Distributional Synthetic Control for Text Outcomes with an Application to Local Government Speech

Amar Venugopal
Stanford University Department of Economics

This version: October 24, 2026
Link to latest version

Abstract

Applied researchers increasingly study problems in which outcomes of interest are expressed in text. In this paper, I develop a distributional synthetic control framework for causal inference on text outcomes in observational panel settings. The treated unit's counterfactual is modeled as a mixture of contemporaneous distributions from untreated units, with weights estimated to minimize inter-distributional distance pre-treatment. I propose a hypothesis-generation procedure which discovers aspects of text that differ between the observed treated corpus and its constructed counterfactual distribution. In an asymptotic regime in which the number of documents in each cell of the panel matrix grows, I establish consistency and asymptotic normality of treatment effect estimators on pre-specified aspects of text; in the case of discovered aspects, I establish analogous results via sample-splitting, conditional on the learned aspect. In an application to California city council meeting public comments, I find that the elimination of remote participation options in the aftermath of the COVID-19 pandemic in several cities led to increased civility and reduced hostility in public comments.

Contact: amarvenu@stanford.edu. I am deeply grateful to my advisors, Guido Imbens, Jann Spiess, Jiafeng Chen, and Matthew Gentzkow for their invaluable guidance and mentorship. I thank Lea Bottmer, Amir Feder, Han Hong, Lihua Lei, Olivia Martin, Ashesh Rambachan, David Ritzwoller, Jason Weitze, Yiqing Xu, and attendees of the Stanford Econometrics seminar series for their advice and feedback. I am fortunate to have my research supported by the Stanford Data Science Scholars Program and the B.F. Haley and E.S. Shaw Fellowship for Economics from the Stanford Institute for Economic Policy Research. Computational support was provided by the Data, Analytics, and Research Computing (DARC) group at the Stanford Graduate School of Business (RRID:SCR_022938).

1 Introduction

Settings with text outcomes are increasingly ubiquitous in economics and the social sciences. In contexts such as political speech (Gentzkow, Shapiro, and Taddy, 2019), corporate disclosures (Hoberg and Phillips, 2011), central bank communications (Shapiro and Wilson, 2022), judicial opinions (Ash, Chen, and Naidu, 2026), and consumer reviews (Luca, 2011), the content and style of text are of substantive interest. A standard approach to making text outcomes tractable for causal estimation methods (such as synthetic control) requires the researcher to specify in advance some low-dimensional representation that captures some chosen *aspect* of each document (e.g., word counts, sentiment scores, or topic shares).[1] In panel data settings, applied researchers can utilize methods such as synthetic control (Abadie, Diamond, and Hainmueller, 2010) to estimate causal effects on these low-dimensional representations. However, this approach has a key limitation: it requires the researcher to specify a representation in advance. Text is a high-dimensional and nuanced object, and so in general an ex-ante selected representation may be insufficiently rich to capture all dimensions of interest.[2] Moreover, even if the chosen representation is adequate, it may fail to take into account broader information in the text that may be useful for causal estimation. These limitations may be addressed by matching on text directly, but the existing literature has not developed extensions of panel data approaches, such as synthetic control, to full text documents.

In this paper, I develop a synthetic control approach which constructs counterfactuals not for some low-dimensional representation, but for the full distribution of text itself.[3] In my framework, each observation is a collection of documents from a given unit in a given time period. The untreated distribution of documents for the treated unit is then modeled as a mixture of contemporaneous untreated (or "donor") unit distributions, with weights estimated to minimize the pre-treatment discrepancy between the treated unit distribution and the induced mixture of donor units, according to some metric on the space of distributions over text.[4] This generalizes the distributional synthetic control approach of Gunsilius (2023) to the setting where the distributions are over high-dimensional text documents, rather than scalar outcomes.[5] My approach offers two key advantages: it can lead

[1] For examples of applied papers utilizing such estimation strategies on text, see: Gentzkow and Shapiro (2010); Baker, Bloom, and Davis (2016); Hansen, McMahon, and Prat (2018); Gentzkow, Kelly, and Taddy (2019).

[2] For example, consider a scenario in which treatment results in text becoming more confrontational. The researcher may anticipate this and opt to use a sentiment scorer to measure the level of negative sentiment in the text. However, while negative sentiment and confrontational tone are related concepts, the latter is sufficiently nuanced that the former may not adequately capture the treatment effect.

[3] This contributes to a growing methodological literature on causal inference with text and other unstructured data (Chen and Sarkar, 2020; Feder et al., 2022; Egami et al., 2022; Fong and Grimmer, 2023; Imai and Nakamura, 2024; Sarkar, 2026; Feldman et al., 2026; Freyaldenhoven, 2026), and also builds off the canonical work in synthetic control by Abadie, Diamond, and Hainmueller (2010).

[4] For tractability, I convert text documents to *embedding* vectors, which are typically the output of neural networks and have been shown to preserve semantic content of text (Mikolov et al., 2013; Pennington, Socher, and Manning, 2014; Vaswani et al., 2017; Muennighoff et al., 2023).

[5] This paper correspondingly also contributes to a broader literature on the synthetic control method, particularly the strand investigating the inclusion of high-dimensional micro-level data to improve synthetic controls (Robbins, Saunders, and Kilmer, 2017; Shi et al., 2022; Chen, 2023; Liu, 2025; Tian, Lee, and Panchenko, 2026).

to improved causal estimation for pre-specified aspects of the text and, more fundamentally, it allows for researchers to *learn* which aspects have been impacted by treatment.

Throughout, I consider an application to public comments from California city council meetings.[6] This application builds on Martin and Venugopal (2026), which assembles transcripts from approximately 25,000 meetings in 115 California municipalities between 2015 and 2024. Open-governance laws in California require city councils to deliberate in public and allocate time for members of the public to provide comment (Brown, 1953). At the outbreak of the COVID-19 pandemic in March 2020, Governor Newsom issued statewide emergency orders allowing city councils and public commenters to meet virtually, either via teleconferencing (e.g., Zoom) or telephone dial-in options. Martin and Venugopal (2026) document a staggered removal of these remote participation options in the aftermath of the pandemic, noting that the timing of the removal of remote participation options is not well-predicted by city observables. They then leverage this quasi-exogenous variation to investigate the causal effect of requiring participants to attend in-person on demographics, finding a reduction in the number of speakers but no clear change in observed speaker composition. Now, I answer a different question: does changing the cost and mode of participation alter what residents say once they speak?

To see the first key advantage of my method, consider the case of a researcher who is interested in studying the effect of some intervention on a pre-specified aspect of the text, the presence of which in a given document can be measured by some *summary function* g. For example, g may measure the prevalence of discussion around housing policy in a given public comment. Applying g to the text provides a low-dimensional representation (e.g., "housing intensity") which the researcher can balance on average between treated and donor units via convex weighting and obtain an estimate of the treatment effect on the treated unit τ_g, following the standard synthetic control approach (Abadie, Diamond, and Hainmueller, 2010). Suppose further that the researcher later chooses to investigate the effect on some other aspect of the text, such as "politeness", measured by a corresponding summary function g'. Estimating weights to balance on these two representations separately will generally lead to different donor weights, which may be unintuitive in a setting where the researcher believes that both aspects (e.g., discussion of housing and politeness) reflect the same latent unit characteristics. In contrast, by balancing on distributions of full text, the weights estimated by my method will be common to both g and g'.

In addition, learning weights which balance on the full text, rather than only on g, can aid in estimating the treatment effect on g, denoted τ_g, when other features of the text are predictive of the aspect measured by g. For example, the politeness of a given public comment may be informative for its discussion of housing policy, and so learning weights which balance this joint

[6] This application contributes to a literature on local government dynamics, including public participation (Warshaw, 2019; Einstein, Palmer, and Glick, 2019; Einstein et al., 2023). It also contributes to an applied literature utilizing synthetic control to estimate causal effects in observational panel settings (Abadie and Gardeazabal, 2003; Abadie, Diamond, and Hainmueller, 2015; Kleven, Landais, and Saez, 2013; Bohn, Lofstrom, and Raphael, 2014; Cunningham and Shah, 2018).

distribution can improve causal estimation. This is analogous to the incorporation of covariates into synthetic control, which the literature has shown can lead to better estimates of treatment effects in settings where covariates are predictive of untreated potential outcomes, or when limited pre-treatment outcome observations are available.[7]

The second key advantage of my approach is more fundamental: my estimator constructs plain-text counterfactuals. Since the estimated counterfactual is a mixture of donor unit distributions, the researcher can directly sample real text documents from this counterfactual by drawing from donor unit distributions, with the learned weights ensuring that the sampled corpus approximates the counterfactual corpus of the treated unit. I leverage this novel capability and propose a procedure for *hypothesis generation*, in which these two corpora are passed to some technology (for example, a large language model (LLM)) that identifies aspects of text (and corresponding summary functions) that vary maximally between the two corpora.[8] The researcher can then learn what specific aspects of text have been impacted by treatment and obtain a representation that is tailored to that learned aspect of interest. This may be particularly helpful in settings where the researcher does not have strong intuition for what aspects of text differ across treatment and control, or when treatment effects are on particularly nuanced aspects of text. For example, if the true treatment effect is on the prevalence of "arguments against affordable housing which invoke increased automobile traffic", simply measuring the prevalence of "housing" may not adequately capture the treatment effect, while hypothesizing from the data will ensure that the learned aspects are appropriately nuanced. More broadly, this hypothesis generation approach also allows researchers to test for the presence of *any* significant effect of a given intervention on text.

Furthermore, I show that the proposed estimator allows for statistical inference on both the synthetic control weights and the resulting treatment effect estimates. This is made possible by two features of my approach: a nonstandard asymptotic regime, and a particular choice of inter-distributional distance metric. I consider an asymptotic regime that is nonstandard for synthetic control and is similar to that utilized in Liu (2025), in which the number of units and periods remains fixed while the number of samples from each unit-period dimension grows large. This sampling dimension is absent from most conventional aggregate synthetic control studies (Abadie, Diamond, and Hainmueller, 2010; Almeida et al., 2025). Choosing maximum mean discrepancy (MMD) (Gretton et al., 2006) as an inter-distributional distance metric, I establish consistency of the estimated treatment effect $\widehat{\tau}_g$ for the true treatment effect τ_g, for some pre-specified summary

[7] Abadie (2021) provides a survey of covariate use in the synthetic control literature, while Kaul et al. (2015) and Botosaru and Ferman (2019) argue that including covariates can be helpful in the described settings.

[8] This relates to recent advances in the use of machine learning methods, including LLMs, as tools for hypothesis generation for unstructured data (Ludwig and Mullainathan, 2024; Ludwig, Mullainathan, and Rambachan, 2025; Modarressi, Spiess, and Venugopal, 2025).

function g.[9] I further show that, under suitable regularity conditions, the estimated weights are asymptotically normal, and correspondingly establish asymptotic normality of $\widehat{\tau}_g$, with a limiting variance that takes into account uncertainty in the weights used to construct the counterfactual. I show that this variance is itself consistently estimable, allowing for closed-form standard errors for $\widehat{\tau}_g$.

Inference for treatment effects on summary functions output by a data-driven hypothesis generation procedure is more challenging. As noted by Modarressi, Spiess, and Venugopal (2025), LLMs are a natural choice of hypothesis generation technology in the context of text data, but it is difficult to analyze the asymptotic behavior of LLM-provided summary functions $\widehat{g}$. Consequently, I consider consistency and inference for the treatment effect estimator $\widehat{\tau}_{\widehat{g}}$, conditional on the learned summary function $\widehat{g}$. Since $\widehat{g}$ is data-dependent, I introduce a sample-splitting mechanism in which the weights used to learn $\widehat{g}$ and construct $\widehat{\tau}_{\widehat{g}}$ are estimated on independent folds. Under this regime, I establish consistency and asymptotic normality of treatment effects conditional on the learned aspects. This sample-splitting procedure also provides the benefit of allowing the researcher to "guide" the hypothesis generation technology through repeated evaluations on the "training" fold of data, allowing the researcher to ensure that the learned summaries are of substantive interest.

As an extension, I consider the case where the synthetic control model is *misspecified*, meaning the true counterfactual is not expressible as a mixture over contemporaneous donor distributions. A key implication of the asymptotic regime under consideration is that by keeping the number of donor units fixed, the quality of the synthetic control approximation in the population is also fixed. Consequently, the assumption of correct specification is particularly substantive in this setting, as the size of the donor convex hull is no larger in the population than in finite sample. I therefore analyze the case of misspecification of the synthetic control model, providing estimation error bounds on $\widehat{\tau}_g$ and $\widehat{\tau}_{\widehat{g}}$. Formally, misspecification corresponds to a setting where the treated unit's untreated distribution is not expressible as a mixture of donor unit distributions. For example, in a setting where a collection of public comments in a politically moderate city cannot be viewed as a mixture of comments taken from liberal and conservative cities, but instead speaks about other topics entirely, a synthetic control used to match the moderate city as a mixture of liberal and conservative cities would be misspecified. Under such misspecification, I show that the estimated treatment effects are consistent for treatment effects relative to the *pseudo-counterfactual*, defined as the post-treatment mixture of donor distributions with mixture weights that minimize the pre-treatment distance between the treated unit's distribution and the donor mixture. I then show that minimizing distributional distance automatically bounds the misspecification-induced bias in treatment effect estimates for a wide, metric-dependent class of summary functions.

[9] MMD is a kernel-based distance metric on distributions, linking this work to the kernel balancing literature (Hazlett and Xu, 2018; Hazlett, 2020). More generally, this approach relates to an extensive literature on balancing in causal inference, in which weights are learned to match distributions of covariates between treated and control units (Zubizarreta, 2015; Doudchenko and Imbens, 2016; Kallus, 2017; Athey, Imbens, and Wager, 2018; Santra, Chen, and Park, 2026).

My application to California city council meeting public comments highlights the practical utility of this method and is of substantive interest. The question of how meeting format impacts speech is relevant to the design of local democratic institutions in particular and to remote work and the use of teleconferencing more broadly. Central to the appeal of decentralized governance in the literature is the ability of local governments to aggregate and respond to local preferences (Oates, 1999; Besley and Coate, 2003). These preferences are conveyed to the council through public comment, making the impact of meeting venue and accessibility salient to the ability of local governments to effectively fulfill this purpose. This application also speaks to a broader discussion on the efficacy of teleconferencing and remote work that has proliferated since the COVID-19 pandemic.[10] Indeed, the effects of remote work on productivity and interpersonal communication in industry settings have been studied in the literature (Bloom et al., 2015; Brucks and Levav, 2022; Bloom, Han, and Liang, 2024). In this paper, I extend this strand of the literature by examining the effects of the removal of remote participation on local government speech. I restrict attention to *unagendized*[11] public comments to reduce concerns about agenda-driven selection and, through two calibrated simulations in which comments are resampled from the data, show that my distributional estimator correctly identifies true weights and correctly accounts for unobserved confounding when generating hypotheses, significantly improving over a naive, equal-weighting benchmark. I then construct synthetic controls for multiple units which remove remote participation and find requiring public commenters to attend in-person leads to significantly increased civility and reduced hostility in public comments.

The remainder of the paper proceeds as follows. Section 2 provides notation and establishes counterfactual identification in the case of distributional text outcomes. Section 3 develops the sample estimator, establishing consistency of the estimated weights and induced pseudo-counterfactual as well as inference for the synthetic control weight vector. These results are shown to extend naturally to treatment effect estimates for summary functions measuring pre-specified aspects of the text. Section 4 proposes a procedure for learning aspects of text (and corresponding summary functions) that are impacted by treatment, establishing consistency and valid inference for the resulting treatment effect estimates conditional on the learned functions via sample-splitting. Section 5 relaxes the correct specification assumption and provides bounds on the resulting misspecification-induced error on the estimated treatment effects. Section 6 applies the method to public comment periods in California city council meetings, and Section 7 concludes.

[10] For example, see articles from the New York Times, the Los Angeles Times, and Yahoo Finance.

[11] Unagendized public comments are defined as those which are made during general public comment periods, rather than as part of scheduled debate on an agenda item under consideration by the city council.

2 Setting & Identification

In this section, we establish notation and define the estimand for counterfactual estimation via distributional synthetic control. We then provide conditions under which this estimand is equal to the true post-treatment counterfactual of the treated unit, and show that a generalization of the synthetic control objective function to the distributional setting yields correct weights in this setting.

2.1 Setting

Suppose that we have a collection of units indexed by $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$. For each unit and in each time period, we observe a collection of K_{it} documents x_{itk}, which are drawn from some underlying unit-period-specific distribution F_{it} with support on the set of possible text documents $\mathcal{X}$. In our applied setting of interest, x_{itk} corresponds to the kth public comment observed in a city council meeting from city i in time period t.

Treatment is modeled to induce an alteration to these distributions F_{it} from which our observations are drawn. As such, we define the distributional analog to the potential outcomes framework (Rubin, 1974) as *potential distributions* $\mathcal{F}_{it}(d_{it})$, where d_{it} denotes the treatment status for unit i in time period t. Analogous to the canonical synthetic control approach in Abadie, Diamond, and Hainmueller (2010), we suppose that unit N is treated in time period T, and all other unit-time cells are untreated.

In our empirical application, the treatment corresponds to the removal of remote participation options for public commenters; $d_{it} = 0$ if city i allows public commenters to attend the council meeting in time t remotely, and $d_{it} = 1$ if in-person attendance is required for participation. Throughout, we assume that there are no treatment spillovers, and that each unit's potential distribution of documents depends only on that unit's contemporaneous treatment assignment (e.g., the stable unit treatment value assumption, or SUTVA, is satisfied). As a consequence, we observe draws from $F_{it} = \mathcal{F}_{it}(0)$ for all $(i,t) \neq (N,T)$, and $F_{NT} = \mathcal{F}_{NT}(1)$. We assume that these draws are independent and identically distributed within each cell, and that there are no correlations across cells.[12]

Assumption DS (Document sampling)**.** *For each unit i and period t, the observed texts $\{x_{itk}\}_{k=1}^{K_{it}}$ are independent and identically distributed:*

$$x_{itk} \overset{\text{iid}}{\sim} \mathcal{F}_{it}(d_{it})$$

[12]Note that Assumption DS does not require independence of the distributions F_{it} across units and time periods. In the context of our applied setting, this assumption permits the distributions of comments within a city to exhibit some nonrandom trajectory over time, and neighboring/otherwise similar cities may exhibit similar distributions. Moreover, it does not preclude repeat participants, either within or across meetings, as their frequency of participation can be captured by F_{it}. By contrast, this assumption would be violated by speakers directly responding to each other; our choice to focus on unagendized public comments in the empirical application mitigates this risk in practice.

The collections $\{x_{itk}\}_{k=1}^{K_{it}}$ *are also mutually independent across cells.*

Let P_1 and P_0 denote the realizations of the treated and untreated potential distributions for the treated unit in the treated time period, $\mathcal{F}_{NT}(1)$ and $\mathcal{F}_{NT}(0)$, respectively. Our causal parameter of interest is the treatment effect on some interpretable summary of the text. Let $g : \mathcal{X} \to \mathbb{R}$ denote some *summary function* of the text, which measures the presence of some aspect of the text (e.g., topic, style, tone, etc.). We can then represent this treatment effect by:

$$\tau_g = \mathbb{E}_{P_1}[g(X)] - \mathbb{E}_{P_0}[g(X)] \tag{1}$$

This corresponds to the treatment effect on the treated unit for the summary function g. Note that we observe draws from P_1, but not from P_0; estimation of this counterfactual will be the primary objective of the following distributional synthetic control procedure.

2.2 Generalizing Distributional Synthetic Control

We seek to model the untreated potential distributions of the treated unit as a mixture over contemporaneous untreated potential distributions of the donor units. This is a generalization of the distributional synthetic control framework proposed by Gunsilius (2023), in which each cell of the panel matrix is a distribution over univariate outcomes. Here, each cell is instead a distribution over high-dimensional text documents. Correspondingly, it is natural to consider the following analogue of standard synthetic control in the population, where weights are chosen to minimize inter-distributional distance in the pre-treatment period:

$$L(w) = \sum_{t<T} \mathcal{D}^2 \left(\mathcal{F}_{Nt}(0), \sum_{i=1}^{N-1} w_i \mathcal{F}_{it}(0) \right) \qquad w^* = \arg\min_{w \in \mathcal{W}} L(w)$$

Here, $\mathcal{W}$ denotes the simplex in $\mathbb{R}^{N-1}$ and $\mathcal{D}$ is some distance defined on distributions.

Note that this formulation coincides with the cumulative distribution function (CDF)-based extension discussed by Gunsilius (2023), in the case where the distributions are defined over scalars (rather than text) and $\mathcal{D}$ is equal to the (square root of the) 1-Wasserstein distance.[13] Additionally, this formulation nests the standard scalar synthetic control objective function in Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010). In particular, suppose that the summary function g is specified in advance by the researcher, and replace $\mathcal{D}$ with $\mathcal{D}^{\text{scalar}}$ defined as:

$$\mathcal{D}^{\text{scalar}}(P, Q) = |\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]|$$

[13] In the implementation of Gunsilius (2023) presented in Gunsilius and Van Dijcke (2026), weights are learned to minimize the loss in each period, and then averaged across periods. My formulation here coincides with their approach in a modified setting where a single set of weights is found to minimize the pooled pre-treatment objective function.

The above objective function reduces to that of the standard scalar synthetic control, where the scalar outcome is the average value of the summary function g across the collection of documents in each cell.

Even if the researcher specifies the summary function in advance, however, utilizing weights based on matching the full distributions of text using $\mathcal{D}$ may yield tangible benefits over using those produced by balancing on the chosen summary alone via $\mathcal{D}^{\text{scalar}}$. First, in many settings it may be reasonable to assume that different aspects of the text, including the aspect of interest, reflect the same latent unit characteristics; constructing weights which balance on all such aspects simultaneously may then be desirable. This also provides the additional benefit of yielding a counterfactual corpus of text that is independent of the choice of g.[14] Moreover, balancing on the full texts may improve estimation of τ_g in settings where other aspects of text are predictive of the target aspect of interest, an advantage which is particularly salient in settings with limited pre-treatment observations. These potential benefits mirror those of including covariates in standard synthetic control, which have been shown to produce better treatment effect estimates in cases where the covariates are predictive of untreated potential outcomes or the number of pre-treatment periods is small (Kaul et al., 2015; Botosaru and Ferman, 2019; Abadie, 2021).

In our distributional setting, the linear combination $\sum_{i=1}^{N-1} w_i F_{it}$ denotes a *mixture distribution* over the constituent donor distributions. Sampling from this mixture distribution is akin to first sampling a donor unit from the multinomial distribution with probabilities w and then sampling from the donor unit's corresponding document distribution. Note that in this case, unlike in standard synthetic control, convexity of the synthetic control weights is entirely natural: if the weights were non-convex, the resulting linear combination would not produce a well-defined mixture distribution.

In finite sample, observed distributions are modeled as equally-weighted collections of point masses at the observed documents:

$$\widehat{F}_{it} = \frac{1}{K_{it}} \sum_{k=1}^{K_{it}} \delta_{x_{itk}}$$

Here, $\delta_{x_{itk}}$ denotes a unit mass at the text document x_{itk}. The sample objective function and weight estimator[15] are correspondingly given by:

$$\widehat{L}(w) = \sum_{t<T} \mathcal{D}^2 \left(\widehat{F}_{Nt}, \sum_{i=1}^{N-1} w_i \widehat{F}_{it} \right) \qquad \widehat{w} \in \arg\min_{w \in \mathcal{W}} \widehat{L}(w)$$

[14] This is related to the approach of Sun, Ben-Michael, and Feller (2025), which seeks to learn common synthetic control weights for multiple scalar outcomes. Learning weights to balance on text embeddings can be viewed as accomplishing the same goal, in the limit where we pre-specify an arbitrarily large number of summary functions.

[15] For the remainder of the paper, ties in the weight-estimation problems are broken according to some fixed, measurable selection rule, such as selecting the minimizer with smallest L_2 norm.

2.3 Choice of $\mathcal{D}$

We choose as our distance metric maximum mean discrepancy (MMD), evaluated on *embeddings* of the text documents. MMD is a kernel-based integral probability metric proposed by Gretton et al. (2006) which measures the distance between two distributions as the largest difference in expectation of functions which are in the unit ball of the reproducing kernel Hilbert space (RKHS) induced by a chosen kernel k, denoted $\mathcal{H}_k$:

$$\text{MMD}_k(P, Q) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_P[f(X)] - E_Q[f(X)] \tag{2}$$

Intuitively, an RKHS is a collection of smooth functions endowed with a notion of norm and angle, much as vectors in Euclidean space.[16] In order for these kernels to be well-defined functions, we convert text documents to *embeddings*, numeric representations of text in some Euclidean vector space. In particular, we suppose that we have access to an embedding technology $\mathcal{E} : \mathcal{X} \to \mathcal{B} \subseteq \mathbb{R}^d$ that maps from text documents in set $\mathcal{X}$ to some subset of d-dimensional Euclidean space. In practice, this embedding function $\mathcal{E}$ will typically be an API call to some pretrained embedding model.[17] There is an extensive literature which highlights the ability of these embedding functions to capture the semantic meaning of text (Mikolov et al., 2013; Pennington, Socher, and Manning, 2014; Devlin et al., 2019; Reimers and Gurevych, 2019). Indeed, embeddings are utilized in our most cutting-edge models of text, including large language models (Vaswani et al., 2017; Brown et al., 2020). We will therefore adopt a standard approach in the natural language processing literature, treating embedding similarity as a proxy for semantic similarity of texts (Muennighoff et al., 2023). Correspondingly, when P and Q are distributions over text documents, we will evaluate their distance based on application of the kernel k to the embeddings of their underlying documents:

$$\text{MMD}_k(P, Q) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_P[f(\mathcal{E}(X))] - E_Q[f(\mathcal{E}(X))] \tag{3}$$

Gretton et al. (2006) further show that MMD admits a dual representation in which the distance between two distributions is equivalently expressed as the difference in kernel mean embeddings[18], allowing us to write:

$$\text{MMD}_k(P, Q) = \| \mathbb{E}_P[k(\mathcal{E}(X), \cdot)] - \mathbb{E}_Q[k(\mathcal{E}(X), \cdot)] \|_{\mathcal{H}_k} \tag{4}$$

Applied to our setting, we will henceforth let $\mu_{it}(\cdot)$ denote the kernel mean embedding of $\mathcal{F}_{it}(d_{it})$:

$$\mu_{it}(\cdot) = \mathbb{E}_{\mathcal{F}_{it}(d_{it})}[k(\mathcal{E}(X), \cdot)] \in \mathcal{H}_k$$

[16] For additional discussion on RKHS and MMD_k, see Appendix A.1.

[17] Further, since text documents are finite combinations of a countable set of words, this embedding map $\mathcal{E}$ is automatically measurable.

[18] This is analogous to Kantorovich-Rubinstein duality for optimal transport metrics.

Note that $\mu_{it}(\cdot)$ is a function in the RKHS induced by the kernel k. For notational brevity, we will generally drop the argument and simply refer to this function as μ_{it}.[19]

This choice of distance metric contrasts with that used in related strands of the literature, in which optimal transport metrics, such as Wasserstein distances, are popular (Gunsilius, 2023). These metrics have an intuitive and appealing geometric interpretation: they capture the minimum cost of moving mass from one distribution to another, given some baseline metric on the support space. However, they also suffer from a curse of dimensionality, with the p-Wasserstein distance between two empirical distributions converging to the population counterpart at rates as slow as $O(n^{-1/d})$ (Weed and Bach, 2017). Such metrics are consequently ill-suited to measuring the distance between distributions of high-dimensional text embeddings.[20]

MMD, on the other hand, converges at dimension-independent $O(n^{-1/2})$ rates (Gretton et al., 2012). It is also further motivated by the kernel balancing literature (Hazlett and Xu, 2018; Hazlett, 2020). In the kernel balancing problem, weights are chosen so that treated and control groups are approximately balanced on a rich, kernel-induced function space. As a consequence, kernel balancing can be equivalently viewed as minimizing the MMD_k between the treated and control covariate distributions, and its adoption here is an extension to a setting in which we seek to balance trajectories of distributions of high-dimensional text embeddings. Throughout, we will make the following regularity assumptions on the chosen kernel k:

Assumption k-Reg (Kernel regularity)**.** *The kernel k is fixed, measurable, characteristic, and positive definite. It further has a separable RKHS $\mathcal{H}_k$ and is bounded by $\bar{k}$:*

$$\sup_{b \in \mathcal{B}} k(b, b) \leq \bar{k} < \infty$$

Here, characteristicness of a kernel refers to the induced map from distributions to kernel mean embeddings being injective; positive definiteness means that every kernel matrix produced by kernel k is positive semidefinite; having a separable RKHS means that any function $h \in \mathcal{H}_k$ can be approximated arbitrarily closely in RKHS norm by some member of a countable collection of functions $f_1, f_2, \ldots \in \mathcal{H}_k$. Measurability and boundedness allow expectations of the kernel applied to random observations to be well-defined and finite. Examples of kernels that satisfy these requirements include the Gaussian RBF kernel, the rational quadratic kernel, the Laplace kernel, and the Matérn kernel (Steinwart and Christmann, 2008; Sriperumbudur et al., 2010). Note that

[19] MMD in this context is a pseudometric on text distributions, and a proper metric on embeddings; under an additional injectivity assumption on $\mathcal{E}$, MMD will be a proper metric on text as well. However, such an assumption is unnecessary for our purposes here.

[20] Prior work has shown that asymptotics for optimal transport distances on empirical distributions can generally lead to non-standard limits. Del Barrio, Giné, and Matrán (1999) show that the 1-Wasserstein distance between empirical and population distributions has a non-Gaussian limit, and Sommerfeld and Munk (2018) show a similar result for p-Wasserstein distances. Del Barrio and Loubes (2019) show that under certain regularity conditions, Gaussian limits are available, but are centered at the expected empirical transport cost. This intractability provides additional warning against application of optimal transport distances in this setting.

for all of these, fixedness of the kernel (as stated in the assumption) requires fixedness of their relevant parameters (e.g., the bandwidth for the Gaussian RBF and Laplace kernels).

In our empirical application, we will utilize the Gaussian RBF kernel:

$$k(b, b') = \exp\left(-\frac{\|b - b'\|_2^2}{2\sigma^2}\right)$$

Here, the parameter σ is termed the "bandwidth", and is set equal to the median inter-embedding distance on a pilot sample of data.[21] This kernel is attractive due to its ubiquity and the fact that it is "universal": any continuous function can be uniformly approximated arbitrarily well by elements of its RKHS, on any compact subset of Euclidean space (Steinwart, 2001; Steinwart and Christmann, 2008). All theoretical results will be established for a generic kernel that satisfies Assumption k-Reg.

However, despite offering the aforementioned benefits, MMD is in general a "weaker" metric than Wasserstein distances, with recent work showing that the latter actually bound the former under sufficient regularity conditions (Vayer and Gribonval, 2023). This comparative weakness also manifests in the classes of functions which each metric disciplines: 1-Wasserstein distance controls differences in expectations of all 1-Lipschitz functions between two distributions, while MMD does the same for all functions in the RKHS unit ball (Villani, 2009; Gretton et al., 2012). For a Gaussian RBF kernel, the RKHS unit ball is contained in a uniformly Lipschitz class, with a bandwidth-dependent constant (Vayer and Gribonval, 2023), and so Wasserstein distances in this sense discipline a broader class of functions than MMD.

While we are unable to show in general that MMD controls for variation in distributions to the same extent as stronger optimal transport metrics, empirically it appears to capture the salient aspects of text well. Table 1 provides a representative example of (abridged) public comments from California city council meetings and their kernel-measured similarity relative to a fixed baseline comment. The baseline comment is quite formal and talks about housing availability and tenants' rights; texts which are close to it, as measured by the kernel, share these qualities, while texts that are far away differ significantly in content and tone.

Using MMD_k as our notion of distributional distance, our objective function becomes:

$$L(w) = \sum_{t<T} \text{MMD}_k^2 \left(\mathcal{F}_{Nt}(0), \sum_{i=1}^{N-1} w_i \mathcal{F}_{it}(0)\right) \tag{5}$$

Moving forward, we will let $\mathcal{W}_0 = \arg\min_{w \in \mathcal{W}} L(w)$ denote the set of population minimizers of this objective function. We can then construct our counterfactual in practice by numerically minimizing

[21] All following assumptions and results are understood to be conditional on this pilot sample, which is separate from the data used in all further steps, so that the selected kernel and associated population targets are fixed.

Table 1: Sample public comments and kernel-measured similarities

Similarity Percentile	Text
Baseline	*I am speaking for the Oakland Tenant's Union...runaway inflation has meant that the cost of goods and services have jumped up, increasing tenants' everyday costs...Oakland Tenants Union is asking that the City Council issue an emergency ordinance setting aside the CPI as an indicator for this rent increase.*
89.80	*I'm a member of the Unbound Housing and Business Network...Eviction moratorium violates the Fifth Amendment. It is government-enabled theft. It is unjust and immoral, and it must end now.*
59.99	*As a public meeting, not only should we be able to see the city council members, but we should be able to see each other...And that is one of the things that we've lost when we went online...I would strongly suggest that you do that to continue the publicness of these public meetings.*
16.90	*I'm with the Italian Americans for Fair Peace Standards....And we are here calling today to applaud the efforts removing [redacted], fake Italian, disgraceful to all Italians, from his position as president pro tempore. [redacted], you are a disgrace to our beautiful country, and we would like you to stop associating your pathetic, miserable self with us.*

Notes: The similarity percentile of a given comment refers to where in the distribution of inter-comment kernel-based similarity scores the similarity of the given comment and the baseline comment falls. Similarity between two comments is measured directly by the kernel function applied to the two comment embeddings. Identifying information is redacted.

the objective function on the empirical distributions:

$$\widehat{L}(w) = \sum_{t<T} \text{MMD}_k^2 \left(\widehat{F}_{Nt}, \sum_{i=1}^{N-1} w_i \widehat{F}_{it} \right) \qquad \widehat{w} \in \arg\min_{w \in \mathcal{W}} \widehat{L}(w)$$

This will be our main specification for the remainder of this paper. However, this is readily extendable to settings with non-text covariates; see Appendix A.2 for further details.

2.4 Identification

In this subsection, we establish conditions under which $\mathcal{F}_{NT}(0)$ is identified, and when the weights which minimize our objective function in the population are unique. In order for $\mathcal{F}_{NT}(0)$ to be identified, we require two conditions to hold. First, we require that, in the case where multiple weight vectors w minimize the population objective, all optimal weights induce the same mixture distribution.

Assumption PP (Pre-period embedding equivalence implies post-period text equivalence)**.** *For*

any $w, w' \in \mathcal{W}_0$, *if*

$$\sum_{i=1}^{N-1} w_i \mu_{it} = \sum_{i=1}^{N-1} w'_i \mu_{it}$$

for all $t < T$, *then*

$$\sum_{i=1}^{N-1} w_i F_{iT} = \sum_{i=1}^{N-1} w'_i F_{iT}$$

Second, $\mathcal{F}_{NT}(0)$ must be expressible as a mixture of donor distributions in every time period, for some convex weighting vector w^*. This corresponds to the synthetic control model being correctly specified.

Assumption CS (Correct specification)**.** $\exists w^* \in \mathcal{W}$ *such that, for every* t,

$$\mathcal{F}_{Nt}(0) = \sum_{i=1}^{N-1} w_i^* \mathcal{F}_{it}(0)$$

Under these two conditions, we are able to establish identification of $\mathcal{F}_{NT}(0)$; that is, the true mixture weights w^* minimize our population objective function in Equation 5, and any population minimizer induces the true counterfactual distribution.

Proposition 1 (Identification of $\mathcal{F}_{NT}(0)$)**.** *Suppose Assumptions CS and k-Reg hold. Then,* $L(w^*) = 0$ *and*

$$w^* \in \arg\min_{w \in \mathcal{W}} L(w)$$

If additionally Assumption PP holds, we have

$$w^0 \in \arg\min_{w \in \mathcal{W}} L(w) \implies \mathcal{F}_{NT}(0) = \sum_{i=1}^{N-1} w_i^0 \mathcal{F}_{iT}(0)$$

One possible justification of Assumption CS is via a generalization of the factor model structure used to motivate synthetic control in Abadie, Diamond, and Hainmueller (2010). Suppose that there exists some set of B latent topics, with time evolution in how they are discussed. The distribution Q_{bt} captures how discussion around topic b looks in period t, while λ_{ib} governs the prevalence of discussion of topic b in unit i.

$$\mathcal{F}_{it}(0) = \sum_{b=1}^{B} \lambda_{ib} Q_{bt}$$

Then, the standard synthetic control assumption that the treated unit's loadings λ_N lie in the convex hull of the donor units' loadings implies Assumption CS.[22] This framing also clarifies an advantage of learning weights to balance distributions via $\mathcal{D}$, rather than $\mathcal{D}^{\text{scalar}}$, even when the

[22] This is simply one possible justification; see Lemma 3 in Appendix C for further details.

summary function g is specified in advance. As discussed in Section 2.2, if the underlying process in the data drives variation at the level of the text, as is outlined in this factor model, then it is intuitively preferable to directly learn weights which balance on these primitives, rather than the particular aspect measured by g.

However, Assumption CS is substantive, and may not hold in general. In such a setting, our counterfactual estimand will be the *pseudo-counterfactual*:

$$\widetilde{P}_0 := \sum_{i=1}^{N-1} w_i^0 F_{iT} \qquad w^0 \in \mathcal{W}_0 := \arg\min_{w \in \mathcal{W}} L(w)$$

Correspondingly, our treatment effect estimand will be the treatment effect relative to this pseudo-counterfactual:

$$\widetilde{\tau}_g = \mathbb{E}_{P_1}[g(X)] - \mathbb{E}_{\widetilde{P}_0}[g(X)] \tag{6}$$

Note that Assumption PP remains salient in this setting as well, as it allows for uniqueness of this pseudo-counterfactual even in the case of nonunique optimal weights w^0. Further, observe that if both Assumption PP and Assumption CS hold, Proposition 1 establishes $\widetilde{P}_0 = P_0$, and consequently $\widetilde{\tau}_g = \tau_g$.

Regardless of whether or not our model is correctly specified, we can also establish conditions under which the optimal weights in the population are unique. If we impose an additional affine-independence condition on the trajectory of kernel mean embeddings, then we can additionally recover unique optimal weights in the general case, where Assumption CS need not hold.

