Proximal Causal Inference on Quantile Synthteic Control

Zhenxiao Chen*

February 7, 2026

Abstract

This paper extends the synthetic control method (SCM) to a distributional setting, where the object of interest is the quantile treatment effect on the treated (QTT) at a given quantile level τ. The proposed framework provides a flexible and credible tool for policy evaluation in distributional contexts. Unlike the traditional synthetic control literature, we assume that, for each unit and time period, we have access to the entire distribution of the outcome variable of interest, where the quantiles are assumed to be driven linearly by latent common factors with unit- and quantile-specific factor loadings. The key identification strategy for the QTT at τ first defines the τ-quantile synthetic control (τ-QSC) as a weighted combination of donor quantiles that matches the factor loadings of the treated unit at τ, and then contrast the observed outcome for the τth quantile of the treated with the τ-QSC. This approach avoids the often infeasible requirement of a near-perfect pre-treatment fit and offers a more interpretable and realistic alternative to existing methods.

One challenge for this identification strategy is that the factor loadings are unobserved. We address it by leveraging tools from the proximal causal inference literature, treating quantiles from non-donor units as proxies for the latent confounders. This innovation transforms a set of infeasible moment conditions into feasible ones and enables identification of both the τ-QSC and the τ-QTT. We then develop a GMM-based estimation and inference procedure that accommodates endogenous treatment assignment and weakly dependent time series. Simulation evidence confirms the robustness and efficiency of the proposed method relative to standard SCM.

*Department of Economics, University of Pennsylvania

1 Introduction

Evaluating policy impacts is of substantial interest and practical importance for social scientists and policymakers. Such evaluations are rigorously formalized within the potential outcome framework of causal inference, in which the causal effect of a policy (or a treatment in causal inference terminology) on a target entity is defined as the difference between its observed outcome of interest and the corresponding counterfactual outcome—the outcome that would have occurred had the entity's treatment status been reversed (Rubin [1974]). In practice, many policies are implemented at an aggregate level, targeting broad units such as regions, cities, districts, or institutions rather than individuals. These policies often involve large-scale interventions—such as tax reforms, education restructuring, infrastructure investments, or public health initiatives—that affect entire populations or systems. Given their scope and potential to generate far-reaching economic and social consequences, understanding their causal impact—particularly on the treated unit itself—is of central importance to researchers and decision-makers. However, a key challenge in evaluating such policies is that they are rarely implemented through randomized experiments, due to ethical and logistical constraints. This absence of randomization complicates causal identification, as treated and control units may differ systematically in both observed and unobserved characteristics, introducing potential confounding bias. As a result, estimating the causal effect of aggregate-level policies typically necessitates the adoption of identification assumptions tailored to the aggregate nature of the intervention and the structure of the data.

Given a panel dataset in which aggregate units are observed over time, a common—but potentially problematic—identification strategy is to select a control unit that is unaffected by the policy but closely resembles the treated unit in terms of pre-treatment characteristics, and to use it as a proxy for the treated unit's counterfactual outcome in the post-treatment period. However, in aggregate-level settings, the number of available control units is often limited by design, making it difficult to find a sufficiently comparable match—particularly in the presence of unobserved confounders. This limitation can undermine the credibility of

the counterfactual estimate and, by extension, the validity of the overall policy evaluation.

The synthetic control method (SCM), developed in a series of seminal papers by Abadie and co-authors, addresses this challenge by leveraging the insight that a convex combination of multiple control units can more accurately approximate the treated unit's pre-treatment characteristics than any single control unit alone. This approach is grounded in the hypothesis that the data-generating processes of aggregate units follow a linear factor model, in which outcomes are driven by a set of unobserved, time-varying common factors—an assumption often supported by the observed co-movement in outcome trajectories across units. Within this modeling framework, the key identification assumption is that there exists a convex combination of control units—termed the synthetic control—that closely reproduces the trajectory of the treated unit over an extended pre-treatment period. The underlying intuition is that if the synthetic control can closely replicate the untreated behavior of the treated unit for a sufficiently long time prior to the intervention, it can serve as a credible proxy for the counterfactual outcome in the post-treatment period.

Praised by Susan Athey and Guido Imbens as "arguably the most important innovation in the policy evaluation literature in the last 15 years," SCM has been widely applied in empirical studies to evaluate aggregate policies (Athey and Imbens [2017]). In recent years, it has been used to examine the effects of right-to-carry laws (Donohue et al. [2017]), legalized prostitution (Cunningham and Shah [2017]), immigration policy (Bohn et al. [2014]), and other major policy issues. SCM has also served as the primary analytical tool in prominent debates on the effects of immigration (Peri and Yasenov [2015]) and minimum wages (Allegretto et al. [2018], Jardim et al. [2017], Neumark and Shirley [2021], Reich et al. [2017]). Beyond economics, synthetic controls have been adopted in political science (Abadie [2015]) and environmental science (Serra-Burriel et al. [2021]).

Alongside its widespread empirical applications, recent years have also witnessed important theoretical advancements extending the SCM beyond its original framework. Traditionally designed for panel data structures at the aggregate level (e.g., states or countries),

SCM has now been adapted to more complex and granular settings. For instance, when data are disaggregated, Abadie and L'Hour [2021] addresses the issue of non-uniqueness in constructing synthetic controls by introducing a penalized estimator. This approach minimizes discrepancies not only between the treated unit and its synthetic counterpart but also among the contributing donor units, thereby improving robustness and interpretability in settings with high dimensionality. In parallel, researchers have increasingly sought to evaluate policy impacts not just on mean outcomes but across the entire distribution of outcomes, particularly when micro-level data within each aggregate unit are available. In such cases, SCM has been extended to a few distributional frameworks. Notably, Gunsilius [2023] and Chen [2020] develop methodologies that allow researchers to estimate the (partial) counterfactual distribution of outcomes, enabling a more nuanced assessment of heterogeneity in treatment effects. These developments are particularly relevant for policy evaluations where understanding the distributional consequences—such as inequality, tail risk, or quantile-specific impacts—is as important as understanding average effects.

This project focuses on a relevant extension of SCM in a specific distributional setting previously studied by Chen [2020]. The distributional setting considered here assumes that the quantiles of outcome distributions across units follow a linear factor model, driven by common time-varying latent factors and characterized by unit- and quantile-specific factor loadings. To distinguish this particular setting from other distributional settings in the SCM literature, I refer to it as the quantile synthetic control framework. The primary object of interest is the quantile treatment effect on the treated (QTT) at some quantile τ of interest, and the project aims to construct a coherent framework for its identification, estimation, and inference.

Several considerations motivate this effort. First, the existing identification strategy for QTT within the quantile synthetic control framework relies on assumptions that are excessively stringent. In particular, the leading approach requires that the target quantile of the treated unit be closely—if not perfectly—approximated by the corresponding quantile

synthetic control (QSC) at the target quantile level τ over the entire pre-treatment period. In practice, achieving this level of precision is rarely feasible, especially in settings with long pre-treatment histories or when the treated unit displays distinctive, idiosyncratic dynamics that are difficult to replicate. Moreover, the notion of a "near-perfect fit"—whether in conventional or distributional SCM frameworks—is heuristic and lacks a formal, widely accepted definition. There is no established criterion for determining what constitutes a sufficiently close match between the treated unit and its synthetic control, which introduces ambiguity into the identification strategy and undermines the interpretability and robustness of empirical findings.

Second, inference procedures in existing distributional SCM applications typically rely on placebo tests, which are only valid under strong assumptions—most notably, random treatment assignment or full exchangeability of units. These assumptions are often difficult to justify in real-world policy settings, where interventions are rarely randomized and treated units frequently differ systematically from their untreated counterparts. As noted by Hahn and Shi [2017] and Ferman and Pinto [2017], even in the conventional SCM framework, achieving exchangeability effectively requires strong normality assumptions on the error terms, and placebo tests are prone to size distortions when treatment assignment is nonrandom. These concerns naturally extend to distributional settings, and specfically, to the quantile synthetic control framework. As a result, inference on the distributional effects of a policy at target quantiles based on placebo tests may yield misleading conclusions—especially when treatment assignment is endogenous or when treated and control units are not directly comparable—underscoring the need for alternative, more robust inference procedures.

