Submitted to *Econometrica*

DISTRIBUTIONAL SYNTHETIC CONTROLS

F. F. GUNSILIUS
Department of Economics, University of Michigan, Ann Arbor

The method of synthetic controls is a fundamental tool for evaluating causal effects of policy changes in settings with observational data. In many settings where it is applicable, researchers want to identify causal effects of policy changes on a treated unit at an aggregate level while having access to data at a finer granularity. This article proposes an extension of the synthetic controls estimator that takes advantage of this additional structure and provides nonparametric estimates of the heterogeneity within the aggregate unit. The idea is to replicate the quantile function associated with the treated unit by a weighted average of quantile functions of the control units. This estimator relies on the same mathematical theory as the changes-in-changes estimator and can be applied in both repeated cross-sections and panel data with as little as a single pre-treatment period. It also provides a unique counterfactual quantile function for any type of distribution.

KEYWORDS: causal inference, comparative case studies, heterogeneous treatment effects, quantile functions, Wasserstein distance, synthetic controls.

1. INTRODUCTION

The method of synthetic controls, introduced in Abadie and Gardeazabal (2003) and Abadie et al. (2010), has become a main tool for estimating causal effects in comparative case studies with aggregate interventions and a limited number of large units. It is designed for settings where some units are subject to a policy intervention and others are not. The respective outcomes of interest are measured in each population before and after the policy intervention, potentially for many periods. The control units are used to account for unobserved trends in the outcome over time that are unrelated to the effect of the policy intervention. The insight is that an optimally weighted average of the available potential controls, the synthetic control unit, often provides a more appropriate comparison than a single control unit alone (Abadie, 2021).

The original method of synthetic controls is designed for settings with aggregate scalar- or vector-valued quantities where linear regression approaches are not applicable because of data limitations (Abadie, 2021). Researchers and policy makers are frequently interested in estimating the causal impacts of interventions on aggregate units while having access to data at a finer granularity, however. A classical example is assessing the effects of minimum wage policies, where the intervention is at the state level, but researchers have access to individual-level data within a state (e.g. Card and Krueger, 1994, 2000, Neumark and Wascher, 2000, Dube, 2019). These additional data could be used to estimate heterogeneous treatment effects of the causal effect of the policy change on the population within a state.[1]

We develop an extension of the synthetic controls estimator that can take advantage of the additional information in the data and estimate heterogeneous treatment effects. The idea is to replicate the quantile function associated with the treated unit by a weighted average of quantile

F. F. Gunsilius: ffg@umich.edu

I want to thank Alberto Abadie, Meng Hsuan Hsieh, Philippe Rigollet, Kaspar Wüthrich, and three anonymous referees for helpful discussions and comments, as well as Siyun He for excellent research assistance. Support through a MITRE research award from the University of Michigan is gratefully acknowledged. All errors are mine.

[1] An example for the relevance of estimating heterogeneous treatment effects is Ropponen (2011), who utilizes the additional data structure by applying the changes-in-changes estimator by Athey and Imbens (2006) to estimate the heterogeneous treatment effects of minimum wage changes on employment levels. This allows the author to disentangle estimates of aggregate causal effects in Card and Krueger (1994) and Neumark and Wascher (2000).

functions of the control units, and to use this weighted average to construct the counterfactual quantile function of the treated unit had it not received treatment. From the quantile function one can obtain other quantities of interest such as Lorenz curves (Gastwirth, 1971) or interquartile ranges. We also provide a permutation test analogous to the one developed in Abadie et al. (2010) for the classical synthetic control method.

The proposed method provides a synthetic control unit at the level of interest, that is, the aggregate level. For every control unit, it finds an optimal weight for the entire quantile function. This setting is of practical relevance, as the method can be applied in panel-data settings where individuals cannot be followed over time. The goal is to identify the shape of the counterfactual distribution, not treatment effects for individuals within an aggregate unit. By contrast, linear point-wise approaches like Chen (2020), or approaches that (i) decompose distributions into bins and (ii) match these bins between the different distributions, are local: they obtain different weights for each quantile and hence obtain weight functions for each individual unit instead of one set of weights at the aggregate level. As a result, these synthetic controls methods are sensitive to the choice of quantiles or bins; in particular, they require the assumption that the optimal weights for each point on the quantile curve or in each bin are the same or at least similar within a given state (e.g. Assumption 1 (ii) in Chen (2020)), something that can be difficult to satisfy in practice.

Focusing on replicating quantile functions allows for the replication of the support of the distribution of the treated unit, even in settings where the supports are non-nested. This allows for interpolation, analogous to the classical method of synthetic controls.[2] It also guarantees a unique counterfactual quantile function, irrespective of whether the distributions are continuous, discrete, or mixed.

2. A SYNTHETIC CONTROLS ESTIMATOR USING QUANTILE FUNCTIONS

The setup and notation for the proposed method are analogous to the classical synthetic controls approach (Abadie and Gardeazabal 2003, Abadie, Diamond, and Hainmueller 2010, Abadie 2021).

We have data on a set of $J+1$ units, where the first unit $j=1$ is the treated unit and $j=2,\ldots,J+1$ are the potential control units. These units are observed over T time periods, where $T_0 < T$ denotes the last time period prior to the treatment intervention in unit $j=1$. We call $t \leq T_0$ the pre-intervention- or pre-treatment periods and $t > T_0$ the post-intervention- or post-treatment periods. In the following, we assume there is only one post-intervention period, i.e. $T = T_0 + 1$, as the extension to several post-treatment periods is straightforward.