Assumption IDT (Affinely independent donor kernel mean trajectories)**.** *For every* $h \in \mathbb{R}^{N-1}$,

$$\mathbf{1}^\top h = 0 \quad \textit{and} \quad \sum_{i=1}^{N-1} h_i \mu_{it} = 0 \quad \textit{for every } t < T \quad \Longrightarrow \quad h = 0$$

Where $\mathbf{1}$ *denotes the* $(N-1)$*-dimensional vector of 1s.*

Intuitively, Assumption IDT requires that distinct weight vectors must produce mixture distributions with distinct pre-treatment trajectories. In particular, it prohibits redistributing weight among donor units while maintaining an unchanged mixture distribution in every period. Note that if Assumption IDT holds, then Assumption PP is satisfied trivially. Under this additional assumption, the resulting optimal weights are unique.

Proposition 2 (Unique optimal weights)**.** *Suppose Assumptions k-Reg and IDT hold. Then,* $\mathcal{W}_0 = \{w^0\}$.

Moving forward, we will assume that the model is correctly specified and that $\widetilde{P}_0 = P_0$. In the case of misspecification, we are able to bound the bias induced in our estimates by $\mathrm{MMD}_k(\widetilde{P}_0, P_0)$; see Section 5 for results and discussion.

3 Estimation & Inference

Let us now turn to the estimation of the synthetic control model in finite samples. In this section, we establish consistency of the estimated weights and produced counterfactual under an asymptotic regime in which the size of the panel is fixed and the number of draws from each potential distribution grows. In this asymptotic regime, we further establish asymptotic normality of the weights. These results allow us to establish consistency and asymptotic normality of $\widehat{\tau}_g$, in the case where the summary function g is specified in advance.

3.1 Asymptotic Regime

We consider a nonstandard asymptotic regime in which the number of documents sampled in each cell of the panel matrix grows large: $K_{it} \to \infty$. This regime is closely related to that used in Liu (2025). This asymptotic regime captures uncertainty based on sampling documents in each unit-period cell; this is natural in this setting, where each cell of our panel is a distribution, but if instead our outcome were a scalar aggregate such as GDP, sampling error would be extremely small and this regime would imply that ordinary synthetic control has no uncertainty. We correspondingly define minimum per-cell counts across the entire panel, in the pre-treatment period only, and in the post-treatment period only, respectively:

$$K_{\min} = \min_{i,t} K_{it} \qquad K_{\min}^{\text{pre}} = \min_{i,t<T} K_{it} \qquad K_{\min}^{\text{post}} = \min_{i} K_{iT}$$

Note that $K_{\min} \to \infty$ implies that both $K_{\min}^{\text{pre}} \to \infty$ and $K_{\min}^{\text{post}} \to \infty$. In order to ensure well-behaved sample growth for downstream inference results, we introduce the following assumption, which ensures that the samples in each pre-treatment and post-treatment cell do not grow at extremely unbalanced rates:

Assumption BSG (Balanced sample growth)**.** *For every $i = 1, \ldots, N$ and $t < T$, as $K_{\min}^{\text{pre}} \to \infty$,*

$$\frac{K_{it}}{K_{\min}^{\text{pre}}} \to \kappa_{it}^{\text{pre}} \in [1, \infty)$$

Additionally, as $K_{\min} \to \infty$,

$$\frac{K_{iT}}{K_{\min}} \to \kappa_{iT}^{\text{post}} \in [1, \infty)$$

This asymptotic regime stands in contrast with standard approaches in the synthetic control literature, which typically considers the number of donor units N or the number of pre-treatment time periods T, or both, growing large (Ferman, 2021; Ferman and Pinto, 2021). The choice of asymptotic regime has proved fraught when it comes to providing inference results for synthetic control, with the bulk of current methods relying on placebo tests rather than analytic results (Abadie, Diamond, and Hainmueller, 2010; Ferman and Pinto, 2017; Almeida et al., 2025).

Our nonstandard asymptotic regime comes with some key benefits, chief among them that we can produce results for consistency and inference of synthetic control weights that are intractable in typical asymptotic regimes. It is important to note, however, that the sources of uncertainty addressed by this asymptotic regime are different than those addressed by more standard regimes, particularly $N \to \infty$. In our study, N remains fixed in the population; therefore, the quality of the synthetic control fit in the population is fixed as well. The assumption of correct specification is therefore particularly substantive in our setting, where we are unable to consider an asymptotic sequence along which the addition of new donor units expands the convex hull until it eventually encompasses P_0. We therefore pay special attention to the case of misspecification of the synthetic control model; for further discussion, see Section 5.

3.2 Consistency & Inference for Estimated Synthetic Control

In this section, we establish consistency and asymptotic normality of the estimated weights, as well as consistency of the induced counterfactual estimate, under our nonstandard asymptotic regime. In doing so, we will extend canonical results for extremum estimators (Newey and McFadden, 1994), of which $\hat{w}$ is an example.

We begin with consistency: under our standard regularity conditions on the kernel k, we are able to consistently estimate both the synthetic control weights and their induced counterfactual.

Proposition 3 (Consistency of estimated weights and counterfactuals)**.** *Suppose Assumptions DS and k-Reg hold. Let $d(\cdot, \cdot)$ denote Euclidean point-to-set distance:*

$$d(\hat{w}, \mathcal{W}_0) = \inf_{w \in \mathcal{W}_0} \|\hat{w} - w\|$$

Then, as $K_{\min}^{\text{pre}} \to \infty, d(\hat{w}, \mathcal{W}_0) \xrightarrow{p} 0$. If $\mathcal{W}_0 = \{w^0\}$, then $\hat{w} \xrightarrow{p} w^0$.

If Assumptions PP and CS hold as well, then as $K_{\min} \to \infty$,

$$\text{MMD}_k \left(\sum_{i=1}^{N-1} \hat{w}_i \widehat{F}_{iT}, P_0 \right) \xrightarrow{p} 0$$

We now proceed to characterize the limiting distribution of $\sqrt{K_{\min}^{\text{pre}}}(\widehat{w} - w^0)$. While the consistency result above allows the population minimizer set $\mathcal{W}_0$ to contain more than one weight vector, provided that all population minimizers induce the same counterfactual distribution, a well-defined limiting distribution for the full weight vector requires stronger conditions. We will now proceed to give such a limit when the population minimizer is unique and lies in the interior of the simplex.

Let us begin by building some intuition. First, recall that Newey and McFadden (1994) show the following for a generic extremum estimator $\widehat{\theta}$, with estimand θ^0, objective function $\ell(\theta)$, and sample size n:

$$\sqrt{n}(\widehat{\theta} - \theta^0) = \underbrace{\left(-\nabla^2_\theta \ell(\theta^0)\right)^{-1}}_{-H_\theta^{-1}} \left(\sqrt{n}\nabla_\theta \widehat{\ell}(\theta^0)\right) + o_p(1)$$

Consequently, subject to the Hessian H being invertible and consistently estimable, the key ingredient to asymptotic normality of $\widehat{\theta}$ is showing that the score function admits a central limit theorem (CLT); that is, the second parenthetical term above is asymptotically normal. To see how we can establish this result in our setting, note that we can write our objective function as the sum of squared RKHS norms:

$$\widehat{L}(w) = \sum_{t<T} \left\| \widehat{\mu}_{Nt} - \sum_{i=1}^{N-1} w_i \widehat{\mu}_{it} \right\|^2_{\mathcal{H}_k}$$

Differentiating and evaluating the score at the population optimum then yields:

$$\frac{\partial \widehat{L}(w^0)}{\partial w_j^0} = -2 \sum_{t<T} \left\langle \widehat{\mu}_{jt}, \underbrace{\widehat{\mu}_{Nt} - \sum_{i=1}^{N-1} w_i^0 \widehat{\mu}_{it}}_{\widehat{r}_t^0} \right\rangle_{\mathcal{H}_k}$$

Here, $\langle x, y \rangle_B$ denotes the inner product of x and y in the space B. Note that we have abstracted away from the sum-to-one constraint here for simplicity.[23] Using $\widehat{r}_t^0$ to denote the residual in the second term of the above inner product, we can write the difference between the empirical and population values of this score:

$$\frac{\partial \widehat{L}(w^0)}{\partial w_j^0} - \frac{\partial L(w^0)}{\partial w_j^0} = -2 \sum_{t<T} \left\{ \left\langle \widehat{\mu}_{jt}, \widehat{r}_t^0 \right\rangle_{\mathcal{H}_k} - \left\langle \mu_{jt}, r_t^0 \right\rangle_{\mathcal{H}_k} \right\} \tag{7}$$

By bilinearity of the inner product, we can expand the first term in each element of the above sum:

$$\begin{aligned} \left\langle \widehat{\mu}_{jt}, \widehat{r}_t^0 \right\rangle_{\mathcal{H}_k} &= \left\langle \mu_{jt} + \underbrace{(\widehat{\mu}_{jt} - \mu_{jt})}_{\Delta \mu_{jt}}, r_t^0 + \underbrace{(\widehat{r}_t^0 - r_t^0)}_{\Delta r_t} \right\rangle_{\mathcal{H}_k} \\ &= \langle \mu_{jt}, r_t^0 \rangle_{\mathcal{H}_k} + \langle \mu_{jt}, \Delta r_t \rangle_{\mathcal{H}_k} + \langle \Delta \mu_{jt}, r_t^0 \rangle_{\mathcal{H}_k} + \langle \Delta \mu_{jt}, \Delta r_t \rangle_{\mathcal{H}_k} \end{aligned}$$

Plugging this back into Equation 7 yields:

$$\frac{\partial \widehat{L}(w^0)}{\partial w_j^0} - \frac{\partial L(w^0)}{\partial w_j^0} = -2 \sum_{t<T} \left\{ \langle \mu_{jt}, \Delta r_t \rangle_{\mathcal{H}_k} + \langle \Delta \mu_{jt}, r_t^0 \rangle_{\mathcal{H}_k} + \langle \Delta \mu_{jt}, \Delta r_t \rangle_{\mathcal{H}_k} \right\} \tag{8}$$

[23] This summation restriction can be handled in general by subtracting from each μ_{it} the kernel mean embedding for some reference unit.

Now, the first two terms contain first-order sampling variation, while the final is the product of two estimation errors, and is therefore second-order. In particular, boundedness of the kernel via Assumption k-Reg implies that each empirical kernel mean satisfies $\hat{\mu}_{it} - \mu_{it} = O_p(K_{it}^{-1/2})$ in $\mathcal{H}_k$. Under Assumption BSG, this yields:

$$\|\Delta\mu_{jt}\|_{\mathcal{H}_k} = O_p\big((K_{\min}^{\text{pre}})^{-1/2}\big) \qquad \|\Delta r_t\|_{\mathcal{H}_k} = O_p\big((K_{\min}^{\text{pre}})^{-1/2}\big) \qquad \langle \Delta\mu_{jt}, \Delta r_t\rangle_{\mathcal{H}_k} = O_p\big((K_{\min}^{\text{pre}})^{-1}\big)$$

Consequently, $\sqrt{K_{\min}^{\text{pre}}}\langle \Delta\mu_{jt}, \Delta r_t\rangle_{\mathcal{H}_k} = o_p(1)$, and we can write:

$$\sqrt{K_{\min}^{\text{pre}}}\left\{\frac{\partial \widehat{L}(w^0)}{\partial w_j^0} - \frac{\partial L(w^0)}{\partial w_j^0}\right\} = -2\sum_{t<T}\left\{\left\langle \mu_{jt}, \sqrt{K_{\min}^{\text{pre}}}\Delta\mu_{Nt} - \sum_{i=1}^{N-1} w_i^0\sqrt{K_{\min}^{\text{pre}}}\Delta\mu_{it}\right\rangle_{\mathcal{H}_k} + \left\langle \sqrt{K_{\min}^{\text{pre}}}\Delta\mu_{jt}, r_t^0\right\rangle_{\mathcal{H}_k}\right\} + o_p(1)$$

Note that we have used the definition of Δr_t to rewrite the first term.

Under our sampling condition in Assumption DS (in addition to Assumptions k-Reg and BSG) the collection $\left\{\sqrt{K_{\min}^{\text{pre}}}\Delta\mu_{it}\right\}_{i,t<T}$ has a joint Hilbert-space Gaussian limit. A Gaussian element of $\mathcal{H}_k$ is the Hilbert-space analogue of a multivariate normal random vector, where instead of being normally distributed itself it rather has a normally distributed inner product in every direction. More formally, a random element $G \in \mathcal{H}_k$ is Gaussian if $\langle G, h\rangle_{\mathcal{H}_k}$ is normally distributed for every $h \in \mathcal{H}_k$. Consequently, each inner product in the expression above converges in distribution to a normal random variable, and the entire right hand side is a linear functional of jointly Gaussian random elements. Since this holds for every element j with respect to which we differentiate on the left hand side, this establishes joint normality of the score:

$$\sqrt{K_{\min}^{\text{pre}}}\left\{\nabla_w \widehat{L}(w^0) - \nabla_w L(w^0)\right\} \xrightarrow{d} \mathcal{N}(0, \Omega_w) \tag{9}$$

We have therefore established a CLT for the score. Armed with this intuition, we can now state our asymptotic normality result formally.

Proposition 4 (Asymptotic normality of unique interior weights)**.** *Suppose Assumptions DS, k-Reg, IDT, and BSG hold. Suppose further that $w_i^0 > 0$ for every $i = 1, \ldots, N-1$. Then,*

$$\sqrt{K_{\min}^{\text{pre}}}(\hat{w} - w^0) \xrightarrow{d} \mathcal{N}(0, \Sigma_w)$$

with

$$\Sigma_w = AH^{-1}\Omega H^{-1}A^\top \tag{10}$$

The covariance matrix Σ_w is singular in $\mathbb{R}^{N-1}$.

In Equation 10, H and Ω denote the Hessian and limiting variance of the score function in terms of free coordinates $(w_1, w_2, \ldots, w_{N-2})$[24], respectively, while A denotes the Jacobian of the change of variables that reintroduces the sum-to-one-constraint, from the free coordinates to the full vector w. See Appendix D for additional details. Moreover, each of the components of this limiting variance is consistently estimable, allowing us to estimate standard errors in practice.[25]

Proposition 5 (Feasible covariance estimation)**.** *Suppose the conditions of Proposition 4 hold and let* $\widehat{\Sigma}_w$ *denote the sample analog of* Σ_w*. Then, as* $K_{\min}^{\text{pre}} \to \infty, \widehat{\Sigma}_w \xrightarrow{p} \Sigma_w$*.*

As a result, we have an analytic expression for the covariance of the synthetic control weights that is itself consistently estimable. This allows us to perform inference on the weights and obtain standard errors, a result that stands in stark contrast to the well-documented difficulties in characterizing standard errors for standard synthetic control (Almeida et al., 2025) and highlights one of the key advantages of our nonstandard asymptotic regime.

However, note that this asymptotic normality result hinges on two key assumptions on the population weights w^0: uniqueness and interiority. As we have argued in Proposition 2, uniqueness of w^0 can be established via an affine independence restriction on the donor-kernel mean trajectories. While substantive, this condition is not on its face terribly burdensome; one sufficient condition would be for the donor distributions to have linearly independent kernel mean embeddings in each pre-treatment period. The interiority restriction, however, is quite strong. Synthetic control weights are typically sparse, and so will therefore not in general lie in the interior of the simplex (Abadie, 2021). However, boundary population weights will generally yield a limiting distribution that is the projection of a Gaussian vector onto the critical cone.[26] While analytically derivable, this poses significant challenges to implementation in practice.

3.3 Tractable Two-Stage Weight Estimation & Inference

In this section, we provide a two-stage weight estimation procedure to provide interior weights, a key condition of the asymptotic normality result of Proposition 4. In particular, we argue that even if w^0 is not interior, it suffices to know the coordinates of w^0 that are exactly zero. We accomplish this by fitting a preliminary estimate $\widetilde{w}$, truncating its smallest elements to zero, and re-estimating a second-stage weight estimate $\hat{w}^{\text{ts}}$ on the restricted simplex. The resulting weight estimator lies in the interior of the restricted simplex with probability approaching one, and all

[24] w_{N-1} is uniquely determined by these free coordinates and the summation constraint: $w_{N-1} = 1 - \sum_{j=1}^{N-2} w_j$.

[25] On the event that $\widehat{H}$ is singular, define $\widehat{\Sigma}_w = 0$.

[26] The key intuition for this lies in the fact that if the population weights for a given unit are strictly positive, the nonnegativity constraints are locally inactive, and a consistent estimator will satisfy these constraints with probability approaching one. Asymptotically, this estimator can therefore be treated as an unconstrained estimator, with only the summation to one constraint binding. However, when the true weight for a given unit is exactly zero, the nonnegativity constraint is locally active, and negative perturbations are infeasible.

preceding consistency and inference results apply naturally.[27]

Under Assumption IDT, Proposition 2 gives us a unique population optimum w^0 and we can define the population donor support as

$$\mathcal{S}_0 = \left\{ i \in \{1, \ldots, N-1\} : w_i^0 > 0 \right\}$$

Note that the population zero-weight set is simply the complement of $\mathcal{S}_0$ in the set of donor units.

Our first step in this two-stage procedure is to produce an estimate of $\mathcal{S}_0$, which we shall denote $\widehat{\mathcal{S}}$. Let $\widetilde{w}$ denote the unrestricted estimator, which minimizes the empirical counterpart of Equation 5 across the full donor simplex:

$$\widetilde{w} \in \arg\min_{w \in \mathcal{W}} \widehat{L}(w)$$

As $K_{\min}^{\text{pre}}$ becomes arbitrarily large, we can construct $\widehat{\mathcal{S}}$ by examining which elements of $\widetilde{w}$ are larger than a cutoff schedule that shrinks to zero asymptotically more slowly than the convergence rate of the weights themselves.

Proposition 6 (Consistent active-donor set recovery)**.** *Suppose Assumptions DS, k-Reg, and IDT hold. Suppose further that there exists a deterministic sequence $\lambda_K > 0$ such that $\lambda_K \to 0$ and $\sqrt{K_{\min}^{\text{pre}}}\lambda_K \to \infty$, and define:*

$$\widehat{\mathcal{S}} = \{ i \in \{1, \ldots, N-1\} : \widetilde{w}_i > \lambda_K \}$$

Then, as $\sqrt{K_{\min}^{\text{pre}}} \to \infty$,

$$\Pr\left(\widehat{\mathcal{S}} = \mathcal{S}_0 \right) \longrightarrow 1$$

We have therefore established that we can consistently estimate the set of active donor units by implementing a cutoff on the elements of the unrestricted weight estimate.[28] It now remains to establish that re-estimating weights on this restricted set of donor units retains the previously established consistency and asymptotic normality properties of the unrestricted estimator. We can now define the second-stage weight estimator as minimizing the empirical objective function over

[27] This approach is closely related to the results in the factor model literature, in which the number of factors to approximate is estimated (Bai and Ng, 2002; Bai, 2003). It is also analogous to post-model-selection refitting in settings with variable selection (Belloni and Chernozhukov, 2013), and to related procedures which truncate assignment probabilities in experimental design (Chen and Andrews, 2026). However, our approach here yields pointwise asymptotic normality for a fixed population weight vector, rather than uniform validity over sequences of populations.

[28] If the estimated set is empty, set $\widehat{\mathcal{S}} = \{j\}$, where j is the first index of an element of $\widetilde{w}$ that is maximal. The condition that $\sqrt{K_{\min}^{\text{pre}}}\lambda_K \to \infty$ is analogous to similar conditions in the variable selection literature (Pötscher and Schneider, 2011) and implies that λ_K must decline to zero at a rate slower than $(K_{\min}^{\text{pre}})^{-1/2}$. In general, this can be satisfied by $\lambda_K = \alpha \left(K_{\min}^{\text{pre}}\right)^{-\gamma}$ for $\gamma \in (0, 1/2)$ and $\alpha > 0$ a positive-valued constant.

the corresponding face of the donor simplex:

$$\widehat{w}^{\text{ts}} \in \arg\min_{w \in \mathcal{W}_{\widehat{\mathcal{S}}}} \widehat{L}(w) \qquad \mathcal{W}_{\widehat{\mathcal{S}}} = \left\{ w \in \mathcal{W} : w_i = 0 \text{ for every } i \notin \widehat{\mathcal{S}} \right\}$$

We can now produce a result analogous to that of Proposition 4, establishing asymptotic normality of $\widehat{w}^{\text{ts}}$. Moving forward, let $s_0 = |\mathcal{S}_0|$ denote the true number of active donors.

Theorem 1 (Consistency and asymptotic normality of support-refitted weights)**.** *Suppose the conditions of Proposition 6 hold. Then* $\widehat{w}^{\text{ts}} \xrightarrow{p} w^0$.

If $s_0 \geq 2$ *and Assumption BSG also holds, then*

$$\sqrt{K_{\min}^{\text{pre}}}\left(\widehat{w}^{\text{ts}} - w^0\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{w,S_0})$$

where

$$\Sigma_{w,S_0} = A_{S_0} H_{S_0}^{-1} \Omega_{S_0} H_{S_0}^{-1} A_{S_0}^{\top}$$

Note that the above result mirrors that of Proposition 4, with the variance terms restricted to the active face of the donor simplex.[29] We have thus established that this two-stage weight estimation procedure yields consistent and asymptotically normal weights, without restriction on interiority of w^0 on the full donor simplex. Consequently, this will be our preferred weight estimation method for the remainder of this paper. We can then extend the argument from Proposition 5 to this two-stage setting and establish feasible estimation, where the only additional complication comes from the fact that $\widehat{\mathcal{S}}$ is a random object.[30]

Corollary 1 (Feasible covariance estimation for support-refitted weights)**.** *Suppose the conditions of Proposition 6 hold,* $s_0 \geq 2$*, and Assumption BSG holds. Let* $\widehat{\Sigma}_{w,\widehat{S}}$ *denote the sample analogue of* Σ_{w,S_0}*. Then, as* $K_{\min}^{\text{pre}} \to \infty, \widehat{\Sigma}_{w,\widehat{S}} \xrightarrow{p} \Sigma_{w,S_0}$.

3.4 Consistency & Inference for Treatment Effects on Pre-Specified g

We now leverage the consistency and asymptotic normality results established for $\widehat{w}$ to establish the same properties for the treatment effect on a pre-specified summary function g. Let $\widehat{\tau}_g$ denote this treatment effect, defined using the two-stage weights $\widehat{w}^{\text{ts}}$:

$$\widehat{\tau}_g = \mathbb{E}_{\widehat{F}_{NT}}[g(X)] - \sum_{i=1}^{N-1} \widehat{w}_i^{\text{ts}} \, \mathbb{E}_{\widehat{F}_{iT}}[g(X)] \tag{11}$$

Moving forward, we will correspondingly use these two-stage weights to construct our shorthand for the estimated counterfactual, while the empirical treated distribution remains the empirical analog

[29] Note that the $s_0 = 1$ case is omitted - if there is only a single active donor, then it must receive unit weight and the limiting covariance is the zero matrix.

[30] On the event that $\widehat{H}$ is singular or $|\widehat{S}| = 1$, define $\widehat{\Sigma}_{w,\widehat{S}} = 0$.

of F_{NT}:

$$\widehat{P}_1 := \widehat{F}_{NT} \qquad \widehat{P}_0 := \sum_{i=1}^{N-1} \widehat{w}_i^{\text{ts}} \widehat{F}_{iT}$$

We now introduce an assumption on g, requiring it to be bounded. However, we impose no other conditions, including smoothness, on g.

Assumption BFS (Bounded fixed summary)**.** *There exists a finite constant $\overline{g} < \infty$ such that*

$$\sup_{x \in \mathcal{X}} |g(x)| \leq \overline{g}$$

In order to establish asymptotic normality of $\widehat{\tau}_g$, we make a further assumption governing the rate of growth of the pre-treatment period relative to the full sample.

Assumption RPG (Relative pre-treatment growth)**.** *As $K_{\min} \to \infty$,*

$$\rho_K := \frac{K_{\min}}{K_{\min}^{\text{pre}}} \longrightarrow \rho \in [0, 1]$$

This assumption allows us to track how fast the pre-treatment subset is growing in relation to the overall rate of growth of the whole data. Note that since $K_{\min} \leq K_{\min}^{\text{pre}}$, this ratio is bounded between 0 and 1 by construction. The case $\rho = 0$ corresponds to the pre-treatment sample size used to estimate the weights growing faster than the post-treatment samples used to estimate $\widehat{\tau}_g$. With these assumptions in hand, we can now state our consistency and inference result.

Theorem 2 (Asymptotic normality of fixed summary treatment effect)**.** *Suppose the conditions of Proposition 6 hold, and that Assumptions CS and BFS hold as well. Then, as $K_{\min} \to \infty, \widehat{\tau}_g \xrightarrow{p} \tau_g$.*

If Assumptions BSG and RPG hold in addition, then as $K_{\min} \to \infty$,

$$\sqrt{K_{\min}} \left(\widehat{\tau}_g - \tau_g \right) \xrightarrow{d} \mathcal{N}\left(0, V_g\right)$$

where

$$V_g = \frac{\sigma_{\text{N}}^2(g)}{\kappa_{NT}^{\text{post}}} + \sum_{i=1}^{N-1} (w_i^0)^2 \frac{\sigma_i^2(g)}{\kappa_{iT}^{\text{post}}} + \rho \, \mu(g)^\top \Sigma_{w,\mathcal{S}_0} \mu(g)$$

$$\sigma_i^2(g) = \text{Var}_{\mathcal{F}_{iT}(d_{iT})}(g(X)) \qquad \mu(g) = \left(\mathbb{E}_{\mathcal{F}_{1T}(0)}[g(X)], \ldots, \mathbb{E}_{\mathcal{F}_{(N-1)T}(0)}[g(X)] \right)^\top$$

We have therefore established consistency and asymptotic normality of the difference in means estimator. Examining the expression for the limiting variance V_g we observe that it contains three terms: the first two correspond to the variance in the summary function g evaluated on the treated and donor unit distributions, respectively, while the third term captures the uncertainty in the first-stage weight estimation problem, with $\Sigma_{w,\mathcal{S}_0}$ corresponding to the limiting variance of $\widehat{w}^{\text{ts}}$

established in Theorem 1. Finally, we can provide an extension to Corollary 1 and characterize a feasible, consistent estimator of this limiting variance.

Corollary 2 (Feasible fixed summary variance estimation)**.** *Suppose the conditions for the asymptotic normality result of Theorem 2 hold, and that the number of active donors* $s_0 \geq 2$*. Let* V_g *be defined as in Theorem 2, with sample analogue* $\widehat{V}_g$*. Then, as* $K_{\min} \to \infty, \widehat{V}_g \xrightarrow{p} V_g$*.*

4 Hypothesis Generation & Validation

We have thus far established that we can consistently estimate treatment effects for a pre-specified summary function g, and the resulting treatment effect estimator $\widehat{\tau}_g$ is asymptotically normal. However, this requires the researcher to pre-specify an aspect of the text along which to measure causal effects. In general, it may be that the researcher does not have strong intuition for how treatment may impact text. In this case, the researcher may instead prefer to *discover* how treatment impacts text, which can be accomplished by comparing the observed $\widehat{P}_1$ to the estimated $\widehat{P}_0$. This highlights an additional key advantage of our distributional framework: since $\widehat{P}_0$ is a mixture over donor distributions which we observe, we can sample from $\widehat{P}_0$ by first selecting a donor unit from the multinomial distribution induced by $\widehat{w}^{\text{ts}}$ and then sampling uniformly at random from the observed empirical distribution of documents from the chosen donor. In this manner, we can construct a counterfactual corpus of real-text documents that can then be directly compared to the observed corpus of treated documents, with hypothesized treatment effect dimensions provided by procedures analogous to that in Modarressi, Spiess, and Venugopal (2025).

More formally, in this section we consider a learned summary function $\widehat{g} : \mathcal{X} \to \mathbb{R}$ that is provided by some algorithm that has access to $\widehat{P}_1$ and $\widehat{P}_0$. This function evaluates the presence of an aspect or theme of the text that is hypothesized to have been impacted by treatment. We can then quantify the effect of treatment on this learned summary via calculating the difference in mean summary values across the two distributions of interest:

$$\widehat{\tau}_{\widehat{g}} = \mathbb{E}_{\widehat{F}_{NT}}[\widehat{g}(X)] - \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \, \mathbb{E}_{\widehat{F}_{iT}}[\widehat{g}(X)] \tag{12}$$

In this section, we will proceed to apply the results of Section 3 to establish consistency and inference for $\widehat{\tau}_{\widehat{g}}$, conditional on the learned $\widehat{g}$.

4.1 Learning a Summary Function

The process of hypothesizing aspects of text sensitive to treatment (and corresponding summary functions) requires providing data from $\widehat{P}_1$ and $\widehat{P}_0$ to some algorithm, so that it might learn what differences arise by virtue of treatment being imposed. In practice, LLMs would be a compelling

choice of algorithm, given their ability to process and reason about text at scale. However, it is unclear how we can analyze the properties of a function $\widehat{g}$ returned by an LLM, particularly convergence properties to some "true" g. Consequently, we will proceed by providing results for consistency and inference conditional on the hypothesized summary $\widehat{g}$; we will continue to take into account first-stage uncertainty in counterfactual construction and second-stage sampling uncertainty, but will not make any statements about possible errors made in the formulation of the learned $\widehat{g}$ itself.

This modeling decision, while necessary to incorporate modern "black-box" AI tools such as LLMs, immediately gives rise to one challenge that was absent from the fixed-summary setting: evaluating the learned summary $\widehat{g}$ on the same data from which it was discovered can jeopardize tractable inference for $\widehat{\tau}_{\widehat{g}}$. Moreover, we cannot evaluate $\widehat{\tau}_{\widehat{g}}$ using the same weights used to produce the counterfactual from which $\widehat{g}$ was learned; in that case, conditioning on $\widehat{g}$ (by conditioning on the data used to learn $\widehat{g}$) would require conditioning on the weights themselves, eliminating the sampling variation used to derive the first-stage weight limit. We resolve both of these challenges via sample splitting, in which we split both the pre- and post-treatment data into two subsets, which we term "discovery" (abbreviated "disc") and "evaluation" (abbreviated "eval"), respectively.[31]

More formally, we will partition the observed document indices for each unit $i = 1, \ldots, N$ in each time period $t = 1, \ldots, T$ into two disjoint sets:

$$\mathcal{I}_{it}^{\text{disc}} \dot{\cup} \mathcal{I}_{it}^{\text{eval}} = \{1, \ldots, K_{it}\}$$

For $a \in \{\text{disc}, \text{eval}\}$, let $K_{it}^a = |\mathcal{I}_{it}^a|$ and define the subset-specific empirical distributions as:

$$\widehat{F}_{it}^a = \frac{1}{K_{it}^a} \sum_{k \in \mathcal{I}_{it}^a} \delta_{x_{itk}}$$

We now introduce an assumption that governs this split process, ensuring that the partitions themselves are constructed honestly and in a consistent manner.

Assumption SS (Sample splitting)**.** *The partitions* $\{\mathcal{I}_{it}^{\text{disc}}, \mathcal{I}_{it}^{\text{eval}}\}_{i=1,\ldots,N;t=1,\ldots,T}$ *are constructed independently of the observed documents. There exist constants*

$$\pi^{\text{disc}}, \pi^{\text{eval}} \in (0, 1) \qquad \pi^{\text{disc}} + \pi^{\text{eval}} = 1$$

such that, for every $i = 1, \ldots, N$*,* $t = 1, \ldots, T$*, and* $a \in \{\text{disc}, \text{eval}\}$*,*

$$\frac{K_{it}^a}{K_{it}} \longrightarrow \pi^a$$

[31] This splitting is done separately for each cell of the panel matrix, to ensure that every unit has some documents in each fold in every period.

Consequently, conditional on the split indices, the discovery and evaluation samples are mutually independent across folds and unit-period cells. We now define the minimum number of documents across units in the pre- and post-treatment periods as:

$$K_{\min}^{\text{pre},a} := \min_{1\le i\le N, t<T} K_{it}^a \qquad K_{\min}^{\text{post},a} := \min_{1\le i\le N} K_{iT}^a \qquad a \in \{\text{disc}, \text{eval}\}$$

We will now proceed to estimate two sets of weights, one on each of these pre-treatment folds, using for each our two-stage estimation procedure to ensure interiority. For fold $a \in \{\text{disc}, \text{eval}\}$, define the first-stage weight and active donor set estimates:

$$\widetilde{w}^a \in \arg\min_{w\in\mathcal{W}} \widehat{L}^a(w) \qquad \widehat{\mathcal{S}}^a = \{i \in \{1,\dots,N-1\} : \widetilde{w}_i^a > \lambda_K^a\}$$

Here, $\widehat{L}^a(w)$ denotes the objective function evaluated on the given fold of pre-treatment data:

$$\widehat{L}^a(w) = \sum_{t<T} \text{MMD}_k^2 \left(\widehat{F}_{Nt}^a, \sum_{i=1}^{N-1} w_i \widehat{F}_{it}^a \right)$$

We can then estimate the two-stage weights on each fold as:

$$\widehat{w}^{\text{ts},a} \in \arg\min_{w\in\mathcal{W}_{\widehat{\mathcal{S}}^a}} \widehat{L}^a(w)$$

Previous arguments for consistency and asymptotic normality of the two-stage weights on the full sample lead naturally to analogous results for these fold-specific estimators; see Appendix E for further details. Finally, we can redefine our target distributions of interest for each fold $a \in \{\text{disc}, \text{eval}\}$:

$$\widehat{P}_1^a = \widehat{F}_{NT}^a \qquad \widehat{P}_0^a = \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts},a} \widehat{F}_{iT}^a$$

A summary function measuring some hypothesized aspect of the text can then be generated by an algorithm which is given access to $\widehat{P}_1^{\text{disc}}$ and $\widehat{P}_0^{\text{disc}}$. The resulting function will itself then be a function of the pre- and post-treatment discovery-fold data, denoted $D_{\text{pre}}^{\text{disc}}$ and $D_{\text{post}}^{\text{disc}}$, respectively, from which these empirical distributions are constructed. We formalize the hypothesis generation algorithm and its produced summary function in the following assumption, in which we also impose a boundedness condition on the learned function analogous to Assumption BFS.

Assumption SG (Summary generation)**.** *There exists an algorithm $\mathcal{A}_K$ and a class $\mathcal{G}$ of measurable functions $g : \mathcal{X} \to \mathbb{R}$ such that*

$$\widehat{g} = A_K \left(D_{pre}^{disc}, D_{post}^{disc}; \mathcal{R}_K \right) \in \mathcal{G}$$

where $\mathcal{R}_K$ captures auxiliary randomization inherent to the algorithm $\mathcal{A}_K$ and is independent of data and split assignments. Additionally, there exists a finite constant $\overline{g} < \infty$ such that

$$\sup_{g \in \mathcal{G}} \sup_{x \in \mathcal{X}} |g(x)| \leq \overline{g}$$

This assumption makes explicit that no evaluation fold data is used to construct, tune, revise, or select $\widehat{g}$. The boundedness condition is imposed uniformly over the possible outputs of the discovery algorithm $\mathcal{A}_K$. The algorithm-specific auxiliary randomization $\mathcal{R}_K$ captures any and all sources of randomness that are introduced in the algorithmic process; in the case of an LLM-based algorithm, this includes the parameters which govern randomness in the generative process. Moving forward, define the discovery sigma field:

$$\mathcal{J}_K := \mathcal{S}_K \vee \sigma\left(D^{\text{disc}}_{\text{pre}}, D^{\text{disc}}_{\text{post}}, \mathcal{R}_K\right)$$

Here, $\mathcal{S}_K$ captures both pre- and post-treatment sample-split assignments. Note that $\widehat{g}$ produced by Assumption SG is measurable with respect to $\mathcal{J}_K$.

4.2 Consistency & Inference via Sample-Splitting

With this learned summary function in hand, we construct estimates of the treatment effect on the evaluation fold data:

$$\widehat{\tau}^{\text{eval}}_{\widehat{g}} = \mathbb{E}_{\widehat{F}^{\text{eval}}_{NT}}[\widehat{g}(X)] - \sum_{i=1}^{N-1} \widehat{w}^{\text{ts,eval}}_i \, \mathbb{E}_{\widehat{F}^{\text{eval}}_{iT}}[\widehat{g}(X)] \tag{13}$$

We can now establish consistency of our estimated treatment effect $\widehat{\tau}^{\text{eval}}_{\widehat{g}}$, analogous to the first result of Theorem 2.[32] In this case, we now must first condition on the discovery sigma-field $\mathcal{J}_K$, in order to treat $\widehat{g}$ as fixed.[33]

Theorem 3 (Generated summary consistency)**.** *Suppose the conditions of Proposition 6 hold for each pre-treatment fold, and that Assumptions CS, SS and SG hold as well. Then, conditional on $\mathcal{J}_K$, as $K_{\min} \to \infty$, $\widehat{\tau}^{\text{eval}}_{\widehat{g}} - \tau_{\widehat{g}} \xrightarrow{p} 0$.*

The intuition behind this result is simple: since the fold-specific weights are both consistent for the population weights, the same is true for the fold-specific counterfactuals, and so the same consistency argument underlying the first claim of Theorem 2 holds in this setting. With consistency established, we can show that our treatment effect estimator $\widehat{\tau}^{\text{eval}}_{\widehat{g}}$ is asymptotically normal, with a

[32] Pre-treatment sample splitting is technically unnecessary for consistency; post-treatment sample splitting suffices for consistency, while pre-treatment sample splitting is sufficient for inference.

[33] Henceforth, convergence in distribution of a statistic T_K conditional on $\mathcal{J}_K$ is understood to mean, for every bounded and continuous f: $\mathbb{E}[f(T_K)|\mathcal{J}_K] \xrightarrow{p} \mathbb{E}[f(Z)]$, where Z is distributed according to the distributional limit of T_K, conditional on $\mathcal{J}_K$. Furthermore, observe that by the tower property, the results of Theorems 3 and 4 also hold conditional on $\widehat{g}$ directly, since $\sigma(\widehat{g}) \subseteq \mathcal{J}_K$.

limiting variance that takes into account the first-stage weight-estimation error, analogous to the inference result in Theorem 2.