This project seeks to address these gaps by proposing identification conditions for the QTT that are both more realistic and empirically relevant compared to those in the existing literature, and by introducing inference methods that are robust to violations of random assignment and exchangeability. These contributions aim to expand the applicability of distributional SCM in empirical research while strengthening its theoretical foundation. Specif-

ically, the project defines the QSC by matching on the unobserved factor loadings associated with the target quantile of the treated unit, in the spirit of Ferman and Pinto [2021]. This approach relaxes the requirement that the QSC must closely replicate the corresponding quantile trajectory of the treated unit during the pre-treatment period—an often unrealistic condition in practice. However, this relaxation introduces new challenges. In contrast to Chen [2020], identification in the quantile synthetic control framework under this approach is no longer straightforward, as it relies on unobserved components of the model. To address this issue, the project draws on techniques from the proximal causal inference literature. As suggested by Shi et al. [2021], under the assumed factor structure, quantiles can be interpreted as noisy proxies for latent factors, which represent the principal source of confounding. The partial information embedded in the quantiles of units not used to construct the QSC can potentially be leveraged to facilitate identification of both the QSC and the QTT, by transforming an otherwise infeasible set of moment conditions into a feasible one. Estimation of the QSC and QTT, as well as inference for the latter, then proceeds naturally via the Generalized Method of Moments (GMM), under standard time-series assumptions on the quantile processes.

The remainder of the paper is organized as follows. Section 2 reviews the literature relevant to this study. Section 3 introduces the unconventional data structure and presents the quantile synthetic control framework. Section 4 outlines the identification strategy by leveraging proxies in the framework. Section 5 discusses the estimation and inference procedures inspired by GMM.

2 Literature Review

This paper contributes to several strands of literature. First and foremost, it falls within the broader literature on the identification strategies underpinning synthetic control methods. Broadly speaking, two main identification approaches have emerged in this literature. The

first, formalized in the seminal work of Abadie et al. [2010], requires that the synthetic control unit closely mimics the behavior of the treated unit over the entire pre-treatment period. In this framework, the synthetic control is constructed as a convex combination of control units. Under regularity conditions imposed on a factor model, it can then be shown that the average treatment effect on the treated (ATT) is identified as the number of pre-treatment periods tends to infinity.

The second identification strategy, first introduced by Ferman and Pinto [2021], defines the synthetic control as a weighted combination of control units whose aggregate factor loadings match those of the treated unit. This project adopts the perspective of Ferman and Pinto [2021] and extends their definition of a synthetic control unit to a distributional setting, which we refer to as the quantile synthetic control framework. One advantage of this definition is that it avoids the need to specify an ad hoc criterion for what constitutes a "near-perfect" match, making it conceptually more natural.

Moreover, the identification condition in Abadie et al. [2010]—that the number of pre-treatment periods grows large—can be interpreted as implicitly requiring the synthetic control to match the treated unit's factor loadings. This makes the Ferman and Pinto [2021] condition strictly weaker, while sidestepping the need to rely on asymptotic identification as the number of time periods increases. However, a limitation of the Ferman and Pinto [2021] approach is that it does not establish direct identification of the synthetic control itself.

This paper addresses that gap by leveraging techniques from the proximal inference literature to augment the factor loading matching condition and formally establish identification of the synthetic control in the distributional setting. Once the QSC is identified, it follows immediately that the QTT is also identified.

Second, this paper contributes to the subfield of generalized synthetic control methods in distributional settings. Two key papers have pioneered this area. Gunsilius [2023] constructs the entire counterfactual distribution for the treated unit by leveraging the distributions of control units. Their identification strategy requires that, in the pre-treatment period, the

distribution of the treated unit is matched by a convex combination of the distributions of selected control units. This approach is referred to as the distributional synthetic control framework.

In contrast, Chen [2020] takes a more parametric route by imposing a factor structure on the quantiles of each unit at each time point. Under this framework, the QSC for the treated unit at a given quantile τ is identified as a weighted combination of the $\tau-$quantiles of selected controls that matches the $\tau-$quantile of the treated unit throughout the pre-treatment period. This methodology is referred to as the quantile synthetic control framework.

This project builds on the framework introduced by Chen [2020], as it targets the QTT at specific quantiles—rather than the full distribution—making the QSC approach more flexible. Unlike the distribution-matching approach in Gunsilius [2023], this framework does not require full distributional alignment, which can be unnecessarily restrictive when the policy effect is only of interest at certain points in the outcome distribution.

An additional advantage of the Chen [2020] framework is that the imposed structure facilitates the use of standard estimation and inference procedures. As a result, it avoids reliance on placebo-based inference, which has been shown to be problematic when treatment assignment is nonrandom. The contribution of this paper is to further relax the identification condition in the QSC framework by adopting the factor loading matching perspective of Ferman and Pinto [2021]. It also proposes a complementary inference procedure based on the GMM, which remains valid under nonrandom treatment assignment.

Finally, this paper relates to the proximal inference literature, particularly the work of Shi et al. [2021]. Their study focuses on the traditional SC setting, where the parameter of interest is the ATT. They make the key observation that unit-level outcomes serve as rough proxies for latent factors and therefore contain partial information that can be exploited for identification. This paper extends their proximal perspective on factor models to the QSC framework, in which each quantile of the units can similarly be viewed as a noisy measurement of the common latent factors. This information—often overlooked in

distributional analyses—can be used to transform otherwise infeasible moment conditions into feasible counterparts, thereby establishing identification of the QSC. The identification of the QTT then follows immediately.

3 Model

Suppose we observe $N + 1$ units over $T = T_0 + T_1$ periods of time. The treated unit is labeled as 0, and the control units are collected in $[N] = 1, \ldots, N$. The number of pre-treatment periods is T_0 and the number of post-treatment periods is T_1. The asymptotic regime considered in this project is that both T_0 and T_1 tend to infinity while the number of units is fixed. This assumption on the dataset is not foreign and has been adopted by several papers in the conventional synthetic control framework (Li [2020], Shi et al. [2021], Liu et al. [2023], etc.).

In particular, this project considers a scenario where the number of units is small. This has real-world implications, as aggregate units are by definition limited in number. The assumption of a long pre-treatment period is standard and necessary for consistent estimation of the synthetic control unit. The long post-treatment period assumption is used to consistently estimate the causal effects; however, it could be relaxed by employing the conformal inference approach in Chernozhukov et al. [2021]. For each unit i and at each time t, we observe the entire quantile function of a certain outcome, denoted by $q_{i,t}(\tau)$ for $\tau \in (0, 1)$. For example, in the context of minimum wage studies, $q_{i,t}(\tau)$ may represent the earnings distribution of restaurant workers, which is typically estimable with reasonable accuracy using readily available dataset . The availability of full distributional information distinguishes the setup of this project from most existing studies in the synthetic control framework.

We adopt potential outcomes notation and denote by $\left(q^I_{i,t}(\tau), q^N_{i,t}(\tau)\right)$ the quantiles of unit i under treatment and no treatment, respectively. The parameter of interest is the QTT at a given quantile τ and post-treatment period t, defined as $q^I_{i,t}(\tau) - q^N_{i,t}(\tau)$. For expositional

clarity, we assume this quantity is nonstochastic. Alternatively, if the QTT is treated as a random variable, one could instead focus on identifying, estimating, and making inference on its expectation, $\mathbb{E}\left[q_{i,t}^I(\tau) - q_{i,t}^N(\tau)\right]$. We assume the observed quantiles are realizations of the potential quantiles under the assigned treatment status:

Assumption 1 (Consistency) *For the treated unit* 0, $q_{0,t}(\tau) = q_{0,t}^N(\tau)$ *for all* $t \leq T_0$, *and* $q_{0,t}(\tau) = q_{0,t}^I(\tau)$ *for all* $t \geq T_0 + 1$. *For any control unit* $i \in [N]$, $q_{i,t}(\tau) = q_{i,t}(\tau)^N$ *for all* t.

Each quantile follows a linear factor model, similar to Chen [2020], featuring a set of common factors driving all quantiles of different units. The effect of the factor loadings depends on the unit and the quantile. In other words, the factor loadings are unit-and-quantile specific.