2.1. *The causal model in the classical setting*

The classical method of synthetic controls focuses on an aggregated outcome Y_{jt} that is observed for each unit $j=1,\ldots,J$ over the time periods $t=1,\ldots,T$. We denote by $Y_{jt,N}$ the outcome of group j that would have been observed at time t in the absence of the intervention; analogously, we denote by $Y_{jt,I}$ the outcome of group j that would have been observed at time t if the unit was exposed to the treatment at time $t > T_0$. The standard assumption in this setting is that the intervention has no effect on the outcome before the implementation period, so that we have $Y_{jt,N} = Y_{jt,I}$ for all units j and all pre-intervention periods $t \leq T_0$.

The key quantity to estimate in the classical setting is $Y_{1t,N}$, the outcome of the treatment unit had it not received the treatment in the post-intervention periods. Based on this, one defines

[2] On the flip side, this also means that it can be susceptible to interpolation bias, just as the classical method.

the effect $\alpha_{jt} = Y_{jt,I} - Y_{jt,N}$ of the intervention for unit j at time t, so that one can write the observable outcome in terms of the counterfactual notation as $Y_{jt} = Y_{jt,N} + \alpha_{jt} D_{jt}$, where $D_{jt} = 1$ if $j = 1$ and $t > T_0$ and 0 otherwise.

The goal in the synthetic controls literature is to estimate the treatment effect on the treated group in the post-treatment period, i.e. $\alpha_{1t} = Y_{1t,I} - Y_{1t,N} = Y_{1t} - Y_{1t,N}$ for $t > T_0$. The fact that $Y_{jt,N} = Y_{jt}$ for $t > T_0$ and $j = 2, \dots, J+1$ implies the treatment effect on the treated unit can be estimated by a weighted average

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} \lambda_j^* Y_{jt} = Y_{1t,I} - \sum_{j=2}^{J+1} \lambda_j^* Y_{jt,N},$$

where $\{\lambda_j^*\}_{j=2,\dots,J}$ is an optimal set of weights.

2.2. *The causal model in the distributional setting*

The distributional setting is analogous to the classical setting, but with the quantile function $F^{-1}_{Y_{jt}}$ of Y_{jt} as the quantity of interest. The quantile function is defined as

$$F^{-1}(q) \coloneqq \inf_{y \in \mathbb{R}} \{F(y) \geq q\}, \quad q \in (0,1),$$

where $F(y)$ is the corresponding cumulative distribution function. A common setting for this is if the researcher is interested in a policy change at an aggregated level and has access to data at a finer granularity. An example is data on household income when the policy change of interest is at the state-level, such as in analyses of changes in the minimum wage (e.g. Dube, 2019). In this case, it is possible to estimate the quantile function of household income within each state.

The goal is to estimate the counterfactual quantile function $F^{-1}_{Y_{1t,N}}$ of the treated unit had it not received treatment by an optimally weighted average of the control quantile functions $F^{-1}_{Y_{jt}}$, $j = 2, \dots, J+1$, i.e.,

$$F^{-1}_{Y_{1t,N}}(q) = \sum_{j=2}^{J+1} \lambda_j^* F^{-1}_{Y_{jt}}(q) \quad \text{for all } q \in (0,1).$$

The assumptions on the model are hence different from classical assumptions in nonlinear econometric panel data: for each time period t and unit j, the counterfactual measures $P_{Y_{jt,N}}$ are generated from a latent distribution $P_{U_{jt}}$, which evolves over time to account for trends. We show in Appendix A that the proposed synthetic controls method identifies the correct counterfactual distribution if these latent dynamics of $P_{U_{jt}}$ as well as the connection between the counterfactual measures $P_{Y_{jt,N}}$ and the latent measures $P_{U_{jt}}$ are linear.

At an abstract level, this model can therefore be written as

$$P_{Y_{jt,N}} = h_t \sharp P_{U_{jt}} \quad \text{for} \quad P_{U_{jt}} = g_t \sharp P_{U_{j(t-1)}},$$

where $P_{Y_{jt}}$ is the probability measure corresponding to the probability distribution $F_{Y_{jt}}$, $h_t(u) \equiv \alpha_t + \beta_t \cdot u$ and $g_t(u) \equiv \gamma_t + \delta u$ are linear functions, and $h_t \sharp P_{U_{jt}}$ and $g_t \sharp P_{U_{jt}}$ denote the pushforward measure of $P_{U_{jt}}$ via h_t and g_t.[3] Note that the measures are defined at the aggregate level j and do not need individual-level data to be defined in principle.

[3]The pushforward measure of P_{U_j} via h_t is defined as $P_{Y_{jt,N}}(A) = P_{U_j}(h_t^{-1}(A))$ for all (Borel-) sets A, where $h^{-1}(A)$ denotes the pre-image of the function h.