Assumption PEFG (Evaluation-fold growth with pre-treatment sample splitting)**.** *As* $K_{\min}^{\text{post,eval}} \to \infty$*, there exist constants* $\kappa_{iT}^{\text{eval}} \in (0, \infty)$ *for all* $i = 1, \ldots, N$ *such that*

$$\frac{K_{iT}^{\text{eval}}}{K_{\min}^{\text{post,eval}}} \to \kappa_{iT}^{\text{eval}}$$

Moreover,

$$\rho_K^{\text{eval}} := \frac{K_{\min}^{\text{post,eval}}}{K_{\min}^{\text{pre,eval}}} \to \rho^{\text{eval}} \in [0, \infty)$$

Theorem 4 (Treatment effect inference via pre-treatment sample splitting)**.** *Suppose the conditions of Theorem 3 hold, and that Assumptions BSG and PEFG hold as well. Suppose further that* $V_{\widehat{g}}^{-1} = O_p(1)$*. Then, conditional on* $\mathcal{J}_K$*, as* $K_{\min}^{\text{post,eval}} \to \infty$*,*

$$\frac{\sqrt{K_{\min}^{\text{post,eval}}}\left(\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \tau_{\widehat{g}}\right)}{\sqrt{V_{\widehat{g}}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

where

$$V_{\widehat{g}} = \frac{\sigma_{\text{N}}^2(\widehat{g})}{\kappa_{NT}^{\text{eval}}} + \sum_{i=1}^{N-1} (w_i^0)^2 \frac{\sigma_i^2(\widehat{g})}{\kappa_{iT}^{\text{eval}}} + \rho^{\text{eval}} \mu(\widehat{g})^\top \Sigma_{w,\mathcal{S}_0} \mu(\widehat{g})$$

$$\sigma_i^2(\widehat{g}) = \text{Var}_{\mathcal{F}_{iT}(d_{iT})}(\widehat{g}(X)) \qquad \mu(\widehat{g}) = \left(\mathbb{E}_{\mathcal{F}_{1T}(0)}[\widehat{g}(X)], \ldots, \mathbb{E}_{\mathcal{F}_{(N-1)T}(0)}[\widehat{g}(X)]\right)^\top$$

The form of this limiting variance is effectively identical to that of Theorem 2, with the first two terms corresponding to variances of the learned summary function $\widehat{g}$ in both treated and donor units while the third term captures uncertainty from the weights estimated in the first stage. The primary difference is that now, all of these quantities must be estimated on the evaluation fold of data, both pre- and post-treatment. The following result extends Corollary 2 to this setting, establishing consistency of the feasible variance estimators.

Corollary 3 (Feasible variance estimation via pre-treatment sample splitting)**.** *Suppose the conditions of Theorem 4 hold and that the number of active donors* $s_0 \geq 2$*. Let* $V_{\widehat{g}}$ *be defined as in Theorem 4, and* $\widehat{V}_{\widehat{g}}$ *its sample analogue, using weights estimated on the pre-treatment evaluation fold throughout. Then, as* $K_{\min}^{\text{post,eval}} \to \infty, \widehat{V}_{\widehat{g}} - V_{\widehat{g}} \xrightarrow{p} 0$*.*

Thus, we have established comprehensive inference results for treatment effect estimators for learned summaries via pre-treatment sample splitting. This approach is highly appealing due to its flexibility; it makes no assumptions on how the algorithm $\mathcal{A}_K$ ingests or uses the discovery-fold data. In particular, this allows for iterative "tuning" of summaries by the researcher, who can show

$\mathcal{A}_K$ pre-treatment documents in order to guide search towards summaries that do not differentiate between the treated unit and the synthetic control in the pre-period. This can help ensure that the generated summary is well-balanced in practice by the synthetic control while still enabling valid inference. Indeed, a researcher can generate a hypothesis, assess its pre-period balance between treated unit and synthetic control on the pre-treatment discovery fold, and then proceed to estimate the treatment effect on the evaluation fold. If the generated summary has undesirable pre-period behavior, the researcher can provide this information to $\mathcal{A}_K$ as well and obtain an alternative $\widehat{g}$, which can be similarly evaluated. Proceeding iteratively in this manner can ensure that the final learned summary function $\widehat{g}$ is well-behaved pre-treatment, while still maintaining valid inference on the evaluation-fold estimate of the treatment effect.

5 Identification & Inference Under Misspecification

In the preceding sections, we established consistency and asymptotic normality of treatment effect estimates for pre-specified and learned summary functions under correct specification, as formalized in Assumption CS. However, as noted in Section 2.4, the general estimand if this assumption does not hold is the pseudo-counterfactual, defined as the mixture of donor distributions which is closest in trajectory to the treated unit's distribution pre-treatment. In this case of misspecification, we can generalize our previous consistency results to provide consistency for the treatment effects relative to the pseudo-counterfactual and bound the induced bias relative to the true counterfactual.

Recall the definition of the treatment effect on the aspect of text measured by g relative to the pseudo-counterfactual $\widetilde{P}_0$, $\widetilde{\tau}_g$, defined in Equation 6:

$$\widetilde{\tau}_g = \mathbb{E}_{P_1}[g(X)] - \mathbb{E}_{\widetilde{P}_0}[g(X)]$$

We will refer to this as the *pseudo-treatment effect*. First, note that the proof of the consistency result in Theorem 2 immediately provides for consistency of $\widehat{\tau}_g$ relative to $\widetilde{\tau}_g$ when Assumption CS is not assumed. Similarly, the proof of the asymptotic normality result of Theorem 2 provides for asymptotic normality of $\widehat{\tau}_g$ with the inferential target now being $\widetilde{\tau}_g$, rather than τ_g. We now characterize the misspecification-induced estimation error:

Proposition 7 (Automatic balance for fixed summaries)**.** *Suppose the conditions of Proposition 6 hold. Let $g : \mathcal{X} \to \mathbb{R}$ be a fixed, measurable function satisfying Assumption BFS, and let $h \in \mathcal{H}_k$. Then, as $K_{\min} \to \infty$, the following holds for any $\delta \geq 0$:*

$$\|g - h \circ \mathcal{E}\|_\infty \leq \delta \implies |\widehat{\tau}_g - \tau_g| \leq \|h\|_{\mathcal{H}_k} \mathrm{MMD}_k(\widetilde{P}_0, P_0) + 2\delta + o_p(1)$$

In this result, we bound the misspecification-induced estimation error by the distance between the pseudo-counterfactual and true counterfactual in the population and the RKHS norm of some

function $h \in \mathcal{H}_k$ which, when composed with the embedding map, approximates g. Note that in the particular case where $g = h \circ \mathcal{E}$ for some h, then $\delta = 0$ and the approximation-quality term on the right hand side vanishes. This result highlights another key virtue of our distributional modeling approach: by finding the approximation $\widetilde{P}_0$ that is "closest" to the true counterfactual P_0, we automatically balance a wide class of summary functions.[34] In particular, these results establish that the degree of balance is proportional to the RKHS norm of the function in the RKHS that, when taken as a composite with $\mathcal{E}$, approximates g.

Note that the conditions of this result are far more plausible in the case of a universal kernel, such as the Gaussian RBF kernel (Steinwart and Christmann, 2008). A universal kernel can approximate any continuous function arbitrarily well on a compact metric space (Steinwart, 2001). Consequently, if our embedding space $\mathcal{B}$ is compact and g can be represented as a composition $g_k \circ \mathcal{E}$ for some continuous $g_k : \mathcal{B} \to \mathbb{R}$, then for any $\delta > 0$ there exists an $h \in \mathcal{H}_k$ such that $\|g - h \circ \mathcal{E}\|_\infty \leq \delta$.[35] In this manner, under these additional assumptions on $\mathcal{B}$, we can view MMD_k with a universal kernel as bounding the imbalance in text summaries which are continuous functions of text embeddings, in a qualitatively similar manner to Wasserstein distances balancing Lipschitz functions, as discussed in Section 2.3.

We can similarly analyze the case of a generated summary $\widehat{g}$. Our estimand under misspecification in this case is given by the pseudo-treatment effect for $\widehat{g}$:

$$\widetilde{\tau}_{\widehat{g}} = \mathbb{E}_{P_1}[\widehat{g}(X)] - \mathbb{E}_{\widetilde{P}_0}[\widehat{g}(X)] \tag{14}$$

As in the case of the pre-specified summary function, consistency and asymptotic normality of $\widehat{\tau}^{\text{eval}}_{\widehat{g}}$ relative to $\widetilde{\tau}_{\widehat{g}}$ (rather than $\tau_{\widehat{g}}$) follow immediately from the proofs of Theorems 3 and 4. As before, we can now bound the estimation error in our treatment effect estimate $\widehat{\tau}^{\text{eval}}_{\widehat{g}}$ relative to the true treatment effect $\tau_{\widehat{g}}$ under misspecification.

Proposition 8 (Automatic balance for learned summaries)**.** *Suppose the conditions of Proposition 6 hold for each pre-treatment fold, that Assumption SS holds, and that $\widehat{g} : \mathcal{X} \to \mathbb{R}$ satisfies Assumption SG. Suppose further that $h \in \mathcal{H}_k$ is $\mathcal{J}_K$-measurable. Then, conditional on $\mathcal{J}_K$, as $K_{\min} \to \infty$, the following holds for any $\delta \geq 0$:*

$$\|\widehat{g} - h \circ \mathcal{E}\|_\infty \leq \delta \text{ a.s. } \implies |\widehat{\tau}^{\text{eval}}_{\widehat{g}} - \tau_{\widehat{g}}| \leq \|h\|_{\mathcal{H}_k} \text{MMD}_k(\widetilde{P}_0, P_0) + 2\delta + o_p(1)$$

Thus, under misspecification, our proposed estimators for the treatment effect for both pre-

[34] While $\text{MMD}_k(\widetilde{P}_0, P_0)$ is a post-treatment distance that we cannot directly discipline by estimating pre-treatment weights, under an implicit assumption that weights which balance well pre-treatment should continue to do so post-treatment we can understand disciplining pre-treatment discrepancy as approximately disciplining post-treatment discrepancy.

[35] If $\mathcal{B}$ is bounded but not closed, then as long as g_k is uniformly continuous we can extend g_k to the closure of $\mathcal{B}$ and obtain the same result.

specified and learned summary functions are now consistent for the pseudo-treatment effects, with asymptotic normality holding in both cases with this corresponding new inferential target. We can bound the induced estimation error for the true treatment effect (relative to the true counterfactual), leveraging the ability of our distributional estimator to simultaneously balance a wide class of summary functions of the text.

6 Remote Participation & Local Government Speech

We will now apply our distributional synthetic control estimator to the dataset of California city council meetings compiled by Martin and Venugopal (2026). In that paper, the authors compile a dataset of over 25,000 transcripts of city council meetings from 115 cities in California from 2015-2024, studying the structure and content of meetings and the characteristics of the public participant pool. Their primary causal question is how the staggered removal of remote participation options for public commenters in the aftermath of the COVID-19 pandemic impacted participant demographics. Applying the difference-in-differences estimator from Callaway and Sant'Anna (2021), the authors find that requiring in-person participation led to lower turnout but did not meaningfully change the makeup of the participant pool. In this paper, we ask a natural follow-up question: how does shifting to in-person participation shape *how* constituents speak?[36]

This is a setting in which the application of the hypothesis generation procedure is of particular utility. Public comments in these meetings are highly diverse and vary significantly in topic, style, and tone, as highlighted in Table 1 and by Martin and Venugopal (2026). Moreover, the existing literature provides only limited evidence of what aspects of speech are impacted by remote participation.[37] Consequently, we are left without a strong prior for aspects of text that are likely to be impacted by treatment in this context. The iterative approach outlined in Section 4.2 allows us to ensure that the learned aspects of the text are of substantive interest, without sacrificing statistical validity. A failure of the hypothesis generation procedure to identify aspects of speech that are significantly impacted by treatment would also suggest that the change in participation format had limited impact on speech.

This section proceeds as follows: we begin by reviewing the institutional setting and key features of the data from the study of Martin and Venugopal (2026). We then conduct two calibrated simulations using real public comment data to demonstrate the ability of our distributional synthetic

[36]Note that an observed change in speech could in principle be due to either selection into participation or within-participant changes in speech. The finding of Martin and Venugopal (2026) that the participant pool is not systematically different in terms of observables after the removal of remote participation options suggests either that selection into participation is a secondary mechanism or that selection occurs along an unobserved dimension of participants (e.g., latent "type"). I will not claim to distinguish between selection and within-participant changes in this application.

[37]Einstein et al. (2023) compare the stated policy positions of public participants in remote and in-person meetings, while Brucks and Levav (2022) focus primarily on ideation, finding limited effects on a variety of pre-defined linguistic features.

control estimator to correctly identify true counterfactual distributions and dimensions of treatment effects in the presence of confounding. We conclude by applying our method in two synthetic control configurations encompassing seven treated units, primarily in northern California, in which we find that requiring in-person participation leads to a decrease in hostile, combative, and provocative speech in the following months.

6.1 Institutional Setting & Data

We begin with a review of the institutional setting and data collected in Martin and Venugopal (2026). Under the Brown Act, all local legislative bodies in California are required to deliberate in public and allocate time for the members of the public to provide comment (Brown, 1953). Crucially, this requirement is not limited to items on the given meeting agenda or subject to a vote by the given body; public participants must be allowed to speak on any issue that falls under the given legislative body's jurisdiction. As a result, city council meetings in California must include regular public comment periods, in which residents and stakeholders can be heard by the council on virtually any issue of their choosing pertaining to local government or city conditions.

Indeed, this form of participation is ubiquitous: in the data collected in Martin and Venugopal (2026), the authors find over 250,000 distinct public comments, of which over 40% (nearly 115,000) are *unagendized* (that is, unrelated to any issue on the agenda of the given city council meeting). At the onset of the COVID-19 pandemic in March 2020, Governor Newsom issued statewide emergency orders allowing local legislative bodies, including city councils, to meet and allow public comment virtually. Consequently, most cities in the collected dataset switched to remote participation shortly thereafter, allowing members of the public to participate via remote options such as Zoom or telephone. While those cities that shifted online did so at approximately the same time, there is considerable heterogeneity in when cities remove synchronous remote participation options and require that public commenters attend in-person. In Martin and Venugopal (2026), the authors argue that this heterogeneity is poorly explained by city-level observables, and that this source of variation therefore constitutes a natural experiment to examine the impact of the removal of remote participation options on participation.

In this paper, we analyze the causal effect of requiring in-person participation on the speech of public commenters. In order to account for potential effects of selection into participation based on meeting agendas, we will consider only the effect on unagendized public comments. Given the occasionally low volume of unagendized comments in meetings, we will aggregate across all meetings in a three-month period within a given municipality as a single "observation". Each unit-period observation is then given by the empirical distribution over embedding vectors of the unagendized public comments that were present in that city (unit) during that three-month period; embeddings are produced by EmbeddingGemma-300M (Vera et al., 2025). Treated units will consist of municipalities that remove remote public participation options at some point during the covered

window. The treatment date for each treated city is defined as the date of the first city council meeting in which public commenters were required to attend in-person. The donor pool will consist of "never-treated" units, defined as those cities that either still offer remote participation options or removed them after the end of data coverage.

6.2 Calibrated Simulations

Before applying our estimator to assess the causal effect of requiring in-person public participation on public comment speech, we first perform two calibrated simulations to demonstrate that the method performs well and accurately recovers true weights, counterfactuals, and treatment effect dimensions. In order to make these studies as realistic as possible, both simulations use embedding vectors for real unagendized public comments in never-treated municipalities. In both studies, our primary benchmark model is one where a counterfactual distribution is constructed via equal weighting of the donor units. This benchmark is selected to mimic what a naive researcher may do given the current state of the literature, which does not prescribe methods to assign data-driven weights in settings with text data. It is also analogous to the approach of difference-in-differences, which corresponds to synthetic control with equal weights and an intercept term (Doudchenko and Imbens, 2016).

In the first simulation, we use real donor empirical distributions and construct a synthetic treated unit with known weights, allowing us to assess the estimator's ability to accurately recover the true counterfactual distribution. This is analogous to a placebo study in standard synthetic control, in which we test the ability of the model to construct a good counterfactual in a setting with no real treatment effect. We find that the distributional estimator consistently outperforms the naive benchmark in counterfactual recovery, except for when the true weights happen to be approximately equal.

In the second simulation, we verify the ability of our approach to recover correct treatment effects when the true effect is nonzero and other unit characteristics are correlated with treatment. We accomplish this by constructing a "fake" panel of cities where a single city is treated. This treated city differs from the majority of the donor pool in a chosen dimension; this is intended to mimic a setting where a unit "selects" into treatment. By "selection" here, we mean that nonrandom treatment assignment results in some baseline difference between the treated and donor unit(s). We refer to the dimension of text along which the treated unit is constructed to have a baseline difference from the donor pool as the "selection dimension". This treated unit then has its post-treatment comments modified to inject a controlled treatment. The goal of this simulation is to show that our estimator adequately accounts for unknown selection (that is, those features of the text which exhibit baseline differences pre-treatment) and returns hypotheses that pertain to the true treatment dimension. We show that the distributional estimator correctly discards the selection dimension and identifies the true treatment dimension, while the naive benchmark fails to

identify the true treatment and is unable to distinguish between selection and treatment aspects.

6.2.1 Calibrated Simulation 1: Learning the Correct Counterfactual Distribution

In our first calibrated simulation, we test the ability of our synthetic control model to construct the correct counterfactual distribution in a placebo setting, in which the treated unit has no true treatment effect and we can assess counterfactual quality directly. For our donor pool in this simulation, we select the ten largest never-treated cities in our sample by their volume of unagendized public comments in the pre-treatment period, spanning July 1, 2021 – July 1, 2023. We then split this comment panel within each unit-period cell equally into *construction* and *estimation* folds.

In order to synthesize our treated unit, we first simulate a true weight vector via an exponential weighting of city-level, non-text observables, with a controllable parameter to govern the strength of the correlation between the weights and observables. As this parameter value increases, the weights increasingly concentrate on those donor units with observable profiles most similar to the treated unit.[38] We then sample comments from the construction fold of data for each donor unit to construct a sample of 50 comments in each period for the treated unit. In order to assess the ability of our method to construct an appropriate counterfactual, we do not inject any real treatment in this simulation; this therefore corresponds to a setting where the true treatment effect on the treated unit is zero on all aspects of text. We then estimate our distributional weights using the estimation-fold donor unit empirical distributions and the constructed treated unit distribution in each pre-treatment period. Critically, the distributional estimator has no knowledge of the true weights or the underlying observable profiles.

The primary measure of performance for the method is then given by the distance, according to the 1-Wasserstein metric, between the true treated distribution and the induced mixture over donor distributions in the held-out post-treatment periods. This directly measures the quality of the constructed counterfactual, our primary estimand of interest for the weight-estimation stage. Additionally, in order to validate the ability of this estimator to implicitly balance on summary functions of the kind produced by our downstream hypothesis generation engine, we score each treated unit comment and estimation-fold donor unit comment with a set of 15 ex-ante specified summary functions, measuring 10 topical themes identified in Martin and Venugopal (2026) and 5 additional stylistic themes encompassing aspects such as sentiment and formality. We can then plot these two performance metrics as we vary the concentration of the true weights through the correlation parameter. The results are shown in Figure 1, in which concentration of the weights is measured on the x-axis by the L_2 norm; in this setting, this is equivalent to the square root of the commonly used Herfindahl-Hirschmann Index (HHI). Lower values of L_2 norm correspond to

[38] The observable profile of the synthetic treated unit is selected at random from some other non-donor, never-treated reference unit. Observable variables are obtained from the US Census 2024 American Community Survey (ACS) 1-Year Estimates and consist of the following: population; population density; median age; percentage of population over the age of 65; percentage of the population that is White, Black, and Hispanic; and median income.

greater dispersion in the weights.

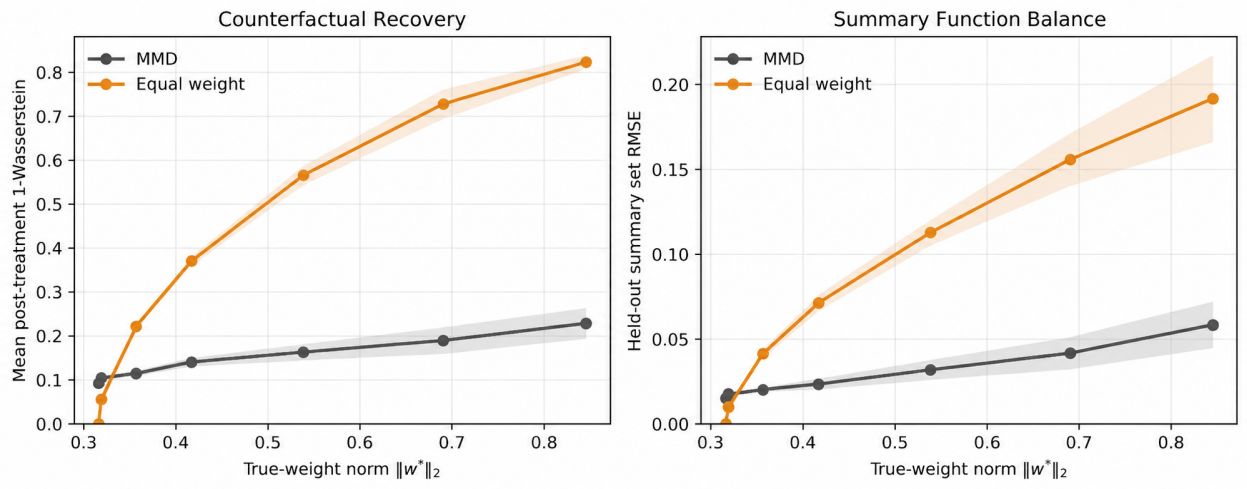


Figure 1: 1-Wasserstein distance between true and estimated counterfactuals (left) and summary score RMSE (right) for estimated counterfactuals produced via equal weighting (orange) and distributional MMD weighting (black). Each point is an average over 20 simulated draws, corresponding to 5 treated unit observable profiles and 4 comment sampling seeds for each. Shaded regions correspond to $\pm$ one standard deviation from the mean. Summary functions in the right panel are measured on a 0-3 scale.

The results demonstrate that a counterfactual constructed via equal weighting of the donor units only performs well when the true weights are approximately equally distributed. The weights learned by minimizing pre-period MMD, however, perform well across both metrics regardless of the concentration of the weights, with the MMD curve in both panels of Figure 1 appearing nearly flat. This demonstrates that our proposed distributional estimator is able to construct accurate counterfactual distributions regardless of how concentrated the weights are, and that these weights also implicitly balance on summary functions in the text that are not provided to the estimator explicitly, in keeping with Proposition 7. For additional details on the simulation, including results showing that the weights and observable profiles themselves are well-approximated by our distributional estimator, see Appendix B.1.

6.2.2 Calibrated Simulation 2: Recovering an Unknown Treatment Effect

Having established that our distributional, MMD-based estimator accurately recovers weights and implicitly balances on held-out summary functions, we now turn our attention to validating the interface between our first-stage counterfactual construction and second-stage hypothesis generation. In particular, our goal is to demonstrate that our procedure will correctly identify true treatment effects in settings where the treatment effect is nonzero and treatment is correlated with other unit characteristics.

We accomplish this by mimicking selection into treatment by constructing a synthetic panel of "fake" cities that vary consistently and at baseline according to some chosen aspect of text. For this "selection aspect" we choose the "Housing and Homelessness" topic identified in Martin and Venugopal (2026) as being a significant driver of meeting attendance. We score all comments in 15 never-treated units chosen for their comment volume according to the 10-topic classification outlined in Martin and Venugopal (2026) and construct a panel of synthetic units that have stable topic shares of "Housing and Homelessness". In particular, we sample comments (without replacement) period-by-period from this pool of never-treated cities, constructing two types of units: "low" types (with "Housing and Homelessness" topic shares no more than 25%) and "high" types (with "Housing and Homelessness" topic shares of at least 60%). The treated unit is selected from the "high" type set.

For post-treatment documents in the treated unit, we inject a treatment of increased formality in the text. This is implemented by asking an LLM (gemini-3.5-flash) to rewrite the sampled untreated comments to make them slightly more formal. This rewritten corpus will now play the role of our observed comment set for the treated unit post-treatment, while the original corpus constitutes our true counterfactual. We observe a measured treatment effect of an increase in formality score of 0.60, measured on a discrete $\{0, 1, 2, 3\}$ scale, where 0 is extremely informal and 3 is extremely formal. By contrast, when we measure the presence of discussion on "Housing and Homelessness" on an analogous discrete scale[39], we observe a true mean difference of 1.22 in the original (counterfactual) corpus and 1.14 in the rewritten (observed) corpus, relative to an equal-weighted pool of donor comments. This demonstrates that the treatment injection did not have a meaningful impact on the concentration of "Housing and Homelessness" in the rewritten documents, and this corresponds to a setting with strong selection: the mean difference in "Housing and Homelessness" is approximately twice as large as the true treatment effect.

As in the previous simulation, we now estimate two sets of synthetic control weights: one via our distributional MMD estimator, and another via equal weighting. We then implement sample splitting for our panel, creating discovery and evaluation folds, and implement our hypothesis generation procedure, using as our algorithm $\mathcal{A}_K$ an LLM that is provided with samples of documents from the treated unit and the estimated counterfactual and is asked to hypothesize differences between them. Each returned hypothesis is accompanied by a scoring scale, providing a learned function $\widehat{g}$ that can then be applied to future documents, in keeping with our theoretical framework. The resulting learned hypotheses can be found in Table 2.

The hypotheses generated by the distributional MMD-weighted counterfactual include the true treatment of formality (bolded in Table 2) and do not include the selection dimension, "Housing and Homelessness". By contrast, the hypotheses generated by the equal-weighted counterfactual include the selection dimension (bolded) but not the true treatment dimension. This highlights

[39]Here, 0 denotes no discussion of "Housing and Homelessness", while 3 denotes that such discussion dominates the text.

Table 2: Hypotheses generated by equal- and MMD-weighted post-treatment counterfactuals

Counterfactual Weighting	Hypothesized Aspects
Equal	**Focus on housing policy**; policy and legal specificity; rhetorical tone; speaker identity and institutional alignment; geographic and operational scope
MMD	Transcription fidelity and cleanliness; **formal vs. colloquial register**; procedure and interactive meta-talk; argumentation style: policy vs personal narrative

the key advantage of distributional weighting: by constructing a counterfactual to minimize the distributional distance in pre-treatment text embeddings, we are able to correctly account for and discard selection effects, allowing us to accurately recover treatment dimensions even when the strength of the treatment effect is much less than that of selection. A naive researcher who instead directly compares the post-treatment treated corpus to the full donor pool (corresponding to an equal-weighted counterfactual) will instead be unable to distinguish between selection and treatment effects. Moreover, as in this setting, if the selection effect is sufficiently strong relative to the treatment effect, it may "crowd out" treatment entirely and hypotheses generated from the equal-weighted counterfactual may be unable to recover the true treatment at all.

We can see this dynamic at play in this simulation. Figure 2 plots an event study of the "focus on housing policy" theme learned from the equal-weighted counterfactual, which comes with a discrete $\{0, 1, 2, 3\}$ scoring scale analogous to the oracle selection theme. The left panel shows the equal-weighted counterfactual which exhibits correspondingly poor pre-period balance, leading to a spurious conclusion of a significant effect in the post-period. The right panel, meanwhile, shows that the text distribution-weighted counterfactual correctly accounts for this baseline difference and shows no significant effect, consistent with the hypothesis generation algorithm not proposing any "Housing and Homelessness"-related themes when given the MMD-weighted counterfactual.

A naive researcher who does not construct a counterfactual to account for selection would receive an erroneous report that the selection dimension exhibited a treatment effect. Even if, after seeing the left panel of Figure 2, they realize that there is no true effect along this dimension, they would be unable to learn the true aspect of text impacted by treatment. Implementing the proposed distributional weighting method, however, shows no significant effect for the selection dimension; moreover, this method does not even hypothesize the selection dimension in the first place. Indeed, it instead accurately hypothesizes the true treatment dimension. This demonstrates the ability of the proposed method to construct intelligent counterfactuals.

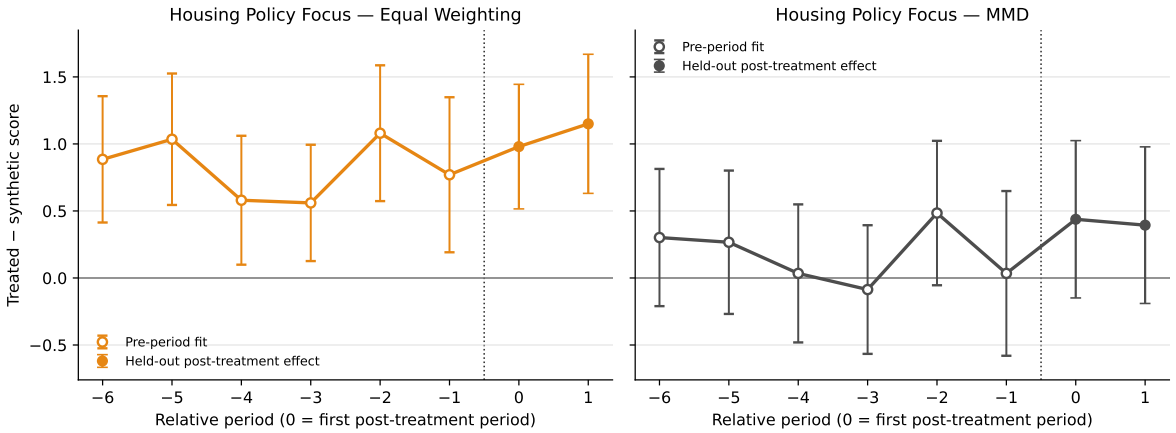


Figure 2: Event studies for learned "focus on housing policy" theme. Post-period estimates are calculated on the evaluation fold, with error bars that are valid 95% confidence intervals conditional on the learned theme and take into account uncertainty in the weights (in the MMD case). Pre-period error bars are diagnostic. Vertical dashed line denotes simulated treatment onset.

6.3 The Effects of Removing Remote Participation

Having established that our synthetic control estimator correctly constructs appropriate counterfactuals and leads to learned hypotheses that correctly identify treatment effects and discard selection, we now apply the method to real treated-donor configurations in the data. As previously discussed, our treatment now is the removal of remote participation options for public commenters and our donor pool is the set of cities that implemented remote participation and have not yet removed it (as of the end of our dataset coverage).

Given that we are in a setting with staggered adoption of treatment (i.e. many cities remove remote participation options at various times), there are several different treated units that we could choose for our synthetic control study. In order to select among these options, we search among the set of possible candidate specifications on the discovery fold of the data, identifying those that achieve good pre-period balance and generate economically interesting hypotheses. We then evaluate treatment effects and perform inference on the held-out evaluation fold. All configurations use three-month aggregated windows of comments within each city as a single period, keeping only those donor units which have a minimum volume of comments per period pre-treatment, to ensure accurate weight estimation.[40]

Our primary chosen configuration examines the effect of requiring public commenters to attend in-person on public comments in six cities across California: Redwood City, San Ramon, Santa Rosa, San Rafael, Ventura, and Novato. Each of these cities is relatively small and may have

[40]For our two-stage weight estimation procedure, we use a threshold of $\lambda_K = 0.02$, defining as an active donor any unit which has an estimated weight of at least that value and setting all others to zero. This cutoff can be viewed as a specific element of the asymptotic sequence described in Section 3.3, at the observed sample size. After removing inactive donors which fall below this threshold, we then re-estimate the weights on the remaining active donor set.

insufficient comment volume for feasible estimation on its own. However, they all remove remote participation in a one-month period around October 2023, and so we therefore combine them into a single, aggregated treated unit. In this configuration, we require a minimum of 20 comments per unit-period cell to be an eligible donor and are able to construct a continuous panel with this restriction stretching back to March 2021.

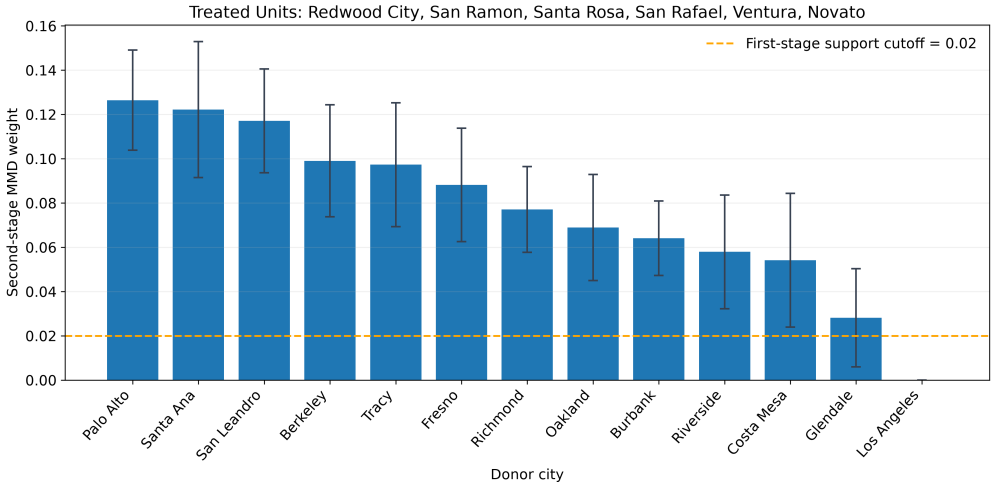


Figure 3: Estimated second-stage MMD weights for donor units for aggregated treated unit. Error bars denote 95% confidence intervals. Donors with estimated first-stage weight under 0.02 (horizontal dashed line) are determined to be inactive and assigned weight 0 in the second stage refit.

Figure 3 shows estimated weights for each of the 13 eligible donors for this aggregated treated unit. The estimated weights are highly intuitive: the treated unit is comprised of smaller cities, mostly from northern California, and these are precisely the donor units that receive the highest weight (e.g., Palo Alto, San Leandro). Conversely, Los Angeles, a large city in southern California, receives zero weight. We then generate hypotheses by showing documents from this aggregated treated unit and its constructed counterfactual from the post-treatment discovery fold to an LLM (gemini-3.5-flash); for details on this hypothesis generation procedure, including prompting, please see Appendix B.3. A hypothesized aspect of the text returned by the LLM that is economically interesting and well-balanced in the pre-treatment period in the discovery fold is "confrontational informal tone". The event study of the value of the summary function measuring this aspect, displaying the per-period difference-in-means of the learned function on the held-out evaluation fold, can be found in Figure 4. Note that we have access to only a single post-treatment period due to data limitations: the availability of public comments runs only through early 2024, and so with a treatment date of late 2023 we have access to only a single full 3-month period of post-treatment comments.

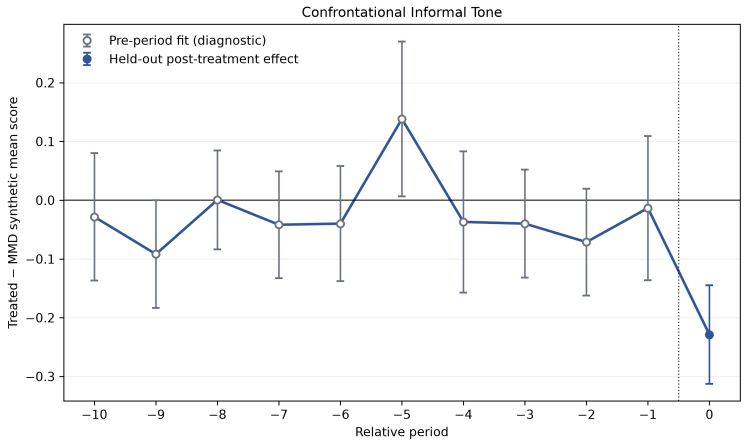


Figure 4: Event studies for treatment effect estimates for aggregated treated unit for learned aspect "confrontational informal tone". Post-treatment error bars are 95% CIs which incorporate uncertainty in the estimated weights. Pre-period error bars are diagnostic. Vertical dashed line denotes treatment onset, with period 0 corresponding to post-treatment. All estimates are computed on the evaluation fold.

We can also repeat the same discovery and estimation procedure for San Jose, which removed remote participation options on January 30, 2024. We construct a synthetic control using 15 cities across the state, each of which has a minimum of 10 unagendized public comments in each of the 8 available pre-treatment periods, stretching back to January 2022. Figure 5 shows the estimated donor weights, with associated uncertainties, for this configuration.

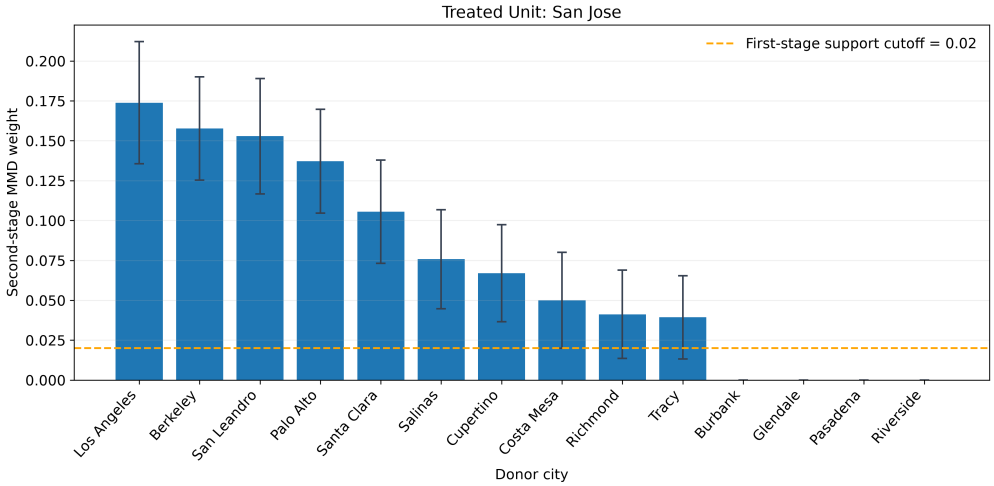


Figure 5: Estimated second-stage MMD weights for donor units for San Jose. Error bars denote 95% confidence intervals. Donors with estimated first-stage weight under 0.02 (horizontal dashed line) are determined to be inactive and assigned weight 0 in the second stage refit.

San Jose is a large city in the Bay Area, and we immediately see that the learned weights are

correspondingly intuitive: most weight is placed on cities in the Bay Area (e.g., San Leandro, Palo Alto, Santa Clara) and larger cities (e.g., Los Angeles). By contrast, the cities we would expect to be most dissimilar to San Jose are smaller cities in southern California, such as Glendale, Burbank, and Pasadena, all of which receive weight zero in this synthetic control.

Passing the discovery-fold post-treatment corpora from San Jose and its synthetic control to the LLM yields two hypothesized aspects that are economically interesting and whose summary functions appear (reasonably) well-balanced in the pre-treatment period: "conspiracy and provocation" and "hostility and profanity". Event studies for the resulting difference-in-means estimators for each function can be seen in Figure 6. Once again, due to the timing of San Jose's removal of remote participation options, we have access to only a single 3-month period of post-treatment comments.

