function is, which will be a quantity that we control in constructing the actual bias bound, but
which will remain constant in a given analysis and need not be estimated in order to choose the
number of dimensions $P$ to balance on.

Let $V_1$ be rows of $V$ corresponding to treated units, and $V_0$ the rows of $V$ corresponding to
control units. Suppose then we choose the vector of weights $w_0$ on the control units, and $w_1$ on
treated units. Because we target the ATT, every element of \( w_1 \) is set equal to 1 over the number of treated units. The bias of the ATT due to approximation, denoted \( \text{bias}_{w} \), is then

\[
\text{bias}_{w} = E[Y_0|G_i = 1] - E[Y_0|G_i = 0] = (w_1'V_1 - w_0'V_0)d = (w_1'V_1 - w_0'V_0)AV'c
\]

Note that the first term, \( (w_1'V_1 - w_0'V_0) \), is the imbalance on the eigenvectors under a given weighting scheme, and the second term, \( AV'c \), determines how much each imbalance matters towards producing bias. To obtain a worst-case bound on this bias when we do not know \( c \) (or \( d \)), we must instead control some related quantity. We impose control over only the Hilbert norm of the regression function, \( c'Kc \), as this controls how wildly the regression function is allowed to vary. Suppose we restrict the function to those with norm \( c'Kc \leq \gamma \). It is only if one desires an actual numerical estimate of the worst-case bias that a choice of \( \gamma \) would be required.\(^8\) We are then interested in the worst-case bias due to the approximation, \( \text{bias bound} \), given by

\[
\text{bias bound} = \sup_{c'Kc \leq \gamma} |(w_1'V_1 - w_0'V_0)AV'c|
\]

Letting \( z = c'K^{1/2}\gamma^{-1} \), the above can be rewritten as

\[
\sqrt{\gamma} \sup_{z'z \leq 1} |(w_1'V_1 - w_0'V_0)AV'K^{-1/2}z|
\]

which by Cauchy-Schwarz gives

\[
\text{bias bound} \leq \sqrt{\gamma} ||(w_1'V_1 - w_0'V_0)AV'K^{-1/2}||_2 \\
\leq \sqrt{\gamma} ||(w_1'V_1 - w_0'V_0)A^{1/2}||_2
\]

The form of this worst-case bound is informative. First, the \( L_2 \) norm of the regression function \((\sqrt{\gamma})\) controls the overall scale of potential bias. Second, the imbalance on the eigenvectors of \( K \) after weighting, \((w_1'V_1 - w_0'V_0)\), enters directly. Third and most illuminating, the impact of

\(^8\)A reasonable choice of \( \gamma \) can be made, for example by actually regressing the outcome among control units on their pre-treatment outcomes, using the same kernelized specification (e.g. using Kernel Regularized Least Squares, Hainmueller and Hazlett 2014).
imbalance on each eigenvector is scaled by the square root of the corresponding eigenvalue. This suggests that our approach of achieving balance on the first $P$ eigenvectors is a reasonable strategy for quickly minimizing worst case bias. Because the matrix $K$ typically has a few large eigenvalues, then many very small ones, it is usually possible to achieve fine balance on enough eigenvectors such that the remaining eigenvalues carry a tiny fraction of the total variation in $K$.

This bound is effectively a measure of remaining imbalance on $K$ as it can impact ATT estimates, and we use it as the target of optimization when choosing how many eigenvectors to seek balance on. For optimization purposes, $\gamma$ does not vary, hence, it can be ignored. The procedure thus chooses $P$ so as to minimize the worst-case bias in the ATT that remains due to the approximate nature of balancing, regardless of the true norm of the regression function or the actual coefficient values.

In short, while feature expansion incurs a cost, the cost is manageable. Although the balance achieved on the new bases is usually approximate, we are still able to bound a worst-case bias due to this approximation.

### 3.3. Extensions

We discuss two extensions to the trajectory balancing approach. First, we incorporate time-varying pre-treatment covariates in a balancing procedure. Second, we allow an intercept shift prior to balancing.

**Pre-treatment covariates.** Consistent with other work (e.g., ADH 2010), we consider the possibility that users would like to explicitly ensure balance on pre-treatment, time-invariant covariates as well. The argument for including covariates should typically be that they account for potential confounders; hence, the identifying assumption is more credible after their inclusion. Denote $X_i$ as a vector of time-invariant covariates for unit $i$. Assumption 1 (conditional ignorability) then becomes:

$$Y_{it}^0 \perp G_i | X_i, Y_{i, pre}^0, \forall t > T_0,$$
which implies $E[Y^0_{i,t} | X_i, Y_{i,pre}] = f(X_i, Y_{i,pre}; \theta_t)$, for all $t > T_0$. Therefore, if we add $X_i$ to the $Y_{i,pre}$ being conditioned upon in Assumptions 2–5 and in both the mean balancing and kernel balancing estimators, the unbiasedness property follows.

With the kernel estimator, by including both the covariates and the pre-treatment outcomes in $\phi(\cdot)$, this feature expansion includes the full range of interactions between the two. In practice, it means that the kernel matrix $K$ is formulated using vectors $[\tilde{X}_i, Y_i]$ as arguments to the Gaussian kernel in Equation (3.2), where $\tilde{X}_i$ is a rescaled version of the covariates with mean zero and variance one. Again, weights are chosen to achieve mean balance on the rows, $K_i$, as in Equation (5). Approximate mean balance on this form of $K_i$ can be thought of as ensuring that the joint-distribution of covariates and pre-treatment outcomes is approximately balanced.

In general, adding covariates will increase dimensionality, making balance more difficult to achieve. Moreover, if any treated units take values on the covariates more extreme than that of control units, it can make even the best available weights perform poorly. In the first empirical example below (Truex 2014), this is not an issue in the observed data, though through bootstrap repetitions we notice that it could easily become an issue had key control units not been sampled. The addition of covariates makes effective weights more difficult to find in the second example (ADH 2010), which we use to motivate a demeaning procedure to which we now turn.

**Intercept shift and demeaning.** When the number of control units is small or the outcome trajectories of the treated units do not lie in the convex hull of those of the control units, finding a set of weights that significantly reduce the imbalance between the treatment and control groups can be difficult. The intuition is simple: when data are sparse and time-invariant individual heterogeneity (i.e., unit intercept) is strong, for at least some of the treated units, it is hard to find control units that share both the same level and the same dynamics, a failure of Assumption 5. This problem gets worse when covariates are added to the balancing scheme.

One option to alleviate this problem is to allow an intercept shift for each unit. Specifically, before reweighting, the average outcome from period 1 to period $T_0$ is subtracted from the original outcome for each unit. This demeaning procedure ensures the outcome of each treated or control unit is of mean zero in the pre-treatment period. After an intercept shift, we only consider imbalance
in dynamics but not levels, therefore, it gives opportunities to more control units to contribute substantially to the estimation of the average treated counterfactual.