In practice, the counterfactual measures $P_{Y_{jt,N}}$ and $P_{Y_{jt,I}}$ are usually generated based on data available at a finer granularity. We model this by assuming we observe independent and identically distributed individual draws $Y_{ijt,N}$ for each unit j. This leads to the causal model

$$Y_{ijt,N} = \alpha_t + \beta_t U_{ijt}, \quad \text{for } U_{ijt} = \gamma_t + \delta_t U_{ij(t-1)}, \tag{1}$$

where for each $j = 1, \ldots, J+1$ and $t \leq T$, $U_{ij(t-1)}$ are independent and identically distributed draws from the unobservable distribution $F_{U_{jt}}$.[4]

This causal model is different from classical nonlinear panel data models in that it shifts the focus from the individual level i to the aggregate level j. In particular, we do not assume that the functions $h(t, \cdot)$ are strictly increasing and continuous in the unobservable U. Therefore, the proposed method does not identify individual effects at the level i within a quantile function, but only identifies the entire counterfactual distribution $P_{Y_{1t,N}}$ via its quantile function.

Having this causal model also means that the method is applicable panel data where the individuals i can change over time within a unit j. In a stylized schooling example, where the unobservable variable is ability and the outcome of interest is wages earned after graduation, this means that we can can allow for observing different students within a class over time. The causal effect of interest is at the classroom level: we identify the counterfactual quantile function of outcomes in each classroom j, but do not claim to identify causal effects for a specific individual i within a classroom.

3. IMPLEMENTATION OF THE METHOD

The optimal weights $\vec{\lambda}_t^* \in \Delta^J$ in every time period $t \leq T_0$ in the synthetic controls estimator are obtained in such a way that the corresponding weighted average of quantile functions of the control units is "as close as possible" to the treated unit. To quantify this mathematically, we introduce a distance on the set of all quantile functions. We choose the 2-Wasserstein distance[5] (Villani, 2003, section 2.2) for this purpose because it reduces the problem of finding the optimal weights λ^* to a simple regression problem. The 2-Wasserstein distance, denoted $W_2(P_1, P_2)$, between two probability measures P_1 and P_2 with finite second moments is defined as (Villani, 2003, Theorem 2.18)

$$W_2(P_1, P_2) = \left(\int_0^1 \left| F_1^{-1}(q) - F_2^{-1}(q) \right|^2 dq \right)^{1/2},$$

where F_1^{-1} and F_2^{-1} are the quantile functions corresponding to P_1 and P_2, respectively.

3.1. *Details of the implementation*

To obtain the optimal weights $\vec{\lambda}_t^* \in \Delta^J$ in each pre-intervention period $t \leq T_0$, we compute

$$\vec{\lambda}_t^* = \underset{\vec{\lambda} \in \Delta^J}{\operatorname{argmin}} \int_0^1 \left| \sum_{j=2}^{J+1} \lambda_j F_{Y_{jt}}^{-1}(q) - F_{Y_{1t}}^{-1}(q) \right|^2 dq. \tag{2}$$

[4] This model only introduces explicit randomness at the individual level i, not the aggregate j. One could introduce randomness at the aggregate level by assuming that there exists random measures $\tilde{P}(\omega)$ on some probability space which in expectation coincide with the measures $P_{Y_{jt}}$. However, for practical purposes these two forms of randomness cannot be distinguished, so that we only model randomness at the level i and via the fact that the $P_{U_{jt}}$ are unobserved and unspecified.

[5] Also called Mallows distance, Monge-Kantorovich distance, or earth mover's distance.

Mathematically, the weighted sum of the quantile functions is a barycenter (Agueh and Carlier, 2011), or Fréchet mean (Fréchet, 1948), in the 2-Wasserstein space. Therefore, the method formally consists of finding the optimal weights $\vec{\lambda}^* \equiv (\lambda_2^*, \ldots, \lambda_{J+1}^*)$ such that the corresponding barycenter $\sum_{j=2}^{J+1} \lambda_j^* F_{Y_{jt}}^{-1}(q)$ is as close as possible to the target $F_{Y_{1t}}^{-1}$ in the 2-Wasserstein space. One could also replace the unit simplex Δ^J by the set $\Sigma^{J-1} = \left\{ \vec{\lambda} \in \mathbb{R}^J : \vec{\lambda}^\top \vec{1} = 1 \right\}$ of all weights that sum to unity.[6] This would allow for extrapolation beyond the unit simplex, just as in the classical method (Abadie et al., 2015), and would not change the implementation of the method.

The optimization (2) is a convex problem for the weights $\vec{\lambda}_t^*$ with a unique solution. In practice, one can approximate the integral by randomly sampling a large number M of independent draws $\{V_m\}_{m=1}^M$ from the uniform distribution on the unit interval $V_m \sim U[0,1]$ and solving

$$\vec{\lambda}_t^* = \underset{\vec{\lambda} \in \Delta^J}{\operatorname{argmin}} \frac{1}{M} \sum_{m=1}^{M} \left| \sum_{j=2}^{J+1} \lambda_{jt} F_{Y_{jt}}^{-1}(V_m) - F_{Y_{1t}}^{-1}(V_m) \right|^2 .$$

So if the quantile functions $F_{Y_{jt}}^{-1}$ are known, one can construct an artificial sample $\tilde{Y}_{jtm} = F_{Y_{jt}}^{-1}(V_m)$ indexed by the sample size m chosen by the researcher. One can then write the last expression as a linear regression constrained to the unit simplex, i.e.