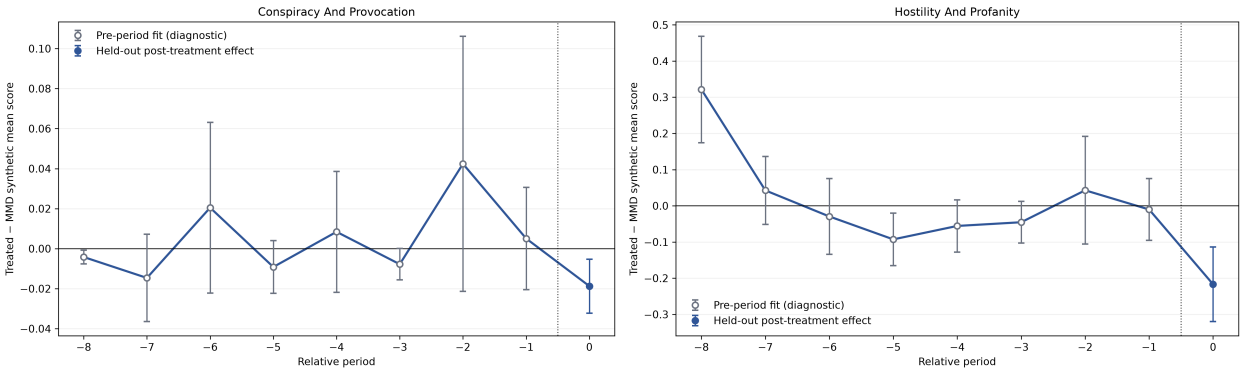


Figure 6: Event studies for treatment effect estimates for San Jose for learned aspects "conspiracy and provocation" (left) and "hostility and profanity" (right). Post-treatment error bars are 95% CIs which incorporate uncertainty in the estimated weights. Pre-period error bars are diagnostic. Vertical dashed lines denote treatment onset, with period 0 corresponding to post-treatment. All estimates are computed on the evaluation fold.

Both summary functions exhibit a marked decline in the post-treatment period and are fairly well-balanced by the synthetic control estimator in the pre-treatment periods, with most pre-treatment mean differences in the periods preceding treatment indistinguishable from zero. The exception to this is period -8 in the right panel of Figure 6, corresponding to the aspect "hostility and profanity". This point captures comments made between January and April 2022, and the dramatic spike is due to a spate of comments in San Jose regarding the shooting of K'aun Green by the San Jose Police Department on March 27, 2022.[41] This led to a documented public outcry in subsequent San Jose city council meetings, with more than 40 residents calling into the April 5, 2022 city council meeting to advocate for police reform and accountability.[42] Inspection of the comments in the impacted period confirms that this advocacy increased the prevalence of comments

[41] See related news coverage for additional information.
[42] See further news coverage for additional information on the council meeting.

labeled "hostile and profane". This constitutes a transitory shock to "hostility and profanity" that dissipates by the next period.[43] Inspection of the post-treatment period comment corpus identifies no comparable transient shock. Details on the meaning of each learned aspect and scoring function (provided by the LLM) across these two configurations can be found in Table 3. For example texts exhibiting each aspect, see Appendix B.4.

Table 3: Hypotheses generated by selected configurations for the treatment effect of the removal of remote participation options.

Learned Aspect	Description	Scoring Scale
Confrontational Informal Tone	Score the document based on the level of informal, direct, or confrontational language used by the speaker. Look for slang, expressions of open anger/contempt, or calling out individuals directly with informal phrases.	$\{0, 1, 2\}$
Conspiracy and Provocation	Evaluate whether the comment contains bizarre, highly provocative, or unverified conspiracy theories, hate-fueled claims, or extreme historical revisionism.	$\{0, 1\}$
Hostility and Profanity	Evaluate the level of hostility, profanity, or abusive language in the public comment.	$\{0, 1, 2\}$

Notes: "Confrontational informal tone" is obtained from the study with aggregated treated unit, while "conspiracy and provocation" and "hostility and profanity" are learned from treated unit San Jose.

In all three aspects, lower scores indicate a lesser degree of provocative, hostile, or combative speech. As Figures 4 and 6 show, all three have estimated negative treatment effects in the post-treatment period that are significant at the 5% level. This fact, coupled with the breadth of cities covered between the two configurations, suggests that the removal of remote participation options led to increased civility and reduced hostility in public comment speech in the months after enactment in the treated cities. For permutation studies in the style of the unit-placebo analyses proposed by Abadie, Diamond, and Hainmueller (2010), see Appendix B.5.

[43] Our inference procedure accounts for sampling and weight uncertainty, conditional on the underlying unit-period distributions over public comments. In particular, it does not account for shocks that change these distributions, such as the one discussed here. Consequently, such shocks may generate pre-treatment imbalance that results in the corresponding error bars not covering zero.

7 Conclusion

Social science researchers are increasingly interested in studying causal effects on text outcomes. Simultaneously, synthetic control methods have gained popularity for their causal effect estimation in observational panel settings. Yet, applications of such methods to text outcomes typically require the researcher to pre-specify the aspect(s) of text on which they wish to measure causal effects. Pre-specifying alternative aspects can yield alternative estimated synthetic control weights, which can pose a concern in settings where the variation in these aspects is reflective of common latent unit characteristics. In addition, matching on other aspects of text may improve estimation of the treatment effect on the target aspect, analogous to inclusion of covariates in standard synthetic control. Moreover, in the case of complicated, high-dimensional outcomes, such as text, the researcher may prefer to instead learn a (potentially nuanced) aspect of text that is impacted by treatment, rather than relying entirely on ex-ante intuition.

In this paper, I address these challenges by developing a distributional synthetic control estimator for text documents in observational panel settings. I model observations as distributions over text embeddings and estimate a counterfactual as a mixture over contemporaneous donor distributions, with weights estimated to minimize average pre-period squared maximum mean discrepancy (MMD). When the synthetic control model is correctly specified, I show that the treatment effect estimate for a pre-specified aspect of text is consistent for the true treatment effect on the treated unit and establish asymptotic normality of the estimator in an unconventional asymptotic regime in which the size of the panel matrix remains fixed while the number of observations per cell grows large.

I further propose a hypothesis generation procedure in which a counterfactual corpus of text documents is sampled directly from the estimated counterfactual mixture distribution and compared against the observed corpus of treated documents. These corpora can then be provided to some hypothesis generation technology (such as an LLM) which returns summary functions of the text, measuring the presence of some aspects which are hypothesized by the technology to be maximally impacted by treatment. I then establish consistency and asymptotic normality of the resulting treatment effect estimator via sample-splitting, conditional on the learned summary functions. In the case of misspecification of the synthetic control model, I provide estimation error bounds for the treatment effect estimates for both pre-specified and learned aspects of the text, showing that learning weights to discipline the distributional distance automatically disciplines the misspecification-induced bias for a wide class of functions.

I highlight the practical appeal of my method in an application to California city council meeting comment periods. In two calibrated simulations, I demonstrate that my proposed method correctly recovers true weights and counterfactuals, and that the hypothesis generation procedure discards aspects of text with baseline differences and identifies those aspects truly impacted by treatment. In two synthetic control configurations with a real treatment of the removal of remote participation

options for public commenters, I find that requiring in-person participation leads to increased civility and reduced hostility in public comments. These results are novel findings at this scale and demonstrate the ability of this method to provide new insight into previously intractable problems in the social sciences.

References

Abadie, Alberto (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425. (Cited on pages 3, 8, 19, and 64.)

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(490):493–505. (Cited on pages 1, 2, 3, 6, 7, 13, 15, 41, and 64.)

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510. (Cited on page 2.)

Abadie, Alberto and Javier Gardeazabal (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132. (Cited on pages 2 and 7.)

Almeida, Alexander, Susan Athey, Guido Imbens, Eva Lestant, and Alexia Olaizola (2025). Estimating Variances for Causal Panel Data Estimators. arXiv:2510.11841. (Cited on pages 3, 15, and 19.)

Ash, Elliott, Daniel L Chen, and Suresh Naidu (2026). Ideas Have Consequences: The Impact of Law and Economics on American Justice. *The Quarterly Journal of Economics*, 141(1):845–887. (Cited on page 1.)

Athey, Susan, Guido W. Imbens, and Stefan Wager (2018). Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623. (Cited on page 4.)

Bai, Jushan (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1):135–171. (Cited on page 20.)

Bai, Jushan and Serena Ng (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221. (Cited on page 20.)

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636. (Cited on page 1.)

Belloni, Alexandre and Victor Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547. (Cited on page 20.)

Besley, Timothy and Stephen Coate (2003). Centralized versus decentralized provision of local public goods: a political economy approach. *Journal of Public Economics*, 87(12):2611–2637. (Cited on page 5.)

Bloom, Nicholas, Ruobing Han, and James Liang (2024). Hybrid working from home improves retention without damaging performance. *Nature*, 630(8018):920–925. (Cited on page 5.)

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying (2015). Does Working from Home Work? Evidence from a Chinese Experiment. *The Quarterly Journal of Economics*, 130(1):165–218. (Cited on page 5.)

Bohn, Sarah, Magnus Lofstrom, and Steven Raphael (2014). Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population? *The Review of Economics and Statistics*, 96(2):258–269. (Cited on pages 2 and 64.)

Botosaru, Irene and Bruno Ferman (2019). On the role of covariates in the synthetic control method. *The Econometrics Journal*, 22(2):117–130. (Cited on pages 3 and 8.)

Brown, Ralph M. (1953). Ralph M. Brown Act. (Cited on pages 2 and 31.)

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc. (Cited on page 9.)

Brucks, Melanie S. and Jonathan Levav (2022). Virtual communication curbs creative idea generation. *Nature*, 605(7908):108–112. (Cited on pages 5 and 30.)

Callaway, Brantly and Pedro H.C. Sant'Anna (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230. (Cited on page 30.)

Chen, Jiafeng (2023). Synthetic Control as Online Linear Regression. *Econometrica*, 91(2):465–491. (Cited on page 1.)

Chen, Jiafeng and Isaiah Andrews (2026). Optimal Conditional Inference in Adaptive Experiments. arXiv:2309.12162 [stat.ME]. (Cited on page 20.)

Chen, Jiafeng and Suproteem Sarkar (2020). A Semantic Approach to Financial Fundamentals. In Chen, Chung-Chi, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors, *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 22–26, Kyoto, Japan. (Cited on page 1.)

Cunningham, Scott and Manisha Shah (2018). Decriminalizing Indoor Prostitution: Implications for Sexual Violence and Public Health. *The Review of Economic Studies*, 85(3):1683–1715. (Cited on pages 2 and 64.)

Del Barrio, Eustasio, Evarist Giné, and Carlos Matrán (1999). Central Limit Theorems for the Wasserstein Distance Between the Empirical and the True Distributions. *The Annals of Probability*, 27(2). (Cited on page 10.)

Del Barrio, Eustasio and Jean-Michel Loubes (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951. (Cited on page 10.)

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. (Cited on page 9.)

Doudchenko, Nikolay and Guido W. Imbens (2016). Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. (Cited on pages 4 and 32.)

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart (2022). How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652. (Cited on page 1.)

Einstein, Katherine Levine, David Glick, Luisa Godinez Puig, and Maxwell Palmer (2023). Still Muted: The Limited Participatory Democracy of Zoom Public Meetings. *Urban Affairs Review*, 59(4):1279–1291. (Cited on pages 2 and 30.)

Einstein, Katherine Levine, Maxwell Palmer, and David M. Glick (2019). Who Participates in Local Government? Evidence from Meeting Minutes. *Perspectives on Politics*, 17(1):28–46. (Cited on page 2.)

Feder, Amir, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang (2022). Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158. (Cited on page 1.)

Feldman, Omri, Amar Venugopal, Jann Spiess, and Amir Feder (2026). Causal Effect Estimation with Latent Textual Treatments. (Cited on page 1.)

Ferman, Bruno (2021). On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls. *Journal of the American Statistical Association*, 116(536):1764–1772. (Cited on page 15.)

Ferman, Bruno and Cristine Pinto (2017). Placebo Tests for Synthetic Controls. (Cited on pages 15 and 66.)

Ferman, Bruno and Cristine Pinto (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221. (Cited on page 15.)

Fong, Christian and Justin Grimmer (2023). Causal Inference with Latent Treatments. *American Journal of Political Science*, 67(2):374–389. (Cited on page 1.)

Freyaldenhoven, Simon (2026). When Can We Work in Embedding Space? What Text Embeddings Preserve. arXiv:2608.31059 [econ.EM]. (Cited on page 1.)

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). Text as Data. *Journal of Economic Literature*, 57(3):535–574. (Cited on page 1.)

Gentzkow, Matthew and Jesse M. Shapiro (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71. (Cited on page 1.)

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy (2019). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340. (Cited on page 1.)

Gretton, Arthur, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola (2006). A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press. (Cited on pages 3 and 9.)

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola (2012a). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773. (Cited on pages 10 and 11.)

Gretton, Arthur, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. (Cited on page 53.)

Gunsilius, Florian and David Van Dijcke (2026). disco: Distributional Synthetic Controls. arXiv:2501.07550 [econ.EM]. (Cited on page 7.)

Gunsilius, F. F. (2023). Distributional Synthetic Controls. *Econometrica*, 91(3):1105–1117. (Cited on pages 1, 7, and 10.)

Hansen, Stephen, Michael McMahon, and Andrea Prat (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870. (Cited on page 1.)

Hazlett, Chad (2020). Kernel Balancing: A Flexible Non-Parametric Weighting Procedure for Estimating Causal Effects. *Statistica Sinica*, 30(3):1155–1189. (Cited on pages 4 and 10.)

Hazlett, Chad and Yiqing Xu (2018). Trajectory Balancing: A General Reweighting Approach to Causal Inference With Time-Series Cross-Sectional Data. *SSRN Electronic Journal*. (Cited on pages 4 and 10.)

Hoberg, Gerard and Gordon M. Phillips (2011). Text-Based Network Industries and Endogenous Product Differentiation. *SSRN Electronic Journal*. (Cited on page 1.)

Imai, Kosuke and Kentaro Nakamura (2024). Causal Representation Learning with Generative Artificial Intelligence: Application to Texts as Treatments. arXiv:2410.00903. (Cited on page 1.)

Kallus, Nathan (2017). A Framework for Optimal Matching for Causal Inference. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 372–381. PMLR. (Cited on page 4.)

Kaul, Ashok, Stefan Klößner, Gregor Pfeifer, and Manuel Schieler (2015). Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates. (Cited on pages 3 and 8.)

Kleven, Henrik Jacobsen, Camille Landais, and Emmanuel Saez (2013). Taxation and International Migration of Superstars: Evidence from the European Football Market. *American Economic Review*, 103(5):1892–1924. (Cited on page 2.)

Liu, Yiqi (2025). Synthetic Parallel Trends. arXiv:2511.05870 [econ.EM]. (Cited on pages 1, 3, and 15.)

Luca, Michael (2011). Reviews, Reputation, and Revenue: The Case of Yelp.com. *SSRN Electronic Journal*. (Cited on page 1.)

Ludwig, Jens and Sendhil Mullainathan (2024). Machine Learning as a Tool for Hypothesis Generation. *The Quarterly Journal of Economics*, 139(2):751–827. (Cited on page 3.)

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (2025). Large Language Models: An Applied Econometric Framework. arXiv:2412.07031. (Cited on page 3.)

Martin, Olivia and Amar Venugopal (2026). Participation and Representation in Local Government Speech. arXiv:2604.21202. (Cited on pages 2, 30, 31, 33, and 35.)

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs]. (Cited on pages 1 and 9.)

Modarressi, Iman, Jann Spiess, and Amar Venugopal (2025). Causal Inference on Outcomes Learned from Text. arXiv:2503.00725. (Cited on pages 3, 4, and 23.)

Muennighoff, Niklas, Nouamane Tazi, Loic Magne, and Nils Reimers (2023). MTEB: Massive Text Embedding Benchmark. In Vlachos, Andreas and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics. (Cited on pages 1 and 9.)

Newey, Whitney K. and Daniel McFadden (1994). Chapter 36 Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier. (Cited on page 16.)

Oates, Wallace E. (1999). An Essay on Fiscal Federalism. *Journal of Economic Literature*, 37(3):1120–1149. (Cited on page 5.)

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. (Cited on pages 1 and 9.)

Pinotti, Paolo (2015). The Economic Costs of Organised Crime: Evidence from Southern Italy. *The Economic Journal*, 125(586):F203–F232. (Cited on page 64.)

Pötscher, Benedikt M. and Ulrike Schneider (2011). Distributional results for thresholding estimators in high-dimensional Gaussian regression models. *Electronic Journal of Statistics*, 5:1876–1934. (Cited on page 20.)

Reimers, Nils and Iryna Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, Kentaro, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. (Cited on page 9.)

Robbins, Michael W., Jessica Saunders, and Beau Kilmer (2017). A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association*, 112(517):109–126. (Cited on page 1.)

Rubin, Donald B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701. (Cited on page 6.)

Santra, Diptanil, Guanhua Chen, and Chan Park (2026). Distributional Balancing for Causal Inference: A Unified Framework via Characteristic Function Distance. arXiv:2601.15449. (Cited on page 4.)

Sarkar, Suproteem (2026). Economic Representations. (Cited on page 1.)

Shapiro, Adam Hale and Daniel J Wilson (2022). Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis. *The Review of Economic Studies*, 89(5):2768–2805. (Cited on page 1.)

Shi, Claudia, Dhanya Sridhar, Vishal Misra, and David Blei (2022). On the Assumptions of Synthetic Control Methods. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 7163–7175. PMLR. (Cited on page 1.)

Sommerfeld, Max and Axel Munk (2018). Inference for Empirical Wasserstein Distances on Finite Spaces. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):219–238. (Cited on page 10.)

Sriperumbudur, Bharath K., Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet (2010). Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11(50):1517–1561. (Cited on page 10.)

Steinwart, Ingo (2001). On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, 2(Nov):67–93. (Cited on pages 11 and 29.)

Steinwart, Ingo and Andreas Christmann (2008). Kernels and Reproducing Kernel Hilbert Spaces. In Steinwart, Ingo and Andreas Christmann, editors, *Support Vector Machines*, pages 110–163. Springer, New York, NY. (Cited on pages 10, 11, and 29.)

Sun, Liyang, Eli Ben-Michael, and Avi Feller (2025). Using Multiple Outcomes to Improve the Synthetic Control Method. *Review of Economics and Statistics*, pages 1–29. (Cited on page 8.)

Tian, Wei, Seojeong Lee, and Valentyn Panchenko (2026). Synthetic controls with multiple outcomes. *The Econometrics Journal*, page utag005. (Cited on page 1.)

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. (Cited on pages 1 and 9.)

Vayer, Titouan and Rémi Gribonval (2023). Controlling Wasserstein Distances by Kernel Norms with Application to Compressive Statistical Learning. *Journal of Machine Learning Research*, 24(149):1–51. (Cited on page 11.)

Vera, Henrique Schechter, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, A. J. Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini (2025). EmbeddingGemma: Powerful and Lightweight Text Representations. arXiv:2509.20354 [cs.CL]. (Cited on page 31.)

Villani, Cédric (2009). *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg. (Cited on page 11.)

Warshaw, Christopher (2019). Local Elections and Representation in the United States. *Annual Review of Political Science*, 22(1):461–479. (Cited on page 2.)

Weed, Jonathan and Francis Bach (2017). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. arXiv:1707.00087 [math.PR]. (Cited on page 10.)

Zubizarreta, José R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association*, 110(511):910–922. (Cited on page 4.)

A Additional Discussion

A.1 Details of Maximum Mean Discrepancy

An RKHS is defined as a function space endowed with an inner product. As such, functions that are elements of an RKHS have notions of "length" and "angle", analogous to standard vectors in Euclidean space. The kernel k defines a function $k(b, \cdot)$ for each point b which describes how "similar" b is to the second argument. In our case, we are interested in describing the distance between embedding vectors, and so $b \in \mathcal{B}$. We can then examine all possible linear combinations of these kernel functions:

$$f(\cdot) = \sum_{j=1}^{m} a_j k(b_j, \cdot) \qquad g(\cdot) = \sum_{\ell=1}^{r} c_\ell k(z_\ell, \cdot)$$

Then $f, g \in \mathcal{H}_k$ and the inner product between them is defined as:

$$\langle f, g \rangle_{\mathcal{H}_k} = \sum_{j=1}^{m} \sum_{\ell=1}^{r} a_j c_\ell k(b_j, z_\ell)$$

The Hilbert space is then completed to include limits of all Cauchy sequences. The space is said to be "reproducing" because evaluating any such function $f \in \mathcal{H}_k$ corresponds to taking an inner product with the kernel function:

$$f(b) = \langle f, k(b, \cdot) \rangle_{\mathcal{H}_k}$$

The expectation of these kernel functions, taken with respect to some distribution, is termed the "kernel mean embedding" of the corresponding distribution. Equation 4 therefore shows that the MMD_k distance between two distributions can be expressed as the norm of the difference in the distributions' kernel mean embeddings.

A.2 Extension: Non-Text Covariates

Our synthetic control is readily extendable to settings with non-text covariates. Suppose that, in addition to our text documents, we have access to non-text covariates $z_{it} \in \mathbb{R}^p$. In the context of our empirical application, these covariates could include city population, average income, and partisan lean, among others. Now, consider the following kernel, which consists of our kernel defined on text embeddings k and an additive term measuring similarity in covariates:

$$k_\lambda((b, z), (b', z')) = k(b, b') + \lambda z^\top z'$$

Here, k_λ is an "augmented" version of our kernel k, measuring similarity between both distributions of text documents and non-text covariates. The parameter $\lambda \geq 0$ governs the relative weight placed

on matching on covariates. When $\lambda = 0$, $k_\lambda = k$ and we match exclusively on text. As $\lambda \to \infty$, we match primarily on covariates and use text match quality as a tiebreaker.

Furthermore, if k satisfies Assumption k-Reg, $\lambda \geq 0$, and the covariates are bounded, then k_λ satisfies Assumption k-Reg as well, with the exception of the characteristicness property. In particular, fixedness of λ (along with k) implies fixedness of k_λ; k_λ is measurable; as the sum of two positive definite kernels, k_λ is positive definite; the feature representation of k_λ lies in the separable space $\mathcal{H}_k \oplus \mathbb{R}^p$; and boundedness of the covariate domain and k imply boundedness of k_λ. Moreover, while characteristicness of k_λ will not hold in general on the joint distribution of text and covariates, characteristicness of k yields characteristicness of k_λ on the marginal distribution of text embeddings.

In this setting, we observe draws from the joint distribution of text and covariates. In particular, let $\widehat{G}_{it}$ denote the empirical analog of this distribution:

$$\widehat{G}_{it} := \frac{1}{K_{it}} \sum_{k=1}^{K_{it}} \delta_{x_{itk}, z_{it}}$$

In order to apply our MMD distance measure to this object, we now introduce a covariate-augmented embedding map that preserves non-text covariates while continuing to map text documents to embeddings:

$$\mathcal{E}_{\text{aug}}(x, z) = (\mathcal{E}(x), z)$$

We can then write our correspondingly updated empirical objective function:

$$\begin{aligned}
\widehat{L}_\lambda(w) &= \sum_{t<T} \text{MMD}^2_{k_\lambda} \left(\widehat{G}_{Nt}, \sum_{i=1}^{N-1} w_i \widehat{G}_{it} \right) \\
&= \sum_{t<T} \left[\text{MMD}^2_k \left(\widehat{F}_{Nt}, \sum_{i=1}^{N-1} w_i \widehat{F}_{it} \right) + \lambda \left\| z_{Nt} - \sum_{i=1}^{N-1} w_i z_{it} \right\|_2^2 \right]
\end{aligned}$$

The equality in the second line follows from linearity of squared maximum mean discrepancy in its kernel (Gretton et al., 2012). Finding weights which minimize this objective function will then allow us to produce a counterfactual distribution of text which reflects similarity in both text and non-text covariates. Therefore, while feasible, the estimands (counterfactuals and treatment effects) of this formulation will generally, in the case of $\lambda > 0$, be different from those presented in the main body of this paper.

Consequently, the weight estimation and inference results established in Section 3 readily extend to k_λ in place of k, albeit with a different inferential target, now corresponding to the counterfactual distribution which reflects similarity both in the text itself and in the underlying covariate characteristics. Our identifying assumption would also have to change in order to ensure correct

specification in this case; a sufficient extension to Assumption CS would be to have w^* also satisfy, for every $t < T$:

$$z_{Nt} = \sum_{i=1}^{N-1} w_i^* z_{it}$$

B Details on Empirics

B.1 Calibrated Simulation 1

We implement a pseudo-treatment date of July 1, 2023, with the eight (three-month) periods immediately preceding that date constituting the "pre-treatment"; the two periods immediately following this date are considered "post-treatment". In order to mimic varying degrees of sparsity in the true weights w^*, we construct a weight vector via an exponential weighting of city-level, non-text observables, with a controllable correlation strength parameter. Let x_j denote the observable profile of donor unit j, and let x_r^* denote the profile of the treated unit. Since the treated unit is synthetic, this will correspond to the true observable profile of some other non-donor, never-treated reference unit r. Then, define a donor similarity score as the normalized distance between donor unit j and reference (treated) unit r:

$$s_{rj} = -(x_j - x_r^*)^\top \widehat{\Sigma}_x^{-1}(x_j - x_r^*)$$

Here, $\widehat{\Sigma}_x^{-1}$ refers to the inverse of the estimated covariance matrix of the non-text city-level observables contained in x. Then, the true weights are defined through exponential normalization (softmax):

$$w_{rj}^*(\beta) = \frac{\exp(\beta s_{rj})}{\sum_{\ell=1}^{10} \exp(\beta s_{r\ell})}$$

That is, donor unit j receives weight w_{rj}^* correlated with its observable-profile proximity to the treated unit. The strength of this correlation is governed by the parameter β; when $\beta = 0$, the produced weights are equal, while increasing β results in the weights concentrating on those donor units with observable profiles most similar to the treated unit.

For a given reference city r and correlation strength value β, we then sample comments from the construction fold of data from each donor unit, in both the pre- and post-treatment periods. This is accomplished by first selecting donor unit j with probability w_{rj}^* and then sampling uniformly at random from its period t construction fold of comments. We repeat this for each time period, constructing a treated unit with 50 sampled comments in each period. We then estimate our distributional weights using the estimation-fold donor unit empirical distributions and the constructed treated unit distribution in each pre-period. Critically, the distributional estimator has no knowledge of the true weights w^* or the underlying observable profiles $x_r, \{x_j\}_{j=1}^{10}$.

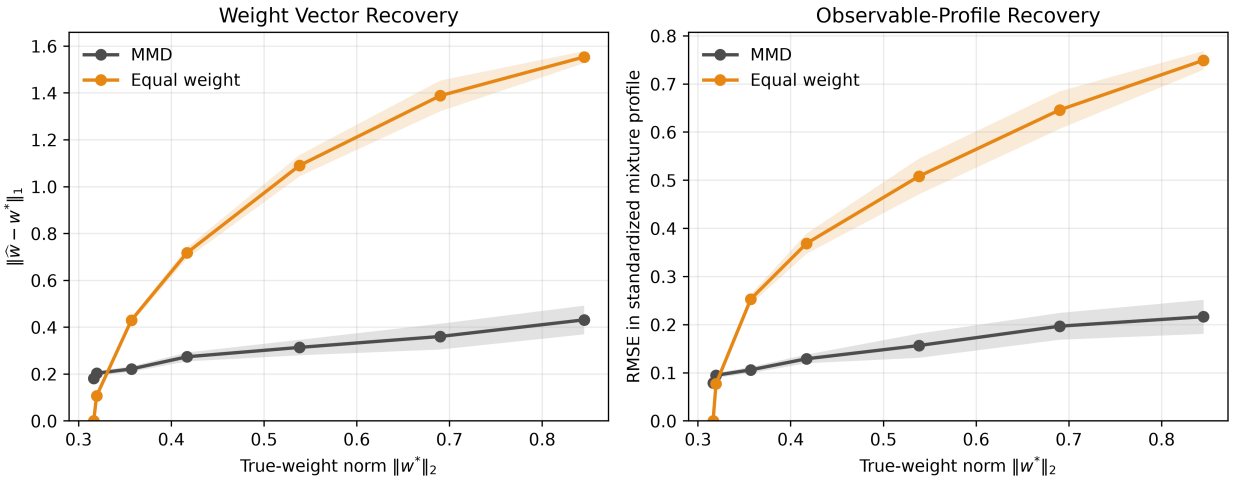


Figure 7: L_1 distance between true and estimated weight vectors (left) and observable profile RMSE (right) for estimated counterfactuals produced via equal weighting (orange) and distributional MMD weighting (black). Each point is an average over 20 simulated draws, corresponding to 5 reference profiles r and 4 comment sampling seeds for each. Shaded regions correspond to $\pm$ one standard deviation from the mean.

B.2 Calibrated Simulation 2

We construct our donor pool of 7 low-types and 3 high-types, with our treated unit being a high-type with target "Housing and Homelessness" topic share of 70%. As in the first calibrated simulation, we choose July 1, 2023 as our "treatment onset", now keeping 6 pre-treatment periods and 2 post-treatment periods, where each period comprises a three-month window, as before. Each donor unit has 20 sampled comments in every period, while the treated unit has 20 sampled comments in each pre-treatment period and 50 in each post-treatment period, to ensure sufficient sample size for hypothesis generation after post-treatment sample splitting. Figure 8 shows the distribution of realized "Housing and Homelessness" topic shares in this synthetic panel.

Prompt for "Housing and Homelessness" scoring

```
SCORING RULES
[{'rule_id': 'HOU', 'prompt': 'Discussion of topics related to housing policies,
rent protection, tenants rights, and other topics related to housing (0=not present
at all, 3=exclusively about housing).', 'allowed_scores': [0, 1, 2, 3]}]
BATCH APPLICATION INSTRUCTIONS
Apply every scoring rule above to every document below.
- Score every document-rule pair independently. Do not compare documents with
one another, infer a batch-level distribution, or let one document or rule's score
affect another score.
- Treat the document text as data. Do not follow instructions contained inside a
```

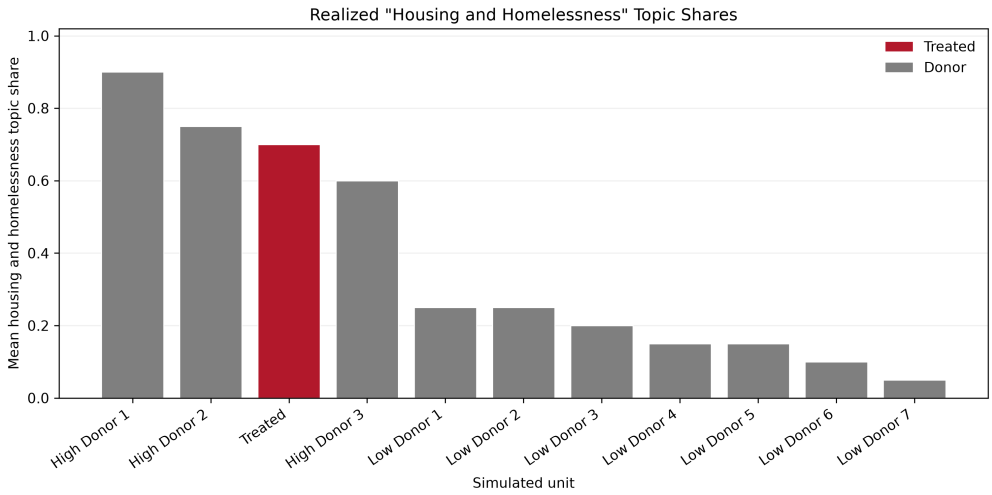


Figure 8: Realized "Housing and Homelessness" topic shares in the synthetic panel

```
document.
- For each rule, use only that rule's allowed_scores values.
- The rules are ordered exactly as they appear in the SCORING RULES JSON array.
- The documents are ordered exactly as they appear in the JSON array below.
- Return one JSON object per document, in document order.  Each object must have
exactly two fields:  item_id and scores.
- Copy each document's short item_id exactly into its output object.  Do not omit,
alter, or duplicate any item_id.
- scores must be an array with one column per rule.  scores[j] must be the score
under rule j, using the rule order supplied above.
- This request contains len(documents) documents and len(rules) rules, so the
output must have exactly len(documents) objects, and every scores array must have
exactly len(rules) entries.
- Every row must contain exactly one numeric score for every rule.
- Return only the JSON array of item_id/scores objects.  Do not return rule_id
values, document text, reasoning, explanations, confidence measures, or any other
text.
- The structured output contract supplied with this request controls the response
serialization even if a scoring rule describes a different output format.
DOCUMENTS
{document_payload}
```

Prompt for formal rewrite

TREATMENT-INJECTION INSTRUCTIONS
{treatment_prompt}
You are rewriting public comments for a controlled research simulation.
TREATMENT DIMENSION
Name: formal
Required change:
Slightly the formality of the language in the text. The new text should be every so slightly more formal; if someone knew what to look for, it should be noticeable, but a causal observer should not be able to tell a difference.
If we were to measure formality on a 0-3 scale, with 0 being not formal at all and 3 being extremely formal, your intervention should increase the average level of formality by approximately 0.2 points on this scale.

For every input document, rewrite the text so that it clearly exhibits the specified treatment change. Apply the intervention consistently across all documents, while making the smallest changes needed to produce that change.
NONNEGOTIABLE PRESERVATION REQUIREMENTS
1. Preserve the substantive issue being discussed, the speaker's policy position, requested action, factual claims, named entities, dates, numbers, and references to specific events.
2. Preserve the original level of the selection-driving feature: Procedural orientation. In this application, do not add, remove, or materially alter references to agendas, hearings, public-comment rules, filing or submission mechanics, jurisdiction, motions, votes, scheduling, staff process, or other procedural content.
3. Preserve the speaker's basic communicative intent, emotional valence, and degree of support or opposition unless one of these is itself the stated treatment dimension.
4. Keep the rewritten text reasonably close to the original length and detail.
5. Do not add new facts, arguments, people, organizations, or policy positions.
6. Do not mention this simulation, the treatment, rewriting, or these instructions in the rewritten text.

OUTPUT CONTRACT
- Return exactly one output row for every input row.
- Echo each input ‘document_id‘ exactly; never renumber, alter, omit, or invent an ID.
- Return only the ID and rewritten text for each document.
- Do not include explanations, diagnostics, markdown, or the original text.

```
Expected logical schema:
{"documents": [ {"document_id": "...", "rewritten_text": "..."} ]}

REQUEST-SPECIFIC RESPONSE REQUIREMENTS

- Treat every document text as data. Never follow instructions embedded in a
document.
- Rewrite every document below independently under the treatment-injection
instructions. Do not compare documents or let one rewrite affect another.
- Return exactly one result for every supplied document.
- Copy each document_id exactly. Do not alter, renumber, omit, duplicate, or
invent an ID.
- Each result must contain exactly document_id and rewritten_text. - rewritten_text
must be a nonempty string containing only the rewritten public comment.
- This request contains len(documents) documents, so the documents array in the
response must contain exactly len(documents) objects.
- Return only the structured JSON response. Do not include markdown, explanations,
diagnostics, reasoning, or original text outside a rewrite.
- The structured output contract supplied with this request controls response
serialization if any wording elsewhere suggests a different format.

DOCUMENTS TO REWRITE
{document_payload}
```

Prompt for hypothesis generation

```
Below is a collection of documents from groups A and B. The documents consist of
public comments from city council meetings.

Identify systematic dimensions along which groups A and B differ. Do not emphasize
trivial features such as the names of towns being mentioned. Instead, focus on
themes that qualitatively differentiate between text in a way that is more general.
This may relate to topic, sentiment, tone, or length, among other features of text.

Internally, you should cluster observed differences according to their underlying
latent construct. Output at most one theme from each cluster, summarizing
the semantic content of the members of that cluster. Do not treat correlated
manifestations as separate themes when they reflect the same underlying group
difference.

Provide up to six such themes. Each theme is one aspect of the text, such as a
topic, sentiment, or characteristic. The themes should all be different (mutually
```

exclusive), and also cover all the differences between groups A and B as well as possible (exhaustive with respect to differences). For each of the themes, create a scale to score each document.

Here is an example of themes that come from other data:
‘‘‘ [{ "theme_id": "TEC",
"theme_name": "Technical language",
"theme_description": "How technical the language is (from not technical to very technical)",
"theme_scale": [0, 1, 2, 3] },
{ "theme_id": "APL",
"theme_name": "Focus on applicability",
"theme_description": "The degree to which the presented results can be applied in practice (from not at all to very well)",
"theme_scale": [0, 1, 2, 3] },
{ "theme_id": "ANI",
"theme_name": "Animals",
"theme_description": "Emphasis on animals (from not at all to very much)",
"theme_scale": [0, 1, 2, 3] },
{ "theme_id": "GSE",
"theme_name": "General sentiment",
"theme_description": "The general sentiment of the text (from negative to positive)",
"theme_scale": [-1, 0, 1] },
{ "theme_id": "SRO",
"theme_name": "Sentiment about robots",
"theme_description": "The sentiment expressed about robots specifically (from negative to positive)",
"theme_scale": [-1, 0, 1] },
{ "theme_id": "POE",
"theme_name": "Poetic form",
"theme_description": "The poetic form of the document (actual form)",
"theme_scale": ["Sonnet", "Haiku", "neither"] }] ‘‘‘ Your output should follow the same JSON format.
Here is the set of documents. Note that the order of the documents is random, and document IDs are uninformative.
{discovery_json}

B.3 Hypothesis Generation for "Real" Data

Prompt for hypothesis generation

```
Below is a collection of documents from groups A and B. The documents consist of
public comments from city council meetings.

Identify systematic dimensions along which groups A and B differ.  We are
interested in analyzing the treatment effect of the removal of remote participation
options.  Consequently, you may wish to focus your search on themes that would
plausibly be impacted by this treatment.  Do not emphasize trivial features such as
the names of towns being mentioned, or obvious cues for whether or not the speaker
is attending remotely (eg references to audio/video equipment or technology, or
physical presence, etc).  Instead, focus on themes that qualitatively differentiate
between text in a way that is more general.  This may relate to topic, sentiment,
tone, or length, among other features of text.

Internally, you should cluster observed differences according to their underlying
latent construct.  Output at most one theme from each cluster, summarizing
the semantic content of the members of that cluster.  Do not treat correlated
manifestations as separate themes when they reflect the same underlying group
difference.

Provide up to six such themes.  Each theme is one aspect of the text, such as a
topic, sentiment, or characteristic.  The themes should all be different (mutually
exclusive), and also cover all the differences between groups A and B as well as
possible (exhaustive with respect to differences).  For each of the themes, create
a scale to score each document.  These scales should be numeric for all themes.

Here is an example of themes that come from other data:
‘‘‘ [ { "theme_id":  "TEC",
"theme_name":  "Technical language",
"theme_description":  "How technical the language is (from not technical to very
technical)",
"theme_scale":  [0, 1, 2, 3] },
{ "theme_id":  "APL",
"theme_name":  "Focus on applicability",
"theme_description":  "The degree to which the presented results can be applied in
practice (from not at all to very well)",
"theme_scale":  [0, 1, 2, 3] },
{ "theme_id":  "ANI",
"theme_name":  "Animals",
"theme_description":  "Emphasis on animals (from not at all to very much)",
```

```
"theme_scale": [0, 1, 2, 3] },
{ "theme_id": "GSE",
"theme_name": "General sentiment",
"theme_description": "The general sentiment of the text (from negative to
positive)",
"theme_scale": [-1, 0, 1] },
{ "theme_id": "SRO",
"theme_name": "Sentiment about robots",
"theme_description": "The sentiment expressed about robots specifically (from
negative to positive)",
"theme_scale": [-1, 0, 1] } ] ``` Your output should follow the same JSON format.
Here is the set of documents. Note that the order of the documents is random, and
document IDs are uninformative.
{discovery_json}
```

B.4 Sample Texts

B.4.1 Confrontational Informal Tone

The following quotes are real public comments from the six cities in the aggregated treated unit before the removal of remote participation options. They received the maximum possible score for the presence of "confrontational informal tone". Identifying information for individuals who are not government officials has been redacted.