Allowing an intercept shift gives more room for successful solutions to be found, but it does not come without cost. First, the conditional ignorability assumption needs to be replaced by the following “parallel trends” assumption:

**Assumption 6 (Parallel Trends)**

\[
E[Y_0^t - Y_{0i}^t | \dot{Y}_{0i,pre}] = E[Y_0^t - Y_{0i}^t | \dot{Y}_{0i,pre}, G_i], \quad \forall t, s \in \{1, 2, \cdots, T\}
\]

where \(\dot{Y}_{0i,pre} = (\dot{Y}_{0i1}, \dot{Y}_{0i2}, \cdots, \dot{Y}_{0iT_0})\) is a \((1 \times T_0)\) vector demeaned pre-treatment outcomes, in which \(\dot{Y}_{0it} = Y_{0it} - \bar{Y}_{0i,pre}\) and \(\bar{Y}_{0i,pre} = \sum_{t=1}^{T_0} Y_{0it}\).

Assumption 6 says that, once the demeaned pre-treatment dynamics are being conditioned on, the average \(Y_0\) of the treated units would follow a parallel path of that of control units. This assumption is analogous to the “parallel trends” assumption in the two-period DID framework (Abadie 2005). Assumption 6 implies that:

\[
E[\dot{Y}_{0it} | \dot{Y}_{i,pre}] = f(\dot{Y}_{i,pre}; \theta_t), \text{ for all } t > T_0.
\]

Hence, we replace Assumption 4 with: \(E[\dot{Y}_{0it} | \dot{Y}_{i,pre}] = \phi(\dot{Y}_{i,pre}/\theta_t)\).

Intuitively, this means that any two pre-treatment trajectories that vary only by a vertical translation (an intercept shift) carry the same information, which may not be true under the most general circumstances.\(^9\) However, most commonly used methods, such as DID and fixed effects models, are invariant to vertical shifts when no non-linear transformation of the outcome is attempted. With DID, for example, only within-unit changes matter; with unit fixed effects, each unit receives it’s own independent intercept shift, with the same consequence. Doudchenko and Imbens (2016) considers forbidding intercept shifts an unnecessary restriction under many circumstances.

\(^9\)For example, a country’s annual GDP growth rate falls from 7% to 6% (an absolute change of 1 percentage point) is likely very different in kind from one that drops from 2% to 1%, with the same absolute percentage point change. To treat them as similar by allowing vertical translations thus may create inappropriate counterfactuals. Therefore, whenever it is possible, we prefer not to impose this additional assumption.
3.4. Quantifying Uncertainties

The question of variance estimation under this procedure remains open to further research. When the number of treated units is very small or one as in many applications of the SCM, then the treated outcome has no well defined variance. While it may be possible to construct standard errors on the (weighted) average outcome among the controls (see below), the treated unit’s outcome has no standard error or confidence interval, and thus the uncertainty of the ATT is undefined. In such cases investigators may avoid classical inference altogether, and instead seek to estimate a large number of placebo tests in which that units by construction should not have an treatment effect, and thus whose results collectively constitute a null distribution of no effect. The estimated effect can then be compared to these. For example, treatment can be reassigned to different units, much as in Fisherian permutation inference.\(^\text{10}\) Such approaches can be pursued with this method as well.

The trajectory balancing method introduced in this paper tolerates many more treated units than the original SCM, opening the door to inference on the ATT estimates. One approach would be to regard the weights as fixed and apply them as a pre-processing step. Any average taken with these weights can then be given the usual weighted standard error. However, the user may also reasonably worry about the uncertainty due to the choice of weights. A closed form solution to this problem would be highly desirable, but we are not yet aware of such an option. In its place, we propose a bootstrap procedure that mimics the intended re-sampling experiment and fully replicates the multiple sources of uncertainty:

1. Re-sample both the treatment and control group with replacement, preserving the time-series of each unit. \(N_{tr}\) and \(N_{co}\) are kept as fixed.
2. Estimate the weights, allowing the number of dimensions chosen by the approximation routine to vary as it is data-driven
3. Estimate the desired target quantity, such as the ATT at a given time period or an average of ATTs over the post-treatment periods
4. Use the variation in the estimated quantities across trials to construct confidence intervals, either by directly examining percentiles or taking the standard deviation (error) and applying normal theory.

\(^\text{10}\)See, for example, ADH (2010) and more recently Robbins et al. (2017).
Note that it is easiest to think of the data in “wide” format, where each unit takes a row and the outcomes at different times are spread across columns, as are any covariates. In this format, the bootstrap is a simple one: each row in the treatment group will have the same probability of being sampled with replacement; so is each row in the control group. If one thinks instead of a “long” format, in which a given unit takes multiple rows representing each time point, then this would amount to a “block bootstrap” on the unit level. The validity of the above procedure requires further validation, and with hopes that it may some day be replaced by a faster, potentially closed-form alternative.

4. A Simulated Example

In this section, we provide a simulated example to illustrate situations in which the kernel balancing method may outperform mean balancing. We focus on the advantage of using the kernel method in terms of ensuring balance on higher-order features with fewer pre-treatment periods.

The main shortcoming of mean balancing, in intuitive terms, is that there may be many ways to average together the trajectories of control units to produce a weighted average that “looks like” that of the treated at each time point, but these solutions can give substantial weight to units whose trajectories do not look like those of the treated. Because \( \phi(Y_{i}^{pre}) \) encodes higher-order features such as “curvature” and “low frequency oscillation” in the history of pre-treatment outcome, obtaining balance on \( \phi(Y_{i}^{pre}) \) ensures similarity in these qualities.\(^{11}\)

Consider a dataset of \( N = 150 \) countries of 24 years \( T \in \{1, 2, ..., 24\} \), and a simulated outcome we label as \( GDP \). A treatment (e.g. the onset of some public policy) begins in Year 15. Further, we imagine two “types” of countries in this stylized example. Suppose that 100 of these countries

\(^{11}\)It would be possible to instead use matching on the pre-treatment histories, using some form of distance over the entire pre-treatment vector in order to establish what control units are sufficiently close to each treated unit (e.g. Imai, Kim and Wang 2018). Instead, here we take the weighting approach. Among other reasons, one benefit of doing so is that we can analyze the solution using the approach above, in which we first establish a feature space in which the outcomes are linear, and second establish equal means on all these features.
have an outcome that is volatile, with cyclical noise, and no long-term growth, i.e.,

\[ GDP_{it} = 5 + a_i \sin(2\pi t) + b_i \cos(2\pi t) + .1\varepsilon_{it} \]

\[ \varepsilon_{it} \sim N(0,1), \quad a_i, b_i \sim U(-1,1) \]

These outcomes are thus periodic, but with a phase that differs between units depending on \( a_i \) and \( b_i \). Some countries may have low values of both \( a_i \) and \( b_i \), producing more stable outcomes, but not steady growth. The remaining 50 countries have slow-but-steady growth at a rate of 4% per year:

\[ GDP_{it} = 4 + c_i 1.04^t + .1\varepsilon_{it} \]

\[ \varepsilon_{it} \sim N(0,1), \quad c_i \sim U(0.9,1.1) \]

We note that this is a case of an IFE with three factors: each country takes some linear combination of the sine, cosine, and steady growth functions. Finally, treatment is assigned only to stable countries, with one quarter of them assigned to treatment at random. For simplicity, let us assume the treatment effect is exactly zero, and so the values of GDP represent both the treatment and non-treatment potential outcomes. As a result, if we find a good counterfactual for the treated by weighting the control units properly, it should match exactly the post-treatment outcomes of the treated units. We will know our solution generates bias if we see an apparent effect in the post-treatment period.