$$\vec{\lambda}_t^* = \underset{\vec{\lambda} \in \Delta^J}{\operatorname{argmin}} \frac{1}{M} \sum_{m=1}^{M} \left| \sum_{j=2}^{J+1} \lambda_{jt} \tilde{Y}_{jtm} - \tilde{Y}_{1tm} \right|^2 = \underset{\vec{\lambda} \in \Delta^J}{\operatorname{argmin}} \| \tilde{\mathbb{Y}}_t \vec{\lambda}_t - \vec{\tilde{Y}}_{1t} \|_2^2, \tag{3}$$

where $\tilde{\mathbb{Y}}_t$ is the $m \times J$-matrix with entry $\tilde{Y}_{jtm}$ at position (m, j), $\vec{\tilde{Y}}_{1t}$ is the vector of elements $\tilde{Y}_{1tm}$ for $m = 1, \ldots, M$, and $\| \cdot \|_2$ is the Euclidean norm on $\mathbb{R}^M$. Since we approximate the integral by simulations from a uniform distribution we can make the approximation error as small as desired by choosing a sufficiently large number of simulation samples. By the convexity and continuity of the objective function in $\vec{\lambda}$, the optimal weights obtained from the approximation converge to the optimal weights of the integral expression as $M \to \infty$ (Newey and McFadden, 1994).

In practice, the quantile functions $F_{Y_{jt}}^{-1}$ are not known and have to be estimated from the data via empirical quantile functions $\hat{F}_{Y_{jtn}}^{-1}(q)$ of the sample $\{Y_{ijt}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, J+1$. One way to do this is via order statistics: $\hat{F}_{Y_{jtn}}^{-1}(q) = Y_{n(k)}$ where k is chosen such that $(k-1)/n < q < k/n$ and $Y_{n(k)}$ are the order statistics of the data sample $\{Y_{ijt}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, J+1$.

This approach works for any type of distributions, regardless of whether they are absolutely continuous, discrete, or mixed, as long as they have finite second moments—otherwise the Wasserstein distance can be infinite and the problem becomes trivial.[7] We can then compute the optimal weights $\vec{\lambda}^*$ as a weighted average of the weights $\vec{\lambda}_t^*$ over all pre-intervention periods, i.e.

$$\vec{\lambda}^* = \sum_{t \le T_0} w_t \vec{\lambda}_t^*, \qquad \text{for } w_t \ge 0 \text{ and } \sum_{t \le T_0} w_t = 1.$$

[6] $\vec{1} = \{1, 1, \ldots, 1\} \in \mathbb{R}^J$ denotes the unit vector and $a^\top b$ denotes the inner product in $\mathbb{R}^J$.

[7] All results hold for the p-Wasserstein distance, $p \ge 1$. For $p = 1$ we only need to require that all distributions have a finite first moment. We focus on $p = 2$ because it provides a regression approach.

Arkhangelsky et al. (2021) provide potential choices of weights w_t that can also be used in our case. In our simulations and applications, equal weighs $w_t = \frac{1}{T_0}$ perform well, see Section 5.

At every time point $t > T_0$ in the post-intervention period, we compute the counterfactual quantile function for the treatment unit had it not received the treatment by $F^{-1}_{Y_{1t,N}} = \sum_{j=2}^{J+1} \lambda_j^* F^{-1}_{Y_{jt}}$.

3.2. *Summary of the proposed method*

In summary, the abstract procedure for a data-generating process of the form Y_{ijt}, $i = 1, \ldots, n_j$, $j = 1, \ldots, J+1$, $t = 1, \ldots, T$ is as follows.

Algorithm 1 Distributional synthetic controls

Input: 1. data-generating process Y_{ijt} with $i = 1, \ldots, n_j$, $j = 1, \ldots, J+1$, $t = 1, \ldots, T$
2. weights $\{w_t\}_{t \leq T_0} \subset \Delta^{T_0}$

procedure DSC
  for each time period $t \leq T$ **do**
    for each unit $\iota = 1, \ldots, J+1$ **do**
      estimate the empirical quantile functions $\hat{F}^{-1}_{Y_{\iota tn}}$
    end for
  end for
  for each time period $t \leq T_0$ **do**
    obtain the optimal weights $\vec{\lambda}_t^*$ by solving (2) via the regression (3)
  end for
  obtain the optimal weights $\vec{\lambda}^* = \sum_{t=1}^{T_0} w_t \vec{\lambda}_t^*$ over all $t \leq T_0$
  for each time period $t = T_0 + 1, \ldots, T$ **do**
    obtain the counterfactual quantile function $\hat{F}^{-1}_{Y_{1nt,N}} = \sum_{j=2}^{J+1} \lambda_j^* \hat{F}^{-1}_{Y_{jnt}}$
  end for
end procedure

From the estimated counterfactual quantile function one can obtain other quantities, such as averages $E[Y_{1t,N}] = \int_0^1 F^{-1}_{Y_{1t,N}}(q)dq$, counterfactual Lorenz curves (Gastwirth, 1971), Gini-coefficients, or interquartile ranges. Mirroring the classical setting (Abadie et al., 2010, equation (2)), the proposed method identifies the true counterfactual distribution for the causal model in (1) in the case where there exists a set of weights $\vec{\lambda}^*$, which allows us to perfectly replicate the target in each time period, i.e. $F^{-1}_{Y_{1t}} = \sum_{j=2}^{J+1} \lambda_j^* F^{-1}_{Y_{jt}}$ for all $t \leq T_0$.