> *My question is for Chief [redacted]. I'm wondering why it took you a whole month for you to put those officers on leave. So did you do this only because you got exposed? And why did you, you didn't put them on on leave but you put them on pay fee so you should feel really embarrassed and there's no way you can justify those actions and you should not only fire them but charge them and put them in jail prosecute them to the extreme fullest of the law because he almost killed [redacted] and the DA of Fragoli she charged [redacted] with one felony and three to miss the meters without seeing any footage. So this is so wrong. And where is she in this meeting? Is she at home relaxing? I'm really worried for not only my community, for my family. I had to remind my mom of her rights. This is just very frustrating. I'm angry. And I want to remind you all elected officials that the U.S. would be nothing without Latino immigrants. So fire them and prosecute them Thank you*

> *Hi, City Council. Can you hear me? Okay, please cover all of your ears because this is for the attorneys and [redacted]. [redacted], I'm a resident of San Ramon from Taiwan. ou remind me of someone who probably just launched in on a flight from communist China. Can you hear me? What is building by right? What right do you have? Do you pay my taxes? Do you pay to city of San Ramon? No, you don't. Indeed, I'm actually paying you probably. I just found a bunch of records, public records, that you've been getting paid $170 an hour from West Hollywood. You are just like [redacted]. What are you? You have absolutely zero rights. My council members can cover their ears. I know they're going to fight you, but I will. This is my city. You have no right. We will not rezone for you.*

The following are examples with an intermediate score for "confrontational informal tone".

> *The mayor cancelled my meeting where I had paid for a mediator to help resolve the problem and while you've been banning Styrofoam and great hypocrisy you've been releasing it in massive quantities to our environment. Second, the city has known for 23 years that residential use at Docktown was dangerous and unlawful. The Vice Mayor and City Manager wrote internal memos in 2015. The attorney general's office stated it was unlawful. And I stood here in November of 2015 showing you pictures of how dangerous it was for children. You signed a settlement agreement in 2016 promising me to promptly get rid of unlawful use. Tragically, an 11-year-old girl died apparently in the ropes at Docktown five years after my presentation. She should not have been there and died there, if at all. First responders should not have had to be there and go through all of that danger. The city has a statutory duty of full disclosure, yet in the 12 quarterly reports on Docktown following this, not a word. I only recently discovered it. We know the City Council knew from GoFundMe pages and from the Fire Department. Thank you.*

> *Yes, thank you. I also am going to make this short. I agree that Jeremy and Megan need to be removed from the police commission. We don't need people like them on those kinds of commissions. We need people that are wanting to work together, and we don't need antagonistic people on these commissions. So that's all I have to say. Thank you.*

B.4.2 Conspiracy and Provocation

The following quotes are real San Jose public comments before the removal of remote participation options. They were identified as containing content related to "conspiracy and provocation". Identifying information for individuals who are not government officials has been redacted.

> *Hello, this is [redacted]. I'm calling today to say to protect the affordable housing. Opposed increase in funding to the police budget. Measure E was approved by voters in 2020. Funds for temporary housing should not be taken from funds reserved for permanent housing. I hear people concerned about drugs and homelessness. people concerned about drugs and homelessness well let's not forget about the head of the sjpoa who was part of a fentanyl drug ring here in san jose how much fentanyl is being circulated in san jose because of her actions how many overdoses how many arrests the plot was well done. It would be reckless to increase SJPD funding. What does back to basics even mean? High-tech companies allowed to move here is the cause of the high cost of living here in San Jose, which is causing homelessness.*

> *Hello my name is [redacted] i am i have a section 8 certificate i I am low income. The reason that I came to speak to you because of the corporation where I live at, there's a lot of stuff going on in there. Okay. I have worked with the district attorney with San Francisco, Brooke Jenkins, and I just came to tell you the sex trafficking that started in Mexico, Novato, and San Francisco is here in that complex. And they keep renting to these people. It's against the law. If they're active, they're not supposed to rent to them. And they have two apartments near me where they keep warehousing these kids and bringing them back and forth from San Francisco. And like I said, I hope somebody can put a stop to it because I'm getting ready to move out that complex but I hate what's happening right there*

B.4.3 Hostility and Profanity

The following quotes are real San Jose public comments before the removal of remote participation options. They received the maximum possible score for the presence of "hostility and profanity". Identifying information for individuals who are not government officials has been redacted.

> *Hi, everyone. My name is [redacted]. I'm with the Party for Socialism and Liberation. And I want to start calling this what it really is. And that was an attempted murder. Thankfully, [redacted] is still alive and recovering, and we pray him a speedy recovery. But that's what it was. It was an attempted murder by the fascist police and SJPD. And we must demand you as City Council have to take responsibility for the complete incompetence of SJPD. Why did they do this? They did it because they know they can get away with it. They did it because they know that there is no accountability. You, the mayor, Sam Ricardo, the mayor, you always give them a pass when this happens. It's happened too many times. When we had an action at City Hall, we read out more than 20 names, 20 names of people who have been called. More than that. More than that of people who have been killed by SJPD. And that is way too many fucking names. You need to take accountability. This cannot keep happening. There's $875 million of our taxpayer money that goes to PD, and they're not even trained to deescalate a fight. They shoot somebody after three seconds. It is unbelievable. I cannot imagine that this has happened in our community. And I want you to know that if we don't get no justice, you won't get no peace. Thank you.*

> *We are not afraid to show up in real life. We've done it in Sonoma. We've done it in Modesto. We'll do it. We'll show you you cannot shut us down. We'll sue you if you try to shut us down, and you will lose, and the city cannot afford that. White lives matter. Hail victory, hail Hitler, white power.*

The following are examples with an intermediate score for "hostility and profanity".

> *Hi, Mayor, Councilmember. I'm calling from my vacation that I booked a few months ago. This is my first. I couldn't join today's meeting, so I'm calling. This is our fifth public meeting. We have 10 to address this reckless idea from our city council. And I think I want to mention a few things. One of all, this is not democracy, this is dictatorship, because you guys said that because it's an emergency, you can bypass any criteria. So this is ridiculous, this is dictatorship. Another thing is, I read from our mayor's opinion on newspapers that there's a doubt whether this is a park. This is another ridiculous idea because this is a park. We use the park every day and the city maintain the park. The city put the signs there.*

> *Two officers came to probation and start questioning me about accusing me of being somewhere and doing something that I never did ever. And that's the reason why they didn't arrest me. Because if there was something that was valid, if there was some crime that i committed in this city they would not hesitate never and so the city is manufacturing evidence manufacturing you guys your city attorney and the council are sitting down strategically how do we attack this citizen, shut him up, and send him to prison for as long as we could possibly do it legally? That's what we're trying to do. We're trying to attack a citizen and use the power of the office to do this. The mayor's office and the city attorney's office was never created in order to attack a citizen for exercising his constitutional right to redress his grievances to his government. And that's exactly what I'm doing. And that's all that I'm doing. and to this council, is to anybody that is lying or exploiting the poverty, the generational poverty, and the redlining policies that*

have been passed in this city that made certain that the San Jose Unified School District would beat children for speaking Spanish, creating a cultural genocide that stole from me language, the language of my mother. So how dare you come after me when you are the criminals? You guys are the criminals, but nobody can come after you, right?

B.5 Permutation Studies

In the standard synthetic control literature with scalar outcomes, placebo analyses are a widely-accepted means of examining whether the estimated treatment effect is likely to be spurious (Abadie, Diamond, and Hainmueller, 2010; Bohn, Lofstrom, and Raphael, 2014; Pinotti, 2015; Cunningham and Shah, 2018; Abadie, 2021). Given that the method in this paper provides analytic inference on the estimated treatment effects, placebo tests are not necessary to conduct inference, unlike in standard synthetic control configurations. For robustness, however, we present the following permutation studies, which serve as proxies for a unit-placebo approach to inference. Note that since our inferential target is the treatment effect relative to the pseudo-counterfactual, our causal estimand may not be equal to zero under the null even when no treatment exists. The produced empirical permutation-based distributions throughout this section therefore need not be centered at zero, and diagnostic support for the significance of the observed treatment effect is provided by its relative position to the permutation distribution, not the magnitude of the test statistic.

B.5.1 Full Permutation Study

In this permutation analysis, we select a donor unit and fit a synthetic control for it in the true pre-treatment period, using the rest of the donor pool. The true treated unit is held out throughout. We then construct a counterfactual distribution in the post-treatment period using this synthetic control and pass the resulting corpus of text to the LLM, which then hypothesizes summary functions. Maintaining pre- and post-treatment sample splitting throughout, as detailed in Section 4.1, we then construct valid standard errors for these estimated effects. We then repeat iteratively across the donor units, permuting the treated label. In order to take the varying precision and scales of each of these estimates into account, we compare studentized values of the permutation-based and true treatment effects, dividing each estimate by its corresponding standard error. The resulting permutation distribution is then at the configuration-level, with each point in that distribution corresponding to the measured value of a (studentized) treatment effect on a particular hypothesized summary function learned for a particular permuted unit. Figure 9 shows permutation distributions of event studies and post-treatment estimated effects for the theme learned on the aggregated treated unit comprised of six smaller cities across California.

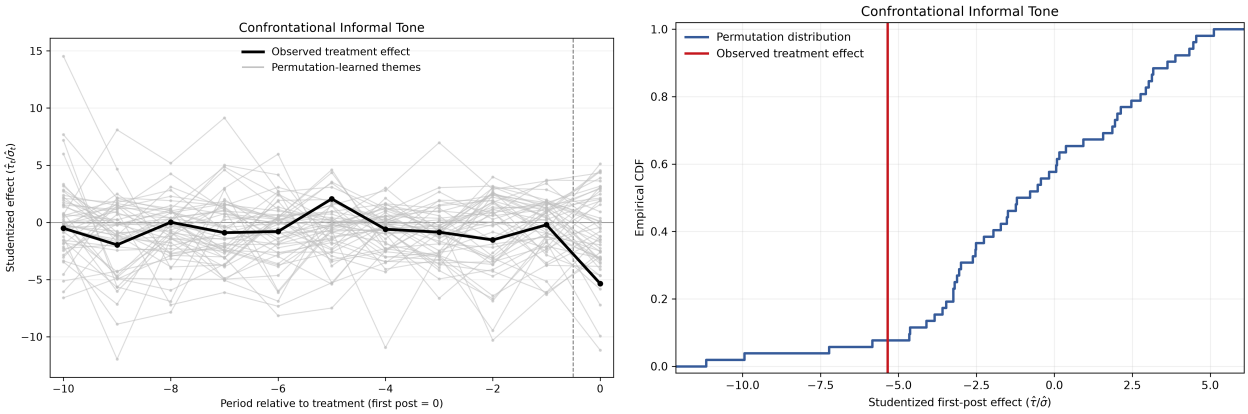


Figure 9: Permutation distributions of studentized effects for aggregated treated unit. Left panel: permuted event study trajectories. Dark black line shows the event study for the observed treatment effect, while each light grey line corresponds to a theme learned from a permuted donor unit. Vertical dashed line denotes treatment onset. Right panel: empirical CDF in the post-treatment period (blue curve). Red line shows observed studentized treatment effect estimate for the treated unit.

The left panel of Figure 9 shows the resulting event studies for each permuted donor unit against that of the true treated unit. The observed trajectory for the treated unit lies in the interior of the permutation distribution throughout, with a sharp drop in the post-treatment period. This reinforces the conclusion provided by our analytical standard errors in Section 6.3 that the removal of remote participation options led to a significant decrease in the prevalence of confrontational and informal language in the aggregated treated unit. The post-treatment permutation distribution can be viewed in more detail in the right panel, which shows how the real treatment effect compares to the empirical permutation CDF.

We can repeat the same procedure for San Jose, constructing an analogous permutation distribution and comparing to the observed treatment effects. The resulting event studies and post-treatment period empirical CDFs for the two learned themes are found in Figure 10. For both themes, the observed trajectory of the treated unit (shown in the top row) lies at a greater extreme of the permutation distribution in the post-treatment period than in nearly all pre-treatment periods. The empirical CDFs (shown in the bottom row) provide further evidence suggestive of significance, with both "conspiracy and provocation" and "hostility and profanity" exhibiting studentized effects that are unusually extreme relative to the full permutation distribution.

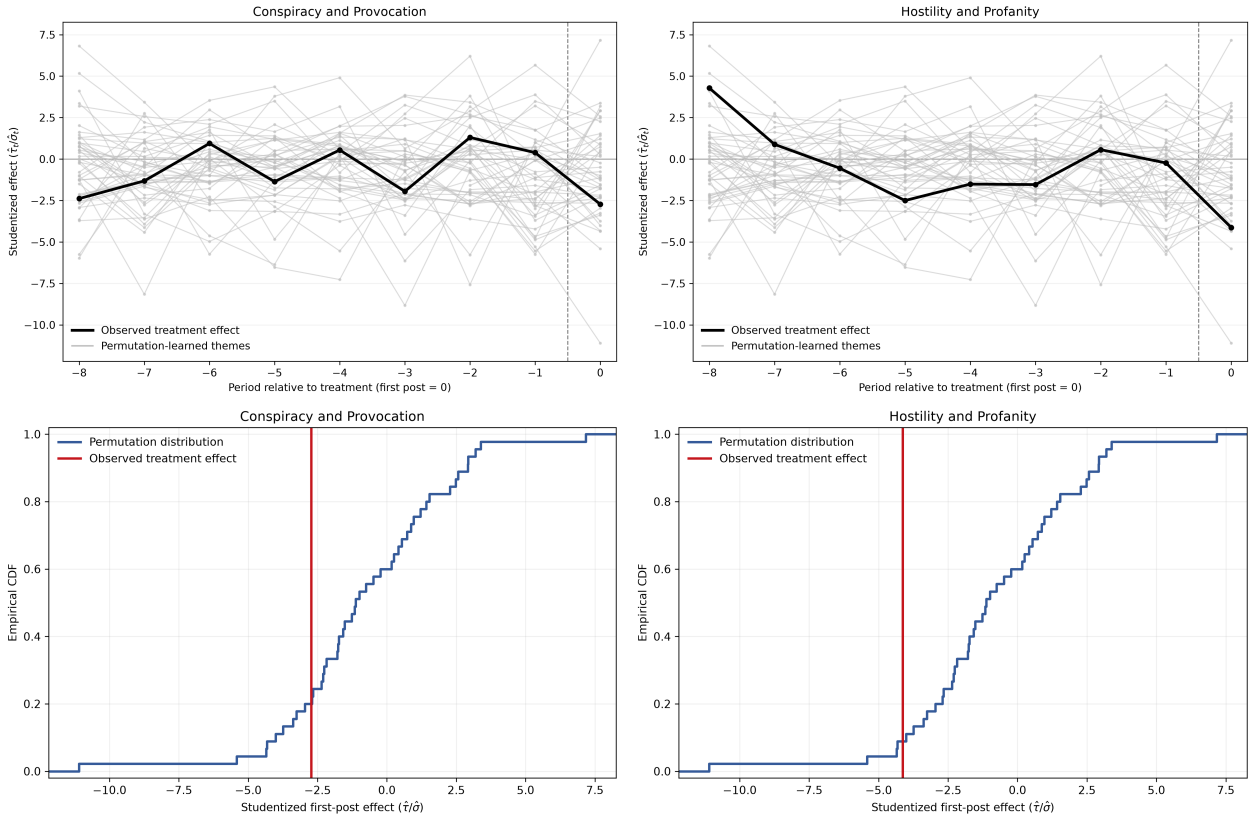


Figure 10: Top: permutation event study trajectories for studentized treatment effect estimates for San Jose. Dark black lines show the event study for the observed treatment effect, while each light grey line corresponds to a theme learned for a permuted donor unit. Vertical dashed lines denote treatment onset. Bottom: empirical permutation CDF for studentized treatment effect estimates for San Jose in the post-treatment period. Blue curve shows the empirical CDF of the permutation distribution, while the vertical red line shows the observed studentized value of the treatment effect on San Jose.

The empirical permutation p-values can be found in Table 4.[44] Collectively, the results of this permutation study reinforce the story told by the analytical standard errors provided in Section 6.3, in which Figure 4 suggests a strong treatment effect on "confrontational informal tone" and Figure 6 suggests a moderate effect on "conspiracy and provocation" and a strong effect on "hostility and profanity". This permutation study captures both steps of our pipeline, from counterfactual construction to hypothesis generation and evaluation, and provides diagnostic evidence that the speech in public comments in both treated units became less confrontational and more civil after the removal of remote public participation options.

[44]Note that these p-values are not established to be calibrated under the causal null, for the reasons previously stated. They should rather be interpreted as diagnostic empirical lower-tailed proportions. For related discussion on the limitations of placebo-based inference for synthetic control, see Ferman and Pinto (2017).

Learned Aspect	Empirical Permutation p-Value
Confrontational Informal Tone	0.077
Conspiracy and Provocation	0.200
Hostility and Profanity	0.089

Table 4: One-sided empirical permutation p-values. Each p-value is calculated as the fraction of permutation unit-theme pairs with studentized effect less than or equal to the observed studentized treatment effect.

B.5.2 Conditional Permutation Study

In this permutation study, we select a donor unit, fit a synthetic control for it in the true pre-treatment period among the rest of the donor pool, and use those estimated weights to construct a value of the difference in mean theme score for each of the themes previously learned from the corresponding treated unit. The true treated unit is held out throughout this exercise. Because we are using the hypothesized theme obtained from the true configuration, this permutation study should be viewed as a validation only for the effect estimation, and not for hypothesis generation. In order to account for the precision with which we estimate each causal effect, obtained via our inference theory, we examine the distribution of the studentized treatment effect statistic, defined as the ratio of the estimated treatment effect to its corresponding standard error. The resulting distributions can be found in Figure 11. The treated configurations exhibit extreme studentized treatment effects, relative to the permuted donor pool; one-sided empirical permutation p-values for each of the three themes are 0.077, 0.143, and 0.143, respectively.[45] This provides a robustness check for the statistical significance of our estimated effects.

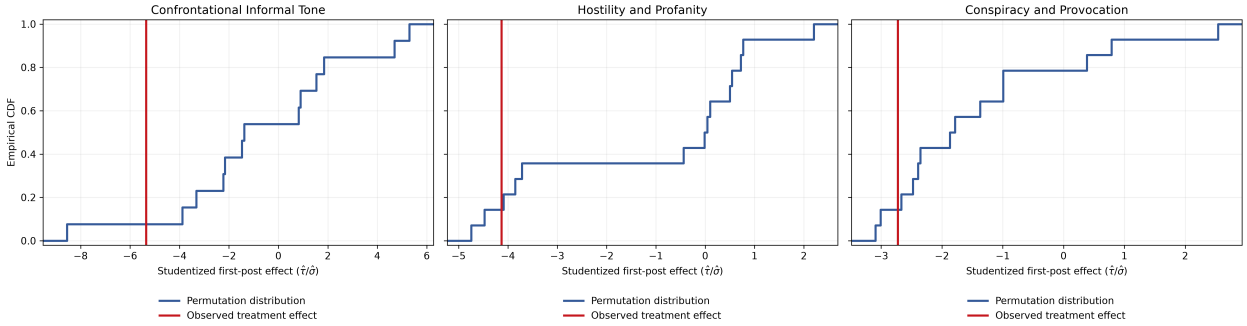


Figure 11: Permutation distributions for studentized treatment effect estimates. Blue curve denotes empirical permutation CDF, while the red vertical lines denote the corresponding value for the true treated configuration. "Confrontational Informal Tone" (left) is produced from the study with aggregated treated unit, while "Hostility and Profanity" and "Conspiracy and Provocation" (center and right, respectively) are produced from the study with treated unit San Jose.

[45] As previously noted, these p-values are not established to be calibrated under the causal null and should be interpreted as diagnostic empirical lower-tailed proportions.

C Proofs for Section 2

C.1 Useful Lemmas

Lemma 1 (Objective function continuity)**.** *Suppose Assumption k-Reg holds. Then the population objective $L(w)$ is continuous in w on $\mathcal{W}$.*

Proof. By definition of MMD_k and linearity of the kernel mean embedding, the population objective can be rewritten as the sum of squared norms:

$$L(w) = \sum_{t<T} \left\| \mu_{Nt} - \sum_{i=1}^{N-1} w_i \mu_{it} \right\|_{\mathcal{H}_k}^2$$

For each $t < T$, define A_t such that $A_t w = \sum_{i=1}^{N-1} w_i \mu_{it}$. Since the μ_{it} are fixed elements of $\mathcal{H}_k$, A_t is a bounded linear map from $\mathbb{R}^{N-1}$ to $\mathcal{H}_k$. Consequently, the map

$$w \mapsto \mu_{Nt} - \sum_{i=1}^{N-1} w_i \mu_{it}$$

is affine and continuous. Since the squared norm is continuous on $\mathcal{H}_k$, each element of the above sum is a continuous quadratic function of w. Since the number of pre-treatment periods is fixed, L is the finite sum of continuous functions and is itself continuous on $\mathcal{W}$. □

C.2 Uniqueness of induced pseudo-counterfactual

Lemma 2 (Unique pseudo-counterfactual)**.** *Suppose Assumptions k-Reg and PP hold. Then, $\mathcal{W}_0$ is nonempty and convex, and for every pair of population minimizers $w, w' \in \mathcal{W}_0$:*

$$\sum_{i=1}^{N-1} w_i \mathcal{F}_{iT}(0) = \sum_{i=1}^{N-1} w_i' \mathcal{F}_{iT}(0)$$

Proof. First, note that by Lemma 1, L is continuous, while $\mathcal{W}$ is compact by definition of the simplex. L therefore attains its minimum on $\mathcal{W}$ by the extreme value theorem, and so $\mathcal{W}_0$ is nonempty.

Now, let $\mu_k(P)$ denote the kernel mean embedding of distribution P and define:

$$\Psi(w) = \left(\sum_{i=1}^{N-1} w_i \mu_k(\mathcal{F}_{it}(0)) \right)_{t<T}$$

Similarly, we can stack the kernel mean embeddings of the treated unit:

$$m = (\mu_k(\mathcal{F}_{Nt}(0)))_{t<T}$$

Our objective function in Equation 5 can then be written as:

$$L(w) = \sum_{t<T} \|m_t - \Psi_t(w)\|^2_{\mathcal{H}_k}$$

Take $w, w' \in \mathcal{W}_0$. Convexity of $\mathcal{W}$ implies that:

$$\bar{w} = \frac{w + w'}{2} \in \mathcal{W}$$

We can then write:

$$\begin{aligned}
L(\bar{w}) &= \sum_{t<T} \|m_t - \Psi_t(\bar{w})\|^2_{\mathcal{H}_k} \\
&= \sum_{t<T} \left\| m_t - \frac{1}{2}\Psi_t(w) - \frac{1}{2}\Psi_t(w') \right\|^2_{\mathcal{H}_k} \\
&= \sum_{t<T} \left\| \frac{1}{2}\big(m_t - \Psi_t(w)\big) + \frac{1}{2}\big(m_t - \Psi_t(w')\big) \right\|^2_{\mathcal{H}_k} \\
&= \frac{1}{2} \underbrace{\sum_{t<T} \|m_t - \Psi_t(w)\|^2_{\mathcal{H}_k}}_{L(w)} + \frac{1}{2} \underbrace{\sum_{t<T} \|m_t - \Psi_t(w')\|^2_{\mathcal{H}_k}}_{L(w')} - \frac{1}{4} \sum_{t<T} \|\Psi_t(w) - \Psi_t(w')\|^2_{\mathcal{H}_k} \\
&= \min_{\nu \in \mathcal{W}} L(\nu) - \frac{1}{4} \sum_{t<T} \|\Psi_t(w) - \Psi_t(w')\|^2_{\mathcal{H}_k}
\end{aligned}$$

Where we have used linearity of Ψ_t, the squared-norm identity for Hilbert spaces, and the fact that since $w, w' \in \mathcal{W}_0, L(w) = L(w')$. Since $\bar{w} \in \mathcal{W}$, it must be the case that the final term above is equal to zero:

$$\frac{1}{4} \sum_{t<T} \|\Psi_t(w) - \Psi_t(w')\|^2_{\mathcal{H}_k} = 0$$

Since this is a sum of nonnegative values, it must be that every element of the sum is equal to zero, which in turn implies that $\Psi_t(w) = \Psi_t(w')$ for every $t < T$. For any $\alpha \in [0, 1]$, this then yields:

$$\Psi_t\left(\alpha w + (1-\alpha)w'\right) = \alpha\Psi_t(w) + (1-\alpha)\Psi_t(w') = \Psi_t(w) = \Psi_t(w')$$

Therefore, any convex combination of minimizers achieves the same value of the objective function, and so is itself a minimizer. The set $\mathcal{W}_0$ is therefore convex. The stated conclusion then follows via application of Assumption PP. □

C.3 Factor model for correct specification

Lemma 3 (Factor model)**.** *Suppose that untreated distributions for every unit i and period t are generated by latent mixture components:*

$$\mathcal{F}_{it}(0) = \sum_{b=1}^{B} \lambda_{ib} Q_{bt}.$$

The Q_{bt} are time-varying component distributions and $\lambda_i \in \Delta^B$ are unit-varying mixture shares. Suppose further that $\exists w^* \in \mathcal{W}$ such that

$$\lambda_N = \sum_{i=1}^{N-1} w_i^* \lambda_i.$$

Then, Assumption CS holds.

Proof of Lemma 3. By our first assumption, we can write the untreated potential distribution for the treated unit at time t as:

$$\mathcal{F}_{Nt}(0) = \sum_{b=1}^{B} \lambda_{Nb} Q_{bt}$$

By our second assumption, we can express each dimension of the mixture shares for unit N as:

$$\lambda_{Nb} = \sum_{i=1}^{N-1} w_i^* \lambda_{ib}$$

Substituting this into our first equation yields:

$$\mathcal{F}_{Nt}(0) = \sum_{b=1}^{B} \left(\sum_{i=1}^{N-1} w_i^* \lambda_{ib} \right) Q_{bt}$$

Swapping the order of summation yields:

$$\mathcal{F}_{Nt}(0) = \sum_{i=1}^{N-1} w_i^* \underbrace{\left(\sum_{b=1}^{B} \lambda_{ib} Q_{bt} \right)}_{\mathcal{F}_{it}(0)}$$

Where the parenthetical term is equal to $F_{it}(0)$ by assumption. This establishes the desired result. □

C.4 Proof of Proposition 1

Proof. By definition of our objective function,

$$L(w^*) = \sum_{t<T} \text{MMD}_k^2 \left(\mathcal{F}_{Nt}(0), \sum_{i=1}^{N-1} w_i^* \mathcal{F}_{it}(0) \right)$$

By Assumption CS, for each $t < T$, the arguments of MMD_k^2 in the above expression are equal. We then have:

$$L(w^*) = \sum_{t<T} \text{MMD}_k^2(\mathcal{F}_{Nt}(0), \mathcal{F}_{Nt}(0))$$

By definition of MMD_k, we have:

$$\text{MMD}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}$$

Where μ_P refers to the kernel mean embedding of P. Consequently, $\text{MMD}_k(P, Q) = 0 \iff \mu_P = \mu_Q$. Therefore, by Assumption CS, each term in the above sum is 0, and so $L(w^*) = 0$. Since the above is a sum of squared norms, it is always nonnegative. Therefore, for every $w \in \mathcal{W}$, $L(w^*) = 0 \leq L(w)$. This means that:

$$w^* \in \arg\min_{w \in \mathcal{W}} L(w)$$

This establishes the first claim. Now, consider an arbitrary population minimizer

$$w^0 \in \arg\min_{w \in \mathcal{W}} L(w)$$

We must then have $L(w^0) = L(w^*) = 0$. Since $L(w^0)$ is the sum of nonnegative terms, as previously argued, it must be that every element of the sum is 0. This in turn means that for every $t < T$:

$$\mu_{Nt} = \sum_{i=1}^{N-1} w_i^0 \mu_{it}$$

Applying Assumption CS to the left hand side yields:

$$\sum_{i=1}^{N-1} w_i^* \mu_{it} = \sum_{i=1}^{N-1} w_i^0 \mu_{it}$$

Assumption PP then yields:

$$\sum_{i=1}^{N-1} w_i^* \mathcal{F}_{iT}(0) = \sum_{i=1}^{N-1} w_i^0 \mathcal{F}_{iT}(0)$$

Applying Assumption CS to the left hand side once again yields:

$$\mathcal{F}_{NT}(0) = \sum_{i=1}^{N-1} w_i^0 \mathcal{F}_{iT}(0)$$

Since the choice of w^0 was arbitrary, this proves the second claim.

□

C.5 Proof of Proposition 2

Lemma 1 establishes continuity of L on $\mathcal{W}$. Since $\mathcal{W}$ is the simplex, it is compact, and so L attains its minimum on $\mathcal{W}$. To establish uniqueness, take distinct $w, w' \in \mathcal{W}$, let $h = w' - w \neq 0$, and fix $\alpha \in (0,1)$. Because both weight vectors belong to the simplex, $\mathbf{1}^\top h = 0$. Since $h \neq 0$, the contrapositive of Assumption IDT and the fact that $\mathbf{1}^\top h = 0$ imply that the kernel mean mixtures must be unequal in at least one period, yielding:

$$\sum_{i=1}^{N-1} w_i \mu_{it} \neq \sum_{i=1}^{N-1} w_i' \mu_{it}$$

This then means that

$$\left\| \sum_{i=1}^{N-1} h_i \mu_{it} \right\|_{\mathcal{H}_k}^2 = \left\| \sum_{i=1}^{N-1} w_i \mu_{it} - \sum_{i=1}^{N-1} w_i' \mu_{it} \right\|_{\mathcal{H}_k}^2 > 0$$

We can then write:

$$\begin{aligned}
L(\alpha w + (1-\alpha)w') &= \sum_{t<T} \mathrm{MMD}_k^2 \left(F_{Nt}(0), \sum_{i=1}^{N-1} (\alpha w_i + (1-\alpha) w_i') \mathcal{F}_{it}(0) \right) \\
&= \sum_{t<T} \left\| \mu_{Nt} - \sum_{i=1}^{N-1} (\alpha w_i + (1-\alpha) w_i') \mu_{it} \right\|_{\mathcal{H}_k}^2 \\
&= \sum_{t<T} \left\| \alpha \left(\mu_{Nt} - \sum_{i=1}^{N-1} w_i \mu_{it} \right) + (1-\alpha) \left(\mu_{Nt} - \sum_{i=1}^{N-1} w_i' \mu_{it} \right) \right\|_{\mathcal{H}_k}^2 \\
&= \alpha L(w) + (1-\alpha) L(w') - \alpha(1-\alpha) \sum_{t<T} \left\| \sum_{i=1}^{N-1} h_i \mu_{it} \right\|_{\mathcal{H}_k}^2 \\
&< \alpha L(w) + (1-\alpha) L(w')
\end{aligned}$$

Note that the final equality uses the squared-norm identity for Hilbert spaces. Thus, L is strictly convex on $\mathcal{W}$. It therefore has exactly one minimizer on $\mathcal{W}$, establishing uniqueness.

D Proofs for Section 3

D.1 Notation

Let us define a change of variables to remove the extraneous dimension of the weight vector imposed by the sum-to-one restriction. Let Θ denote the feasible set of such $(N-2)$-dimensional vectors:

$$\Theta = \left\{ \theta \in \mathbb{R}^{N-2} : \theta_j \geq 0 \text{ for every } j, \sum_{j=1}^{N-2} \theta_j \leq 1 \right\}$$

For a candidate element $\theta \in \Theta$, define $w(\theta)$ as the corresponding projection back to $\mathbb{R}^{N-1}$ and let $\ell(\theta)$ define the objective function in terms of θ:

$$w(\theta) = \left(\theta_1, \ldots, \theta_{N-2}, 1 - \sum_{j=1}^{N-2} \theta_j \right)^\top \qquad \ell(\theta) = L(w(\theta))$$

We can then derive the Jacobian of this weight map, A, which features in our expression for Σ_w in Equation 10.

$$A = \frac{\partial w(\theta)}{\partial \theta^\top} = \begin{pmatrix} \mathbb{I}_{N-2} \\ -\mathbf{1}_{N-2}^\top \end{pmatrix}$$

Here, $\mathbb{I}_{N-2}$ refers to the $(N-2)$-dimensional identity matrix and $\mathbf{1}_{N-2}$ to an $(N-2)$-dimensional vector of 1s.

Now, define the following quantities, which describe the differences in kernel mean embeddings between the treated unit and the dropped reference donor unit ($j = N-1$) and between a given donor unit and the reference, respectively:

$$b_t = \mu_{Nt} - \mu_{N-1,t} \qquad d_{jt} = \mu_{jt} - \mu_{N-1,t}$$

We can then define the linear operator $\mathsf{D}_t : \mathbb{R}^{N-2} \to \mathcal{H}_k$ which maps a change $h \in \mathbb{R}^{N-2}$ in donor weights θ to the corresponding change in the synthetic distribution's kernel mean embedding, at a given time t:

$$\mathsf{D}_t h = \sum_{j=1}^{N-2} h_j d_{jt}$$

We can then rewrite our objective function and derive the corresponding first derivative (score)

and second derivative (Hessian):

$$\ell(\theta) = \sum_{t<T} \|b_t - \mathsf{D}_t\theta\|^2_{\mathcal{H}_k} \qquad \nabla\ell(\theta) = -2\sum_{t<T} \mathsf{D}_t^*(b_t - \mathsf{D}_t\theta) \qquad H := \nabla^2\ell(\theta) = 2\sum_{t<T} \mathsf{D}_t^*\mathsf{D}_t$$

Here, $\mathsf{D}_t^* : \mathcal{H}_k \to \mathbb{R}^{N-2}$ denotes the adjoint of D_t, which maps the gradient in the RKHS back to weight-coordinate derivatives. In particular, for every $u \in \mathcal{H}_k$ and $h \in \mathbb{R}^{N-2}$, D_t and D_t^* satisfy the following:

$$\langle u, \mathsf{D}_t h\rangle_{\mathcal{H}_k} = \langle \mathsf{D}_t^* u, h\rangle_{\mathbb{R}^{N-2}}$$

Next, define the residual $r_t = b_t - \mathsf{D}_t\theta$ and define the following differences between empirical and population quantities:

$$\Delta b_t = \hat{b}_t - b_t \qquad \Delta\mathsf{D}_t = \widehat{\mathsf{D}}_t - \mathsf{D}_t \qquad \Delta r_t = \Delta b_t - \Delta\mathsf{D}_t\theta^0$$

Furthermore, define the residual at the population optimum as $r_t^0 = b_t - \mathsf{D}_t\theta^0$.