We use both mean balancing and kernel balancing to find weights for the controls. Figure 1 shows the pre-treatment period, with the trajectories of the most heavily-weighted control units (those accounting for 90% of the total weight), as selected by mean balancing and kernel balancing. Mean balancing finds a good match on the pre-treatment period, but does so by spreading its weight among a combination of volatile and stable-growing types of countries. By contrast, kernel balancing emphasizes almost entirely stable-growing types of countries exactly because it takes into account higher order features.

Next, we investigate the consequences of these weighting choices. To illustrate the benefits of using the kernel method, we consider three scenarios that differ only in the amount of available pre-treatment data: 2 years, 6 years, 10 years, or 14 years. Figure 2 shows balance on pre- and
Figure 1. Heavily Weighted Control Units

Note: Controls chosen by each method in simulated example. (Left): Pre-treatment time trends for the control units that receive the top 90% of the total weight, under mean balancing. Many high-volatility type units are included, in addition to some more stable-growing units. (Right): Pre-treatment time trends for the most heavily weighted controls chosen by kernel balancing, also accounting for 90% of the total weight. This approach weights almost exclusively the stable-growing units, which are more similar to the treated units on features such as variance or frequency content.

post-treatment periods for each. In all 4 cases, as expected, both methods can achieve good (mean) balance on the pre-treatment trends. However, when there are only 2 pre-treatment periods (upper-left panel), mean balancing fails to predict the average trajectory of the treated units because it chooses relatively equal weights on control units, giving too much weight to the high volatility units. As a result, the growth experienced by the treated units is lost in the noise. The number of constraints solved by the balancing procedure is too small to produce an average counterfactual that captures that growth of the treated. However, the kernel based approach correctly selects the low-volatility units that are more similar to the treated units. This is because it solves for more constraints and ensure balance on higher order features of the pre-treatment outcome. As a result, the predicted counterfactual remains well matched to the average of treated units in the post-treatment period. As the number of periods grows, kernel balancing maintains its performance while mean balancing continues to select control units poorly until approximately 14 time periods are available (bottom-right panel).

Finally, to gain a deeper insight into what is being balanced upon, we examine the principal components of the kernel matrix. An important feature of the outcome variable in each unit’s pre-treatment period is its volatility. Figure 3 shows that the first principal component of \( \mathbf{K} \) is highly informative as to the variance in each country’s pre-treatment history. The horizontal axis gives
Figure 2. Mean Balancing vs. Kernel Balancing: A Simulated Example

Note: Pre- and post-treatment trends as determined by mean balancing and kernel balancing, varying the number of pre-treatment periods available. The true treatment effect is zero in every post-treatment period, hence an unbiased estimator should show no “gap” between the treatment and constructed control outcomes. In each panel, the solid red line shows the average GDP of the treated units. Under mean balancing, the weighted outcome among controls (dashed blue line) is an excellent match on the pre-treatment trends. However, when the number of pre-treatment periods is smaller, this does not imply good “post-treatment balance”. That is, the control group constructed by mean balancing weights is not comparable to the treated group on its future non-treatment outcomes. By contrast, using kernel balancing weights, the weighted control group (dashed black line) follows the treated group closely. While noisier with only two pre-treatment periods, this approach achieves a reasonably good counterfactuals despite small number of pre-treatment periods, and correctly estimates an ATT near 0 at each time point. Only once the number of pre-treatment periods grows to approximately 14 (bottom right) does the mean balancing approach begin to catch up.

Each country’s log of variance in pre-treatment outcomes, and the vertical axis is each country’s value on the first principal component of $K$. The empty gray circles show these values for each control unit and the red filled circles represent the treated. Furthermore, the filled black circles of varying size show the location of control units again, but scaled by the weights found by kernel balancing. Figure 3 shows that these weights are largest in the low-variance region occupied by the treated units, indicating that the method properly upweights those units whose variance are similar to that of the treated and that almost no weight is given to units further afield. Note that variance is a non-linear function of the pre-treatment history, and hence, not automatically balanced on by simply achieving mean balance.
Figure 3. Pre-treatment variance and the first component of $K$

Note: Analysis of first component of $K$. Empty circles show control units, whose pattern reveals a strong relationship between the log of variance of each country’s pre-treatment outcomes and its value on the first eigenvector or component of $K$. Filled red circles show the location of the treated units. The filled black circles again show the control units, but sized according to the weight that trajectory balance assigns them. The choice of weights focuses heavily on control units similar to the treated on variance, and thus on the first component of $K$.

5. Empirical Examples

In this section, we demonstrate the applicability of trajectory balancing to several empirical examples. We explore two examples that have previously been analyzed using entropy balancing and the synthetic control method, Truex (2014) and ADH (2010). They not only employ different estimation strategies but also vary in important features such as the number of pre-treatment periods and the numbers of treated and control units.

5.1. Return to Office in China’s National People’s Congress (Truex 2014)

First, we investigate the data from Truex (2014), who studied the return to office of China’s “rubber-stamp” parliament, the National People’s Congress (NPC). Truex (2014) uses a firm-level annual dataset spanning from 2005 to 2010. There are 48 treated firms, whose CEOs started to hold seats in the NPC in 2008, and 948 controls, whose CEOs never held seats in the NPC during this time window. Hence, the first three periods (2005–2007) are pre-treatment. The author uses entropy balancing to achieve equal means on the three pre-treatment outcomes and two time-invariant
covariates (i.e., the mean balancing approach described here), including share of state ownership and each firm’s total revenue in 2007. Using a DID model with a reweighted sample, he finds that a seat in the NPC occupied by a company’s CEO will lead to an increase in return to assets (ROA) of 3 to 4 percentage points.

We re-examine the data using both the mean balancing and kernel balancing methods to re-weight the control firms. As in the original paper, we include all three pre-treatment outcome measures and the two covariates. Figure 4 shows the ATT estimates for the effect of an NPC seat on ROA. The ATT estimates based on the two reweighting methods turn out to be very similar to each other. Using mean balancing, we find that a firm’s CEO taking a seat in the NPC, on average increased the firm’s ROA by 2.1 percentage points in the three-year period with a 95% confidence interval of (0.9, 3.2). The estimate is 1.6 percentage points with a 95% confidence interval of (0.3, 2.6) using kernel balancing, suggesting that the original finding is robust to adjusting for higher-order features in the outcome trajectories.