3.3. *A placebo permutation test*

In order to perform inference on the estimated causal effect we use a placebo permutation test as in the classical setting (Abadie et al., 2010). The idea is to apply the synthetic controls estimator to each control unit by pretending this control unit is the treated one. If there is an actual treatment effect only in the treatment group post-intervention, then the estimated effect for the actual treatment unit should be among the most extreme. Algorithm 2 provides the pseudo-code for the placebo permutation test. The obtained probability p_t provides the probability of observing a difference between the observable $F^{-1}_{Y_{1t}}$ and the estimated counterfactual $F^{-1}_{Y_{1t,N}}$ given all permutations of the treatment and control units.

Algorithm 2 Placebo test

Input: 1. quantile functions $F^{-1}_{Y_{jt}}$, $j = 1, \ldots, J+1$, $y = 1, \ldots, T$
2. weights $\{w_t\}_{t \leq T_0} \subset \Delta^{T_0}$

procedure PERMUTATION INFERENCE FOR CAUSAL EFFECTS AT TIMES $t \geq T_0$
  for each unit $\iota = 1, \ldots, J+1$ **do**
    for each time period $t = 1, \ldots, T_0$ **do**
      obtain and record the optimal weights $\lambda^*_{st,\iota}$ using (2)
    end for
    compute the overall optimal weights $\lambda^*_{s,\iota} = \sum_{t \leq T_0} w_t \lambda^*_{st,\iota}$
    for each time period $t = 1, \ldots, T$ **do**
      compute 2-Wasserstein barycenter using the weights $\lambda^*_{s,\iota}$ to obtain $F^{-1}_{Y_{\iota t,N}}$
      record the squared distances $d^2_{\iota t} := \int_0^1 \left| F^{-1}_{Y_{\iota t,N}}(q) - F^{-1}_{Y_{\iota t}}(q) \right|^2 dq$
    end for
  end for
  for each time period $t = 1, \ldots, T$ **do**
    sort $d_{\iota t}$ decreasingly in ι
    record the rank $r(d_{1t})$, e.g., $r(d_{1t}) = 1$ if d_{1t} is largest
    compute the probability p_t of obtaining a value of d_{1t} as $p_t = \frac{r(d_{1t})}{J+1}$
  end for
end procedure

The difference compared to the classical method is that our outcome of interest, $F^{-1}_{Y_{1t,N}}$, is a functional quantity. Thus, in contrast to the classical setting, we use the 2-Wasserstein distances $d_{\iota t}$ to rank the difference between the observed $F_{Y_{\iota t}}$ and the computed $F_{Y_{\iota t,N}}$. These distances are always non-negative, in contrast to the classical setting (Abadie et al., 2010).

4. COMPARISON TO OTHER APPROACHES

By using quantile functions, the proposed method becomes the an extension of the classical synthetic controls estimator in two ways: first, it can replicate target distributions whose support is not nested in the control distributions, hence allowing for interpolation, as the classical method of synthetic controls; second, it rests on the same mathematical foundation, the p-Wasserstein space, as the changes-in-changes estimator by Athey and Imbens (2006). This makes it the complementary approach to the changes-in-changes estimator in the same way that the synthetic controls estimator is a complementary approach to the classical difference-in-differences estimator. In this section, we formalize these connections. We also introduce an alternative approach that uses averages of cumulative distribution functions and compare the proposed method to it.

4.1. *Relation to the classical synthetic controls method*

The proposed method is an extension of the classical method in a rigorous sense: if we apply it to probability measures supported on one point, i.e., Dirac measures of the form $\delta_y(A)$, taking the value 1 if $y \in A$ and 0 otherwise, then we obtain the same results as the classical method.[8]

[8]To see this note that the quantile function $F_y^{-1}(q)$ corresponding to a Dirac measure $\delta_y(A)$ at a point y is the constant function with value y.

This means the proposed method reduces to the classical estimator when we are only given aggregate values and not distributions.

Note that this does not imply a relation between the proposed method and the classical method applied to moments of a given distribution, such as averages. The proposed estimator provides different optimal weights than the classical estimator applied to averages. Intuitively, the reason for this is that the distributional synthetic controls method finds optimal weights that replicate all moments of the target distribution as closely as possible. In contrast, applying the classical method to averages of the distributions will find the optimal weights based on replicating the first moment. It follows from Jensen's inequality that the optimal weights can replicate the average at least as well as whole distributions.

4.2. *Relation to the changes-in-changes estimator*

The proposed method is the complementary approach to the changes-in-changes estimator in the sense that the latter is also based on the p-Wasserstein space. Recall the approach of the changes-in-changes estimator: in the case of only one pre- and one post-intervention period, as well as only one control group, it constructs the counterfactual distribution $F_{Y_{11,N}}$ as $F_{Y_{10}}(F_{Y_{00}}^{-1}(F_{Y_{01}}))$, where the first index is with respect to the treatment group (0 for control and 1 for treatment) and the second index is with respect to the time period (0 is pre-intervention and 1 is post-intervention).