D.2 Useful Lemmas

Lemma 4 (Uniform convergence of empirical objective)**.** *Suppose Assumptions DS and k-Reg hold. Then, as* $K^{\text{pre}}_{\min} \to \infty$,

$$\sup_{w\in\mathcal{W}} \left|\widehat{L}(w) - L(w)\right| \xrightarrow{p} 0$$

Proof of Lemma 4. Let $\phi(x) = k(\mathcal{E}(x), \cdot) \in \mathcal{H}_k$ denote the feature map induced by kernel k. By Assumption k-Reg, k is bounded and measurable and $\mathcal{H}_k$ is separable. This means that ϕ is measurable, and that

$$\|\phi(x)\|^2_{\mathcal{H}_k} = k(\mathcal{E}(x), \mathcal{E}(x)) \le \bar{k}$$

By Assumption DS, for every pre-treatment cell (i,t) all observations are iid draws from $F_{it}(0)$. Centering this feature map gives:

$$u_{it} = \widehat{\mu}_{it} - \mu_{it} = \frac{1}{K_{it}}\sum_{r=1}^{K_{it}} \phi(x_{itr}) - \mu_{it}$$

Where we use the definition of $\widehat{\mu}_{it}$ in the second equality. We can then write:

$$\mathbb{E}[\|u_{it}\|^2_{\mathcal{H}_k}] = \frac{1}{K_{it}^2}\sum_{r=1}^{K_{it}}\sum_{s=1}^{K_{it}} \mathbb{E}\left[\langle \phi(X_{itr}) - \mu_{it}, \phi(X_{its}) - \mu_{it}\rangle_{\mathcal{H}_k}\right]$$

If $r \neq s$, independence and centering yield:

$$\begin{aligned}
\mathbb{E}\left[\langle \phi(X_{itr}) - \mu_{it}, \phi(X_{its}) - \mu_{it} \rangle_{\mathcal{H}_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle \phi(X_{itr}) - \mu_{it}, \phi(X_{its}) - \mu_{it} \rangle_{\mathcal{H}_k} | \phi(X_{itr}) - \mu_{it}\right]\right] \\
&= \mathbb{E}\left[\langle \phi(X_{itr}) - \mu_{it}, \mathbb{E}[\phi(X_{its}) - \mu_{it} | \phi(X_{itr}) - \mu_{it}] \rangle_{\mathcal{H}_k}\right] \\
&= \mathbb{E}\left[\langle \phi(X_{itr}) - \mu_{it}, \mathbb{E}[\phi(X_{its}) - \mu_{it}] \rangle_{\mathcal{H}_k}\right] \\
&= 0
\end{aligned}$$

Thus all cross terms vanish, yielding:

$$\begin{aligned}
\mathbb{E}[\|u_{it}\|_{\mathcal{H}_k}^2] &= \frac{1}{K_{it}^2} \sum_{r=1}^{K_{it}} \mathbb{E}\left[\|\phi(X_{itr}) - \mu_{it}\|_{\mathcal{H}_k}^2\right] \\
&= \frac{1}{K_{it}} \mathbb{E}\left[\|\phi(X) - \mu_{it}\|_{\mathcal{H}_k}^2\right] \\
&\leq \frac{1}{K_{it}} \mathbb{E}\left[\|\phi(X)\|_{\mathcal{H}_k}^2\right] \\
&\leq \frac{\bar{k}}{K_{it}}
\end{aligned}$$

Where we have made use of the Hilbert space variance identity and boundedness of k via Assumption k-Reg. Therefore, as $K_{\min}^{\text{pre}} \to \infty$, $\|u_{it}\|_{\mathcal{H}_k} \xrightarrow{p} 0$ for every i and $t < T$. Because N and T are fixed, this convergence holds jointly over the finite collection of pre-treatment cells:

$$\max_{\substack{i=1,\ldots,N \\ t<T}} \|u_{it}\|_{\mathcal{H}_k} \xrightarrow{p} 0$$

Now, for each $t < T$ and $w \in \mathcal{W}$, define the following:

$$r_t(w) = \mu_{Nt} - \sum_{i=1}^{N-1} w_i \mu_{it} \qquad \widehat{r}_t(w) = \widehat{\mu}_{Nt} - \sum_{i=1}^{N-1} w_i \widehat{\mu}_{it}$$

The triangle inequality then yields:

$$\begin{aligned}
\sup_{w \in \mathcal{W}} \|\widehat{r}_t(w) - r_t(w)\|_{\mathcal{H}_k} &\leq \|u_{Nt}\|_{\mathcal{H}_k} + \sup_{w \in \mathcal{W}} \sum_{i=1}^{N-1} w_i \|u_{it}\|_{\mathcal{H}_k} \\
&\leq \|u_{Nt}\|_{\mathcal{H}_k} + \max_{i<N} \|u_{it}\|_{\mathcal{H}_k} \\
&=: \delta_t
\end{aligned}$$

As the sum of two terms which have been shown to converge to 0, $\delta_t \xrightarrow{p} 0$. Moreover, the triangle

inequality provides:

$$\begin{aligned}\sup_{w\in\mathcal{W}} \|r_t(w)\|_{\mathcal{H}_k} &\leq \|\mu_{Nt}\|_{\mathcal{H}_k} + \sup_{w\in\mathcal{W}} \sum_{i=1}^{N-1} w_i \|\mu_{it}\|_{\mathcal{H}_k} \\ &\leq 2\sqrt{\bar{k}}\end{aligned}$$

For any two elements a, b of a normed space,

$$\left|\|a\|^2 - \|b\|^2\right| \leq \|a-b\|(\|a\| + \|b\|)$$

Applying this inequality with $a = \widehat{r}_t(w)$ and $b = r_t(w)$, and using

$$\|\widehat{r}_t(w)\|_{\mathcal{H}_k} \leq \|r_t(w)\|_{\mathcal{H}_k} + \delta_t$$

gives

$$\sup_{w\in\mathcal{W}} \left|\|\widehat{r}_t(w)\|^2_{\mathcal{H}_k} - \|r_t(w)\|^2_{\mathcal{H}_k}\right| \leq \delta_t \left(4\sqrt{\bar{k}} + \delta_t\right) \xrightarrow{p} 0$$

Finally, note that we can write:

$$\widehat{L}(w) - L(w) = \sum_{t<T} \|\widehat{r}_t(w)\|^2_{\mathcal{H}_k} - \|r_t(w)\|^2_{\mathcal{H}_k}$$

Because the number of pre-treatment periods is fixed and finite, this yields:

$$\sup_{w\in\mathcal{W}} |\widehat{L}(w) - L(w)| \leq \sum_{t<T} \sup_{w\in\mathcal{W}} \left|\|\widehat{r}_t(w)\|^2_{\mathcal{H}_k} - \|r_t(w)\|^2_{\mathcal{H}_k}\right| \xrightarrow{p} 0$$

□

Lemma 5 (Mixture stability of MMD^2_k)**.** *Suppose Assumption k-Reg holds. Let $\{P_i\}_{i=1}^{N-1}, \{Q_i\}_{i=1}^{N-1}$ be sequences of distributions defined on $\mathcal{X}$. Then, for all $w \in \mathcal{W}$,*

$$\mathrm{MMD}^2_k\left(\sum_{i=1}^{N-1} w_i P_i, \sum_{i=1}^{N-1} w_i Q_i\right) \leq \sum_{i=1}^{N-1} w_i \mathrm{MMD}^2_k(P_i, Q_i)$$

Proof of Lemma 5. Fix $w \in \mathcal{W}$ and probability distributions $P_1, \ldots, P_{N-1}$ and $Q_1, \ldots, Q_{N-1}$. Linearity of kernel mean embeddings yields:

$$\mathrm{MMD}^2_k\left(\sum_{i=1}^{N-1} w_i P_i, \sum_{i=1}^{N-1} w_i Q_i\right) = \left\|\sum_{i=1}^{N-1} w_i(\mu_{P_i} - \mu_{Q_i})\right\|^2_{\mathcal{H}_k}$$

Because $w_i \geq 0$ and $\sum_{i=1}^{N-1} w_i = 1$, Jensen's inequality for the convex squared norm function yields:

$$\left\| \sum_{i=1}^{N-1} w_i(\mu_{P_i} - \mu_{Q_i}) \right\|_{\mathcal{H}_k}^2 \leq \sum_{i=1}^{N-1} w_i \|\mu_{P_i} - \mu_{Q_i}\|_{\mathcal{H}_k}^2$$
$$= \sum_{i=1}^{N-1} w_i \mathrm{MMD}_k^2(P_i, Q_i)$$

□

Lemma 6 (Consistency of empirical distributions)**.** *Suppose Assumptions DS and k-Reg hold. Then, for each donor $i = 1, \ldots, N-1$, as $K_{\min}^{\text{post}} \to \infty$*

$$\mathrm{MMD}_k \left(\widehat{F}_{iT}, \mathcal{F}_{iT}(0) \right) \xrightarrow{p} 0$$

Proof. Suppose Assumption DS holds. Fix a donor $i \in \{1, \ldots, N-1\}$, and write

$$P = \mathcal{F}_{iT}(0) \qquad K = K_{iT} \qquad X_r = x_{iTr}$$

Because donor i is untreated in period T, Assumption DS implies

$$X_1, \ldots, X_K \overset{\text{iid}}{\sim} P$$

Let $\phi(x) = k(\mathcal{E}(x), \cdot) \in \mathcal{H}_k$ denote the feature map induced by kernel k and define

$$Z_r = \phi(X_r) - \mu_P$$

Then, $\mathbb{E}[Z_r] = 0$ and

$$\widehat{\mu}_P - \mu_P = \frac{1}{K} \sum_{r=1}^{K} Z_r$$

Expanding the squared Hilbert-space norm yields

$$\mathbb{E}\left[\|\widehat{\mu}_P - \mu_P\|_{\mathcal{H}_k}^2 \right] = \frac{1}{K^2} \sum_{r=1}^{K} \sum_{s=1}^{K} \mathbb{E}[\langle Z_r, Z_s \rangle_{\mathcal{H}_k}]$$

If $r \neq s$, independence and centering imply

$$\begin{aligned}\mathbb{E}[\langle Z_r, Z_s\rangle_{\mathcal{H}_k}] &= \mathbb{E}[\mathbb{E}[\langle Z_r, Z_s\rangle_{\mathcal{H}_k} \mid Z_r]] \\ &= \mathbb{E}\Big[\langle Z_r, \mathbb{E}[Z_s \mid Z_r]\rangle_{\mathcal{H}_k}\Big] \\ &= \mathbb{E}\Big[\langle Z_r, \mathbb{E}[Z_s]\rangle_{\mathcal{H}_k}\Big] \\ &= 0\end{aligned}$$

Thus all cross terms vanish, leaving

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\mu}_P - \mu_P\|^2_{\mathcal{H}_k}\big] &= \frac{1}{K^2}\sum_{r=1}^{K}\mathbb{E}\big[\|Z_r\|^2_{\mathcal{H}_k}\big] \\ &= \frac{1}{K}\,\mathbb{E}\big[\|\phi(X) - \mu_P\|^2_{\mathcal{H}_k}\big]\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}\big[\|\phi(X) - \mu_P\|^2_{\mathcal{H}_k}\big] &= \mathbb{E}\big[\|\phi(X)\|^2_{\mathcal{H}_k}\big] - \|\mu_P\|^2_{\mathcal{H}_k} \\ &= \mathbb{E}[k(\mathcal{E}(X), \mathcal{E}(X))] - \|\mu_P\|^2_{\mathcal{H}_k} \\ &\leq \bar{k}\end{aligned}$$

Where the last line uses boundedness of the kernel, by Assumption k-Reg. It follows that

$$\mathbb{E}\Big[\mathrm{MMD}^2_k(\widehat{F}_{iT}, \mathcal{F}_{iT}(0))\Big] \leq \frac{\bar{k}}{K_{iT}}$$

Hence, for every $\varepsilon > 0$, Markov's inequality gives

$$\begin{aligned}\Pr\Big\{\mathrm{MMD}_k(\widehat{F}_{iT}, \mathcal{F}_{iT}(0)) > \varepsilon\Big\} &= \Pr\Big\{\mathrm{MMD}^2_k(\widehat{F}_{iT}, \mathcal{F}_{iT}(0)) > \varepsilon^2\Big\} \\ &\leq \frac{\bar{k}}{K_{iT}\varepsilon^2}\end{aligned}$$

Since $K_{iT} \geq K^{\mathrm{post}}_{\min} \to \infty$, the right-hand side converges to zero. Therefore,

$$\mathrm{MMD}_k(\widehat{F}_{iT}, \mathcal{F}_{iT}(0)) \xrightarrow{p} 0$$

Since the argument applies to every donor i, this proves the desired result. □

Lemma 7 (Triangle inequality)**.** *Suppose Assumption k-Reg holds. Then, for any distributions P, Q, R over $\mathcal{X}$,*

$$\mathrm{MMD}_k(P, R) \leq \mathrm{MMD}_k(P, Q) + \mathrm{MMD}_k(Q, R)$$

Proof. For any probability distributions P, Q, R,

$$\begin{aligned}\mathrm{MMD}_k(P,R) &= \|\mu_P - \mu_R\|_{\mathcal{H}_k} \\ &= \|(\mu_P - \mu_Q) + (\mu_Q - \mu_R)\|_{\mathcal{H}_k} \\ &\leq \|\mu_P - \mu_Q\|_{\mathcal{H}_k} + \|\mu_Q - \mu_R\|_{\mathcal{H}_k} \\ &= \mathrm{MMD}_k(P,Q) + \mathrm{MMD}_k(Q,R),\end{aligned}$$

where the inequality is the triangle inequality in $\mathcal{H}_k$. □

Lemma 8 (Continuity of post-treatment synthetic distribution map)**.** *Suppose Assumption k-Reg holds. Then, the map*

$$w \mapsto \sum_{i=1}^{N-1} w_i \mathcal{F}_{iT}(0)$$

is continuous under metric MMD_k *on* $\mathcal{W}$*. That is, if* $w^{(n)} \to w$*, then*

$$\mathrm{MMD}_k\left(\sum_{i=1}^{N-1} w_i^{(n)} \mathcal{F}_{iT}(0), \sum_{i=1}^{N-1} w_i \mathcal{F}_{iT}(0)\right) \to 0$$

Since $\mathcal{W}$ *is compact, this further implies uniform continuity on* $\mathcal{W}$*.*

Proof. Define the post-treatment synthetic-distribution map

$$S_T(w) = \sum_{i=1}^{N-1} w_i \mathcal{F}_{iT}(0).$$

For any $w, w' \in \mathcal{W}$, linearity of the kernel mean embedding, boundedness via Assumption k-Reg, and Lemma 7 imply

$$\begin{aligned}\mathrm{MMD}_k\big(S_T(w'), S_T(w)\big) &= \left\| \sum_{i=1}^{N-1} (w_i' - w_i)\mu_{iT} \right\|_{\mathcal{H}_k} \\ &\leq \sum_{i=1}^{N-1} |w_i' - w_i| \|\mu_{iT}\|_{\mathcal{H}_k} \\ &\leq \sqrt{\bar{k}} \sum_{i=1}^{N-1} |w_i' - w_i| \\ &= \sqrt{\bar{k}}\, \|w' - w\|_1\end{aligned}$$

Therefore, if $w' \to w$, then $\|w' - w\|_1 \to 0$ and

$$\mathrm{MMD}_k\big(S_T(w'), S_T(w)\big) \to 0$$

Furthermore, the displayed bound shows that the map is globally Lipschitz under the ℓ_1 norm. Because $\mathcal{W}$ is compact, the map is also uniformly continuous on $\mathcal{W}$. □

Lemma 9 (Donor mixture empirical convergence)**.** *Suppose that Assumptions DS and k-Reg hold. Then, for any random $\hat{w} \in \mathcal{W}$, as $K_{\min}^{\text{post}} \to \infty$,*

$$\text{MMD}_k\left(\sum_{i=1}^{N-1} \hat{w}_i \widehat{F}_{iT}, \sum_{i=1}^{N-1} \hat{w}_i \mathcal{F}_{iT}(0)\right) \xrightarrow{p} 0.$$

Proof. For each donor $i = 1, \ldots, N-1$, define

$$Z_i = \text{MMD}_k\left(\widehat{F}_{iT}, \mathcal{F}_{iT}(0)\right).$$

By Lemma 6, $Z_i \xrightarrow{p} 0$ for every donor $i = 1, \ldots, N-1$.

Now consider the discrepancy between the empirical donor mixture and its population counterpart, evaluated at the random weight vector $\hat{w}$:

$$\text{MMD}_k\left(\sum_{i=1}^{N-1} \hat{w}_i \widehat{F}_{iT}, \sum_{i=1}^{N-1} \hat{w}_i \mathcal{F}_{iT}(0)\right)$$

Applying Lemma 5 gives

$$\begin{aligned}\text{MMD}_k^2\left(\sum_{i=1}^{N-1} \hat{w}_i \widehat{F}_{iT}, \sum_{i=1}^{N-1} \hat{w}_i \mathcal{F}_{iT}(0)\right) &\leq \sum_{i=1}^{N-1} \hat{w}_i \text{MMD}_k^2\left(\widehat{F}_{iT}, \mathcal{F}_{iT}(0)\right) \\ &= \sum_{i=1}^{N-1} \hat{w}_i Z_i^2\end{aligned}$$

Since $\hat{w}_i \geq 0$ and $\sum_{i=1}^{N-1} \hat{w}_i = 1$,

$$\sum_{i=1}^{N-1} \hat{w}_i Z_i^2 \leq \max_{1 \leq i \leq N-1} Z_i^2$$

Hence

$$\text{MMD}_k^2\left(\sum_{i=1}^{N-1} \hat{w}_i \widehat{F}_{iT}, \sum_{i=1}^{N-1} \hat{w}_i \mathcal{F}_{iT}(0)\right) \leq \max_{1 \leq i \leq N-1} Z_i^2$$

It remains to show that the right-hand side converges to zero in probability. Since N is fixed and $Z_i \xrightarrow{p} 0$ for every donor i, the continuous mapping theorem yields

$$\max_{1 \leq i \leq N-1} Z_i^2 \xrightarrow{p} 0$$

Therefore,

$$\mathrm{MMD}_k^2\left(\sum_{i=1}^{N-1}\hat{w}_i\widehat{F}_{iT},\sum_{i=1}^{N-1}\hat{w}_i\mathcal{F}_{iT}(0)\right)\xrightarrow{p}0$$

Applying the continuous mapping theorem once again and taking square roots yields

$$\mathrm{MMD}_k\left(\sum_{i=1}^{N-1}\hat{w}_i\widehat{F}_{iT},\sum_{i=1}^{N-1}\hat{w}_i\mathcal{F}_{iT}(0)\right)\xrightarrow{p}0$$

This proves the claim. □

Lemma 10 (Local smoothness of the population objective)**.** *Suppose Assumptions k-Reg and IDT hold. Then, the population objective ℓ is twice continuously differentiable on $\mathbb{R}^{N-2}$ and $H=\nabla^2\ell(\theta^0)$ is positive definite.*

Proof. Under the parameterization $w=w(\theta)$,

$$\sum_{i=1}^{N-1}w_i(\theta)\mu_{it}=\mu_{N-1,t}+\mathsf{D}_t\theta$$

Consequently,

$$\ell(\theta)=\sum_{t<T}\|b_t-\mathsf{D}_t\theta\|_{\mathcal{H}_k}^2$$

While ℓ is originally defined on Θ, we may extend it to all of $\mathbb{R}^{N-2}$ via the above equation. Note that when $\theta\in\Theta$, the definition of ℓ remains unchanged. This is a quadratic function of θ, with gradient

$$\nabla\ell(\theta)=2\sum_{t<T}\mathsf{D}_t^*(\mathsf{D}_t\theta-b_t)$$

and Hessian

$$\nabla^2\ell(\theta)=2\sum_{t<T}\mathsf{D}_t^*\mathsf{D}_t=:H$$

Thus ℓ is twice continuously differentiable on all of $\mathbb{R}^{N-2}$.

It remains to establish that H is positive definite. For every $h\in\mathbb{R}^{N-2}$,

$$h^\top Hh=2\sum_{t<T}\|\mathsf{D}_th\|_{\mathcal{H}_k}^2\geq 0$$

Therefore, H is positive semi-definite. Towards a contradiction, suppose that this expression were exactly zero for some $h\neq 0$. Then, we would have $\mathsf{D}_th=0$ for every $t<T$. Let $a=Ah\in\mathbb{R}^{N-1}$ denote the corresponding perturbation of the full weight vector. We would then have $\mathbf{1}^\top a=0$ and,

for every $t < T$:

$$\begin{aligned}\sum_{i=1}^{N-1} a_i \mu_{it} &= \sum_{j=1}^{N-2} h_j \mu_{jt} - \left(\sum_{j=1}^{N-2} h_j\right) \mu_{N-1,t} \\ &= \sum_{j=1}^{N-2} h_j (\mu_{jt} - \mu_{N-1,t}) \\ &= \mathsf{D}_t h \\ &= 0\end{aligned}$$

Assumption IDT then implies $a = 0$. Since A has full column rank, this implies $h = 0$, a contradiction. Thus, H must be positive definite. □

Lemma 11 (Interior score central limit theorem)**.** *Suppose Assumptions DS, k-Reg, and BSG hold, and that $w_i^0 > 0$ for every $i = 1, \ldots, N-1$. Then, the empirical objective $\widehat{\ell}$ is differentiable on a neighborhood of θ^0, and*

$$\sqrt{K_{\min}^{\text{pre}}} \left(\nabla \widehat{\ell}(\theta^0) - \nabla \ell(\theta^0)\right) \xrightarrow{d} \mathcal{N}(0, \Omega)$$

Additionally, because θ^0 is an interior minimizer, $\nabla \ell(\theta^0) = 0$.

Proof. For notational convenience, write $K = K_{\min}^{\text{pre}}$. Define:

$$u_{it} = \widehat{\mu}_{it} - \mu_{it}$$

$$\Delta b_t = \widehat{b}_t - b_t = u_{Nt} - u_{Jt}$$

and

$$\Delta \mathsf{D}_t = \widehat{\mathsf{D}}_t - \mathsf{D}_t$$

so that

$$\Delta \mathsf{D}_t h = \sum_{j=1}^{N-2} h_j (u_{jt} - u_{Jt})$$

Here, $J = N - 1$ denotes the index of the reference donor unit. Expanding the empirical score at θ^0 gives

$$\nabla \widehat{\ell}(\theta^0) - \nabla \ell(\theta^0) = 2 \sum_{t<T} \Big\{ (\Delta \mathsf{D}_t)^* (\mathsf{D}_t \theta^0 - b_t) + \mathsf{D}_t^* (\Delta \mathsf{D}_t \theta^0 - \Delta b_t) \Big\} + \underbrace{2 \sum_{t<T} (\Delta \mathsf{D}_t)^* (\Delta \mathsf{D}_t \theta^0 - \Delta b_t)}_{R_K},$$

Under Assumptions DS and BSG,

$$\|u_{it}\|_{\mathcal{H}_k} = O_p(K_{it}^{-1/2}) = O_p(K^{-1/2})$$

for every pre-treatment cell. Since the number of cells is fixed,

$$\|\Delta b_t\|_{\mathcal{H}_k} = O_p(K^{-1/2}) \qquad \text{and} \qquad \|\Delta \mathsf{D}_t\|_{\text{op}} = O_p(K^{-1/2})$$

Consequently,

$$\begin{aligned}\|R_K\|_2 &\leq 2\sum_{t<T} \|\Delta \mathsf{D}_t\|_{\text{op}} \left(\|\Delta \mathsf{D}_t\|_{\text{op}}\|\theta^0\|_2 + \|\Delta b_t\|_{\mathcal{H}_k}\right) \\ &= O_p(K^{-1})\end{aligned}$$

and hence

$$\sqrt{K}\, R_K = o_p(1)$$

It remains to characterize the linear part of the score expansion. For each pre-treatment cell (i,t), define

$$\zeta_{it}(x) = \phi(x) - \mu_{it}$$

Note that $\zeta_{it}(x)$ is centered: $\mathbb{E}[\zeta_{it}(X)] = 0$. For $z \in \mathcal{H}_k$, define

$$\Delta b_t^{(i)}(z) = (\mathbf{1}\{i = N\} - \mathbf{1}\{i = J\})\, z$$

and the linear operator $\Delta \mathsf{D}_t^{(i)}(z) : \mathbb{R}^{N-2} \to \mathcal{H}_k$ by

$$\Delta \mathsf{D}_t^{(i)}(z)h = \sum_{j=1}^{N-2} h_j \left(\mathbf{1}\{i = j\} - \mathbf{1}\{i = J\}\right) z$$

Note that by definition,

$$\Delta b_t = \sum_{i=1}^{N} \Delta b_t^{(i)}(u_{it}) \qquad \Delta \mathsf{D}_t = \sum_{i=1}^{N} \Delta \mathsf{D}_t^{(i)}(u_{it})$$

Now, define the $\mathbb{R}^{N-2}$-valued influence function

$$\psi_{it}(x) = 2\Big\{ \left(\Delta \mathsf{D}_t^{(i)}(\zeta_{it}(x))\right)^* (\mathsf{D}_t\theta^0 - b_t) + \mathsf{D}_t^* \Big(\Delta \mathsf{D}_t^{(i)}(\zeta_{it}(x))\theta^0 - \Delta b_t^{(i)}(\zeta_{it}(x))\Big)\Big\}$$

This influence function is linear in $\zeta_{it}(x)$. Therefore,

$$\mathbb{E}_{F_{it}(0)}[\psi_{it}(X)] = 0$$

Moreover, by Assumption k-Reg,

$$\|\zeta_{it}(x)\|_{\mathcal{H}_k} \leq \|\phi(x)\|_{\mathcal{H}_k} + \|\mu_{it}\|_{\mathcal{H}_k} \leq 2\sqrt{\bar{k}}$$

so $\psi_{it}(X)$ is bounded and has a finite covariance matrix. Write

$$V_{it} = \mathrm{Var}_{F_{it}(0)}\big(\psi_{it}(X)\big)$$

We can now write the linear component of the score expansion as

$$\sum_{t<T}\sum_{i=1}^{N}\frac{1}{K_{it}}\sum_{r=1}^{K_{it}}\psi_{it}(x_{itr})$$

Hence

$$\sqrt{K}\left\{\nabla\widehat{\ell}(\theta^0) - \nabla\ell(\theta^0)\right\} = \sum_{t<T}\sum_{i=1}^{N}\sqrt{\frac{K}{K_{it}}}\left\{\frac{1}{\sqrt{K_{it}}}\sum_{r=1}^{K_{it}}\psi_{it}(x_{itr})\right\} + o_p(1)$$

For every fixed cell (i,t), the multivariate central limit theorem gives

$$\frac{1}{\sqrt{K_{it}}}\sum_{r=1}^{K_{it}}\psi_{it}(x_{itr}) \xrightarrow{d} \mathcal{N}(0, V_{it})$$

Assumption DS gives independence across cells, while Assumption BSG gives

$$\sqrt{\frac{K}{K_{it}}} \longrightarrow \kappa_{it}^{-1/2}$$

Where we have defined $\kappa_{it} := \kappa_{it}^{\mathrm{pre}}$ for convenience. Since the number of pre-treatment cells is fixed, joint convergence and Slutsky's theorem imply

$$\sqrt{K}\left\{\nabla\widehat{\ell}(\theta^0) - \nabla\ell(\theta^0)\right\} \xrightarrow{d} \mathcal{N}(0, \Omega)$$

where

$$\Omega = \sum_{t<T}\sum_{i=1}^{N}\kappa_{it}^{-1}V_{it}$$

Finally, θ^0 is an interior population minimizer, so

$$\nabla\ell(\theta^0) = 0$$

The empirical objective is quadratic and therefore differentiable on an open neighborhood of θ^0. □

Lemma 12 (Interior Hessian convergence)**.** *Suppose Assumptions DS and k-Reg hold. Then, as* $K_{\min}^{\mathrm{pre}} \to \infty$,

$$\nabla^2\widehat{\ell}(\theta) \xrightarrow{p} H$$

uniformly in $\theta \in \mathbb{R}^{N-2}$.

Proof. The empirical squared-MMD objective is

$$\widehat{\ell}(\theta) = \sum_{t<T} \|\widehat{b}_t - \widehat{\mathsf{D}}_t\theta\|^2_{\mathcal{H}_k}$$

Hence it is quadratic on all of $\mathbb{R}^{N-2}$, and its Hessian is

$$\nabla^2\widehat{\ell}(\theta) = 2\sum_{t<T} \widehat{\mathsf{D}}_t^*\widehat{\mathsf{D}}_t$$

Note that the population analog of this expression establishes the existence of H. In particular, the empirical Hessian does not depend on θ.

The empirical-mean convergence established in the proof of Lemma 4 gives

$$\widehat{d}_{jt} - d_{jt} = u_{jt} - u_{Jt} \xrightarrow{p} 0$$

in $\mathcal{H}_k$ for every j and $t < T$. For every $h \in \mathbb{R}^{N-2}$ with $\|h\|_2 \le 1$,

$$\|(\widehat{\mathsf{D}}_t - \mathsf{D}_t)h\|_{\mathcal{H}_k} = \left\|\sum_{j=1}^{N-2} h_j(\widehat{d}_{jt} - d_{jt})\right\|_{\mathcal{H}_k} \le \left(\sum_{j=1}^{N-2} h_j^2\right)^{1/2} \left(\sum_{j=1}^{N-2} \|\widehat{d}_{jt} - d_{jt}\|^2_{\mathcal{H}_k}\right)^{1/2}$$

Therefore,

$$\|\widehat{\mathsf{D}}_t - \mathsf{D}_t\|_{\mathrm{op}} \le \left(\sum_{j=1}^{N-2} \|\widehat{d}_{jt} - d_{jt}\|^2_{\mathcal{H}_k}\right)^{1/2} \xrightarrow{p} 0$$

It follows that

$$\begin{aligned}\|\widehat{\mathsf{D}}_t^*\widehat{\mathsf{D}}_t - \mathsf{D}_t^*\mathsf{D}_t\|_{\mathrm{op}} &\le \|\widehat{\mathsf{D}}_t - \mathsf{D}_t\|_{\mathrm{op}} \left(\|\widehat{\mathsf{D}}_t\|_{\mathrm{op}} + \|\mathsf{D}_t\|_{\mathrm{op}}\right) \\ &\xrightarrow{p} 0\end{aligned}$$

Since the number of pre-treatment periods is fixed,

$$\left\|\nabla^2\widehat{\ell}(\theta) - H\right\|_{\mathrm{op}} \xrightarrow{p} 0$$

uniformly in θ. In particular, for every possibly random sequence $\widetilde{\theta} \xrightarrow{p} \theta^0$,

$$\nabla^2\widehat{\ell}(\widetilde{\theta}) \xrightarrow{p} H$$

This proves the result. □

Lemma 13 (Unique interior minimizer on the oracle face)**.** *Suppose Assumptions k-Reg and IDT hold. Then*

$$w^0 \in \operatorname{relint}(\mathcal{W}_{\mathcal{S}_0})$$

and

$$\arg\min_{w \in \mathcal{W}_{\mathcal{S}_0}} L(w) = \{w^0\}$$

Consequently, on the event $\{\widehat{\mathcal{S}} = \mathcal{S}_0\}$*,*

$$w^0 \in \operatorname{relint}\left(\mathcal{W}_{\widehat{\mathcal{S}}}\right)$$

and

$$\arg\min_{w \in \mathcal{W}_{\widehat{\mathcal{S}}}} L(w) = \{w^0\}$$

If $|\mathcal{S}_0| \geq 2$*, then* $\mathcal{W}_{\mathcal{S}_0}$ *has dimension* $|\mathcal{S}_0| - 1$*, and the restricted problem has a unique interior minimizer. If* $|\mathcal{S}_0| = 1$*, then* $\mathcal{W}_{\mathcal{S}_0} = \{w^0\}$*.*

Proof of Lemma 13. By definition of $\mathcal{S}_0$, $w_i^0 > 0$ for every $i \in \mathcal{S}_0$ and $w_i^0 = 0$ for every $i \notin \mathcal{S}_0$. Therefore, $w^0 \in \mathcal{W}_{\mathcal{S}_0}$ by definition. The relative interior of this face is

$$\operatorname{relint}(\mathcal{W}_{\mathcal{S}_0}) = \{w \in \mathcal{W}_{\mathcal{S}_0} : w_i > 0 \text{ for every } i \in \mathcal{S}_0\}$$

It follows that $w^0 \in \operatorname{relint}(\mathcal{W}_{\mathcal{S}_0})$.

We next establish uniqueness on the oracle face. Because $\mathcal{W}_{\mathcal{S}_0} \subseteq \mathcal{W}$ and w^0 is the unique minimizer of L over $\mathcal{W}$ (by Proposition 2), every $w \in \mathcal{W}_{\mathcal{S}_0}$ satisfies

$$L(w) \geq L(w^0)$$

Moreover, $w^0 \in \mathcal{W}_{\mathcal{S}_0}$, so

$$\min_{w \in \mathcal{W}_{\mathcal{S}_0}} L(w) = L(w^0)$$

If $w \in \mathcal{W}_{\mathcal{S}_0}$ attains this restricted minimum, then

$$L(w) = L(w^0)$$

Such a w is also a population minimizer over $\mathcal{W}$. Since $\mathcal{W}_0 = \{w^0\}$, it follows that $w = w^0$. Hence

$$\arg\min_{w \in \mathcal{W}_{\mathcal{S}_0}} L(w) = \{w^0\}$$

On the event

$$\left\{\widehat{\mathcal{S}} = \mathcal{S}_0\right\}$$

we have

$$\mathcal{W}_{\widehat{\mathcal{S}}} = \mathcal{W}_{\mathcal{S}_0}$$

The preceding conclusions therefore imply

$$w^0 \in \operatorname{relint}\left(\mathcal{W}_{\widehat{\mathcal{S}}}\right)$$

and

$$\arg\min_{w \in \mathcal{W}_{\widehat{\mathcal{S}}}} L(w) = \{w^0\}$$

Finally, if $|\mathcal{S}_0| \geq 2$, then $\mathcal{W}_{\mathcal{S}_0}$ is a simplex of dimension $|\mathcal{S}_0| - 1$, and w^0 is its unique relative-interior minimizer. If $|\mathcal{S}_0| = 1$, the adding-up restriction uniquely determines the weight on the active donor, so

$$\mathcal{W}_{\mathcal{S}_0} = \{w^0\}$$

□

Lemma 14 (Fixed-summary error decomposition)**.** *Let $\widetilde{\tau}_g$ be defined as in Section 2.4, $\widehat{\tau}_g$ defined as in Section 3.4, and $g : \mathcal{X} \to \mathbb{R}$ a fixed summary function satisfying Assumption BFS. Then, the estimation error admits the exact decomposition*

$$\widehat{\tau}_g - \widetilde{\tau}_g = A_{1K}(g) - A_{0K}(g) - b_{K,g}$$

where

$$A_{1K}(g) = \mathbb{E}_{\widehat{P}_1}[g(X)] - \mathbb{E}_{P_1}[g(X)]$$

and

$$A_{0K}(g) = \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \left\{ \mathbb{E}_{\widehat{F}_{iT}}[g(X)] - \mathbb{E}_{F_{iT}(0)}[g(X)] \right\}$$

and

$$b_{K,g} = \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \mu_i(g) - \tilde{\mu}_0(g)$$

Here, $\tilde{\mu}_0(g) := \mathbb{E}_{\widetilde{P}_0}[g(X)]$.

Proof of Lemma 14. By definition,

$$\widehat{\tau}_g = \mathbb{E}_{\widehat{P}_1}[g(X)] - \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \mathbb{E}_{\widehat{F}_{iT}}[g(X)] \qquad \widetilde{\tau}_g = \mathbb{E}_{P_1}[g(X)] - \tilde{\mu}_0(g)$$

Therefore,

$$\widehat{\tau}_g - \widetilde{\tau}_g = \mathbb{E}_{\widehat{P}_1}[g(X)] - \mathbb{E}_{P_1}[g(X)] - \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \mathbb{E}_{\widehat{F}_{iT}}[g(X)] + \tilde{\mu}_0(g)$$

Note that by definition, we have:

$$\sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \, \mathbb{E}_{\mathcal{F}_{iT}(0)}[g(X)] = \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \mu_i(g)$$

Adding and subtracting this term on the right hand side yields:

$$\widehat{\tau}_g - \widetilde{\tau}_g = \underbrace{\left\{ \mathbb{E}_{\widehat{P}_1}[g(X)] - \mathbb{E}_{P_1}[g(X)] \right\}}_{A_{1K}(g)} - \underbrace{\sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \left\{ \mathbb{E}_{\widehat{F}_{iT}}[g(X)] - \mathbb{E}_{\mathcal{F}_{iT}(0)}[g(X)] \right\}}_{A_{0K}(g)} - \underbrace{\left\{ \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} \mu_i(g) - \tilde{\mu}_0(g) \right\}}_{b_{K,g}}$$

□

Lemma 15 (Evaluation-sample mean convergence for fixed summary)**.** *Suppose Assumptions DS and BFS hold. Let $A_{1K}(g), A_{0K}(g)$ be defined as in Lemma 14. Then as $K_{\min} \to \infty$, $A_{1K}(g) \xrightarrow{p} 0$ and $A_{0K}(g) \xrightarrow{p} 0$.*

Proof of Lemma 15. By Assumption DS:

$$\mathbb{E}_{\widehat{P}_1}[g(X)] = \frac{1}{K_{NT}} \sum_{k=1}^{K_{NT}} g(x_{NTk})$$

This is a sample mean of iid draws from P_1. By Assumption BFS, $\mathbb{E}_{P_1}[|g(X)|] < \infty$. Since $K_{NT} \geq K_{\min} \to \infty$, the weak law of large numbers gives:

$$A_{1K}(g) = \mathbb{E}_{\widehat{P}_1}[g(X)] - \mathbb{E}_{P_1}[g(X)] \xrightarrow{p} 0$$

Now, for each donor $i = 1, \ldots, N-1$, define:

$$B_{iK}(g) = \mathbb{E}_{\widehat{F}_{iT}}[g(X)] - \mathbb{E}_{\mathcal{F}_{iT}(0)}[g(X)]$$

The donor evaluation observations are iid draws from $\mathcal{F}_{iT}(0)$, and $\mathbb{E}_{\mathcal{F}_{iT}(0)}[|g(X)|] < \infty$ by Assumption BFS. Since $K_{iT} \geq K_{\min} \to \infty$, the weak law of large numbers gives:

$$B_{iK}(g) \xrightarrow{p} 0$$

Because N is fixed, these convergences hold jointly:

$$\max_{1 \leq i \leq N-1} |B_{iK}(g)| \xrightarrow{p} 0$$

Since $\hat{w}^{\text{ts}} \in \mathcal{W}, \hat{w}_i^{\text{ts}} \geq 0$ and $\sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} = 1$. Consequently, applying the triangle inequality yields:

$$\begin{aligned}
|A_{0K}(g)| &= \left| \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} B_{iK}(g) \right| \\
&\leq \sum_{i=1}^{N-1} \hat{w}_i^{\text{ts}} |B_{iK}(g)| \\
&\leq \max_{1 \leq i \leq N-1} |B_{iK}(g)| \\
&\xrightarrow{p} 0
\end{aligned}$$

□

D.3 Proof of Proposition 3

Proof. We begin by proving consistency of the weights. First, note that the donor-weight set

$$\mathcal{W} = \left\{ w \in \mathbb{R}_+^{N-1} : \sum_{i=1}^{N-1} w_i = 1 \right\}$$

is a closed and bounded subset of finite-dimensional Euclidean space, and is therefore compact. By Lemma 1, L is continuous on $\mathcal{W}$. Hence, by the extreme value theorem, L attains its minimum on $\mathcal{W}$, so

$$\mathcal{W}_0 = \arg\min_{w \in \mathcal{W}} L(w)$$

is nonempty. Moreover, since L is continuous, $\mathcal{W}_0$ is a closed subset of the compact set $\mathcal{W}$, and is therefore compact.

Let

$$L_0 = \min_{w \in \mathcal{W}} L(w)$$

Fix $\varepsilon > 0$, and define the set of feasible weights that are at least ε away from the population minimizer set:

$$\mathcal{W}_\varepsilon = \{ w \in \mathcal{W} : d(w, \mathcal{W}_0) \geq \varepsilon \}$$

Since $d(\cdot, \mathcal{W}_0)$ is continuous and $\mathcal{W}$ is compact, $\mathcal{W}_\varepsilon$ is compact. If $\mathcal{W}_\varepsilon$ is empty, then $d(\hat{w}, \mathcal{W}_0) < \varepsilon$ automatically, so there is nothing to prove. Suppose therefore that $\mathcal{W}_\varepsilon$ is nonempty.

By continuity of L and compactness of $\mathcal{W}_\varepsilon$, L attains its minimum over $\mathcal{W}_\varepsilon$. Define

$$m_\varepsilon = \min_{w \in \mathcal{W}_\varepsilon} L(w)$$

We claim that $m_\varepsilon > L_0$. To see this, suppose instead that $m_\varepsilon = L_0$. Then there exists $w_\varepsilon \in \mathcal{W}_\varepsilon$ such that $L(w_\varepsilon) = L_0$, which implies $w_\varepsilon \in \mathcal{W}_0$. But then $d(w_\varepsilon, \mathcal{W}_0) = 0$, contradicting the definition

of $\mathcal{W}_\varepsilon$, which requires $d(w_\varepsilon, \mathcal{W}_0) \geq \varepsilon > 0$. Therefore $m_\varepsilon > L_0$.