Assumption 2 (Linear Factor Model on Quantiles) *In the pretreatment period* $1 \leq t \leq T_0$ *and for any quantile* τ,

$$\begin{cases} q_{0,t}(\tau) = \mu_0(\tau)'\lambda_t + \epsilon_{0,t}(\tau), \\ q_{i,t}(\tau) = \mu_i(\tau)'\lambda_t + \epsilon_{i,t}(\tau), \text{ for any } i \in [N]; \end{cases}$$

In the posttreatment period $T_0 + 1 \leq t \leq T_0 + T_1$ *and for any quantile* τ,

$$\begin{cases} q_{0,t}(\tau) = \delta_t^*(\tau) + \mu_0(\tau)'\lambda_t + \epsilon_{0,t}(\tau), \\ q_{i,t}(\tau) = \mu_i(\tau)'\lambda_t + \epsilon_{i,t}(\tau), \text{ for any } i \in [N], \end{cases}$$

where λ_t *is an* $F \times 1$ *vector of stochastic common factors characterizing the co-movement of quantiles across units; for* $i \in \{0\} \cup [N]$, $\mu_i(\tau)$ *is an* $F \times 1$ *vector of fixed unit- and quantile-specific factor loadings; for* $i \in \{0\} \cup [N]$, $\epsilon_{i,t}(\tau)$ *is an idiosyncratic shock affecting the* τ*-th quantile of unit* i *at time* t *satisfying* $\mathbb{E}[\epsilon_{i,t}(\tau) \mid \lambda_t] = 0$ *; and* $\delta_t^*(\tau)$ *denotes the QTT experienced by the* τ*-th quantile of the treated unit as a result of the intervention.*

Most research in the synthetic control literature relies on model-based assumptions about potential outcomes to justify identification. The use of factor models is motivated by their ability to capture co-movements among variables of interest in high-dimensional settings Bai [2003]. Empirically observed co-movement in outcomes is the original rationale for adopting linear factor models Abadie [2021]. In 2, the availability of an additional dimension—quantiles—places us in an even higher-dimensional environment, where quantiles themselves may exhibit co-movement. This further strengthens the justification for employing linear factor models in this context. 2 could be expended to allow time-varying factors with constant factor loadings across units for each quantile. The feature has been shut off for exposition. Also, note that we have abstracted away from incorporating observed covariates into the model; however, similar results would follow if they were included. For ease of exposition, $\delta_t^*(\tau)$ is assumed to be a constant, but analogous arguments apply for identifying $\mathbb{E}[\delta_t^*(\tau)]$ even when it is stochastic.

Remark 1 *The unit- and quantile-specific factor loadings are treated as fixed population parameters. This choice is motivated by the fact that the number of control units is fixed, and in practice, we often observe the full population of aggregate units such as states or countries. The time-varying factors, by contrast, are modeled as stochastic. For identification purposes, the analysis accommodates both stationary and nonstationary common factors, thereby allowing the model to account for the possibility of growth. However, for estimation, standard assumptions of stationarity and weak dependence must be imposed on the time series of quantiles in order to apply results from the GMM literature.*

Remark 2 *One empirical context that helps motivate the linear factor model on quantiles in 2 involves estimating the distributional effect of a minimum wage increase in California on specific quantiles of the earnings distribution for restaurant workers. In this setting, California is labeled as unit* 0*, while other states are indexed by* $[N]$*. The vector* λ_t *can be interpreted as representing latent common economic shocks. Because each state has its own economic structure, it responds differently to these shocks. Moreover, individuals from*

different socioeconomic backgrounds may experience and react to economic shocks differently, justifying the use of unit- and quantile-specific factor loadings. Finally, the idiosyncratic shocks $\epsilon_{i,t}(\tau)$ can be viewed as disturbances that are localized to workers' labor markets in a given state at time t.

Remark 3 *The DGP described in 2 is similar in spirit to the one studied in Chen [2020], but it differs in several important respects. First, 2 abstracts away from observed covariates, whereas heterogeneous covariates across units are a central feature of the model in Chen [2020]. In their approach, the quantiles are residualized by regressing them on observed covariates to improve pre-treatment fit; identification of their QSC and QTT then proceeds through a perfect pre-treatment fit of these residualized quantiles. While my model currently omits observed covariates for simplicity, it can in principle incorporate them. Residualizing the quantiles in the same manner would not alter the identification strategy, as all arguments would apply analogously to the residualized quantiles.*

A second key distinction lies in the structure of the latent factors. While 2 assumes common latent factors λ_t shared across units, Chen [2020] instead allow for time-varying, unit-specific factors $\lambda_{i,t}$ that affect all quantiles within the same unit. Although this assumption makes their model more flexible in that dimension, their identification strategy requires a strong assumption that $\mathbb{E}[\lambda_{i,t}]$ is constant over time. This assumption rules out identification in settings where latent factors exhibit growth—a feature explicitly allowed in my DGP.

Thus, we do not view the DGP in Chen [2020] as strictly more general. Rather, there is a tradeoff between the assumptions in 2 and those in Chen [2020]. Practitioners should choose between the two based on the characteristics and limitations of their data.

Whenever a model like 2 is written, it is implicitly conditioned on a particular realization of the treatment assignment mechanism. In this realization, unit 0 is treated, while all other units remain untreated. However, from a statistical perspective, both the treatment assignment mechanism and the factor structure should be viewed as random variables—and they may be arbitrarily correlated. Because the factor loadings are fixed, the treatment

assignment mechanism can be arbitrarily correlated with the common factors. For this reason, we refer to λ_T as the ultimate source of unobserved confounding.

This implicit assumption has been largely overlooked in the synthetic control literature, where it is common to assume that treatment is randomly assigned and therefore uncorrelated with the factor structure. However, this assumption is often overly strong, particularly in empirical settings involving large-scale policy interventions. For instance, it is implausible to assume that a policy such as a minimum wage increase is randomly assigned across states. In reality, states endogenously choose whether to implement such policies, making treatment assignment potentially correlated with unobserved factors. One important implication of nonrandom assignment is that the traditional placebo test becomes invalid. A key contribution of this project is the development of an inference procedure that remains valid under nonrandom treatment assignment.

Since the factor structure in the linear factor model for quantiles can be interpreted as unmeasured confounding, it provides intuitive justification for using a synthetic control method to recover the QTT. Consider the τ-quantile of the treated unit and the τ-quantile of any control unit. The difference between them in a post-treatment period takes the form:

$$q_{0,t}(\tau) - q_{i,t}(\tau) = \delta_t^*(\tau) + (\mu_0(\tau) - \mu_i(\tau))\prime\lambda_t + (\epsilon_{0,t}(\tau) - \epsilon_{i,t}(\tau)).$$

The first term is the target of identification—the QTT at quantile τ. The third term captures the difference in idiosyncratic shocks between the treated and a control unit at quantile τ within the same period; this term has mean zero and therefore does not pose a challenge for identification. The second (middle) term, however, is nontrivial. It arises due to the mismatch in factor loadings between the treated unit and the control unit at quantile τ, and it cannot be eliminated because λ_t is not mean zero. In fact, this issue persists even if we compare the τ-quantile of the treated unit with any quantile of any control unit. That is, we cannot directly find a comparable counterpart for the treated unit's τ-quantile that reliably

mimics its counterfactual behavior. The core idea underlying any synthetic control approach is that, when no single comparable unit exists to isolate the causal estimand of interest, we construct a synthetic counterpart using a weighted combination of selected control units. Specifically, the synthetic control unit in the quantile setting, constructed as a weighted average of the τth quantiles of the donors, would match the factor loadings of the τ-quantile of the treated.

4 Identification

Before presenting the identification procedure, it is helpful to introduce additional notation to categorize control units. Let us refer to the control units whose quantiles contribute to the construction of the τ-QSC—the synthetic control unit for the τ-quantile of the treated—as donors, and denote the set of such units by D. Similarly, control units that do not contribute to the construction are referred to as non-donors, with their set denoted by D^C. This notation reflects the empirical reality that not all control units are used to form the τ-QSC. As in Abadie [2015], only a subset of countries—specifically the OECD countries—were considered as potential donors for constructing a synthetic West Germany, and among them, only a few were ultimately selected. In many applications, certain units may be subject to spillover effects or may experience a similar treatment during the same time window as the treated unit. Including such units in the donor pool can lead to biased estimates of the QTT. For this reason, these units should be excluded from D and instead placed in the non-donor pool D^C. With the new notation in place, we are now ready to define the τ-QSC using the following identification condition.