The monotone rearrangement $F_{Y_{00}}^{-1}(F_{Y_{01}})$ is the optimal transport map between F_{00} and F_{01} with respect to the p-Wasserstein distance. In particular, when $F_{Y_{01}}$ is absolutely continuous, one can compute the p-Wasserstein distance between $F_{Y_{01}}$ and $F_{Y_{00}}$ as (Villani, 2003, Remark 2.19 (iv))

$$W_2(P_{Y_{01}}, P_{Y_{00}}) = \left(\int \left| x - F_{Y_{00}}^{-1}(F_{Y_{01}}(x)) \right|^2 dF_{Y_{01}}(x) \right)^{1/2}.$$

Hence the changes-in-changes estimator relies on the same mathematical theory as our proposed method in the form of the p-Wasserstein distance.

Both approaches are complementary and suited for different settings. In a nutshell: the proposed distributional synthetic controls method estimates between units and extrapolates over time, while the changes-in-changes method estimates changes over time and extrapolates between units. This implies different assumptions for each method. The proposed method requires a linear function h in the causal model (1) due to the extrapolation over time, while the changes-in-changes estimator requires that h is strictly increasing and continuous in the unobservable. This fact makes the changes-in-changes estimator invariant to nonlinear increasing transformations. This does not hold for the proposed synthetic controls estimator: if the distributions are transformed in a nonlinear fashion, then the optimal weights obtained will in general be different. On the other hand, synthetic control methods are designed for settings where several control units are available over potentially multiple time periods. It also provides point-identification in the sense of a unique counterfactual distribution for any type of distribution.

4.3. *A synthetic controls method based on cumulative distribution functions*

An alternative for estimating the optimal weights $\vec{\lambda}_t^*$ in time period t is a mixture of distribution functions:

$$\vec{\lambda}_t^* = \underset{\vec{\lambda} \in \Delta^J}{\operatorname{argmin}} \int_{\mathbb{R}} \left| \sum_{j=2}^{J+1} \lambda_j F_{Y_{jt}}(y) - F_{Y_{1t}}(y) \right| dy. \tag{4}$$

The L^1-distance is chosen because it is the 1-Wasserstein distance between the weighted average of the distribution functions of the control units and the distribution function of the target (Villani, 2003, section 2.2). This is not the case for any other $p > 1$, which are defined via quantile functions. In principle, one could use any other distance or divergence as well, but Wasserstein distances have the benefit that they are applicable for measures with potentially different supports. This can be relevant in finite sample settings and in settings where supports are non-nested.

The main difference between a method based on mixtures of quantiles in contrast to mixtures of distribution functions is that the former replicates the target also at points that lie outside of the support of the control distributions, while the latter is confined to the union of the supports of the control distributions. However, solving (4) is preferred if it is known that the distributions are mixtures of other distributions themselves. An example for this is a setting where the treatment group consists of outcomes of both men and women, and the two control units consist of men only and women only. In this setting, a mixture of distributions is sensible. From a computational perspective, the mixture of quantiles approach relies on the 2-Wasserstein distance, which allows for a practical implementation via linear regression, as shown in (3), while the analogous method using mixture of distributions relies on the L^1-distance and is implemented via a convex optimization routine.

Figure 1 illustrates the differences between the two concepts in a theoretical setting with four Gaussians of equal variance. The equally weighted average of the quantile functions is the Gaussian with the same variance as the control distributions and a mean which is the average of the means of the control distributions. The average of the equally weighted distribution function (blue) is the multimodal function supported on the joint support of the control units.

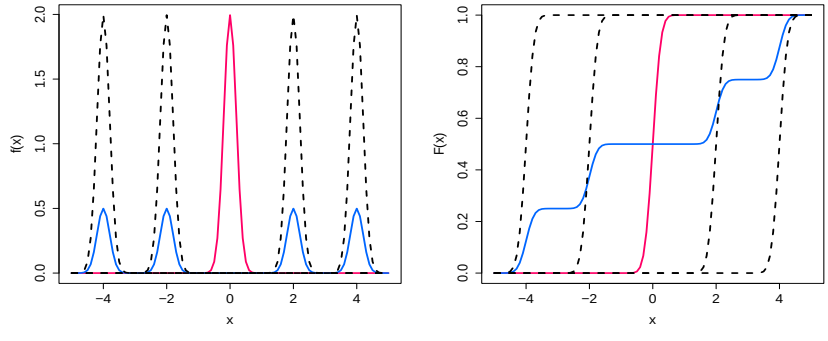

FIGURE 1.—Four Gaussian distributions with equal variance (dashed), the equally weighted average of their quantiles (magenta), and their equally weighted linear mixture (blue). Left: densities, right: cumulative distribution functions.

5. EMPIRICAL APPLICATION

This section illustrates the performance of the proposed method (2) with economic data. We also apply the approach based on mixtures of distribution functions (4) for comparison. We use a subset of the data provided in Dube (2019) on minimum wages in the United States. The data consist of all 50 states and the District of Columbia. The outcome of interest is the distribution of equalized family income from wages and salary, defined as multiples of the federal poverty

threshold as in Dube (2019). We focus on the years $1998 - 2004$ and the state of Alaska as our treated unit, because Alaska increased its minimum wage from $\$\,5.65$ to $\$\,7.15$ in 2003. The 33 other states that did not change their nominal minimum wage during this period are the control units, and $T_0 = 2002$.