**Figure 4. Replicating the Main Result of Truex (2014)**

![Graph showing replication results for Truex (2014) using both mean balancing and kernel balancing methods.](image)

*Note:* The above plot shows the replication results for Truex (2014) using both the mean balancing and kernel balancing methods. The variables to be balanced on include the three pre-treatment outcomes (from 2005 to 2007) and two covariates.

We also observe from Figure 4 that, with mean balancing, the confidence intervals in the pre-treatment period are reduced to the three point estimates. This is because with each bootstrap run, the algorithm is able to achieve exact mean balances on the included variables. In contrast, in this case, applying SCM is not an ideal option for two reasons: (1) the algorithm will have to be repeatedly run for each treated unit; (2) the small number of pre-treatment periods prevents the algorithm from measuring the quality of matches.
with kernel balancing, the confidence intervals on the pre-treatment estimates signal that some proportion of resampled estimates do not find close balance on the pre-treatment outcomes. In general it is the case that mean balance will not be exact under kernel balancing because the moment conditions solved are different and must also attain balance on higher-order functions. However, the result here show a much larger imbalance on some of the bootstrap resamples. This is an informative result and suggests the ability to achieve good balance must depend upon the presence of a small number of units, such that the resampling procedure can create samples where balance is not obtainable. In this case, the difficulty is that a single treated unit has higher revenue in 2007 than any firm among the controls. In the original observed sample, this could be handled by approximate balancing because a single control unit lies relatively close to it. However, with that control unit absent in some bootstrap resamples, good balance cannot be found. In Appendix Figure A1, we show that when this single “non-comparable” treated unit is dropped, kernel balancing achieves excellent mean balance across bootstrap resamples. We provide diagnostic tools for researchers to discover such “non-overlap” issues, allowing a decision about whether to trim treated units before running analyses.

To gain more insights into how these two reweighting schemes may differ, we present Table 1, a balance table using both methods. It shows the differences between the treatment and control groups in the pre-treatment outcomes and covariates before and after reweighting using both mean balancing and kernel balancing. We see that the mean balancing approach is able to achieve exact balance in this case on the pre-treatment outcomes and covariates, while a small amount of mean imbalance remain with kernel balancing. However, with kernel balancing, other functions of these covariates are also well balanced. For instance, we can see that the standard deviations of these variables in the reweighed control group are much close to those in the treatment group than with mean balancing. Graphically, in Figure 5, we plot outcome trajectories of the control units that receives heavy weights from each method and compare them with those of the treated units. The left and right panels of Figure 5 are the outcome trajectories of the top 25% most highly weighted control units using mean balancing (left) and trajectory balancing (right), respectively, while the panel in the middle shows the outcome trajectories of the 48 treated units. Figure 5
shows that, as in the simulation example, trajectory balancing utilizes information of higher order transformations of the pre-treatment outcomes such that the outcome trajectory in the treated and reweighted controls are similar in distribution while mean balancing only ensures balance in means of the outcome measures at each time point. Though this has little impact on the final ATT estimate in this example, the greater similarity of the reweighted controls to the treated units in the pre-treatment period lends greater credibility to the belief that the constructed counterfactual will correctly predict the non-treatment potential outcomes of the treated unit in the post-treatment period—even if the post-treatment outcomes do not happen to be linear in the pre-treatment outcomes.

**Figure 5. Pre-treatment Trajectories of Treated and Heavily Weighted Controls**

<table>
<thead>
<tr>
<th></th>
<th>Treated Mean</th>
<th>Unweighted Controls Mean</th>
<th>Reweighted Controls w/ Mean Balancing Mean</th>
<th>Reweighted Controls w/ Kernel Balancing Mean</th>
<th>Unweighted Controls SD</th>
<th>Reweighted Controls w/ Mean Balancing SD</th>
<th>Reweighted Controls w/ Kernel Balancing SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROA in 2005</td>
<td>0.038</td>
<td>0.022</td>
<td>0.038</td>
<td>0.040</td>
<td>0.067</td>
<td>0.078</td>
<td>0.067</td>
</tr>
<tr>
<td>ROA in 2006</td>
<td>0.051</td>
<td>0.030</td>
<td>0.051</td>
<td>0.052</td>
<td>0.042</td>
<td>0.066</td>
<td>0.046</td>
</tr>
<tr>
<td>ROA in 2007</td>
<td>0.064</td>
<td>0.041</td>
<td>0.064</td>
<td>0.064</td>
<td>0.056</td>
<td>0.160</td>
<td>0.056</td>
</tr>
<tr>
<td>State Ownership</td>
<td>0.325</td>
<td>0.249</td>
<td>0.325</td>
<td>0.323</td>
<td>0.244</td>
<td>0.222</td>
<td>0.240</td>
</tr>
</tbody>
</table>

**Note:** The above plots show the pre-treatment outcome trajectories for the treated units (middle) and top 25% most highly weighted controls using mean balancing (left) and kernel balancing (right). Data are from Truex (2014).
5.2. Tobacco Control Program and Cigarette Sales (ADH 2010)

Finally, we apply the trajectory balancing approach to a classic example of the SCM (ADH 2010). The authors use U.S. state level data to examine the effect of Proposition 99, a tobacco control program the state of California implemented in 1988, including a 25-cent per tax on cigarettes, taxes on other tobacco products, and a ban on cigarette vending machines in certain areas. We are interested in this case for two reasons. First, we intend to compare the performance of trajectory balancing, including both mean balancing and kernel balancing, and the SCM, since it has become increasingly popular in comparative case studies. Second, in the previous case, the number of pre-treatment periods $T_0$ is small and the number of control units $N_{co}$ is relatively large; by contrast, this example has $T_0 = 19$, but with only 39 control units and a single treated unit.

In the original paper, the authors match on only three periods of the pre-treatment outcome (1975, 1980, and 1988), plus four pre-treatment covariates,$^{13}$ reserving the remaining pre-treatment outcomes to select a weighting matrix required by the algorithm. However, in trajectory balancing we include all 19 periods of the pre-treatment outcome (from 1970 to 1988) as well as the four covariates for three reasons: (1) we want to minimize user discretion when choosing pre-treatment variable; (2) no validation period is required to be set aside by our procedure; and (3) by allowing as many periods as possible to enter the kernel matrix, we hope that trajectory balancing can take full advantage of information in the dynamics of each unit.