4.1 Matching on Factor Loadings

Assumption 3 (Matching on Factor Loadings) *For each quantile level $\tau \in (0,1)$, there exists a set of real coefficients $\{\alpha_i^*(\tau)\}_{i \in D \subseteq [N]}$ such that:*

$$\mu_0(\tau) = \sum_{i \in D} \alpha_i^*(\tau)\, \mu_i(\tau)$$

In other words, the τ-QSC is defined as a weighted combination of the τ-quantiles of donor units that exactly matches the factor loadings of the τ-quantile of the treated unit. Since the τ-QSC is fully characterized by the set of weights $\{\alpha_i^*(\tau)\}_{i \in D}$, we use the term τ-QSC interchangeably to refer to either the synthetic control itself or the corresponding weight vector. One direct implication of assumption 3 is:

$$\underbrace{\mu_0(\tau)'\lambda_t}_{\substack{\text{confounder of 0} \\ \text{at } \tau}} = \sum_{i \in D} \alpha_i^*(\tau)\ \underbrace{(\mu_i(\tau)'\lambda_t)}_{\substack{\text{confounder of } i \\ \text{at } \tau}},$$

which ensures that the unmeasured confounder for the treated unit at quantile τ is perfectly matched by the τ-QSC. Accordingly, the τ-QSC can be used to control for unmeasured confounding at quantile τ in the post-treatment period, thereby isolating the QTT at τ.

Remark 4 *Note that the set of weights $\{\alpha_i^*(\tau)\}_{i \in D}$ is not restricted to the simplex, as in Abadie et al. [2010]. In fact, $\{\alpha_i^*(\tau)\}_{i \in D}$ does not need to sum to one—even though, in principle, if we included common time-varying factors with common factor loadings for the τ-quantiles, the set of weights would be constrained to sum to one. Also, $\{\alpha_i^*(\tau)\}_{i \in D}$ may include negative weights or weights greater than one, allowing the τ-QSC to extrapolate outside the convex hull of the τ-quantiles of the donors. The idea is that the τ-quantile of the treated can be systematically different from the τ-quantiles of the donors, making extrapolation necessary to match factor loadings.*

Finally, in Assumption 3, the construction of the τ-quantile of the treated relies solely on the corresponding quantiles of the donors. This restriction can be easily relaxed to allow

different quantiles of the donors to participate in the construction. We focus on matching the same quantiles to align with the approach in Chen [2020], where matching is also done at the same quantile level.

Remark 5 *The definition of the τ-QSC in this project differs from that in Chen [2020] and shares the same spirit as Ferman and Pinto [2021]. The approach adopted here is based on matching factor loadings, whereas Chen [2020] rely on matching the pre-treatment trajectories of the τ-quantiles. We argue that the latter is unrealistic, as achieving a (near) perfect match in observed outcomes is often infeasible in empirical applications—especially over the entire pre-treatment period. Moreover, the latter is heuristic in nature, as there is no universally accepted criterion for what constitutes a (near) perfect match.*

In contrast, the definition used in this project avoids the need to define or achieve a (near) perfect match. It allows for potential discrepancies between the τ-QSC and the τ-quantile of the treated unit due to differences in idiosyncratic shocks. Specifically, in the pretreatment period,

$$q_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau) = \underbrace{\left(\mu_0(\tau) - \sum_{i \in D} \alpha_i^*(\tau)\, \mu_i(\tau) \right)' \lambda_t}_{=0} + \left(\epsilon_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau) \epsilon_{i,t}(\tau) \right)$$

where the first term vanishes by construction of the τ-QSC, and the remaining discrepancy is solely due to idiosyncratic shocks.

Remark 6 *In the post-treatment period,*

$$q_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau) = \delta_t^*(\tau) + \underbrace{\left(\mu_0(\tau) - \sum_{i \in D} \alpha_i^*(\tau)\, \mu_i(\tau) \right)' \lambda_t}_{=0} + \left(\epsilon_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau) \epsilon_{i,t}(\tau) \right)$$

Therefore, once the τ-QSC has been identified, identification of the QTT at τ follows immediately, using posttreatment information.

One issue, however, is that Assumption 3 alone is not sufficient to identify the τ-QSC, since the factor loadings are unobserved. This stands in contrast to Chen [2020], where their version of the τ-QSC is defined in terms of observables in the pre-treatment period and is therefore identified immediately. As a result, additional information is needed to identify the τ-QSC, and consequently, the QTT at τ.

We argue that such information is abundant in the factor model 2, and, perhaps surprisingly, it comes from the non-donors. At first glance, it may seem that the information in the non-donors is either irrelevant or potentially harmful. Indeed, some non-donors may be affected by spillover effects or may have undergone similar interventions, and their inclusion could bias the analysis. However, by adopting the proximal perspective of Shi et al. [2021], it can be shown that the quantiles of the non-donors can, under suitable conditions, be used to facilitate identification. The key insight is that the quantiles of the non-donors serve as proxies for the latent factors, and thus contain partial information about the unmeasured confounder.

4.2 Identification with Proxies

The unobserved confounder poses the ultimate challenge for identifying the τ-QSC and the corresponding QTT. To see this, suppose that λ_t were observable for some pretreatment period $t \leq T_0$. Then,

$$\mathbb{E}\left[q_{0,t} - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau) \;\middle|\; \lambda_t\right] = \mathbb{E}\left[\lambda_t \left(q_{0,t} - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau)\right)\right] = 0.$$

That is, we would have obtained a set of moment conditions for identifying the τ-QSC based on pretreatment information. Unfortunately, two key issues arise in reality. First, these are infeasible moment conditions because λ_t is unobserved. Second, the dimension of λ_t may not be large enough for the system of equations to have a solution, rendering the τ-QSC unidentifiable under these conditions. Nonetheless, this infeasible system provides

useful intuition for constructing a feasible identification strategy. First, we need to replace λ_t with observable variables that contain at least partial information about it. Second, we must have a sufficient number of such observables to obtain a set of moment conditions that at least just-identifies the $\tau-$QSC. Due to the factor structure on the quantiles, these two tasks can be achieved under mild conditions.

By Assumption 2, for any unit $i \in 0 \cup [N]$, we have $q_{i,t}(\tau) = \mu_i(\tau)'\lambda_t + \epsilon_{i,t}(\tau)$. Since $\mu_i(\tau)$ is fixed and $\epsilon_{i,t}(\tau)$ is idiosyncratic noise, the τ-quantile of unit i can be viewed as a noisy measurement of λ_t, and thus contains partial information about the latent confounder. This observation suggests that quantiles are natural candidates to replace λ_t in the moment conditions—and, in fact, there is an abundance of them. For instance, the τ-quantiles of units that experience spillover effects or undergo similar treatments may still serve as rough proxies for λ_t. Additionally, one may also consider quantiles at levels $\tau' \neq \tau$ across various units. Altogether, this highlights that pretreatment information not directly used in constructing the τ-QSC can be potentially repurposed to aid in the identification of the τ-QSC.

Nevertheless, the fact that the quantiles of untreated units contain contaminated information about the latent confounder does not immediately imply they can serve as substitutes for λ_t in the moment conditions. A key condition must be satisfied: in the pretreatment period, these proxy variables must be conditionally independent of the τ-quantiles of the treated unit and the donors, given the latent confounder λ_t. Specifically, for any $j \in D^C$ and $i \in 0 \cup [N]$,

$$q_{j,t}(\tau) \perp q_{i,t}(\tau) \mid \lambda_t, \;\; 0 \leq t \leq T_0.$$

To see why this assumption is useful, observe that:

$$\mathbb{E}\left[q_{0,t} - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau) \;\middle|\; \lambda_t\right] = \mathbb{E}\left[q_{0,t} - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau) \;\middle|\; q_{j,t}(\tau), \lambda_t\right] = 0$$

where the first equality follows from the conditional independence assumption. Integrating

out λ_t on both sides of the second equality yields:

$$\mathbb{E}\left[q_{0,t} - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau) \;\middle|\; q_{j,t}(\tau)\right] = 0.$$

This conditional moment restriction can alternatively be expressed in an unconditional form as:

$$\mathbb{E}\left[q_{j,t}(\tau)\left(q_{0,t} - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau)\right)\right] = 0.$$

Since there are typically multiple non-donors available, we can repeat the above calculation for each non-donor, using different quantiles if necessary. Let Z_t be a $d \times 1$ vector containing quantiles from the non-donors, and suppose $d \geq |D|$. Assume that Z_t serves as a valid proxy for λ_t, and therefore satisfies

Assumption 4 (Existence of Proxies)

$$Z_t \perp\!\!\!\perp \{q_{0,t}(\tau), q_{i \in D,t}(\tau)\} \mid \lambda_t, \quad \textit{for all } t \leq T_0.$$

Then, we have obtained sufficient moment conditions to identify the $\tau-$QSC:

Theorem 1 (Identification of QSC) *Under Assumptions 1, 2, 3 and 4, τ-QSC weights $\{\alpha_i^*(\tau)\}_{i \in D}$ satisfy the conditional moment restrictions:*

$$\mathbb{E}\left[Z_t\left(q_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau)\right)\right] = 0, \quad \textit{for all } t \leq T_0.$$

Note that theorem 1 does not guarantee the uniqueness of the τ-QSC. A sufficient condition for uniqueness would be to impose a full-rank condition on the moment conditions in 1. Additionally, certain completeness assumptions on Z_t could also be imposed to obtain uniqueness. Nevertheless, any τ-QSC that satisfies 1 immediately identifies the QTT at τ based on post-treatment information.