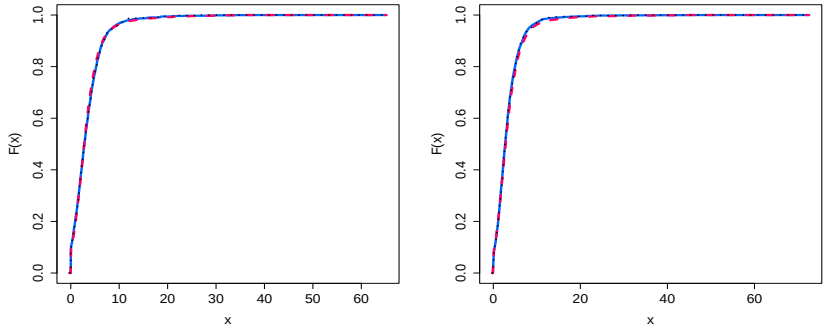

FIGURE 2.—Depiction of the proposed estimator (2) (magenta dashed) and the mixture of distributions estimator (4) (blue) for replicating the target distribution (black dashed) of adjusted family income in 1999 (left) and 2003 (right) in AK.

Figure 2 captures the performance of our proposed method (2) and the mixture of distribution function (4) for replicating the target distribution during the years 1999 and 2003 using an equally weighted average of the optimal weights $\vec{\lambda}^* := \frac{1}{2002-1998} \sum_{t=1998}^{2002} \lambda_t^*$.

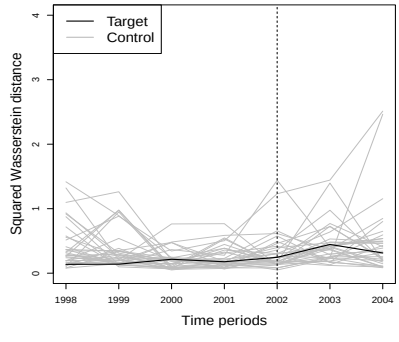

FIGURE 3.—Placebo permutation test following Algorithm 2 with target AK.

Both methods manage to replicate the target well. The reason is that the distribution functions of all units are regular with similar supports and shapes, with very little variation between them. The weights obtained in each approach differ, but are qualitatively similar. They are sparse in both approaches, with around 15 - 20% of control units receiving a weight greater than 1% and about 10% of the control units receiving a weight of greater than 10% in each time period. Taking the uniform average of these weights over all pre-intervention time periods removes the sparsity of the weights, with only one unit receiving more than 10% of the weight for

each method (North Dakota for the proposed method and Ohio for the mixture of distributions) and about $2/3$ of all units receiving a weight of 1% or more. The five control states with the largest weights for the proposed method are: ND (0.11), VA (0.09), NH (0.08), MD (0.08), NE (0.08). The five control states with the largest weights for the method relying on the mixture of distributions are: OH (0.11), MI (0.08), NE (0.08), MD (0.07), IA (0.07). This demonstrates that the two approaches do give different weights in general, which is expected since both define a different metric and definition of average on the space on the space of probability measures in general.

It is also possible to estimate confidence intervals for the distribution functions based on standard resampling techniques. In this example, the 99%-confidence intervals for our proposed method are so tight that they are imperceptible from the replication, so they are not included.

Figure 2 indicates that there is no immediate treatment effect on the equalized family income in Alaska from raising the minimum wage in 2002. Figure 3 depicting the results of a placebo permutation test following Algorithm 2 corroborates this.

6. DISCUSSION

We have developed an extension of the classical synthetic controls estimator that can take advantage of data at a finer granularity within treatment and control units. The idea is to replicate the quantile function of the treated unit in pre-intervention periods by an optimally weighted mixture of the quantile functions of the control units. In post-intervention periods, these weights are used to construct the counterfactual quantile function of the treated unit had it not received the treatment. We also provide a complementary approach based on cumulative distribution functions using the 1-Wasserstein distance.

The proposed method provides one scalar weight for entire quantile functions, hence produces the counterfactual at the aggregate level. This allows researchers to perform causal inference on outcomes of interest beyond averages, such as Lorenz curves or interquartile ranges. Furthermore, it makes the proposed method applicable in relevant practical settings with panel data where individuals cannot be tracked over time. We also provide a placebo permutation test analogous to the classical test introduced in Abadie et al. (2010) and a linear causal model for which our estimator estimates the counterfactual quantile function.

APPENDIX A: FORMAL ARGUMENT FOR THE IDENTIFICATION OF THE CAUSAL EFFECT

This section develops a formal argument that the proposed method identifies the counterfactual quantile function in the population for the linear causal model and there exists a set of weights $\vec{\lambda}^*$ which allows us to replicate the counterfactual quantile function in all pre-intervention time periods, i.e. $F^{-1}_{Y_{1t}} = \sum_{j=2}^{J+1} F^{-1}_{Y_{jt}}$ for all $t \leq T_0$.

To see this, focus first on the case where the unobservables U_j do not change over time. Recall that the optimal weights λ^* obtained over the pre-intervention periods are used in the post-intervention period to construct the counterfactual quantile function of the treated unit had it not received treatment. This implies that the *relative distances* between the quantile functions of the unobservables and the outcomes of the control units cannot change over time. Otherwise, the weights obtained in pre-intevention periods would not be optimal in post-intervention periods. This further implies the functions $h(t, U_j)$ in the causal model need to preserve relative distances. Functions that preserve distances are called *isometries*.