We found that when only balancing on the pre-treatment outcomes, both mean balancing and kernel balancing are able to obtain an excellent improvement in the imbalance and a pre-treatment trajectory for the weighted controls that almost exactly matches that of the treated unit, California; see Figure A2 in the Appendix. However, once covariates are added, neither balancing method is able to find substantial improvements achieve a good alignment of the pre-treatment trends without an intercept shift.$^{14}$ This is because only a handful of states share a similar pre-treatment trajectory in terms on both the dynamics and the level with California. We thus invoke the demeaning

---

$^{13}$They include average values of beer consumption, log income, tobacco retail price, and the share of population aged from 15 to 24 during the period of 1980 to 1988.

$^{14}$This means that only a few dimensions of $K$ can be balanced upon, and the resulting pre-treatment trend for the weighted control is not very well aligned with that of the treated (California).
procedure as described in Section 3, removing the mean of each unit’s pre-treatment outcome.

In Figure 6, we plot the outcomes of the treated unit and heavily weighted control units based on the SCM (left), mean balancing (middle), and kernel balancing (right). In all three plots, the horizontal axis is time, and vertical axis is the outcome variable, *per capita* number of cigarette packs sold during that year. The red dashed line represents the outcome trajectory of California while the gray lines represent the control states that receive weights bigger than 0.03 using each of the three methods.\(^\text{15}\) We see that the three methods differ in significant ways. In particular, kernel balancing puts heavy weights on units that share similar pre-treatment trajectories (e.g. Colorado, Idaho, and Delaware) while both mean balancing and the SCM put heavy weights on units that help ensure mean balance is achieved. Weights generated by mean balancing are much more evenly distributed than those produced by kernel balancing, a consequence of the fact that balancing on trajectories (encoded by higher-order features) is much more selective than balancing on means.

**Figure 6. California Tobacco Control Program**
**Treated and Most Heavily Weighted Controls**

**Note:** The above plots show the outcome trajectories of the treated unit (California, red dashed line) and the control states that have weights bigger than 0.03 (gray) using the SCM (left), mean balancing (middle), and kernel balancing (right). The SCM matches on four time-invariant covariates as well as the outcome (packs of cigarettes) measured in 1975, 1980, and 1988. Both mean balancing and kernel balancing use the same four covariates and the outcome variable for all pre-treatment periods (1970–1988). Line widths of the controls correspond to the assigned weights. The data are from ADH (2010).

Figure 7 plot the counterfactual and treatment effect results from both the SCM and trajectory (kernel) balancing. The left panel shows the actual treated outcome (red solid line) as well as treated counterfactual constructed by the SCM (dark gray, dashed line) and trajectory balancing (black,

\(^\text{15}\)The weights are shown in Table A1 in the Appendix.
solid line). The right panel of Figure 7 shows the gaps between the actual outcome of California and the predicted $Y_0$ using synthetic control (dark gray, dashed line) and kernel balancing (black, solid line). From both plots we see that predicted $Y_0$ from both methods matches the pre-treatment outcome well except for the very early 1970s, and differ only slightly in the post-treatment period. If anything, trajectory balancing provides slightly bigger estimates of the effects of Proposition 99 on tobacco consumption.

**Figure 7. California Tobacco Control Program Stability to Specifications**

Note: The above plots show treated and constructed counterfactual (left) and estimated treatment effect (right) using the SCM and trajectory (kernel) balancing. For the SCM, besides using the default periods (1975, 1980, and 1988) for matching, we repeatedly select alternative pre-treatment periods to be matched on (each time, we randomly select 3 periods from 1970 to 1988), resulting in different constructed counterfactuals and treatment effect estimates; in contrast, trajectory balancing uses all pre-treatment periods. The data are from ADH (2010).

One potential drawback of the SCM is that it asks users to choose the set of pre-treatment variables—and most importantly, what periods of the pre-treatment outcomes—to be matched on. Recent research has shown that cherry picking pre-treatment variables in the SCM may lead to over-rejection of the null hypothesis (Ferman et al. 2017). To examine how severe this problem may be, we randomly select three periods of the pre-treatment to be matched on and conduct synthetic control analyses repeatedly for 100 such choices. The results are also shown in Figure 7. On the left, each gray line represent the trajectory of a synthetic control unit using an alternative combination of pre-treatment outcome measures as matching targets; on the right, each gray line
is a gap plot for each combination. We see that, depending on the set of pre-treatment periods to be matched on, the estimated effect can vary by as much as 10 packs at a given time. In contrast, such discretion is completely removed with trajectory balancing since all pre-treatment periods have been taken into account.

In summary, the above empirical examples illustrate the versatility and robustness of trajectory balancing under different circumstances. It can accommodate TSCS/panel data structure with either short $T_0$ or long $T_0$, and allows for relatively small $N$ if needed. Beyond the theoretical motivation for relaxing the LPO assumption, these examples show that trajectory balancing has advantages over the SCM in two respects: (1) it give heavy weights to control units that share almost the exact same trajectories of that the treated unit and (2) it minimizes user discretion in choosing the variables to be matched/balanced on.

6. Conclusion

In this paper, we propose trajectory balancing, a general reweighting approach to causal inference with TSCS data. We introduce two trajectory balancing estimators: (1) mean balancing and (2) kernel balancing.

The mean balancing estimator reweights the control units such that the treatment group and reweighted control group will have equal means on pre-treatment outcomes and covariates. It directly exploits the LPO assumption built into a variety of existing procedures for causal inference with TSCS data, including those that seek to deal with time varying confounders such as synthetic control method (SCM) and latent factor models (LFMs). However, it avoids a number of practical concerns of the SCM and LFMs. For example, it allows few or many pre-treatment periods, few or many treated units, and ensures convergence utilizing an approximation.

The kernel balancing estimator relaxes the LPO assumption underlying mean balancing procedure by seeking balance on a high-dimensional (kernel) description of the existing features. As a result, the weighted control group is similar to the treated not only on the period-wise means prior to treatment, but on high-order features of the trajectories. Intuitively, kernel balancing ensures
the overall distributions of trajectories are similar between the treatment group and reweighted control group, which is not necessarily the case when mean balancing is applied.

Prior to employing these methods, users may wish to examine their data, particular to assess whether there are treated units that fall outside the range of the control units at least separately on each margin. That is, for each pre-treatment outcome and time-invariant covariate included, it is useful to know if the most extreme values (maximum or minimum) belong to a treated unit, and if so, whether there are one or more control units close to that value. The results may suggest particular a priori strategies: (1) If treated units occupy minimum or maximum values of the pre-treatment outcomes, the demeaning procedure may become necessary to achieve good balance. (2) If treated units occupy minimum or maximum values of the covariates, then the user might (a) reconsider the importance of including those covariates; or (b) trimming of the offending “non-comparable” unit in the treatment group.16