Theorem 2 (Identification of QTT) *Suppose Assumptions 1, 2, 3 and 4 hold. Then for all $t > T_0$,*

$$\delta_t^*(\tau) = \mathbb{E}\left[q_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau)\, q_{i,t}(\tau)\right].$$

It remains to establish under what conditions the quantiles of non-donors serve as legitimate proxies—specifically, when they satisfy the required conditional independence assumption. Given the factor structure imposed on quantiles and the assumption that idiosyncratic shocks are mean-independent of λ_t, it is sufficient to argue that $\epsilon_{j,t}(\tau) \perp \epsilon_{i,t}(\tau)$ for any $j \in D^C$ and any $i \in 0 \cup D$. While this assumption is untestable in practice, as the idiosyncratic shocks are not directly observable, it can be justified on several grounds. First, many studies in the synthetic control literature adopt the assumption of independence across cross-sectional units, typically formalized as i.i.d. error terms at each time period (e.g., Abadie et al. [2010], Ferman and Pinto [2021]). Second, empirical contexts tend to lend support to the plausibility of shock independence. For instance, in studying the distributional effects of minimum wage increases, geographic and institutional separation between the treated state and control states naturally supports the assumption. States typically experience localized economic shocks, distinct labor market conditions, and different policy environments. Even when broader macroeconomic trends are shared, such local heterogeneity helps ensure that residual shocks are primarily idiosyncratic. Third, in empirical applications, the inclusion of rich observed covariates can further mitigate the risk of correlated shocks. By conditioning on a comprehensive set of economic, demographic, and institutional characteristics, much of the systematic variation that might drive cross-unit correlation is absorbed. While no finite set of covariates can fully guarantee independence, a well-specified model greatly enhances the plausibility of this assumption.

5 Estimation and Inference

Inspired by the moment conditions that over- or just-identify the τ-QSC, we propose to estimate and make inference on it using pre-treatment data via GMM. With the estimated τ-QSC, we can estimate residuals in the post-treatment period and apply appropriate time series analysis to recover the trajectory of the QTT at τ.

One important issue that must be addressed is the potential non-uniqueness of the τ-QSC, which could complicate the GMM estimation and inference procedure. To simplify the analysis, this project imposes the assumption that the moment-condition matrix for the τ-QSC in the pre-treatment period is of full row rank, ensuring uniqueness of the solution. We refer readers to Shi et al. [2020] for discussion of settings—particularly in categorical data—where multiple sets of synthetic control weights may exist. More generally, one may define additional criteria for selecting an "optimal" set of synthetic control weights. A related strategy is formally examined in the supplementary material of Shi et al. [2021], which explores such approaches in the context of proximal causal inference.

Another mild cost associated with our approach is the need to impose standard assumptions on the time series of quantiles in order to apply GMM theory. Specifically, the quantiles of all units involved in the moment conditions—including the treated unit, the donors, and the non-donors serving as proxies—must be stationary and weakly dependent. In practice, non-stationarity in the data can often be mitigated through first-differencing or other algebraic transformations. The weak dependence assumption, meanwhile, is commonly adopted in the synthetic control literature to facilitate estimation and inference (e.g., Ferman and Pinto [2021], Chernozhukov et al. [2021], Cattaneo et al. [2021]).

Denote the moment function corresponding to the moment conditions that identify the τ-QSC by

$$U_t(\alpha(\tau)) = Z_t \left(q_{0,t}(\tau) - \sum_{i \in D} \alpha_i(\tau) q_{i,t}(\tau) \right),$$

where $\alpha(\tau) = (\alpha_i(\tau))_{i \in D} \in \mathbb{R}^{|D|}$.

Define the sample moment conditions as

$$m(\alpha(\tau)) = \frac{1}{T_0}\sum_{t=1}^{T_0} U_t(\alpha(\tau)).$$

It can be easily shown that $m(\alpha(\tau))$ has population mean zero. Then, the following theorem establishes the consistency of the GMM estimator for the τ-QSC:

Theorem 3 *Let $\theta = \alpha(\tau)'$ and $\theta^* = \alpha^*(\tau)'$. Let Ω be a $|D| \times |D|$ positive definite weighting matrix. Suppose $\mathbb{E}[U_t(\alpha(\tau))]$ is of low full rank for all $t \leq T_0$ and suppose Assumptions 1, 2, 3, 4 and standard regularity conditions listed in the appendix hold.*

Then the GMM estimator

$$\hat{\theta}^{GMM}(\tau) = \arg\min_{\theta \in \mathbb{R}^{|D|}} m(\theta)'\,\Omega\, m(\theta)$$

converges in probability:

$$\hat{\theta}^{GMM}(\tau) \xrightarrow{p} \theta^*$$

as $T_0, T_1 \to \infty$, with $T_0/T_1 \to \rho \in (0, \infty)$.

With a consistent estimate of the τ-QSC obtained from pretreatment data, we can proceed to estimate the trajectory of the QTT at quantile τ in the posttreatment period. To illustrate this, suppose the true parameter vector θ^* is known—that is, we have a highly accurate estimate of the τ-QSC based on pre-treatment observations. Then, for any $t > T_0$,

$$q_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau) q_{i,t}(\tau) = \delta_t^*(\tau) + \left(\epsilon_{0,t}(\tau) - \sum_{i \in D} \alpha_i^*(\tau)\epsilon_{i,t}(\tau) \right),$$

which closely resembles a standard regression model. This structure suggests that, under a parametric or semiparametric specification of $\delta_t^*(\tau)$—along with assumptions of stationarity and weak dependence for the post-treatment idiosyncratic shocks associated with the τ-quantiles of the treated unit and donors—we can analyze the model using tools from the

time-series literature to conduct estimation and inference on the QTT.

For illustration, this project focuses on a special case where $\delta_t^*(\tau)$ is constant over time and equals $\delta^*(\tau)$. Then a natural estimator for $\delta^*(\tau)$ is

$$\hat{\delta}(\tau) = \frac{1}{T_1} \sum_{t>T_0} \left(q_{0,t}(\tau) - \sum_{i \in D} \hat{\alpha}_i^{GMM}(\tau) q_{i,t}(\tau) \right).$$

The two-step procedure of estimating $\{\alpha_i^*(\tau)\}_{i \in D}$ and subsequently $\delta_t^*(\tau)$ can alternatively be implemented in a single step by stacking their associated moment conditions and performing joint estimation via GMM. As before, denote the moment function corresponding to the joint moment conditions for $\{\alpha_i^*(\tau)\} i \in D$ and $\delta_t^*(\tau)$ by

$$\tilde{U}_t(\alpha(\tau), \delta(\tau)) = \begin{pmatrix} Z_t \left(q_{0,t}(\tau) - \sum_{i \in D} \alpha_i(\tau) q_{i,t}(\tau) \right) \cdot \mathbb{I}\{t \leq T_0\} \\ \left(q_{0,t}(\tau) - \delta(\tau) - \sum_{i \in D} \alpha_i(\tau) q_{i,t}(\tau) \right) \cdot \mathbb{I}\{t > T_0\} \end{pmatrix}$$

and define the corresponding sample moment conditions as

$$\tilde{m}(\alpha(\tau), \delta(\tau)) = \frac{1}{T} \sum_{t=1}^{T} \tilde{U}_t(\alpha(\tau), \delta(\tau)).$$

It is easy to show the sample mean moment conditions have population mean zero and therefore we could perform joint GMM to estimate and do inference on the parameters, particularly $\delta^*(\tau)$.