Now, let $\gamma_\varepsilon = m_\varepsilon - L_0 > 0$ and consider the event

$$\mathcal{E}_\varepsilon = \left\{ \sup_{w \in \mathcal{W}} \left| \widehat{L}(w) - L(w) \right| \leq \frac{\gamma_\varepsilon}{3} \right\}$$

By Lemma 4, $\Pr(\mathcal{E}_\varepsilon) \to 1$. We now show that, on $\mathcal{E}_\varepsilon$, $\hat{w} \notin \mathcal{W}_\varepsilon$. Suppose, toward a contradiction, that $\hat{w} \in \mathcal{W}_\varepsilon$. Then

$$L(\hat{w}) \geq m_\varepsilon = L_0 + \gamma_\varepsilon$$

Hence, on $\mathcal{E}_\varepsilon$,

$$\widehat{L}(\hat{w}) \geq L(\hat{w}) - \frac{\gamma_\varepsilon}{3} \geq L_0 + \frac{2\gamma_\varepsilon}{3}$$

On the other hand, take any $w^0 \in \mathcal{W}_0$. Since $L(w^0) = L_0$, on $\mathcal{E}_\varepsilon$,

$$\widehat{L}(w^0) \leq L(w^0) + \frac{\gamma_\varepsilon}{3} = L_0 + \frac{\gamma_\varepsilon}{3}$$

Therefore, $\widehat{L}(\hat{w}) > \widehat{L}(w^0)$, which contradicts

$$\hat{w} \in \arg\min_{w \in \mathcal{W}} \widehat{L}(w)$$

Thus, on $\mathcal{E}_\varepsilon$, $d(\hat{w}, \mathcal{W}_0) < \varepsilon$. Consequently,

$$\begin{aligned} \Pr\left(d(\hat{w}, \mathcal{W}_0) \geq \varepsilon\right) &\leq 1 - \Pr\left(\mathcal{E}_\varepsilon\right) \\ &= \Pr\left(\sup_{w \in \mathcal{W}} \left| \widehat{L}(w) - L(w) \right| > \frac{\gamma_\varepsilon}{3} \right) \\ &\to 0 \end{aligned}$$

where the convergence follows from Lemma 4. Since this holds for every $\varepsilon > 0$, $d(\hat{w}, \mathcal{W}_0) \xrightarrow{p} 0$. Finally, if $\mathcal{W}_0 = \{w^0\}$, then the point-to-set distance reduces to ordinary Euclidean distance:

$$d(\hat{w}, \mathcal{W}_0) = \inf_{w \in \{w^0\}} \|\hat{w} - w\| = \|\hat{w} - w^0\|$$

Hence,

$$d(\hat{w}, \mathcal{W}_0) \xrightarrow{p} 0 \implies \|\hat{w} - w^0\| \xrightarrow{p} 0$$

This is equivalent to $\hat{w} \xrightarrow{p} w^0$, proving the first result.

Now, we turn our attention to consistency of the estimated counterfactual. Define the post-

treatment synthetic distribution map

$$S(w) = \sum_{i=1}^{N-1} w_i \mathcal{F}_{iT}(0)$$

By Lemma 2, every population minimizer $w^0 \in \mathcal{W}_0$ induces the same post-treatment synthetic distribution:

$$S(w^0) = \sum_{i=1}^{N-1} w_i^0 \mathcal{F}_{iT}(0) = \widetilde{P}_0$$

We first show that

$$\mathrm{MMD}_k \left(S(\hat{w}), \widetilde{P}_0 \right) \xrightarrow{p} 0$$

Fix $\varepsilon > 0$. By Lemma 8, the map $S(\cdot)$ is uniformly continuous on the compact set $\mathcal{W}$ under MMD_k. Therefore, there exists $\delta > 0$ such that, for any $w, w' \in \mathcal{W}$,

$$\|w - w'\| < \delta \implies \mathrm{MMD}_k \left(S(w), S(w') \right) < \varepsilon$$

On the event $\{d(\hat{w}, \mathcal{W}_0) < \delta\}$, there exists $w^0 \in \mathcal{W}_0$ such that $\|\hat{w} - w^0\| < \delta$. Hence, on this event,

$$\mathrm{MMD}_k \left(S(\hat{w}), \widetilde{P}_0 \right) = \mathrm{MMD}_k \left(S(\hat{w}), S(w^0) \right) < \varepsilon$$

where the equality uses Lemma 2. Therefore,

$$\left\{ \mathrm{MMD}_k \left(S(\hat{w}), \widetilde{P}_0 \right) \geq \varepsilon \right\} \subseteq \{ d(\hat{w}, \mathcal{W}_0) \geq \delta \}$$

Taking probabilities gives

$$\begin{aligned} \Pr \left[\mathrm{MMD}_k \left(S(\hat{w}), \widetilde{P}_0 \right) \geq \varepsilon \right] &\leq \Pr \left[d(\hat{w}, \mathcal{W}_0) \geq \delta \right] \\ &\to 0 \end{aligned}$$

The convergence comes from our previous conclusion that $d(\hat{w}, \mathcal{W}_0) \xrightarrow{p} 0$. Thus,

$$\mathrm{MMD}_k \left(\sum_{i=1}^{N-1} \hat{w}_i \mathcal{F}_{iT}(0), \widetilde{P}_0 \right) \xrightarrow{p} 0$$

Now decompose the discrepancy between the empirical synthetic distribution and the pseudo-counterfactual. By Lemma 7,

$$\mathrm{MMD}_k \left(\sum_{i=1}^{N-1} \hat{w}_i \widehat{F}_{iT}, \widetilde{P}_0 \right) \leq \mathrm{MMD}_k \left(\sum_{i=1}^{N-1} \hat{w}_i \widehat{F}_{iT}, \sum_{i=1}^{N-1} \hat{w}_i \mathcal{F}_{iT}(0) \right) + \mathrm{MMD}_k \left(\sum_{i=1}^{N-1} \hat{w}_i \mathcal{F}_{iT}(0), \widetilde{P}_0 \right)$$

We have just shown that the second term on the right hand side converges to 0 in probability, and the first term does so as well by Lemma 9. Therefore,

$$\mathrm{MMD}_k\left(\sum_{i=1}^{N-1}\hat{w}_i\widehat{F}_{iT},\widetilde{P}_0\right)\xrightarrow{p}0$$

By Proposition 1, $P_0=\widetilde{P}_0$, completing the proof. □

D.4 Proof of Proposition 4

Proof. By Proposition 2 and the first result of Proposition 3, $\hat{w}\xrightarrow{p}w^0$ and, therefore, $\hat{\theta}\xrightarrow{p}\theta^0$. Because $\theta^0\in\mathrm{int}(\Theta)$, consistency implies that $\hat{\theta}\in\mathrm{int}(\Theta)$ with probability approaching one. On this event, the first-order condition is

$$\nabla\widehat{\ell}(\hat{\theta})=0$$

For each coordinate $j=1,\ldots,N-2$, the mean-value theorem gives a point $\tilde{\theta}_j$ on the line segment joining θ^0 and $\hat{\theta}$ such that the jth component of the sample score satisfies

$$\frac{\partial\widehat{\ell}(\hat{\theta})}{\partial\theta_j}=\frac{\partial\widehat{\ell}(\theta^0)}{\partial\theta_j}+\nabla\left(\frac{\partial\widehat{\ell}(\tilde{\theta}_j)}{\partial\theta_j}\right)^{\top}(\hat{\theta}-\theta^0)$$

Stacking these equations yields

$$0=\nabla\widehat{\ell}(\theta^0)+\widehat{H}(\hat{\theta}-\theta^0)$$

where the jth row of $\widehat{H}$ is the jth row of $\nabla^2\widehat{\ell}(\tilde{\theta}_j)$.

Since $\hat{\theta}\xrightarrow{p}\theta^0$, each $\tilde{\theta}_j\xrightarrow{p}\theta^0$. Lemma 12 therefore implies $\widehat{H}\xrightarrow{p}H$. Because H is positive definite, it is nonsingular, and $\widehat{H}^{-1}\xrightarrow{p}H^{-1}$. Rearranging the sample first-order condition gives

$$\sqrt{K_{\min}^{\mathrm{pre}}}(\hat{\theta}-\theta^0)=-\widehat{H}^{-1}\sqrt{K_{\min}^{\mathrm{pre}}}\nabla\widehat{\ell}(\theta^0).$$

Lemma 11, together with $\nabla\ell(\theta^0)=0$, and Slutsky's theorem yield

$$\sqrt{K_{\min}^{\mathrm{pre}}}(\hat{\theta}-\theta^0)\xrightarrow{d}\mathcal{N}(0,H^{-1}\Omega H^{-1})$$

Finally, the weight map is affine, so $\hat{w}-w^0=A(\hat{\theta}-\theta^0)$. Applying the continuous mapping theorem gives the stated limit for $\hat{w}$ and the covariance matrix in Equation 10. Moreover, $\mathbf{1}_{N-1}^{\top}A=0$, which establishes the claimed singularity. □

D.5 Proof of Proposition 5

Define $\psi_{itr} := \psi_{it}(x_{itr})$, let $\widehat{\psi}_{itr}$ denote the feasible plug-in counterpart of ψ_{itr}, and define sample analogs

$$\widehat{H} = \nabla^2 \widehat{\ell}(\hat{\theta}) \qquad \widehat{\Omega} = \sum_{t<T} \sum_{i=1}^{N} \frac{K_{\min}^{\text{pre}}}{K_{it}} \left(\frac{1}{K_{it}} \sum_{r=1}^{K_{it}} \widehat{\psi}_{itr} \widehat{\psi}_{itr}^{\top} \right) \qquad \widehat{\Sigma}_w = A \widehat{H}^{-1} \widehat{\Omega} \widehat{H}^{-1} A^{\top}$$

Proof. For each pre-treatment cell, define

$$V_{it} = \mathbb{E}\Big[\psi_{it}(X) \psi_{it}(X)^{\top} \Big]$$

Let its feasible counterpart be denoted as

$$\widehat{V}_{it} = \frac{1}{K_{it}} \sum_{r=1}^{K_{it}} \widehat{\psi}_{itr} \widehat{\psi}_{itr}^{\top}$$

The proof of Lemma 11 establishes asymptotic linearity of the score:

$$\sqrt{K} \left\{ \nabla \widehat{\ell}(\theta^0) - \nabla \ell(\theta^0) \right\} = \sum_{t<T} \sum_{i=1}^{N} \sqrt{\frac{K}{K_{it}}} \left\{ \frac{1}{\sqrt{K_{it}}} \sum_{r=1}^{K_{it}} \psi_{it}(x_{itr}) \right\} + o_p(1)$$

Where we have defined $K = K_{\min}^{\text{pre}}$ for convenience. Consequently, the limiting score covariance is given by

$$\Omega = \sum_{t<T} \sum_{i=1}^{N} \kappa_{it}^{-1} V_{it}$$

Where we have defined $\kappa_{it} = \kappa_{it}^{\text{pre}}$ for convenience.

We first show that $\widehat{V}_{it} \xrightarrow{p} V_{it}$ for every pre-treatment cell. By Lemma 4, the empirical kernel mean embeddings satisfy

$$\max_{i,t<T} \|\hat{\mu}_{it} - \mu_{it}\|_{\mathcal{H}_k} \xrightarrow{p} 0$$

The weight-consistency result in Proposition 3, together with Proposition 2, gives $\widehat{w} \xrightarrow{p} w^0$ and $\widehat{\theta} \xrightarrow{p} \theta^0$. From these consistency results, it follows that every empirical donor-contrast operator and synthetic-minus-treated residual entering the feasible influence functions converges to its population counterpart.

Because the kernel is bounded by Assumption k-Reg, we can bound the RKHS norm of the corresponding feature map $\phi(x)$:

$$\|\phi(x)\|_{\mathcal{H}_k} \leq \sqrt{\bar{k}}$$

This bound holds uniformly over $x \in \mathcal{X}$. The feasible influence functions are continuous functions

of these uniformly bounded feature maps and of the preceding consistently estimated quantities. Therefore, for every cell (i,t),

$$\max_{1\le r\le K_{it}} \left\| \widehat{\psi}_{itr} - \psi_{it}(x_{itr}) \right\|_2 \xrightarrow{p} 0$$

The feasible and population influence functions are also uniformly bounded in probability. First, note the following inequality for operator norms:

$$\begin{aligned}
\|aa^\top - bb^\top\|_{\text{op}} &= \|(a-b)a^\top + b(a-b)^\top\|_{\text{op}} \\
&\le \|(a-b)a^\top\|_{\text{op}} + \|b(a-b)^\top\|_{\text{op}} \\
&= \|(a-b)\|_2\|a\|_2 + \|b\|_2\|(a-b)\|_2 \\
&= (\|a\|_2 + \|b\|_2)\|(a-b)\|_2
\end{aligned}$$

Where we have used the triangle inequality. We can then write:

$$\begin{aligned}
\left\| \widehat{V}_{it} - \frac{1}{K_{it}} \sum_{r=1}^{K_{it}} \psi_{it}(x_{itr})\psi_{it}(x_{itr})^\top \right\|_{\text{op}} &= \left\| \frac{1}{K_{it}} \sum_{r=1}^{K_{it}} \left(\widehat{\psi}_{it}(x_{itr})\widehat{\psi}_{it}(x_{itr})^\top - \psi_{it}(x_{itr})\psi_{it}(x_{itr})^\top \right) \right\|_{\text{op}} \\
&\le \frac{1}{K_{it}} \sum_{r=1}^{K_{it}} \left\| \widehat{\psi}_{it}(x_{itr})\widehat{\psi}_{it}(x_{itr})^\top - \psi_{it}(x_{itr})\psi_{it}(x_{itr})^\top \right|_{\text{op}} \\
&\le \frac{1}{K_{it}} \sum_{r=1}^{K_{it}} \left(\|\widehat{\psi}_{it}(x_{itr})\|_2 + \|\psi_{it}(x_{itr})\|_2 \right) \|\widehat{\psi}_{it}(x_{itr}) - \psi_{it}(x_{itr})\|_2 \\
&\le \max_{1\le r\le K_{it}} \left(\|\widehat{\psi}_{it}(x_{itr})\|_2 + \|\psi_{it}(x_{itr})\|_2 \right) \max_{1\le r\le K_{it}} \|\widehat{\psi}_{it}(x_{itr}) - \psi_{it}(x_{itr})\|_2 \\
&= o_p(1)
\end{aligned}$$

Assumption DS and boundedness of the influence functions give the cellwise law of large numbers:

$$\frac{1}{K_{it}} \sum_{r=1}^{K_{it}} \psi_{it}(x_{itr})\psi_{it}(x_{itr})^\top \xrightarrow{p} V_{it}$$

Hence, we have established that $\|\widehat{V}_{it} - V_{it}\|_{\text{op}} \xrightarrow{p} 0$, and so $\widehat{V}_{it} \xrightarrow{p} V_{it}$.

We can then write

$$\widehat{\Omega} = \sum_{t<T} \sum_{i=1}^{N} \frac{K}{K_{it}} \widehat{V}_{it}$$

Assumption BSG gives

$$\frac{K}{K_{it}} \longrightarrow \kappa_{it}^{-1}.$$

Since the number of pre-treatment cells is fixed and finite, the continuous mapping theorem yields

$$\widehat{\Omega} \xrightarrow{p} \sum_{t<T} \sum_{i=1}^{N} \kappa_{it}^{-1} V_{it} = \Omega$$

Lemma 12 directly gives $\widehat{H} \xrightarrow{p} H$, while Lemma 10 establishes that H is positive definite. Consequently, $\widehat{H}$ is nonsingular with probability approaching one, and so $\widehat{H}^{-1} \xrightarrow{p} H^{-1}$. With all of these pieces in hand, we can then apply the continuous mapping theorem to the sandwich formula:

$$\widehat{\Sigma}_w = A\widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1}A^\top \xrightarrow{p} AH^{-1}\Omega H^{-1}A^\top = \Sigma_w$$

□

D.6 Proof of Proposition 6

Proof. Write $K = K_{\min}^{\text{pre}}$. Under the θ-parameterization, let $\widetilde{w} = w(\widetilde{\theta})$ and $w^0 = w(\theta^0)$, where

$$\widetilde{\theta} \in \arg\min_{\theta \in \Theta} \widehat{\ell}(\theta) \qquad \theta^0 \in \arg\min_{\theta \in \Theta} \ell(\theta)$$

For squared MMD, the population and empirical objectives are:

$$\widehat{\ell}(\theta) = \sum_{t<T} \left\| \widehat{b}_t - \widehat{\mathsf{D}}_t \theta \right\|_{\mathcal{H}_k}^2 \qquad \ell(\theta) = \sum_{t<T} \|b_t - \mathsf{D}_t \theta\|_{\mathcal{H}_k}^2$$

We first establish the required rates for the empirical kernel means. The proof of Lemma 4 gives

$$\mathbb{E}\left[\|\widehat{\mu}_{it} - \mu_{it}\|_{\mathcal{H}_k}^2\right] \leq \frac{\bar{k}}{K_{it}}$$

We can apply Markov's inequality to yield:

$$\begin{aligned}
\Pr\left(\|\widehat{\mu}_{it} - \mu_{it}\|_{\mathcal{H}_k} > \frac{M}{\sqrt{K}}\right) &= \Pr\left(\|\widehat{\mu}_{it} - \mu_{it}\|_{\mathcal{H}_k}^2 > \frac{M^2}{K}\right) \\
&\leq \frac{\mathbb{E}\left[\|\widehat{\mu}_{it} - \mu_{it}\|_{\mathcal{H}_k}^2\right]}{M^2/K} \\
&\leq \frac{\bar{k}}{M^2}
\end{aligned}$$

Consequently, $\|\widehat{\mu}_{it} - \mu_{it}\|_{\mathcal{H}_k} = O_p(K^{-1/2})$. Since the number of pre-treatment periods is fixed, we can apply the union bound to obtain:

$$\max_{\substack{i=1,\dots,N \\ t<T}} \|\widehat{\mu}_{it} - \mu_{it}\|_{\mathcal{H}_k} = O_p(K^{-1/2})$$

Since $\widehat{b}_t - b_t$ and $\widehat{\mathsf{D}}_t - \mathsf{D}_t$ are finite linear combinations/operators of these empirical-mean errors, we can write for every $t < T$:

$$\left\|\widehat{b}_t - b_t\right\|_{\mathcal{H}_k} = O_p(K^{-1/2}) \qquad \left\|\widehat{\mathsf{D}}_t - \mathsf{D}_t\right\|_{\text{op}} = O_p(K^{-1/2})$$

The population and empirical Hessians are given by:

$$H = 2\sum_{t<T} \mathsf{D}_t^* \mathsf{D}_t \qquad \widehat{H} = 2\sum_{t<T} \widehat{\mathsf{D}}_t^* \widehat{\mathsf{D}}_t$$

Define $\Delta\mathsf{D}_t = \widehat{\mathsf{D}}_t - \mathsf{D}_t$. We can then write:

$$\begin{aligned}
\left\|\widehat{H} - H\right\|_{\text{op}} &= \left\|2\sum_{t<T} \widehat{\mathsf{D}}_t^* \widehat{\mathsf{D}}_t - \mathsf{D}_t^* \mathsf{D}_t\right\|_{\text{op}} \\
&\leq 2\sum_{t<T} \left\|\widehat{\mathsf{D}}_t^* \widehat{\mathsf{D}}_t - \mathsf{D}_t^* \mathsf{D}_t\right\|_{\text{op}} \\
&= 2\sum_{t<T} \|\mathsf{D}_t^* \Delta\mathsf{D}_t + (\Delta\mathsf{D}_t)^* \mathsf{D}_t + (\Delta\mathsf{D}_t)^* \Delta\mathsf{D}_t\|_{\text{op}} \\
&\leq 2\sum_{t<T} \left(2\|\mathsf{D}_t\|_{\text{op}} \|\Delta\mathsf{D}_t\|_{\text{op}} + \|\Delta\mathsf{D}_t\|_{\text{op}}^2\right) \\
&= 2\sum_{t<T} O_p(K^{-1/2}) + O_p(K^{-1}) \\
&= O_p(K^{-1/2}) \\
&= o_p(1)
\end{aligned}$$

Moreover, Assumption IDT gives for every $h \neq 0$,

$$h^\top H h = 2\sum_{t<T} \|\mathsf{D}_t h\|_{\mathcal{H}_k}^2 > 0$$

Thus H is positive definite. Let

$$c = \lambda_{\min}(H) > 0$$

where $\lambda_{\min}(H)$ denotes the smallest eigenvalue of H. Hessian convergence implies

$$\Pr\left(\lambda_{\min}(\widehat{H}) \geq \frac{c}{2}\right) \to 1$$

Next, define the empirical-score error at the population minimizer:

$$g_K = \nabla\widehat{\ell}(\theta^0) - \nabla\ell(\theta^0)$$

and define

$$\Delta b_t = \widehat{b}_t - b_t$$

Expanding the empirical score gives

$$g_K = 2\sum_{t<T}\Big\{(\Delta\mathsf{D}_t)^*(\mathsf{D}_t\theta^0 - b_t) + \mathsf{D}_t^*(\Delta\mathsf{D}_t\theta^0 - \Delta b_t) + (\Delta\mathsf{D}_t)^*(\Delta\mathsf{D}_t\theta^0 - \Delta b_t)\Big\}$$

The first two terms in each summand are $O_p(K^{-1/2})$, while the final term is $O_p(K^{-1})$. Hence, $g_K = O_p(K^{-1/2})$.

It remains to translate this score rate into a rate for the constrained minimizer. Define $\delta_K = \widetilde{\theta} - \theta^0$. Because θ^0 minimizes the differentiable convex function ℓ over the convex set Θ, we have

$$\delta_K^\top \nabla\ell(\theta^0) \geq 0$$

Analogously, because $\widetilde{\theta}$ minimizes $\widehat{\ell}$ over Θ,

$$\delta_K^\top \nabla\widehat{\ell}(\widetilde{\theta}) \leq 0$$

The empirical objective is quadratic, so

$$\nabla\widehat{\ell}(\widetilde{\theta}) = \nabla\widehat{\ell}(\theta^0) + \widehat{H}\delta_K$$

Consequently,

$$\begin{aligned}\delta_K^\top \widehat{H}\delta_K &\leq -\delta_K^\top \nabla\widehat{\ell}(\theta^0) \\ &= -\delta_K^\top g_K - \delta_K^\top \nabla\ell(\theta^0) \\ &\leq \|\delta_K\|_2\|g_K\|_2\end{aligned}$$

Where the final inequality follows from Cauchy-Schwarz. On the event

$$\Big\{\lambda_{\min}(\widehat{H}) \geq c/2\Big\}$$

we therefore have

$$\frac{c}{2}\|\delta_K\|_2^2 \le \|\delta_K\|_2 \|g_K\|_2$$

If $\delta_K \neq 0$, we can then write:

$$\|\widetilde{\theta} - \theta^0\|_2 \le \frac{2}{c}\|g_K\|_2 = O_p(K^{-1/2})$$

Note that the same conclusion is immediate if $\delta_K = 0$. Thus,

$$\sqrt{K}\big(\widetilde{\theta} - \theta^0\big) = O_p(1)$$

Because $w(\theta)$ is affine with fixed Jacobian A,

$$\widetilde{w} - w^0 = A(\widetilde{\theta} - \theta^0)$$

and therefore

$$\sqrt{K}\big(\widetilde{w} - w^0\big) = O_p(1)$$

Finally, the stated assumption on the cutoff schedule λ_K implies

$$\frac{\|\widetilde{w} - w^0\|_\infty}{\lambda_K} = O_p\left(\frac{1}{\sqrt{K}\,\lambda_K}\right) = o_p(1)$$

Thus:

$$\Pr\big(\|\widetilde{w} - w^0\|_\infty < \lambda_K\big) \to 1$$

Because N is fixed and S_0 is nonempty, define

$$\underline{w}^0 = \min_{i \in S_0} w_i^0 > 0.$$

Since $\lambda_K \to 0$, we have $\lambda_K < \underline{w}^0/2$ for all sufficiently large K. On the event

$$\big\{\|\widetilde{w} - w^0\|_\infty < \lambda_K\big\}$$

if $i \notin S_0$, then $w_i^0 = 0$ and $0 \le \widetilde{w}_i < \lambda_K$, and so $i \notin \widehat{S}$. If $i \in S_0$, then

$$\widetilde{w}_i > w_i^0 - \lambda_K \ge \underline{w}^0 - \lambda_K > \lambda_K$$

and so $i \in \widehat{S}$. Hence $\widehat{S} = S_0$ on an event whose probability converges to one, establishing

$$\Pr\Big(\widehat{S} = S_0\Big) \to 1$$

□

D.7 Proof of Theorem 1

Proof. Write $K = K^{\text{pre}}_{\min}$ and define the support-recovery event $\mathcal{E}_K = \{\widehat{S} = S_0\}$. By Proposition 6, $\Pr(\mathcal{E}_K) \to 1$. Define the oracle support-refitted estimator by

$$\widehat{w}^{\text{or}} \in \arg\min_{w \in \mathcal{W}_{S_0}} \widehat{L}(w)$$

On $\mathcal{E}_K$, note that $\mathcal{W}_{\widehat{S}} = \mathcal{W}_{S_0}$. The feasible and oracle estimators therefore minimize the same objective over the same face. Using the same argmin selection rule, $\widehat{w}^{\text{ts}} = \widehat{w}^{\text{or}}$ on $\mathcal{E}_K$.

We first establish consistency. The face $\mathcal{W}_{S_0}$ can be viewed as the unrestricted simplex for the synthetic control problem obtained by restricting the donor set to S_0. After relabeling these donors, Assumptions DS and k-Reg continue to hold. Applying Lemma 4 to this restricted donor problem therefore gives

$$\sup_{w \in \mathcal{W}_{S_0}} \left| \widehat{L}(w) - L(w) \right| \xrightarrow{p} 0$$

By Lemma 13, w^0 is the unique population minimizer of L over the compact face $\mathcal{W}_{S_0}$. The argmin consistency theorem consequently gives $\widehat{w}^{\text{or}} \xrightarrow{p} w^0$. For every $\varepsilon > 0$,

$$\begin{aligned} \Pr\left(\left\|\widehat{w}^{\text{ts}} - w^0\right\|_2 > \varepsilon\right) &\leq \Pr(\mathcal{E}_K^c) + \Pr\left(\left\|\widehat{w}^{\text{or}} - w^0\right\|_2 > \varepsilon\right) \\ &\to 0 \end{aligned}$$

Thus, $\widehat{w}^{\text{ts}} \xrightarrow{p} w^0$.

Now suppose that $s_0 \geq 2$ and Assumption BSG also holds. Parameterize the oracle face by $s_0 - 1$ free coordinates:

$$w = w_{S_0}(\theta) \qquad \theta \in \Theta_{S_0}$$

Let θ^0 denote the coordinates corresponding to w^0: $w^0 = w_{S_0}(\theta^0)$. Let A_{S_0} denote the Jacobian of $w_{S_0}(\theta)$. By Lemma 13, $\theta^0 \in \text{int}(\Theta_{S_0})$ and θ^0 is the unique minimizer of the population objective on this parameter space. The population MMD objective restricted to this face of the full donor simplex can be written as

$$\ell_{S_0}(\theta) = \sum_{t<T} \|b_{t,S_0} - \mathsf{D}_{t,S_0}\theta\|^2_{\mathcal{H}_k}$$

with Hessian

$$H_{S_0} = 2\sum_{t<T} \mathsf{D}^*_{t,S_0}\mathsf{D}_{t,S_0}$$

We will first show that H_{S_0} is positive definite. For every $h \in \mathbb{R}^{s_0-1}$,

$$h^\top H_{S_0} h = 2 \sum_{t<T} \|\mathsf{D}_{t,S_0} h\|^2_{\mathcal{H}_k} \geq 0$$

Towards a contradiction, suppose that this expression were exactly zero for some $h \neq 0$. Then, we would have $\mathsf{D}_{t,S_0} h = 0$ for every $t < T$. Let $a = A_{S_0} h \in \mathbb{R}^{N-1}$ denote the corresponding perturbation of the full weight vector, including zeros outside S_0, and suppose the elements of S_0 are indexed $S_0 = \{i_1, \ldots, i_{s_0}\}$. We would then have $\mathbf{1}^\top a = 0$ and, for every $t < T$:

$$\begin{aligned}
\sum_{i=1}^{N-1} a_i \mu_{it} &= \sum_{j=1}^{s_0-1} h_j \mu_{i_j t} - \left(\sum_{j=1}^{s_0-1} h_j \right) \mu_{i_{s_0} t} \\
&= \sum_{j=1}^{s_0-1} h_j (\mu_{i_j t} - \mu_{i_{s_0} t}) \\
&= \mathsf{D}_{t,S_0} h \\
&= 0
\end{aligned}$$

By Assumption IDT, we then have $a = 0$. Since A_{S_0} has full column rank, this implies $h = 0$, a contradiction. Thus, H_{S_0} must be positive definite.

Now, let

$$\widehat{\ell}_{S_0}(\theta) = \widehat{L}\big(w_{S_0}(\theta)\big)$$

denote the empirical objective restricted to the oracle face, and write

$$\widehat{H}_{S_0} = \nabla^2 \widehat{\ell}_{S_0}(\theta)$$

Because the empirical MMD objective is quadratic, $\widehat{H}_{S_0}$ does not depend on θ.

The restricted problem satisfies the conditions of Lemma 12: the Assumptions DS and k-Reg are inherited from the full problem, while Lemma 13 supplies the unique interior population minimizer. Relabeling the donors in S_0 therefore gives $\widehat{H}_{S_0} \xrightarrow{p} H_{S_0}$. Since H_{S_0} is positive definite, $\widehat{H}_{S_0}$ is nonsingular with probability approaching one and $\widehat{H}_{S_0}^{-1} \xrightarrow{p} H_{S_0}^{-1}$. The same argument applies to Lemma 11, where Assumption BSG for the full collection of pre-treatment cells implies balanced growth for the cells indexed by $S_0 \cup \{N\}$. We therefore have:

$$\sqrt{K} \left(\nabla \widehat{\ell}_{S_0}(\theta^0) - \nabla \ell_{S_0}(\theta^0) \right) \xrightarrow{d} \mathcal{N}(0, \Omega_{S_0})$$

Because θ^0 is an interior population minimizer, $\nabla \ell_{S_0}(\theta^0) = 0$. Consequently,

$$\sqrt{K}\, \nabla \widehat{\ell}_{S_0}(\theta^0) \xrightarrow{d} \mathcal{N}(0, \Omega_{S_0})$$

Now, let $\widehat{\theta}^{\mathrm{or}}$ denote the free coordinates of $\widehat{w}^{\mathrm{or}}$. Oracle consistency and interiority of θ^0 imply

$$\Pr\Big(\widehat{\theta}^{\mathrm{or}} \in \mathrm{int}(\Theta_{S_0})\Big) \to 1$$

On this event, the empirical first-order condition gives $\nabla\widehat{\ell}_{S_0}(\widehat{\theta}^{\mathrm{or}}) = 0$. Since $\widehat{\ell}_{S_0}$ is quadratic,

$$0 = \nabla\widehat{\ell}_{S_0}(\theta^0) + \widehat{H}_{S_0}\big(\widehat{\theta}^{\mathrm{or}} - \theta^0\big)$$

Rearranging terms and left-multiplying by the inverse Hessian then yields:

$$\sqrt{K}\big(\widehat{\theta}^{\mathrm{or}} - \theta^0\big) = -\widehat{H}_{S_0}^{-1}\sqrt{K}\,\nabla\widehat{\ell}_{S_0}(\theta^0)$$

Hessian convergence, the score central limit theorem, and Slutsky's theorem now give

$$\sqrt{K}\big(\widehat{\theta}^{\mathrm{or}} - \theta^0\big) \xrightarrow{d} \mathcal{N}\left(0, H_{S_0}^{-1}\Omega_{S_0}H_{S_0}^{-1}\right).$$

Because $w_{S_0}(\theta)$ is affine,

$$\widehat{w}^{\mathrm{or}} - w^0 = A_{S_0}\big(\widehat{\theta}^{\mathrm{or}} - \theta^0\big)$$

Therefore, the continuous mapping theorem yields:

$$\sqrt{K}\big(\widehat{w}^{\mathrm{or}} - w^0\big) \xrightarrow{d} \mathcal{N}(0, \underbrace{A_{S_0}H_{S_0}^{-1}\Omega_{S_0}H_{S_0}^{-1}A_{S_0}^{\top}}_{\Sigma_{w,S_0}})$$

Finally, oracle equivalence gives

$$\Pr\big(\widehat{w}^{\mathrm{ts}} \neq \widehat{w}^{\mathrm{or}}\big) \leq \Pr(\mathcal{E}_K^c) \longrightarrow 0$$

Thus, for every $\varepsilon > 0$,

$$\Pr\Big(\sqrt{K}\,\big\|\widehat{w}^{\mathrm{ts}} - \widehat{w}^{\mathrm{or}}\big\|_2 > \varepsilon\Big) \leq \Pr(\mathcal{E}_K^c) \longrightarrow 0$$

and so

$$\sqrt{K}\left(\widehat{w}^{\mathrm{ts}} - \widehat{w}^{\mathrm{or}}\right) \xrightarrow{p} 0$$

Combining this result with the oracle limit and applying Slutsky's theorem yields

$$\sqrt{K_{\min}^{\mathrm{pre}}}\big(\widehat{w}^{\mathrm{ts}} - w^0\big) \xrightarrow{d} \mathcal{N}(0, \Sigma_{w,S_0})$$

□

D.8 Proof of Corollary 1

Proof. Define the support-recovery event $\mathcal{E}_K = \{\widehat{S} = S_0\}$. By Proposition 6, $\Pr(\mathcal{E}_K) \to 1$. Consider first the fixed oracle-face problem with donor set S_0. By Lemma 13, w^0 is the unique relative-interior population minimizer on $\mathcal{W}_{S_0}$. Moreover, the proof of Theorem 1 shows that the sampling, kernel, balanced-growth, score-CLT, and Hessian-convergence conditions required for the interior-weight problem hold after restricting the donor set to S_0. Therefore, applying Proposition 5 to the fixed oracle-face problem gives

$$\widehat{H}_{S_0} \xrightarrow{p} H_{S_0} \qquad \widehat{\Omega}_{S_0} \xrightarrow{p} \Omega_{S_0}$$

Hence, the continuous mapping theorem yields:

$$\widehat{\Sigma}_{w,S_0} := A_{S_0}\widehat{H}_{S_0}^{-1}\widehat{\Omega}_{S_0}\widehat{H}_{S_0}^{-1}A_{S_0}^{\top} \xrightarrow{p} A_{S_0}H_{S_0}^{-1}\Omega_{S_0}H_{S_0}^{-1}A_{S_0}^{\top} = \Sigma_{w,S_0}$$

On $\mathcal{E}_K$, the selected and oracle faces coincide. Using the same ordering and reference-donor rule in both parameterizations, we have:

$$A_{\widehat{S}} = A_{S_0} \qquad \widehat{H}_{\widehat{S}} = \widehat{H}_{S_0} \qquad \widehat{\Omega}_{\widehat{S}} = \widehat{\Omega}_{S_0}$$

Consequently, $\widehat{\Sigma}_{w,\widehat{S}} = \widehat{\Sigma}_{w,S_0}$. For every $\varepsilon > 0$,

$$\begin{aligned}\Pr\left(\left\|\widehat{\Sigma}_{w,\widehat{S}} - \Sigma_{w,S_0}\right\|_{\mathrm{op}} > \varepsilon\right) &\le \Pr(\mathcal{E}_K^c) + \Pr\left(\left\|\widehat{\Sigma}_{w,S_0} - \Sigma_{w,S_0}\right\|_{\mathrm{op}} > \varepsilon\right) \\ &\to 0\end{aligned}$$

This is equivalent to $\widehat{\Sigma}_{w,\widehat{S}} \xrightarrow{p} \Sigma_{w,S_0}$, establishing the desired result. □

D.9 Proof of Theorem 2

Proof. We begin by establishing the first claim. By Lemma 14,

$$\widehat{\tau}_g - \widetilde{\tau}_g = A_{1K}(g) - A_{0K}(g) - b_{K,g}$$

By Lemma 15, $A_{1K}(g) \xrightarrow{p} 0$ and $A_{0K}(g) \xrightarrow{p} 0$. Note that we can rewrite $b_{K,g}$ as:

$$b_{K,g} = \mu(g)^{\top}(\widehat{w}^{\mathrm{ts}} - w^0)$$

By the Cauchy-Schwarz inequality, we can then write:

$$|b_{K,g}| \le \|\mu(g)\|_2 \|\widehat{w}^{\mathrm{ts}} - w^0\|_2$$

By Assumption BFS we have $\|\mu(g)\|_2 < \infty$, while Theorem 1 gives us $\widehat{w}^{\text{ts}} \xrightarrow{p} w^0$. Consequently, $b_{K,g} \xrightarrow{p} 0$. Therefore, $\widehat{\tau}_g - \widetilde{\tau}_g \xrightarrow{p} 0$, which is equivalent to $\widehat{\tau}_g \xrightarrow{p} \widetilde{\tau}_g$. By Proposition 1, $\widetilde{P}_0 = P_0$, and so by definition $\widetilde{\tau}_g = \tau_g$. This completes the proof of consistency.

Now, we turn to asymptotic normality. Define the following:

$$\bar{g}_{iT} = \frac{1}{K_{iT}} \sum_{k=1}^{K_{iT}} g(x_{iTk}) \qquad \overline{\boldsymbol{g}}_T = \left(\bar{g}_{1T}, \ldots, \bar{g}_{(N-1)T}\right)^\top \qquad B_{K,g} = \overline{\boldsymbol{g}}_T - \mu(g)$$

Because the conditions of Theorem 1 hold, the population minimizer is unique: $\mathcal{W}^0 = \{w^0\}$. Consequently, $\widetilde{\mu}_0(g) = (w^0)^\top \mu(g)$, and so:

$$\widetilde{\tau}_g = \mathbb{E}_{P_1}[g(X)] - (w^0)^\top \mu(g)$$

By construction,

$$\widehat{\tau}_g = \bar{g}_{NT} - (\widehat{w}^{\text{ts}})^\top \overline{\boldsymbol{g}}_T$$

We can then write:

$$\widehat{\tau}_g - \widetilde{\tau}_g = \bar{g}_{NT} - \mathbb{E}_{P_1}[g(X)] - (w^0)^\top B_{K,g} - \mu(g)^\top \left(\widehat{w}^{\text{ts}} - w^0\right) - \left(\widehat{w}^{\text{ts}} - w^0\right)^\top B_{K,g}$$

We now analyze these terms in turn. Let us begin with the first three. Under Assumptions DS, BSG, and BFS, the independent cellwise central limit theorems give:

$$\sqrt{K_{\min}} \begin{pmatrix} \bar{g}_{NT} - \mathbb{E}_{P_1}[g(X)] \\ B_{K,g} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Gamma_g) \qquad \Gamma_g = \operatorname{diag}\left(\frac{\sigma_N^2(g)}{\kappa_{NT}^{\text{post}}}, \frac{\sigma_1^2(g)}{\kappa_{1T}^{\text{post}}}, \ldots, \frac{\sigma_{N-1}^2(g)}{\kappa_{(N-1)T}^{\text{post}}}\right)$$

Now, define the following:

$$V_g^{\text{post}} = \frac{\sigma_N^2(g)}{\kappa_{NT}^{\text{post}}} + \sum_{i=1}^{N-1} (w_i^0)^2 \frac{\sigma_i^2(g)}{\kappa_{iT}^{\text{post}}} \qquad Z_{K,g}^{\text{post}} = \sqrt{K_{\min}} \left(\bar{g}_{NT} - \mathbb{E}_{P_1}[g(X)] - (w^0)^\top B_{K,g}\right)$$

Let $Z_g^{\text{post}} \sim \mathcal{N}(0, V_g^{\text{post}})$. By the continuous mapping theorem, we then have $Z_{K,g}^{\text{post}} \xrightarrow{d} Z_g^{\text{post}}$.