As a general guideline, one workflow may be as follows (see Figure 8 for an illustration). Before applying trajectory balancing, the user would often benefit from examining potential issues of non-overlap. Then, to begin employing trajectory balancing, first users can begin with the weakest set of assumption, employing the kernel balancing approach, without demeaning. The quality of match in the pre-treatment outcomes reveal whether weights could be found that perform well. One should also check the weights to see how heavily the result relies upon a small group of controls. If the balance is poor or the result is heavily weighted towards a few units, the user may wish to impose additional assumptions that can relieve these problems. One option is to allow for the intercept-shift/demeaning. This is most reasonable when one believes that a time-invariant difference in the level of the outcomes for a given unit is not important to the expected trend or trajectory of changes. Alternatively, one could employ the mean balancing option rather than use the Gaussian kernel to see if the resulting weights are more satisfactory. This implies the stricter LPO assumption, but may remain reasonable especially when there are many pre-treatment time points. Recall that with many pre-treatment time points, the mean balancing approach consequently has many moment conditions to satisfy, and solutions that achieve good balance on the pre-treatment period

16We use the term “non-comparable” to refer to (treated) units with values more extreme than control units and which thus create an inability to find comparable control units.
more plausibly maintain balance on $Y^0$ thereafter. By contrast, when there are very few pre-
treatment periods, mean balance implies fewer constraints and is “too easy” to achieve, thus the
kernel balancing approach imposes stronger condition by achieving balance on high-order functions
of the pre-treatment trajectories. While we recommend these general guidelines, further theoretical
and applied work will be useful in better understanding the conditions that recommend one method
over the other.

Compared with the latent factor approach and other model-based methods, our approach re-
quires minimum modeling assumptions. Though we require that the non-treatment potential out-
comes fall in a particular function space determined by the kernel, this is a very general function
space. The procedure never directly fits a model, hence, chances of erroneous extrapolation based
on estimated model parameters is minimized. Second, because we assign an explicit non-negative
weight to each control unit, the method is transparent and easy to understand in terms of the
average counterfactual it constructs.

Compared with the SCM, our approach, and kernel balancing in particular, has several ad-
ditional advantages. First and foremost it achieves balance on (an approximation of) the full
trajectory of pre-treatment outcomes rather than their average levels. Thus, the reweighted control
units are much more similar in their joint distribution of pre-treatment variables to the treated group than can be ensured when using the SCM. Second, our method accommodates various types of data—it works whether the number of the pre-treatment periods is small or big, or the number of treated unit is one or many—thus providing a unified framework. By contrast, the SCM usually requires a long pre-treatment period and is often difficult to extend to cases with multiple treated units. Third, our method is easy to implement and requires minimum user input. Unlike the SCM, it does not require users to specify which pre-treatment outcomes or their higher order interactions to be matched on, thus minimizing the negative effects of research degrees of freedom.

There are several limitations of our approach. First, kernel balancing potentially requires more data than existing methods simply because it attempts to achieve balance on a higher order description of the pre-treatment variables instead of the means only. One way to relieve this is by relaxing assumptions, such as allowing intercept shifts as we did in the second example. A related limitation is that, in practice, it is not possible to achieve balance on all the dimensions of $K$, which form the bases of the assumed function space. Rather, as described, we must instead obtain balance on the best rank $P$ approximation of $K$, where $P$ is selected to minimize an overall imbalance metric. That said, even existing approaches that involve only balance on averages can run into a similar feasibility limitation, which we also handle through the same approximation under mean balancing. Third, the inferential method for trajectory balancing is not fully developed. Future work is needed to address this issue.
References


A. Supplementary Information

A.1. Proofs

A.1.1. DID and two-way fixed effects models imply the LPO assumption

Proof: In DID and two-way fixed effects models, \( Y_{it}^0 = \alpha_i + \xi_t + \varepsilon_{it}, \forall i, t. \) Therefore,

\[
Y_{it}^0 = (Y_{it}^0 - T_0^{-1} \sum_{s=1}^{T_0} Y_{is}^0) + T_0^{-1} \sum_{s=1}^{T_0} Y_{is}^0
\]

\[
= \left[ (\alpha_i + \xi_t + \varepsilon_{it}) - T_0^{-1} \sum_{s=1}^{T_0} (\alpha_i + \xi_s + \varepsilon_{is}) \right] + T_0^{-1} \sum_{s=1}^{T_0} Y_{is}^0
\]

\[
= (\xi_t - T_0^{-1} \sum_{s=1}^{T_0} \xi_s) + (\varepsilon_{it} - T_0^{-1} \sum_{s=1}^{T_0} \varepsilon_{is}) + T_0^{-1} \sum_{s=1}^{T_0} Y_{is}^0 ;
\]

in which the first term on the right-hand-side (RHS), \( \xi_t - \sum_{s=1}^{T_0} \xi_s \) is a time-specific but unit-independent deviation; the second term on the RHS, \( \varepsilon_{it} - T_0^{-1} \sum_{s=1}^{T_0} \varepsilon_{is} \), has expectation of zero, when we assume that \( E[\varepsilon_{it}] = 0 \) for all \( i, t \); and the third term on the RHS is an average of pre-treatment outcomes. Q.E.D.

A.1.2. The structure time-series model considered in Section 2 implies the LPO assumption

Proof: The non-treatment outcome of the model can be written recursively as:

\[
Y_{it}^0 = Y_{i,t-1}^0 + (\varepsilon_{it} - \varepsilon_{i,t-1}) + \xi_{it} + \eta_{it}^\mu
\]

\[
= Y_{i,t-2}^0 + (\xi_{it} + \xi_{i,t-1}) + (\eta_{it}^\mu + \eta_{i,t-1}^\mu) + (\varepsilon_{it} - \varepsilon_{i,t-2})
\]

\[
= \ldots
\]

\[
= Y_{i,T_0}^0 + \sum_{s=T_0+1}^{t} \xi_{is} + \sum_{s=T_0+1}^{t} \eta_{is}^\mu + (\varepsilon_{it} - \varepsilon_{iT_0}), \forall t > T_0, \forall i ;
\]

in which the first term on the RHS, \( Y_{i,T_0}^0 \), is in \( Y_{i,pre}^0 \) and the last term on the RHS, \( \sum_{s=T_0+1}^{t} \eta_{is}^\mu + (\varepsilon_{it} - \varepsilon_{iT_0}) \), has expectation of zero when \( E[\eta_{it}^\mu] = 0 \) and \( E[\varepsilon_{it}] = 0 \) for all \( i, t \). Hence, we only need to consider the second term, \( \sum_{s=T_0+1}^{t} \xi_{is} \). We know that:

\[
\xi_{is} = \rho^{(s-T_0)} \xi_{iT_0} + (s - T_0)(1 - \rho) \kappa + \sum_{p=T_0+1}^{s} \eta_{ip}^\xi
\]
in which the second term on the RHS, \((s - T_0)(1 - \rho)\kappa\), is a constant; the last term, \(\sum_{p=T_0+1}^s \eta_{ip}\), is of zero mean; and the first term can be expressed as:

\[
\xi_{iT_0} = (Y_{iT_0}^0 - Y_{i,T_0-1}^0) - (\varepsilon_{i,T_0} - \varepsilon_{i,T_0-1}) + \eta_{iT_0}
\]

which is a linear combination of two pre-treatment outcomes plus a zero-mean noise. Hence, \(Y_{it}^0\) can be written as a combination of the linear transformation of pre-treatment outcomes, a time-specific but unit-independent coefficient, and a zero-mean noise. Q.E.D.