Theorem 4 (GMM for QTT) *Let $\theta = (\alpha(\tau), \delta(\tau))'$ and $\theta^* = (\alpha^*(\tau), \delta^*(\tau))'$. Let Ω be a $(|D|+1) \times (|D|+1)$ positive definite matrix. Under Assumptions 1, 2, 3, 4, standard regularity conditions listed in the appendix, and the condition that the matrix $\mathbb{E}\left[\tilde{U}_t(\alpha(\tau), \delta(\tau))\right]$ is full rank for all t, the GMM estimator solves*

$$\hat{\theta}^{GMM_J}(\tau) = \arg \min_{\theta \in \mathbb{R}^{|D|+1}} \tilde{m}(\theta)' \, \Omega \, \tilde{m}(\theta)$$

and it satisfies:

$$\sqrt{T}\left(\hat{\theta}^{GMM_J} - \theta\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

as $T_0, T_1 \to \infty$*, with* $T_0/T_1 \to \rho \in (0, \infty)$ *and* Σ *following the standard sandwich formula for GMM asymptotics.*

Theorem 4 presents an inference procedure that accounts for uncertainty arising from time series sampling, in contrast to the state-of-the-art placebo test, where uncertainty stems from treatment assignment, as formalized in Chen [2020]. A key advantage of the inference approach in Theorem 4 is its ability to accommodate endogenous treatment assignment—an appealing feature for many empirical applications in policy evaluation. In contrast, the placebo test requires the treated and control units to be comparable, which is a restrictive assumption when treatment is nonrandomly assigned. Moreover, the GMM-based procedure yields confidence intervals for the parameters of interest, whereas the placebo test is designed primarily for hypothesis testing regarding the presence of a treatment effect. For these reasons, we view the inference procedure in Theorem 4 as a valuable complement to the placebo test, enabling practitioners to conduct valid inference even in settings where treatment assignment is not randomized.

6 Simulation

In this section, we conduct simulation exercises to evaluate the finite-sample performance of our proposed estimation and inference procedures for the QTT under various model specifications. We set the number of pre-treatment and post-treatment periods to be equal, i.e., $T_0 = T_1 \in \{100, 200, 500\}$, and simulate quantile functions for both the treated unit and N control units over $T_0 + T_1$ total time periods. The data are generated according to the following data generating process:

$$\begin{cases} q_{0,t}(\tau) = \delta^*(\tau)\mathbb{I}(t > T_0) + \mu_0(\tau)'\lambda_t + \epsilon_{0,t}(\tau), \\ q_{i,t}(\tau) = \mu_i(\tau)'\lambda_t + \epsilon_{i,t}(\tau) \quad \text{for } i = 1, \ldots, N. \end{cases}$$

The number of latent factors in λ_t, denoted by F, is set to either 5 or 10, and the number of control units is fixed at $N = 2F$. We designate the first half of the control units, indexed by $1, \ldots, F$, as the donor pool D, and the remaining units as the non-donor pool D^C. We consider two scenarios for modeling the factor process λ_t. In the case where λ_t exhibits growth, we generate each factor component as $\lambda_{f,t} \sim N(\log(t), 1)$ for all $f = 1, \ldots, F$, so that λ_t is nonstationary. In the stationary case, we instead draw $\lambda_{f,t} \sim N(0, 1)$ for all $f = 1, \ldots, F$.

The factor loadings of the quantiles for the treated unit are parameterized as $\mu_0(\tau) = \mu_0 \cdot e^\tau$, where $\mu_0 = (1, \ldots, 1)'$. Similarly, the factor loadings for the control units are specified as $\mu_i(\tau) = \mu_i \cdot e^\tau$, where $\mu_i = (0, \ldots, 1, \ldots, 0)$ is a unit vector with the ith entry equal to 1 for unit $i \in D$, and $\mu_j = (0, \ldots, 1, \ldots, 0)$ with the jth entry equal to 1 for unit $j \in D^C$. In other words, $(\mu_1, \ldots, \mu_F) = (\mu_{1+F}, \ldots, \mu_{2F}) = I_{F\times F}$, where $I_{F\times F}$ denotes the $F \times F$ identity matrix. By design, the τ-QSC is simply the set of weights $\{\alpha_i(\tau) = 1\}_{i\in D}$.

For all $i \in 0 \cup [N]$, the idiosyncratic shocks are specified as $\epsilon_{i,t}(\tau) = \epsilon_{it}$, where $\epsilon_{it} \sim N(0, 1)$ or follows an AR(1) process with autocorrelation coefficient 0.2. Thus, the local shocks experienced by each unit i are homogeneous across all quantiles. We consider both the i.i.d. and weakly dependent cases to assess the robustness of our estimator.

Finally, the QTT at quantile τ is assumed to be time-invariant and parameterized as $\delta_t^*(\tau) = e^\tau$. Therefore, the target parameter in our simulation study is $e^{0.5} \approx 1.649$. The goal of this simulation is to evaluate the finite-sample performance of the proposed estimator for the 0.5-QTT and to compare it with the state-of-the-art synthetic control method. We emphasize that the benchmark method is unconstrained, meaning it allows for extrapolation beyond the convex hull of the donor pool in order to match the factor loadings of the treated unit at the median.

We implement the following three methods in estimating the 0.5-QTT:

(1) (SC) The unconstrained synthetic control method implemented based on a linear regression

$$q_{0,t}(0.5) = \alpha_0 + \sum_{i \in D} \alpha_i(0.5) q_{i,t}(0.5) + \delta(0.5) * \mathbb{I}(t > T_0) + u_t, \text{ for } t = 1, \ldots, T.$$

Then, the estimate of $\delta(0.5)$ is the estimate for the 0.5-QTT.

(2) (PI_1) The proposed estimation strategy based on GMM, using the medians of the non-donors as proxies. This method is fully characterized by the joint moment conditions:

$$\mathbb{E}\begin{pmatrix} Z_t(0.5)\left(q_{0,t}(0.5) - \sum_{i \in D} \alpha_i(0.5) q_{i,t}(0.5)\right) \cdot \mathbb{I}\{t \leq T_0\} \\ \left(q_{0,t}(0.5) - \delta(0.5) - \sum_{i \in D} \alpha_i(0.5) q_{i,t}(0.5)\right) \cdot \mathbb{I}\{t > T_0\} \end{pmatrix}$$

where the estimate of $\delta(0.5)$ is the estimate for the 0.5-QTT and $Z_t(0.5)$ is a vector collecting all the medians of the non-donors.

(3) (PI_2) The proposed estimation strategy based on GMM, using the 0.2-quantiles of the non-donors as proxies. This method is fully characterized by the joint moment conditions:

$$\mathbb{E}\begin{pmatrix} Z_t(0.2)\left(q_{0,t}(0.5) - \sum_{i \in D} \alpha_i(0.5) q_{i,t}(0.5)\right) \cdot \mathbb{I}\{t \leq T_0\} \\ \left(q_{0,t}(0.5) - \delta(0.5) - \sum_{i \in D} \alpha_i(0.5) q_{i,t}(0.5)\right) \cdot \mathbb{I}\{t > T_0\} \end{pmatrix}$$

where the estimate of $\delta(0.5)$ corresponds to the estimated 0.5-QTT, and $Z_t(0.2)$ denotes a vector collecting all the 0.2-quantiles of the non-donor units. This method is included to illustrate the broader point that, under a factor structure on the quantiles, there exists an abundance of information that can be leveraged for the identification, estimation, and inference of the QTT.