DEFINITION 1: A map $f : \mathcal{X} \to \mathcal{Y}$ between to metric spaces $(\mathcal{X}, d_x)$ and $(\mathcal{Y}, d_y)$ is an *isometry* if $d_y(f(x), f(x')) = d_x(x, x')$ for all $x, x' \in \mathcal{X}$. We call $f(x)$ a *scaled isometry* if it satisfies $d_y(f(x), f(x')) = \tau d_x(x, x')$ for some $\tau \in (0, +\infty)$.

In the above definition, d_x and d_y are distances on the respective spaces. Since we can allow for functions to only preserve *relative* distances, we only need to require that the functions $h(t, U_j)$ are scaled isometries. Since we work in the 2-Wasserstein space, we require that $h(t, U_j)$ are scaled isometries in the 2-Wasserstein space and hence preserve the 2-Wasserstein distances.

We now show this formally. Without loss of generality let τ_h, the scaling parameter of the isometry $h(t, \cdot)$, be equal to unity. The reason is that the scaling does not affect the relative distance between the respective measures which is needed to determine the counterfactual quantile function. The proof is then straightforward with the definition of isometries. In particular, as $h(t, \cdot)$ is a (surjective) isometry in the 2-Wasserstein space on the real line for all t, it holds by definition that

$$W_2(P_{U_j}, P_{U_i}) = W_2(h_t \sharp P_{U_j}, h_t \sharp P_{U_i}) = W_2(P_{Y_{jt}}, P_{Y_{it}})$$

for $j, i = 1, \ldots, J + 1$. This holds for all time periods t.

This implies that the map $m_{t,t'} \coloneqq h_{t'} \circ h_t^{-1}$ is also an isometry for all t, t', as the composition of (surjective) isometries is a (surjective) isometry. But since isometries retain the weighted averages, this implies that using the weights λ_t^* obtained in the pre-intervention periods is still optimal in the post-intervention periods, so $P_{Y_{jt,N}}$ obtained by the method of distributional synthetic controls provides the correct counterfactual distribution for the model where $h(t, \cdot)$ are scaled isometries.

Linear maps of the form $h(t, U_j) = \alpha_t + \beta_t U_j$ are scaled isometries in the 2-Wasserstein space (Kloeckner, 2010). This implies that for fixed $P_{U_{jt}}$, the optimal weights λ^* obtained via the proposed method stay optimal in post-intervention periods. Therefore, the estimated counterfactual quantile function using these weights is the correct counterfactual quantile function with respect to the causal model (1).

The same argument holds when we assume that the unobservables U_{jt} change linearly over time. Since this change over time is linear the relative distances between unobservables stay the same over time. The same argument above then proves identification for this setting too.

REFERENCES

ABADIE, ALBERTO (2021): "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, 59, 391–425. [1, 2]

ABADIE, ALBERTO, ALEXIS DIAMOND, AND JENS HAINMUELLER (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, 105, 493–505. [1, 2, 6, 7, 11]

——— (2015): "Comparative politics and the synthetic control method," *American Journal of Political Science*, 59, 495–510. [5]

ABADIE, ALBERTO AND JAVIER GARDEAZABAL (2003): "The economic costs of conflict: A case study of the Basque Country," *American Economic Review*, 93, 113–132. [1, 2]

AGUEH, MARTIAL AND GUILLAUME CARLIER (2011): "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, 43, 904 – 924. [5]

ARKHANGELSKY, DMITRY, SUSAN ATHEY, DAVID A. HIRSHBERG, GUIDO W. IMBENS, AND STEFAN WAGER (2021): "Synthetic Difference-in-Differences," *American Economic Review*, 111, 4088–4118. [6]

ATHEY, SUSAN AND GUIDO W. IMBENS (2006): "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 74, 431–497. [1, 7]

CARD, DAVID AND ALAN B. KRUEGER (1994): "Minimum wages and employment: A case study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, 84, 772–793. [1]

——— (2000): "Minimum wages and employment: A case study of the Fast-Food Industry in New Jersey and Pennsylvania: reply," *American Economic Review*, 90, 1397–1420. [1]

CHEN, YI-TING (2020): "A distributional synthetic control method for policy evaluation," *Journal of Applied Econometrics*, 35, 505–525. [2]

DUBE, ARINDRAJIT (2019): "Minimum wages and the distribution of family incomes," *American Economic Journal: Applied Economics*, 11, 268–304. [1, 3, 9, 10]

FRÉCHET, MAURICE (1948): "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'institut Henri Poincaré*, 10, 215–310. [5]

GASTWIRTH, JOSEPH L. (1971): "A general definition of the Lorenz curve," *Econometrica*, 39, 1037–1039. [2, 6]

KLOECKNER, BENOÎT (2010): "A geometric study of Wasserstein spaces: Euclidean spaces," *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9, 297–323. [12]

NEUMARK, DAVID AND WILLIAM WASCHER (2000): "Minimum wages and employment: A case study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment," *American Economic Review*, 90, 1362–1396. [1]

NEWEY, WHITNEY K. AND DANIEL MCFADDEN (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245. [5]

ROPPONEN, OLLI (2011): "Reconciling the evidence of Card and Krueger (1994) and Neumark and Wascher (2000)," *Journal of Applied Econometrics*, 26, 1051–1057. [1]

VILLANI, CÉDRIC (2003): *Topics in Optimal Transportation*, Graduate Studies in Mathematics vol. 58, American Mathematical Society. [4, 8, 9]