**A.1.3. IFE models imply the LPO assumption**

**Proof:** Consider for each unit \(i\), the projection \(Y_{it}^0\) in the pre-treatment period onto the span of \(f_p\), obtaining a set of coefficients:

\[
\hat{\lambda}_i = \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} \sum_{s=1}^{T_0} f_p Y_{ip}
\]

\[
= \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} \sum_{p=1}^{T_0} f_p (f_p' \lambda_i + \xi_p + \varepsilon_{ip})
\]

\[
= \lambda_i + \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} \sum_{p=1}^{T_0} f_p (\xi_p + \varepsilon_{ip})
\]

Note that, because there is no unit fixed effect to act as an intercept in the model, the prediction for \(Y_{it}^0\) is simply \(f'_t \hat{\lambda}_i\). Finally, at some post-treatment time \(t\), we could think of \(Y_{it}^0\) as a combination of the predicted value \(\hat{Y}_{it}^0\) and a residual \(Y_{it}^0 - \hat{Y}_{it}^0\), yielding:

\[
Y_{it}^0 = \hat{Y}_{it}^0 + (Y_{it}^0 - \hat{Y}_{it}^0)
\]

\[
= f'_t \hat{\lambda}_i + f'_t (f_p' \lambda_i + \xi_t + \varepsilon_{it}) - f'_t \left[ \lambda_i + \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} \sum_{p=1}^{T_0} f_p (\xi_p + \varepsilon_{ip}) \right]
\]

\[
= f'_t \hat{\lambda}_i + \xi_t + \varepsilon_{it} - f'_t \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} \sum_{p=1}^{T_0} f_p (\xi_p + \varepsilon_{ip})
\]

\[
= f'_t \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} \sum_{p=1}^{T_0} f_p Y_{is} + (\xi_t + \varepsilon_{it}) - f'_t \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} \sum_{p=1}^{T_0} f_p (\xi_p + \varepsilon_{ip})
\]

\[
= \sum_{s=1}^{T_0} f'_t \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} f_s Y_{is} + \left( \xi_t - \sum_{s=1}^{T_0} f'_t \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} f_s \xi_s \right) + \left( \varepsilon_{it} - \sum_{s=1}^{T_0} f'_t \left( \sum_{p=1}^{T_0} f_p f_p' \right)^{-1} f_s \varepsilon_{is} \right)
\]

which reveals that \(Y_{it}^0\) is a linear combination of the pre-treatment outcomes (the first term on the RHS), plus some time-specific intercept shift (the second term on the RHS) and a zero-mean error term (the last term on the RHS). Q.E.D.
A.1.4. Unbiasedness of the mean balancing estimator

**Proof:** Under Assumptions 2 and 3:

\[
\begin{align*}
\mathbb{E}[\hat{\text{ATT}}_t | Y_{\text{pre}}^0] &= \frac{1}{N_{\text{tr}}} \sum_{G_i=1} \mathbb{E}[Y_{it}^1] - \frac{1}{N_{\text{co}}} \sum_{G_i=0} \mathbb{E}[w_i Y_{it}^0 | Y_{\text{pre}}^0] \\
&= \frac{1}{N_{\text{tr}}} \sum_{G_i=1} \mathbb{E}[Y_{it}^0 + \tau_{it}] - \frac{1}{N_{\text{co}}} \sum_{G_i=0} \mathbb{E}[w_i Y_{it}^0 | Y_{\text{pre}}^0] \\
&= \mathbb{E}[\frac{1}{N_{\text{tr}}} \sum_{G_i=1} \tau_{it}] + \frac{1}{N_{\text{tr}}} \sum_{G_i=1} \mathbb{E}[Y_{it}^0 | Y_{i,\text{pre}}^0] - \sum_{G_i=0} w_i \mathbb{E}[Y_{it}^0 | Y_{i,\text{pre}}^0] \\
\overset{\text{LPO}}{=} & \text{ATT}_t + \frac{1}{N_{\text{tr}}} \sum_{G_i=1} (1 Y_{i,\text{pre}}^0)' \theta_t - \sum_{G_i=0} w_i (1 Y_{i,\text{pre}}^0)' \theta_t \\
\overset{\text{feasibility}}{=} & \text{ATT}_t.
\end{align*}
\]

Q.E.D.
A.2. Additional Results for the Empirical Examples

Figure A1. Dropping One Outlier in Truex (2014)

(a) Results in the main text

(b) Results after dropping one non-comparable unit

Note: The above figures compares the results in the Truex (2014) example in the main text (Figure A1 in Truex (2014)) and the results after we one single unit whose $rev_{2007}$ is 81104.75 (while the maximum of $rev_{2007}$ in the control group is 76180.45) from the treatment group. The substantive findings are similar, but when the unit whose value is too extreme to be supported by the controls (the “non-comparable” unit) is dropped, balance is achieved as expected in the pre-treatment period, resulting in narrow confidence intervals for kernel balancing.
**Figure A2. Trajectory Balancing without Demeaning**

![Figure A2](image)

*Note:* The above plots show the results from trajectory (kernel) balancing without subtracting each unit’s pre-treatment outcome mean. The outcome variable of all pre-treatment periods are included in the balancing scheme, but not the four time-invariant covariates. (a): the outcome trajectories of the treated unit (red, dashed line) and the top 7 control units that receive the largest weights (gray, solid lines). The width of each gray solid line corresponds to the weight of the control unit. We see that the control unit that resembles the treated unit the most in outcome trajectory in the pre-treatment period (Montana) receives the largest weight. (b): the outcome trajectories of the treated unit and the average of reweighted control outcomes (predicted $Y_0$). We see that control units that receive large weights are those sharing a similar trajectory (in terms of both the dynamics and the level) to that of the control unit. Data are from ADH (2010).
<table>
<thead>
<tr>
<th>State</th>
<th>Synthetic Control</th>
<th>Mean Balancing</th>
<th>Kernel Balancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utah</td>
<td>0.34</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Montana</td>
<td>0.25</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Nevada</td>
<td>0.19</td>
<td>0.17</td>
<td>0.00</td>
</tr>
<tr>
<td>Colorado</td>
<td>0.05</td>
<td>0.02</td>
<td>0.45</td>
</tr>
<tr>
<td>New Mexico</td>
<td>0.05</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Idaho</td>
<td>0.03</td>
<td>0.03</td>
<td>0.36</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>0.03</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>Alabama</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Arkansas</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Connecticut</td>
<td>0.00</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>Delaware</td>
<td>0.00</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>Georgia</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Illinois</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Indiana</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Iowa</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Kansas</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Kentucky</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Louisiana</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Maine</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Minnesota</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Mississippi</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Missouri</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Nebraska</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>North Carolina</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>North Dakota</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Ohio</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>South Carolina</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>South Dakota</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Tennessee</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Texas</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Vermont</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Virginia</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>West Virginia</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Wyoming</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
A.3. Two Additional Empirical Examples

We apply trajectory balancing to two additional examples. The first one is a class example on the effect of a job training program from LaLonde (1986). The second one comes from Xu (2017), who studies the effect of Election-Day Registration (EDR) on voter turnout using a linear factor model (LFM).