Table 1: λ_t is Stationary

ϕ	F	T_1	$\hat{\mu}_{\text{PI}_1}$	$\hat{\mu}_{\text{PI}_2}$	$\hat{\mu}_{SC}$	Coverage(PI_1)	Coverage(PI_2)	Coverage(SC)	Length(PI_1)	Length(PI_2)	Length(SC)
0.0	5	100	1.658	1.659	2.070	0.935	0.938	0.566	0.640	0.657	0.462
0.0	5	200	1.644	1.644	2.054	0.943	0.945	0.307	0.454	0.463	0.327
0.0	5	500	1.648	1.648	2.063	0.951	0.952	0.026	0.287	0.293	0.207
0.0	10	100	1.637	1.637	2.140	0.942	0.943	0.658	0.912	0.942	0.631
0.0	10	200	1.663	1.664	2.142	0.951	0.956	0.426	0.637	0.650	0.448
0.0	10	500	1.642	1.641	2.132	0.936	0.939	0.100	0.402	0.408	0.284
0.2	5	100	1.645	1.645	2.051	0.946	0.948	0.594	0.637	0.654	0.462
0.2	5	200	1.643	1.642	2.055	0.944	0.944	0.326	0.453	0.463	0.327
0.2	5	500	1.653	1.653	2.063	0.943	0.944	0.025	0.287	0.293	0.207
0.2	10	100	1.662	1.662	2.142	0.940	0.942	0.664	0.906	0.935	0.630
0.2	10	200	1.634	1.634	2.132	0.942	0.943	0.451	0.638	0.652	0.448
0.2	10	500	1.637	1.637	2.130	0.942	0.946	0.096	0.401	0.407	0.284

Table 2: λ_t is Non-Stationary

ϕ	F	T_1	$\hat{\mu}_{\text{PI}_1}$	$\hat{\mu}_{\text{PI}_2}$	$\hat{\mu}_{SC}$	Coverage(PI_1)	Coverage(PI_2)	Coverage(SC)	Length(PI_1)	Length(PI_2)	Length(SC)
0.0	5	100	1.655	1.655	1.865	0.944	0.944	0.884	0.787	0.796	0.600
0.0	5	200	1.647	1.647	1.814	0.943	0.946	0.879	0.549	0.552	0.417
0.0	5	500	1.649	1.649	1.791	0.948	0.948	0.809	0.340	0.341	0.260
0.0	10	100	1.646	1.646	1.871	0.943	0.948	0.908	1.097	1.122	0.788
0.0	10	200	1.666	1.666	1.827	0.949	0.951	0.900	0.753	0.761	0.549
0.0	10	500	1.641	1.641	1.782	0.942	0.943	0.868	0.465	0.467	0.341
0.2	5	100	1.655	1.655	1.852	0.943	0.943	0.895	0.784	0.793	0.599
0.2	5	200	1.646	1.646	1.814	0.942	0.942	0.869	0.548	0.551	0.417
0.2	5	500	1.654	1.654	1.790	0.950	0.951	0.812	0.341	0.342	0.260
0.2	10	100	1.668	1.667	1.872	0.936	0.940	0.887	1.089	1.113	0.787
0.2	10	200	1.637	1.638	1.818	0.950	0.950	0.899	0.754	0.763	0.548
0.2	10	500	1.637	1.637	1.779	0.947	0.946	0.884	0.465	0.467	0.341

Tables 1 and 2 present the results of the simulation exercises under various specifications, depending on whether λ_t is stationary or nonstationary. The number of Monte Carlo iterations is set to 2,000, and the nominal coverage level for the parameter of interest is 95% across all methods. In all settings, the proposed methods achieve the nominal coverage level, whereas the SC method fails to do so. Moreover, the proposed estimators are approximately unbiased, while SC exhibits noticeable bias, particularly in the stationary case. However, the confidence intervals produced by the proposed methods tend to be more conservative than those produced by SC. This conservatism is expected, as the proxy variables used in the moment conditions function analogously to instrumental variables, introducing additional variability. Despite having shorter confidence intervals, the SC method fails to adequately compensate for its bias, leading to systematic undercoverage. In contrast, the wider intervals of the proposed methods yield valid inference, highlighting the robustness of

proxy-based identification in the presence of latent confounding.

This simulation study also highlights a key feature of the quantile synthetic control framework: the flexibility in the choice of proxies. Both the 0.2-quantiles and the medians of the non-donor units can serve as valid proxies for conducting inference on the QTT. More generally, a wide range of quantiles from the non-donors may be used without compromising the validity of the procedure. Additionally, the proposed method remains robust under both independent and weakly dependent data structures, supporting a broad array of empirical applications. Crucially, the procedure does not rely on any assumption about the randomness of treatment assignment, making it particularly well-suited for evaluating policies that are implemented endogenously—an important advantage in applied settings.

7 Extensions

This section discusses one potential extension of the current framework that may be worth exploring after the submission of the third-year paper.

The biggest criticism of the current approach is that the monotonicity of the counterfactual quantile function cannot be guaranteed, since we construct the counterfactual quantiles point by point. As a result, global properties such as monotonicity may not be maintained. This issue could be avoided by adopting a different representation of the outcome distribution. For example, using the log-density function may circumvent the monotonicity problems that arise in quantile-based representations. Exploring this alternative could enhance both the theoretical appeal and the empirical robustness of the framework, and thus represents a compelling avenue for future research.

In the current quantile synthetic control framework, the modeling approach is highly localized: a separate factor structure is imposed at each quantile level. While this offers flexibility, it may be beneficial to adopt a more "global" perspective on the outcome distribution. Specifically, one could conceptualize the entire distribution function as an element

in an infinite-dimensional functional space, residing within a lower-dimensional subspace spanned by a finite set of basis functions. In this view, the evolution of the distribution over time can be captured by allowing the basis functions themselves to vary with time.

Under such a formulation, any distribution function at a given time point could be expressed as the inner product between a vector of coefficients (analogous to factor loadings) and a set of time-varying basis functions (analogous to latent factors). This functional representation provides a more coherent structure for modeling the dynamics of entire distributions, as opposed to handling each quantile independently. It also naturally preserves important global properties such as smoothness and monotonicity.

Within this framework, the concept of a synthetic control could be redefined in terms of matching the functional factor loadings of the treated unit using a convex or linear combination of those from the donor pool. This generalization would extend the utility of the synthetic control method to a broader class of distributional outcomes and potentially improve both estimation efficiency and interpretability.

8 Conclusion

This paper presents an extension of the standard synthetic control method to a distributional setting. Specifically, for each unit and time period, we observe the corresponding quantile function and impose a linear factor structure on each quantile. The quantiles are assumed to be driven by a set of latent common factors that evolve over time, with factor loadings that are both unit- and quantile-specific.

The primary contribution of this paper is to establish the identification of the quantile treatment effect on the treated at a given quantile level τ. This is achieved by defining the τ-quantile synthetic control as a weighted combination of the τ-quantiles of selected donor units, such that the factor loadings of the treated unit at τ are matched. Compared to existing approaches, this identification strategy is both more realistic and less heuristic, as

it does not require an (often infeasible) near-perfect pre-treatment fit of quantiles, especially when the pre-treatment window is long.

In addition, the paper leverages recent advances from the proximal causal inference literature to establish the identification of both the τ-quantile synthetic control and the τ-quantile treatment effect on the treated. It repurposes typically overlooked information in the quantiles of non-donor units to convert a set of infeasible moment conditions—due to unobserved latent factors—into feasible ones by treating these quantiles as proxies. Finally, the paper proposes GMM-based estimation and inference procedures that are robust to endogenous treatment assignment, offering a flexible and credible alternative to traditional placebo-based methods in applied policy evaluation.

References

Alberto Abadie. Comparative politics and the synthetic control method. *American Journal of Political Science*, 65(4):835–850, 2015. doi: 10.1111/ajps.12592.

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021. doi: 10.1257/jel.20191450. URL https://doi.org/10.1257/jel.20191450.

Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834, 2021. doi: 10.1080/01621459.2021.1971535. URL https://doi.org/10.1080/01621459.2021.1971535. Theory and Methods Special Section on Synthetic Control Methods.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746.

Sylvia Allegretto, Anna Godoey, Carl Nadler, and Michael Reich. The new wave of local minimum wage policies: Evidence from six cities. Technical report, Center on Wage and Employment Dynamics, University of California, Berkeley, 2018. URL https://irle.berkeley.edu/the-new-wave-of-local-minimum-wage-policies-evidence-from-six-cities/. CWED Policy Report, September 6.

Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2017.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

Sarah Bohn, Magnus Lofstrom, and Steven Raphael. Did the 2007 legal arizona workers act reduce the state's unauthorized immigrant population? *The Review of Economics and Statistics*, 96(2):258–269, 2014. doi: 10.1162/REST_a_00429.

Matias D Cattaneo, Yingjie Feng, and Rocio Titiunik. Prediction intervals for synthetic control methods. *arXiv preprint arXiv:1912.07120*, 2021.