A.3.1. Effect of Job Training (LaLonde 1986)

It is useful to know whether this approach can recover an average treatment effect estimate in an observational setting where we nevertheless know the “true” answer. To do so, we consider the National Supported Work (NSW) program which was first used as a benchmark in observational studies by LaLonde (1986) and Dehejia and Wahba (1999), and which has become a routine benchmark for matching and weighting approaches (e.g. Diamond and Sekhon 2005; Iacus et al. 2011; Hainmueller 2012).

Following LaLonde (1986), the treated sample from the NSW experiment is compared to a control sample drawn from a separate, observational sample. Methods of adjustment are tested to see if they accurately recover the treatment effect despite large observable differences between the control sample and the treated sample. Usually this is done on a range of pre-treatment covariates: age, years of education, real earnings in 1974, real earnings in 1975 and a series of indicator variables: Black, Hispanic, and married, and often indicators for being unemployed (having income of $0) in 1974 and 1975, and an indicator for having no highschool degree (fewer than 12 years of education).

Here we instead use this approach in a TSCS setting by considering only the prior outcome measures, earnings in 1974 and 1975. As far as we know, we are the first to show how using only pre-treatment outcome measures in a TSCS allows for estimates in this example that very closely match the experimental benchmark, without need of covariates or addition of an indicator for zero income.17

We use the 185 treated units from NSW, originally selected by Dehejia and Wahba (1999) for the treated sample. The experimental benchmark for this group of treated units is $1794, which is computed by difference-in-means in the original experimental data with these 185 treated units. The control sample is drawn from the Panel Study of Income Dynamics (PSID-1), containing 2490 individuals.

17As found by Dehejia and Wahba (1999), propensity score matching can effectively recover reasonable estimates of the ATT, but are highly sensitive to specification choices in constructing the propensity score model (Smith and Todd 2001). Diamond and Sekhon (2005) use genetic matching to estimate treatment effects with the same treated sample. While matching solutions with the highest degree of balance produced estimates very close to the experimental benchmark, these models included the addition of squared terms and two-way interactions, not to mention the constructed indicators for zero income in 1974 and 1975. Similarly, entropy balancing (Hainmueller 2012) has also been shown to recover good estimates using a similar setup, using a control dataset based on the Current Population Survey (CPS-1), employing all pairwise interactions and squared terms for continuous variables, amounting to 52 covariates.
Figure A3 shows that the ATT of job training on earning in 1978 with different combinations of the reweighting scheme (mean balancing versus kernel balancing) and transformation of the outcome measure or the lack of thereof (logarithmic versus real dollar value). For both mean balance and kernel balance, we use the pre-treatment outcomes (1974 and 1975 income) as the input.

**Figure A3. Lalonde Data with Mean Balancing and Trajectory Balancing**

![Diagram showing estimated ATT of job training with different transformations and balancing methods]

**Note:** The above plots show our replication using the Lalonde data. We use entropy balancing to achieve mean balance on earnings in 1974 and 1975 with dollar values (first estimate) or their logarithmic transformation (second estimate). We apply kernel balancing using dollar values of 1974 and 1975 earnings (third estimate) or their logarithmic transformations (fourth estimate) as input.

Some may argue that investigators should have sufficient substantive knowledge to know what transformations of the variables to work in, such that LPO becomes reasonable. Our general position is not only that authors cannot be expected to know what transformations of the outcome to use, but that giving them this responsibility instead opens the door to abuse through additional researcher degrees of freedom. For example, studying the effect of a job training program on income, it would seem a defensible choice to log those incomes (after adding one to avoid the log of zero) for purposes of estimating weights. And yet, doing so radically alters the results. Figure A3 shows that on the natural scale, mean balance actually does an excellent job and recovers approximately the correct (benchmark) ATT estimate. However, if one takes the reasonable step of logging incomes, the result changes by thousands of dollars. By contrast, trajectory balancing is relatively indifferent to the choice of transformation used, recovering an approximately correct answer in both cases. Because this approach ensures a very rich representation of the pre-treatment trajectory is similar in the treated and in the counterfactual control group, we need not rely on investigators to guess transformations, nor allow them to.
A.3.2. Election Day Registration on Voter Turnout (Xu 2017)

Next, we apply trajectory balancing to the example of EDR and voter turnout in the United States from 1920 to 2012. This example is different from the ones appearing in main text in that the treatment (EDR laws) took effect at different times. Nine states passed EDR laws during this time period. Three states, Maine, Minnesota, and Wisconsin enacted the EDR from 1973 to 1975, which took effect in 1976’s presidential election; three states, Wyoming, Idaho, and New Hampshire, followed in 1994–1995 and the EDR took effect in 1996’s presidential election. In addition, the EDR took effect in Connecticut in 2008 and in Iowa and Montana in 2012. Thirty-eight states that had never passed EDR laws during this period serve as controls.

We apply trajectory balancing (using both mean balancing and kernel balancing allowing intercept shift) for each of four different timings, 1976, 1996, 2008, and 2012. The results are shown in Figure A5. We find that the results from trajectory balancing in general agree with the results from the generalized synthetic control (GSC) method, which is based on a latent factor model (Xu 2017). We then re-align the ATT estimates based on the time passed since the beginning of the treatment and obtain the ATT estimates for all 9 treated states. The results are shown in Figure A4. It is worth noting that the SCM fails to find a solution in 5 of 9 cases.

The above procedures can be easily implemented with the \texttt{tjbal} package.

**Figure A4. Election Day Registration on Voter Turnout: All EDR States**

![Figure A4](image)

**Note:** The above figure shows the ATT of EDR laws on voter turnout using three different methods: (1) the general synthetic control, which is based on a linear factor model, (2) mean balancing, and (3) kernel balancing. The gray bars at the bottom demonstrate the number of treated states at each time period.
**Figure A5. Election Day Registration on Voter Turnout: By EDR Year**

The panels on the left show that the treated average and counterfactual average using three different methods: (1) the general synthetic control, which is based on a linear factor model, (2) mean balancing, and (3) kernel balancing. The panels on the right show the ATT estimates. The data are from Xu (2017).

**Note:** The panels on the left show that the treated average and counterfactual average using three different methods: (1) the general synthetic control, which is based on a linear factor model, (2) mean balancing, and (3) kernel balancing. The panels on the right show the ATT estimates. The data are from Xu (2017).