Yi-Ting Chen. A distributional synthetic control method for policy evaluation. *Journal of Applied Econometrics*, 35(5):505–525, 2020. doi: 10.1002/jae.2778. URL https://doi.org/10.1002/jae.2778.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021. doi: 10.1080/01621459.2021.1920957.

Scott Cunningham and Manisha Shah. Decriminalizing indoor prostitution: Implications for

sexual violence and public health. *The Review of Economic Studies*, 85:1683–1715, 2017.

John J Donohue, Abhay Aneja, and Kyle D Weber. Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis. Working Paper 23510, National Bureau of Economic Research, June 2017. URL http://www.nber.org/papers/w23510.

Bruno Ferman and Cristine Pinto. Placebo tests for synthetic controls. MPRA Paper No. 78079, 2017. URL https://mpra.ub.uni-muenchen.de/78079/.

Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221, 2021. doi: 10.3982/QE1596.

Florian F. Gunsilius. Distributional synthetic controls. *Econometrica*, 91(3):1105–1117, 2023. doi: 10.3982/ECTA18260. URL https://doi.org/10.3982/ECTA18260.

Jinyong Hahn and Ruoyao Shi. Synthetic control and inference. *Econometrics*, 5(4):52, 2017. doi: 10.3390/econometrics5040052. URL https://www.mdpi.com/2225-1146/5/4/52.

Alastair R. Hall. *Generalized Method of Moments*. Oxford University Press, Oxford, UK, 2005.

Ekaterina Jardim, Mark C. Long, Robert D. Plotnick, Eric van Inwegen, Jacob Vigdor, and Hilary Wething. Minimum wage increases, wages, and low-wage employment: Evidence from seattle. NBER Working Paper 23532, National Bureau of Economic Research, 2017. URL https://www.nber.org/papers/w23532.

Kathleen T. Li. Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083, 2020. doi: 10.1080/01621459.2019.1686986.

Jizhou Liu, Eric J. Tchetgen Tchetgen, and Carlos Varjão. Proximal causal inference for synthetic control with surrogates. 2023. URL https://arxiv.org/abs/2308.09527. arXiv preprint arXiv:2308.09527.

David Neumark and Peter Shirley. Myth or measurement: What does the new minimum wage research say about minimum wages and job loss in the united states? Working Paper 28388, National Bureau of Economic Research, January 2021. URL http://www.nber.org/papers/w28388.

Giovanni Peri and Vasil Yasenov. The labor market effects of a refugee wave: Applying the synthetic control method to the mariel boatlift. Working Paper 21801, National Bureau of Economic Research, December 2015. URL http://www.nber.org/papers/w21801.

Michael Reich, Sylvia A. Allegretto, and Anna Godoey. Seattle's minimum wage experience 2015–16. Cwed policy brief, Center on Wage and Employment Dynamics, UC Berkeley, 2017. URL https://irle.berkeley.edu/files/2017/Seattles-Minimum-Wage-Experiences-2015-16.pdf.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Feliu Serra-Burriel, Pedro Delicado, Andrew T. Prata, and Fernando M. Cucchietti. Estimating heterogeneous wildfire effects using synthetic controls and satellite remote sensing. *Remote Sensing of Environment*, 265:112649, 2021. doi: 10.1016/j.rse.2021.112649. URL `https://www.sciencedirect.com/science/article/pii/S0034425721004472`.

Xinyao Shi, Wenguang Miao, Jason C. Nelson, and Eric J. Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540, 2020. doi: 10.1111/rssb.12361.

Xu Shi, Kendrick Qijun Li, Wang Miao, Mengtong Hu, and Eric J. Tchetgen Tchetgen. Theory for identification and inference with synthetic controls: A proximal causal inference framework. *arXiv preprint arXiv:2108.13935*, 2021. URL `https://arxiv.org/abs/2108.13935`.

Appendix

Proof of Theorem 3 and Theorem 4:

Proof. This proof focuses on proving Theorem 4. The proof for Theorem 3 follows similarly. We follow Chapter 3 of Hall [2005] to prove the asymptotic normality of the GMM estimator for the τ-QTT. This proof is very common (Liu et al. [2023], Shi et al. [2021]). To summarize, we have a population moment condition $\mathbb{E}[\tilde{U}_t(\alpha^*(\tau), \delta^*(\tau))] = 0$ with the moment function

$$\tilde{U}_t(\alpha(\tau), \delta(\tau)) = \begin{pmatrix} Z_t \left(q_{0,t}(\tau) - \sum_{i \in D} \alpha_i(\tau) q_{i,t}(\tau)\right) \cdot \mathbb{I}\{t \leq T_0\} \\ \left(q_{0,t}(\tau) - \delta(\tau) - \sum_{i \in D} \alpha_i(\tau) q_{i,t}(\tau)\right) \cdot \mathbb{I}\{t > T_0\} \end{pmatrix}.$$

The parameters $\theta = (\alpha(\tau), \delta(\tau))' \in \mathbb{R}^{|D|+1}$. Let $\mathcal{O}_t = \{q_{0,t}(\tau), q_{i \in D,t}(\tau), Z_t, T_0\} \in \mathbb{R}^{1+|D|+1+r}$ denote the observable vector of random variables. Let $\Theta \subseteq \mathbb{R}^{1+|D|}$ denote the parameter space of θ, and let $O \subseteq \mathcal{R}^{1+|D|+1+r}$ denote the sample space of $\mathcal{O}_t$. Then, $\tilde{U}_t(\alpha(\tau), \delta(\tau)) = \tilde{U}_t(\mathcal{O}_t; \theta)$ is a mapping from $O \times \Theta$ to $\mathbb{R}^r$. We impose the following regularity conditions.

Assumption 5 (Strict Stationary) *The observable vector of random variables $\mathcal{O}_t$ form a strivtly stationary process, such that all expectations of functions of $\mathcal{O}_t$ do not vary in time.*

Assumption 6 (Regularity Conditions for $\tilde{U}_t$) *The function $U_t : O \times \Theta \to \mathbb{R}^r$ where $r < +\infty$ satisfies: (i) it is continuous on Θ for each $\mathcal{O}_t \in O$; (ii) $\mathbb{E}[\tilde{U}_t(\mathcal{O}_t; \theta)]$ exists and is finite for each $\theta \in \Theta$; (iii) $\mathbb{E}[\tilde{U}_t(\mathcal{O}_t; \theta)]$ is continuous on Θ.*

Assumption 7 (Regularity conditions on $\partial \tilde{U}_t(\mathcal{O}_t; \theta)/\partial \theta'$) *(i) $\partial \tilde{U}_t(\mathcal{O}_t; \theta)/\partial \theta'$ exists and in continuous on Θ for each $\mathcal{O}_t \in O$; (ii) The true value of θ does not lie on the boundary of Θ; (iii) $\mathbb{E}[\partial \tilde{U}_t(\mathcal{O}_t; \theta)/\partial \theta']$ exists and is finite.*

Assumption 8 (Properties of the Weighting Matrix) . *The user-specified weighting matrix Ω is a positive semi-definite matrix, possibly depends on data, and converges in probability to a positive definite matrix of constants.*

Assumption 9 (Ergodicity) *The random process $\{\mathcal{O}_t; -\infty < t < +\infty\}$ is ergodic.*

Assumption 10 (Compactness of Θ) *Θ is a compact set.*

Assumption 11 (Domination of $\tilde{U}_t(\mathcal{O}_t; \theta)$) $\mathbb{E}[\sup_{\theta \in \Theta} ||\tilde{U}_t(\mathcal{O}_t; \theta)||] < +\infty$

Assumption 12 (Properties of $G_t(\theta) = T^{-1} \sum_{t=1}^{T} \partial \tilde{U}_t(\mathcal{O}_t; \theta)/\partial \theta'$) *(i) $\mathbb{E}[\partial \tilde{U}_t(\mathcal{O}_t; \theta)/\partial \theta']$ is continuous on some neighborhood N_ϵ of the true value θ^* in Θ; (ii) Uniform convergence of $G_t(\theta)$: $\sum_{\theta \in N_\epsilon} ||G_T(\theta) - \mathcal{E}[\partial \tilde{U}_t(\mathcal{O}_t; \theta)/\partial \theta']|| \xrightarrow{p} 0$.*

Under Assumptions 1-12, we have Theorem 4 by Theorem 3.2 of Hall [2005]. ■