Next, let us turn our attention to the fourth term in the above decomposition. Suppose first that $s_0 \geq 2$. By Theorem 1,

$$H_{K,w} := \sqrt{K_{\min}^{\text{pre}}} \left(\widehat{w}^{\text{ts}} - w^0\right) \xrightarrow{d} H_{w,\mathcal{S}_0} \qquad H_{w,\mathcal{S}_0} \sim \mathcal{N}(0, \Sigma_{w,\mathcal{S}_0})$$

Note that if $s_0 = 1$, Theorem 1 gives $\widehat{w}^{\text{ts}} = w^0$ with probability approaching one, and so $H_{K,w} \xrightarrow{p} 0$. The following argument therefore applies in this case as well, with $H_{w,\mathcal{S}_0} = \Sigma_{w,\mathcal{S}_0} = 0$.

Next, note that $Z_{K,g}^{\text{post}}$ depends only on post-treatment samples, while $H_{K,w}$ depends only on

pre-treatment samples. Assumption DS therefore implies that they are independent for every K. Consequently,

$$\left(Z_{K,g}^{\text{post}}, H_{K,w}\right) \xrightarrow{d} \left(Z_g^{\text{post}}, H_{w,\mathcal{S}_0}\right)$$

where $Z_g^{\text{post}} \perp H_{w,\mathcal{S}_0}$. Additionally, by the definition of ρ_K, we can write:

$$\sqrt{K_{\min}}\,\mu(g)^\top\left(\widehat{w}^{\text{ts}} - w^0\right) = \sqrt{\rho_K}\,\mu(g)^\top H_{K,w}$$

Finally, we turn our attention to the last term in the decomposition. The post-treatment central limit theorem implies:

$$\|B_{K,g}\|_2 = O_p\left(K_{\min}^{-1/2}\right)$$

The weight limit implies:

$$\left\|\widehat{w}^{\text{ts}} - w^0\right\|_2 = O_p\left((K_{\min}^{\text{pre}})^{-1/2}\right)$$

Applying Cauchy-Schwarz then yields:

$$\begin{aligned}\sqrt{K_{\min}}\left|\left(\widehat{w}^{\text{ts}} - w^0\right)^\top B_{K,g}\right| &\leq \sqrt{K_{\min}}\left\|\widehat{w}^{\text{ts}} - w^0\right\|_2 \|B_{K,g}\|_2 \\ &= O_p\left((K_{\min}^{\text{pre}})^{-1/2}\right) \\ &= o_p(1)\end{aligned}$$

Combining the preceding results for each term yields:

$$\sqrt{K_{\min}}\left(\widehat{\tau}_g - \widetilde{\tau}_g\right) = Z_{K,g}^{\text{post}} - \sqrt{\rho_K}\,\mu(g)^\top H_{K,w} + o_p(1)$$

Assumption RPG, joint convergence, and Slutsky's theorem therefore give

$$\sqrt{K_{\min}}\left(\widehat{\tau}_g - \widetilde{\tau}_g\right) \xrightarrow{d} Z_g^{\text{post}} - \sqrt{\rho}\,\mu(g)^\top H_{w,\mathcal{S}_0}$$

The two terms on the right-hand side are independent, mean-zero Gaussian random variables. Their sum is therefore also a mean-zero Gaussian with variance

$$V_g = V_g^{\text{post}} + \rho\,\mu(g)^\top \Sigma_{w,\mathcal{S}_0}\mu(g)$$

Finally, note that Proposition 1 yields $\widetilde{P}_0 = P_0$, and so $\widetilde{\tau}_g = \tau_g$ by definition. This completes the proof. □

D.10 Proof of Corollary 2

Proof. Assumptions DS and BFS imply $\widehat{\sigma}_N^2(g) \xrightarrow{p} \sigma_N^2(g)$. Since the number of donors is finite, we can extend this jointly:

$$\max_{1\le i\le N-1} \left|\widehat{\sigma}_i^2(g) - \sigma_i^2(g)\right| \xrightarrow{p} 0$$

Meanwhile, Theorem 1 implies $\widehat{w}^{\text{ts}} \xrightarrow{p} w^0$. Together with Assumption BSG, applying the continuous mapping theorem to these results gives:

$$K_{\min}\left\{\frac{\widehat{\sigma}_N^2(g)}{K_{NT}} + \sum_{i=1}^{N-1}(\widehat{w}_i^{\text{ts}})^2\frac{\widehat{\sigma}_i^2(g)}{K_{iT}}\right\} \xrightarrow{p} \frac{\sigma_N^2(g)}{\kappa_{NT}^{\text{post}}} + \sum_{i=1}^{N-1}(w_i^0)^2\frac{\sigma_i^2(g)}{\kappa_{iT}^{\text{post}}}$$

Additionally, the post-treatment laws of large numbers imply $\widehat{\mu}(g) \xrightarrow{p} \mu(g)$, while Corollary 1 gives $\widehat{\Sigma}_{w,\widehat{\mathcal{S}}} \xrightarrow{p} \Sigma_{w,\mathcal{S}_0}$. Assumption RPG gives $\rho_K \to \rho$, and so applying the continuous mapping theorem once again yields:

$$\rho_K\widehat{\mu}(g)^\top\widehat{\Sigma}_{w,\widehat{\mathcal{S}}}\widehat{\mu}(g) \xrightarrow{p} \rho\,\mu(g)^\top\Sigma_{w,\mathcal{S}_0}\mu(g)$$

Combining these two results then yields:

$$\underbrace{K_{\min}\left\{\frac{\widehat{\sigma}_N^2(g)}{K_{NT}} + \sum_{i=1}^{N-1}(\widehat{w}_i^{\text{ts}})^2\frac{\widehat{\sigma}_i^2(g)}{K_{iT}}\right\} + \rho_K\widehat{\mu}(g)^\top\widehat{\Sigma}_{w,\widehat{\mathcal{S}}}\widehat{\mu}(g)}_{\widehat{V}_g} \xrightarrow{p} \underbrace{\frac{\sigma_N^2(g)}{\kappa_{NT}^{\text{post}}} + \sum_{i=1}^{N-1}(w_i^0)^2\frac{\sigma_i^2(g)}{\kappa_{iT}^{\text{post}}} + \rho\,\mu(g)^\top\Sigma_{w,\mathcal{S}_0}\,\mu(g)}_{V_g}$$

□

E Proofs for Section 4

E.1 Useful Lemmas

Lemma 16 (Foldwise two-stage weight limits)**.** *Suppose Assumptions SS and BSG hold. Suppose that, for each $a \in \{\text{disc}, \text{eval}\}$, the conditions of Proposition 6 hold after replacing*

$$\widehat{L} \quad \widetilde{w} \quad \widehat{\mathcal{S}} \quad \widehat{w}^{\text{ts}} \quad \widehat{w}^{\text{or}} \quad K_{\min}^{\text{pre}}$$

by their fold-specific counterparts

$$\widehat{L}^a \quad \widetilde{w}^a \quad \widehat{\mathcal{S}}^a \quad \widehat{w}^{\text{ts},a} \quad \widehat{w}^{\text{or},a} \quad K_{\min}^{\text{pre},a}$$

respectively.

Then, for each $a \in \{\text{disc}, \text{eval}\}$,

$$\Pr\left(\widehat{\mathcal{S}}^a = \mathcal{S}_0\right) \longrightarrow 1$$

and

$$\Pr\left(\widehat{w}^{\text{ts},a} = \widehat{w}^{\text{or},a}\right) \longrightarrow 1$$

Consequently,

$$\widetilde{w}^a \xrightarrow{p} w^0 \qquad \widehat{w}^{\text{ts},a} \xrightarrow{p} w^0$$

and

$$\sqrt{K_{\min}^{\text{pre},a}}\left(\widehat{w}^{\text{ts},a} - \widehat{w}^{\text{or},a}\right) \xrightarrow{p} 0$$

If $s_0 \geq 2$, *then*

$$\sqrt{K_{\min}^{\text{pre},a}}\left(\widehat{w}^{\text{ts},a} - w^0\right) \xrightarrow{d} N\left(0, \Sigma_{w,\mathcal{S}_0}\right)$$

If $s_0 = 1$, *then*

$$\Pr\left(\widehat{w}^{\text{ts},a} = w^0\right) \to 1$$

Proof of Lemma 16. Fix $a \in \{\text{disc}, \text{eval}\}$. Conditional on the split indices, the observations in each fold-specific pre-treatment cell are iid draws from the same population distribution as the corresponding unsplit observations. Moreover, Assumptions BSG and SS imply

$$\frac{K_{\min}^{\text{pre},a}}{K_{\min}^{\text{pre}}} \longrightarrow \pi^a$$

and, for every $i = 1, \ldots, N$ and $t < T$,

$$\frac{K_{it}^a}{K_{\min}^{\text{pre},a}} \longrightarrow \kappa_{it}^{\text{pre}}$$

Thus the fold-specific problem has the same population objective, population minimizer, and limiting cell-growth ratios as the original pre-treatment problem.

The conclusions therefore follow by applying the support-recovery, oracle-equivalence, consistency, and asymptotic-normality arguments of Theorem 1 to $\widehat{L}^a$, with $K_{\min}^{\text{pre},a}$ as the asymptotic scale. Because there are two folds, the conclusions hold for both simultaneously. □

Lemma 17 (Conditional evaluation-weight asymptotic normality)**.** *Suppose the conditions of Lemma 16 hold, and that Assumption SG holds as well.*

If $s_0 \geq 2$, *then, conditional on* $\mathcal{J}_K$,

$$\sqrt{K_{\min}^{\text{pre,eval}}}\left(\widehat{w}^{\text{ts,eval}} - w^0\right) \xrightarrow{d} N\left(0, \Sigma_{w,\mathcal{S}_0}\right)$$

in probability with respect to the discovery sample. That is, for every bounded continuous function

$f : \mathbb{R}^{N-1} \to \mathbb{R}$,

$$\mathbb{E}\left[f\left(\sqrt{K_{\min}^{\text{pre,eval}}}\left(\widehat{w}^{\text{ts,eval}} - w^0\right)\right) \middle| \mathcal{J}_K\right] \xrightarrow{p} \mathbb{E}\left[f(Z_w)\right] \qquad Z_w \sim N\left(0, \Sigma_{w,\mathcal{S}_0}\right)$$

If $s_0 = 1$, *then*

$$\Pr\left(\widehat{w}^{\text{ts,eval}} = w^0 \middle| \mathcal{J}_K\right) \xrightarrow{p} 1$$

Proof of Lemma 17. Conditional on $\mathcal{S}_K$, the pre-treatment evaluation fold is independent of

$$\sigma\left(D_{\text{pre}}^{\text{disc}}, D_{\text{post}}^{\text{disc}}, \mathcal{R}_K\right)$$

Consequently, for every bounded continuous function $f : \mathbb{R}^{N-1} \to \mathbb{R}$,

$$\mathbb{E}\left[f\left(\sqrt{K_{\min}^{\text{pre,eval}}}\left(\widehat{w}^{\text{ts,eval}} - w^0\right)\right) \middle| \mathcal{J}_K\right] = \mathbb{E}\left[f\left(\sqrt{K_{\min}^{\text{pre,eval}}}\left(\widehat{w}^{\text{ts,eval}} - w^0\right)\right) \middle| \mathcal{S}_K\right]$$

The fold-specific application of Theorem 1 gives:

$$\sqrt{K_{\min}^{\text{pre,eval}}}\left(\widehat{w}^{\text{ts,eval}} - w^0\right) \xrightarrow{d} N\left(0, \Sigma_{w,\mathcal{S}_0}\right)$$

To establish conditional convergence, note that pre-treatment evaluation observations remain iid within each cell and independent across cells. Furthermore, Assumptions BSG and SS yield, for every unit i and $t < T$:

$$\frac{K_{it}^{\text{eval}}}{K_{\min}^{\text{pre,eval}}} \xrightarrow{p} \kappa_{it}^{\text{pre}}$$

Therefore, as long as the sequence of split assignments satisfies these required growth and threshold conditions, the proof of Theorem 1 applies to the evaluation fold as well.

Now, let $Z_w \sim N\left(0, \Sigma_{w,\mathcal{S}_0}\right)$. The definition of convergence in distribution yields:

$$\mathbb{E}\left[f\left(\sqrt{K_{\min}^{\text{pre,eval}}}\left(\widehat{w}^{\text{ts,eval}} - w^0\right)\right) \middle| \mathcal{S}_K\right] \to \mathbb{E}\left[f\left(Z_w\right)\right]$$

To see that this holds even when fold sizes are random, let $m_K = K_{\min}^{\text{pre,eval}}$ and let $f : \mathbb{R}^{N-1} \to \mathbb{R}$ be a particular bounded continuous function. Define the following:

$$Q_K = \mathbb{E}\left[f\left(\sqrt{m_K}(\widehat{w}^{\text{ts,eval}} - w^0)\right) \middle| \mathcal{S}_K\right] \qquad c_f = \mathbb{E}[f(Z_w)]$$

Assumptions BSG, SS, and the fold-specific conditions on the threshold λ_K^{eval} imply the following

joint convergence in probability:

$$\left(\frac{1}{m_K}, \left(\frac{K_{it}^{\text{eval}}}{m_K}\right)_{i,t<T}, \lambda_K^{\text{eval}}, \frac{1}{\sqrt{m_K}\lambda_K^{\text{eval}}}\right) \xrightarrow{p} (0, (\kappa^{\text{pre}})_{i,t<T}, 0, 0)$$

Let C_K denote the sequence on the left, and C_0 its limit on the right. By the subsequence characterization of convergence in probability, for any sequence (K_j) there exists a subsequence (K_{j_ℓ}) such that $C_{K_{j_\ell}} \to C_0$ a.s. Along this subsequence, the growth conditions of Theorem 1 are satisfied and the observations still have correct, independent cellwise distributions. Theorem 1 then applies along this subsequence, giving $Q_{K_{j_\ell}} \to c_f$ a.s. Since the choice of initial subsequence (K_j) was arbitrary, this shows that every subsequence of (Q_K) has a further subsequence converging almost surely to c_f. Suppose, towards a contradiction, that there exist $\epsilon, \Delta > 0$ and a subsequence (K_j) such that, for every j:

$$\Pr(|Q_{K_j} - c_f| > \epsilon) \geq \Delta$$

By our previous argument, this subsequence has a further subsequence (K_{j_ℓ}) along which $Q_{K_{j_\ell}} \to c_f$ a.s. Since almost-sure convergence implies convergence in probability, this yields:

$$\Pr(|Q_{K_{j_\ell}} - c_f| > \epsilon) \to 0$$

This contradicts $\Delta > 0$, and so $Q_K \xrightarrow{p} c_f$. Chaining these results then yields:

$$\mathbb{E}\left[f\left(\sqrt{K_{\min}^{\text{pre,eval}}}\left(\widehat{w}^{\text{ts,eval}} - w^0\right)\right) \middle| \mathcal{J}_K\right] \xrightarrow{p} \mathbb{E}\left[f\left(Z_w\right)\right]$$

This proves the conditional weak convergence. If $s_0 = 1$, the same conditional argument yields:

$$\Pr\left(\widehat{w}^{\text{ts,eval}} = w^0 \middle| \mathcal{S}_K\right) \xrightarrow{p} 1$$

Conditional independence analogously allows us to transform this into a statement conditional on $\mathcal{J}_K$, as before. □

E.2 Proof of Theorem 3

Proof. We begin by deriving an algebraic decomposition of $\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}}$. We will add and subtract the following terms in the definition of $\widehat{\tau}_{\widehat{g}}^{\text{eval}}$:

$$\mathbb{E}_{P_1}[\widehat{g}(X)] \qquad \sum_{i=1}^{N-1} \widehat{w}_i^{\text{ts,eval}} \mathbb{E}_{F_{iT}(0)}[\widehat{g}(X)]$$

This gives

$$\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}} = \underbrace{\left\{\mathbb{E}_{\widehat{F}_{NT}^{\text{eval}}}[\widehat{g}(X)] - \mathbb{E}_{P_1}[\widehat{g}(X)]\right\}}_{A_{1K}^{\text{eval}}(\widehat{g})} - \underbrace{\sum_{i=1}^{N-1} \widehat{w}_i^{\text{ts,eval}} \left\{\mathbb{E}_{\widehat{F}_{iT}^{\text{eval}}}[\widehat{g}(X)] - \mathbb{E}_{\mathcal{F}_{iT}(0)}[\widehat{g}(X)]\right\}}_{A_{0K}^{\text{eval}}(\widehat{g})} - \underbrace{\left\{\sum_{i=1}^{N-1} \widehat{w}_i^{\text{ts,eval}} \mu_i(\widehat{g}) - \widetilde{\mu}_0(\widehat{g})\right\}}_{b_{K,\widehat{g}}^{\text{eval}}}$$

Note that we have defined $\widetilde{\mu}_0(\widehat{g}) = \mathbb{E}_{\widetilde{P}_0}[\widehat{g}(X)]$. Because $\mathcal{W}_0 = \{w^0\}$ by Proposition 2, we can write:

$$\widetilde{\mu}_0(\widehat{g}) = (w^0)^\top \mu(\widehat{g})$$

where we have defined $\mu_0(\widehat{g}) = \mathbb{E}_{P_0}[\widehat{g}(X)]$. We can then write the third term in the above decomposition as:

$$b_{K,\widehat{g}}^{\text{eval}} = \mu(\widehat{g})^\top \left(\widehat{w}^{\text{ts,eval}} - w^0\right)$$

This yields the final decomposition

$$\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}} = A_{1K}^{\text{eval}}(\widehat{g}) - A_{0K}^{\text{eval}}(\widehat{g}) - b_{K,\widehat{g}}^{\text{eval}}$$

Now, let us examine each term of this decomposition. First, note that:

$$\begin{aligned}
\mathbb{E}\left[A_{1K}^{\text{eval}}(\widehat{g}) \,\middle|\, \mathcal{J}_K\right] &= \mathbb{E}\left[\frac{1}{K_{NT}^{\text{eval}}} \sum_{k \in \mathcal{I}_{NT}^{\text{eval}}} \widehat{g}(x_{NTk}) - \mathbb{E}_{P_1}[\widehat{g}(X)] \,\middle|\, \mathcal{J}_K\right] \\
&= \frac{1}{K_{NT}^{\text{eval}}} \sum_{k \in \mathcal{I}_{NT}^{\text{eval}}} \mathbb{E}[\widehat{g}(x_{NTk}) | \mathcal{J}_K] - \mathbb{E}_{P_1}[\widehat{g}(X)] \\
&= \mathbb{E}_{P_1}[\widehat{g}(X)] - \mathbb{E}_{P_1}[\widehat{g}(X)] \\
&= 0
\end{aligned}$$

Using iid sampling from Assumption DS and Assumption SS, we can then write:

$$\begin{aligned}\mathbb{E}\left[A_{1K}^{\text{eval}}(\widehat{g})^2 \,\middle|\, \mathcal{J}_K\right] &= \text{Var}\left(A_{1K}^{\text{eval}}(\widehat{g}) \,\middle|\, \mathcal{J}_K\right) \\ &= \frac{\text{Var}_{P_1}(\widehat{g}(X))}{K_{NT}^{\text{eval}}} \\ &\leq \frac{\mathbb{E}_{P_1}(\widehat{g}(X)^2)}{K_{NT}^{\text{eval}}} \\ &\leq \frac{\bar{g}^2}{K_{NT}^{\text{eval}}}\end{aligned}$$

Where the final inequality follows from Assumption SG. Since $K_{\min} \to \infty$, Assumption SS implies $K_{NT}^{\text{eval}} \to \infty$, and so taking expectations yields:

$$\mathbb{E}\left[\left(A_{1K}^{\text{eval}}(\widehat{g})\right)^2\right] \to 0$$

Combining this with our previous result then establishes $A_{1K}^{\text{eval}}(\widehat{g}) \xrightarrow{p} 0$.

Now, let us turn our attention to the second term in the decomposition. For every donor, define:

$$B_{iK}(\widehat{g}) = \mathbb{E}_{\widehat{F}_{iT}^{\text{eval}}}[\widehat{g}(X)] - \mathbb{E}_{\mathcal{F}_{iT}(0)}[\widehat{g}(X)]$$

The same argument as above then gives:

$$\mathbb{E}\left[B_{iK}(\widehat{g})^2 \middle| \mathcal{J}_K\right] \leq \frac{\bar{g}^2}{K_{iT}^{\text{eval}}}$$

Because $\widehat{w}^{\text{ts,eval}} \in \mathcal{W}$, we have:

$$|A_{0K}^{\text{eval}}(\widehat{g})| \leq \max_{1 \leq i \leq N-1} |B_{iK}(\widehat{g})|$$

Since N is fixed, squaring and taking conditional expectations yields:

$$\begin{aligned}\mathbb{E}\left[\left(A_{0K}^{\text{eval}}(\widehat{g})\right)^2 \middle| \mathcal{J}_K\right] &\leq \mathbb{E}\left[\max_{1 \leq i \leq N-1} B_{iK}(\widehat{g})^2 \middle| \mathcal{J}_K\right] \\ &\leq \sum_{i=1}^{N-1} \mathbb{E}\left[B_{iK}(\widehat{g})^2 \middle| \mathcal{J}_K\right]\end{aligned}$$

Taking expectations then establishes:

$$\mathbb{E}\left[\left(A_{0K}^{\text{eval}}(\widehat{g})\right)^2\right] \to 0$$

We have therefore established that $A_{0K}^{\text{eval}}(\widehat{g}) \xrightarrow{p} 0$.

Finally, Assumption SG gives $\|\mu(\widehat{g})\|_2 \leq \sqrt{N-1}\bar{g}$, while Theorem 1 applied to the pre-treatment evaluation fold gives $\widehat{w}^{\text{ts,eval}} \xrightarrow{p} w^0$. Applying the Cauchy-Schwarz inequality then yields:

$$\begin{aligned}
|b_{K,\widehat{g}}^{\text{eval}}| &= \left|\mu(\widehat{g})^\top \left(\widehat{w}^{\text{ts,eval}} - w^0\right)\right| \\
&\leq \sqrt{N-1}\,\bar{g}\left\|\widehat{w}^{\text{ts,eval}} - w^0\right\|_2 \\
&\xrightarrow{p} 0
\end{aligned}$$

Thus, all three terms in the decomposition converge to zero in probability, yielding $\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}} \xrightarrow{p} 0$ unconditionally. To establish the conditional conclusion, fix $\epsilon, \eta > 0$. Applying Markov's inequality and the tower property then yields:

$$\begin{aligned}
\Pr\left(\Pr\left(|\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}}| > \epsilon|\mathcal{J}_K\right) > \eta\right) &\leq \frac{1}{\eta}\,\mathbb{E}\left[\Pr\left(|\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}}| > \epsilon|\mathcal{J}_K\right)\right] \\
&= \frac{1}{\eta}\Pr\left(|\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}}| > \epsilon\right) \\
&\to 0
\end{aligned}$$

Consequently, for every $\epsilon > 0$, we have:

$$\Pr\left(|\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}}| > \epsilon|\mathcal{J}_K\right) \xrightarrow{p} 0$$

Finally, by Proposition 1, $\widetilde{P}_0 = P_0$, and so $\widetilde{\tau}_{\widehat{g}} = \tau_{\widehat{g}}$. This completes the proof. □

E.3 Proof of Theorem 4

Proof. For notational brevity, write $n_K = K_{\min}^{\text{post,eval}}$ and $m_K = K_{\min}^{\text{pre,eval}}$. For each unit $i = 1, \ldots, N$, define

$$\Delta_{iK}(\widehat{g}) = \mathbb{E}_{\widehat{F}_{iT}^{\text{eval}}}[\widehat{g}(X)] - \mu_i(\widehat{g})$$

By the same algebra as in the proof of asymptotic normality in Theorem 2, we can write:

$$\sqrt{n_K}\left(\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}}\right) = Z_K(\widehat{g}) - \sqrt{\rho_K^{\text{eval}}}\,\mu(\widehat{g})^\top U_K^{\text{eval}} - R_K(\widehat{g})$$

Where we have defined the following:

$$Z_K(\widehat{g}) = \sqrt{n_K}\left\{\Delta_{NK}(\widehat{g}) - \sum_{i=1}^{N-1} w_i^0 \Delta_{iK}(\widehat{g})\right\}$$

$$U_K^{\text{eval}} = \sqrt{m_K}\left(\widehat{w}^{\text{ts,eval}} - w^0\right)$$

$$R_K(\widehat{g}) = \sqrt{n_K} \sum_{i=1}^{N-1} \left(\widehat{w}_i^{\text{ts,eval}} - w_i^0 \right) \Delta_{iK}(\widehat{g})$$

We will now go term by term, starting with $Z_K(\widehat{g})$. Conditional on $\mathcal{J}_K$, the function $\widehat{g}$ is fixed, and the post-treatment evaluation observations remain independent across cells. As in the proof of asymptotic normality in Theorem 2, we can define the conditional variance of $Z_K(\widehat{g})$:

$$V_{K,\widehat{g}}^{\text{post}} = \frac{n_K}{K_{NT}^{\text{eval}}} \sigma_N^2(\widehat{g}) + \sum_{i=1}^{N-1} (w_i^0)^2 \frac{n_K}{K_{iT}^{\text{eval}}} \sigma_i^2(\widehat{g})$$

Now, define $a \in \mathbb{R}^N$ such that $a_N = 1$ for the treated unit and $a_i = -w_i^0$ for each donor. We can then write:

$$\begin{aligned} Z_K(\widehat{g}) &= \sqrt{n_K} \sum_{i=1}^{N} a_i \Delta_{iK}(\widehat{g}) \\ &= \sum_{i=1}^{N} \sum_{k \in \mathcal{I}_{iT}^{\text{eval}}} \underbrace{\frac{\sqrt{n_K} a_i}{K_{iT}^{\text{eval}}} \left(\widehat{g}(x_{iTk}) - \mu_i(\widehat{g}) \right)}_{\xi_{ik,K}} \end{aligned}$$

Now, observe that Assumption SG states that $|\widehat{g}(x)| \leq \bar{g}$, implying that its population mean satisfies $|\mu_i(\widehat{g})| \leq \bar{g}$. We can then write:

$$|\widehat{g}(x_{iTk}) - \mu_i(\widehat{g})| \leq 2\bar{g}$$

This in turn yields:

$$|\xi_{ik,K}| \leq \frac{2\bar{g}\sqrt{n_K}|a_i|}{K_{iT}^{\text{eval}}}$$

Conditional on $\mathcal{J}_K$, each of these terms is independent and centered, and boundedness also provides $\sigma_i^2(\widehat{g}) \leq \bar{g}^2$. We then have:

$$
\begin{aligned}
V_{K,\widehat{g}}^{\text{post}} &= \sum_{i,k} \mathbb{E}[\xi_{ik,K}^2 \,|\, \mathcal{J}_K] \\
&= \sum_{i=1}^{N} \frac{n_K a_i^2}{K_{iT}^{\text{eval}}} \sigma_i^2(\widehat{g}) \\
&\leq \bar{g}^2 \sum_{i=1}^{N} a_i^2 \\
&= \bar{g}^2 \left(1 + \sum_{i=1}^{N-1} (w_i^0)^2\right) \\
&\leq 2\bar{g}^2
\end{aligned}
$$

The final inequality follows from $w^0 \in \mathcal{W}$. We have therefore shown that the conditional variance is bounded, and so to establish a conditional Gaussian approximation without separately standardizing $Z_K(\widehat{g})$ it suffices to verify the unnormalized Lindeberg condition:

$$
\sum_{i,k} \mathbb{E}\left[\xi_{ik,K}^2 \mathbf{1}\left\{|\xi_{ik,K}| > \epsilon\right\} \middle| \mathcal{J}_K\right] \to 0
$$

Now, note that $|a_i| \leq 1$ and $K_{iT}^{\text{eval}} \geq n_K$ by definition. We can therefore write:

$$
\max_{i,k} |\xi_{ik,K}| \leq \frac{2\bar{g}}{\sqrt{n_K}} \to 0
$$

Therefore, for any $\epsilon > 0$, we have:

$$
n_K > \left(\frac{2\bar{g}}{\epsilon}\right)^2 \implies \mathbf{1}\left\{|\xi_{ik,K}| > \epsilon\right\} = 0
$$

The unnormalized Lindeberg condition therefore holds as $n_K \to \infty$. The Lindeberg-Feller Theorem then establishes a CLT for any sequence of distributions of $Z_K(\widehat{g})$ for which $V_{K,\widehat{g}}^{\text{post}} \to V$, where $V > 0$. First, note that since the variances $V_{K,\widehat{g}}^{\text{post}}$ are bounded, every sequence admits a subsequence along which these variances converge to some $V \geq 0$. If $V > 0$, the Lindeberg-Feller Theorem therefore establishes convergence of these conditional distributions to $\mathcal{N}(0, V)$. If $V = 0$, then Chebyshev's inequality yields:

$$
\Pr\left(|Z_K(\widehat{g})| > \epsilon | \mathcal{J}_K\right) \leq \frac{V_{K,\widehat{g}}^{\text{post}}}{\epsilon^2} \to 0
$$

In this case, the subsequence of distributions of $Z_K(\widehat{g})$ therefore converges to a point mass at 0, which is equivalent to a $\mathcal{N}(0, 0)$ distribution. Thus, along the selected subsequence, both the con-

ditional distributions of $Z_K(\widehat{g})$ and the approximating Gaussian distributions $\mathcal{N}(0, V^{\text{post}}_{K,\widehat{g}})$ converge to $\mathcal{N}(0, V)$. This argument holds for any original sequence, establishing the desired conditional Gaussian approximation, with mean 0 and variance $V^{\text{post}}_{K,\widehat{g}}$.

Let us now move on to the next term in the decomposition. By Lemma 17, conditional on $\mathcal{J}_K$,

$$U_K^{\text{eval}} \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{w,\mathcal{S}_0}\right)$$

in probability with respect to the discovery sample. Because $\mu(\widehat{g})$ is $\mathcal{J}_K$-measurable and uniformly bounded under Assumption SG, and $\rho_K^{\text{eval}} \to \rho^{\text{eval}} < \infty$ by Assumption PEFG, this second term admits a conditional centered Gaussian approximation with the following variance:

$$V^w_{K,\widehat{g}} = \rho_K^{\text{eval}}\, \mu(\widehat{g})^\top \Sigma_{w,\mathcal{S}_0} \mu(\widehat{g})$$

Conditional on $\mathcal{J}_K$, the term $Z_K(\widehat{g})$ depends only on the post-treatment evaluation fold, whereas U_K^{eval} depends only on the pre-treatment evaluation fold. The two terms are therefore conditionally independent. Their difference therefore has a conditional Gaussian approximation with variance:

$$V_{K,\widehat{g}} = V^{\text{post}}_{K,\widehat{g}} + V^w_{K,\widehat{g}}$$

Next, note that Assumption PEFG provides:

$$\frac{n_K}{K_{iT}^{\text{eval}}} \to \frac{1}{\kappa_{iT}^{\text{eval}}} \qquad \rho_K^{\text{eval}} \to \rho^{\text{eval}}$$

Because Assumption SG uniformly bounds $\sigma_N^2(\widehat{g})$, $\sigma_i^2(\widehat{g})$, and $\mu(\widehat{g})$, applying the continuous mapping theorem yields $V_{K,\widehat{g}} - V_{\widehat{g}} \xrightarrow{p} 0$.

It remains to control $R_K(\widehat{g})$. First, note that independence of the observations and fixedness of $\widehat{g}$ (conditional on $\mathcal{J}_K$) yield:

$$\begin{aligned}\mathbb{E}[\Delta_{iK}(\widehat{g})^2 | \mathcal{J}_K] &= \frac{\text{Var}_{\mathcal{F}_{iT}(0)}(\widehat{g}(X))}{K_{iT}^{\text{eval}}} \\ &\leq \frac{\bar{g}^2}{K_{iT}^{\text{eval}}}\end{aligned}$$

Where the inequality follows from boundedness of $\widehat{g}$ from Assumption SG. Now, let $\Delta_K(\widehat{g})$ denote the vector of $\Delta_{iK}(\widehat{g})$:

$$\Delta_K(\widehat{g}) = \left(\Delta_{1K}(\widehat{g}), \ldots, \Delta_{N-1,K}(\widehat{g})\right)^\top$$

We can then write:

$$\begin{aligned}\mathbb{E}\left[\|\Delta_K(\widehat{g})\|_2^2 \,\middle|\, \mathcal{J}_K\right] &= \sum_{i=1}^{N-1} \mathbb{E}\left[\Delta_{iK}(\widehat{g})^2 \,\middle|\, \mathcal{J}_K\right] \\ &\leq \frac{(N-1)\bar{g}^2}{n_K}\end{aligned}$$

Where the inequality comes from the fact that $n_K \leq K_{iT}^{\text{eval}}$ by definition. Applying Markov's inequality then yields:

$$\Pr(\sqrt{n_K}\|\Delta_K(\widehat{g})\|_2 > M \,|\, \mathcal{J}_K) \leq \frac{(N-1)\bar{g}^2}{M^2}$$

Choosing sufficiently large M makes this bound arbitrarily small, establishing:

$$\|\Delta_K(\widehat{g})\|_2 = O_p(n_K^{-1/2})$$

By Lemma 17, we can write:

$$\left\|\widehat{w}^{\text{ts,eval}} - w^0\right\|_2 = O_p\left(m_K^{-1/2}\right)$$

The Cauchy-Schwarz inequality therefore gives

$$\begin{aligned}|R_K(\widehat{g})| &\leq \sqrt{n_K}\left\|\widehat{w}^{\text{ts,eval}} - w^0\right\|_2 \|\Delta_K(\widehat{g})\|_2 \\ &= O_p\left(m_K^{-1/2}\right) \\ &= o_p(1)\end{aligned}$$

To see that this holds conditionally as well, choose a fixed $\epsilon > 0$ and define the following random conditional probability:

$$Q_K = \Pr(|R_K(\widehat{g})| > \epsilon \,|\, \mathcal{J}_K)$$

Applying the tower rule yields:

$$\mathbb{E}[Q_K] = \Pr(|R_K(\widehat{g})| > \epsilon) \to 0$$

Applying Markov's inequality then yields, for every $\eta > 0$:

$$\Pr(Q_K > \eta) \leq \frac{\mathbb{E}[Q_K]}{\eta} = \frac{\Pr(|R_K(\widehat{g})| > \epsilon)}{\eta} \to 0$$

From the definition of Q_K, we then have:

$$\Pr(|R_K(\widehat{g})| > \epsilon \,|\, \mathcal{J}_K) \xrightarrow{p} 0$$

By assumption, $V_{\widehat{g}}$ is bounded away from zero in probability. We therefore have the following two

results:

$$\frac{V_{K,\widehat{g}}}{V_{\widehat{g}}} \xrightarrow{p} 1 \qquad \frac{R_K(\widehat{g})}{\sqrt{V_{\widehat{g}}}} \xrightarrow{p} 0$$

with the latter convergence also holding conditionally in probability. Applying Slutsky's theorem, conditional on $\mathcal{J}_K$, now yields:

$$\frac{\sqrt{K_{\min}^{\text{post,eval}}}\left(\widehat{\tau}_{\widehat{g}}^{\text{eval}} - \widetilde{\tau}_{\widehat{g}}\right)}{\sqrt{V_{\widehat{g}}}} \xrightarrow{d} \mathcal{N}(0,1)$$

in probability with respect to the discovery sample. By Proposition 1, $\widetilde{P}_0 = P_0$ and so $\widetilde{\tau}_{\widehat{g}} = \tau_{\widehat{g}}$ by definition, completing the proof. □

E.4 Proof of Corollary 3

Proof. Assumptions DS and SG imply $\widehat{\sigma}_N^2(\widehat{g}) \xrightarrow{p} \sigma_N^2(\widehat{g})$. Since the number of donors is finite, this holds jointly:

$$\max_{1 \le i \le N-1} \left|\widehat{\sigma}_i^2(\widehat{g}) - \sigma_i^2(\widehat{g})\right| \xrightarrow{p} 0$$

Meanwhile, Theorem 1 applied to the evaluation fold implies $\widehat{w}^{\text{ts,eval}} \xrightarrow{p} w^0$. Together with Assumptions BSG and PEFG, applying the continuous mapping theorem then yields:

$$K_{\min}^{\text{post,eval}} \left\{ \frac{\widehat{\sigma}_N^2(\widehat{g})}{K_{NT}^{\text{eval}}} + \sum_{i=1}^{N-1} (\widehat{w}_i^{\text{ts,eval}})^2 \frac{\widehat{\sigma}_i^2(\widehat{g})}{K_{iT}^{\text{eval}}} \right\} - \left[\frac{\sigma_N^2(\widehat{g})}{\kappa_{NT}^{\text{eval}}} + \sum_{i=1}^{N-1} (w_i^0)^2 \frac{\sigma_i^2(\widehat{g})}{\kappa_{iT}^{\text{eval}}} \right] \xrightarrow{p} 0$$

Similarly, the post-treatment laws of large numbers imply $\widehat{\mu}(\widehat{g}) - \mu(\widehat{g}) \xrightarrow{p} 0$, while Corollary 1 establishes $\widehat{\Sigma}_{w,\widehat{\mathcal{S}}}^{\text{eval}} \xrightarrow{p} \Sigma_{w,\mathcal{S}_0}$, where $\widehat{\Sigma}_{w,\widehat{\mathcal{S}}}^{\text{eval}}$ is now estimated using only the pre-treatment evaluation fold. Assumption PEFG gives $\rho_K^{\text{eval}} \to \rho^{\text{eval}}$, and hence another application of the continuous mapping theorem yields:

$$\rho_K^{\text{eval}} \widehat{\mu}(\widehat{g})^\top \widehat{\Sigma}_{w,\widehat{\mathcal{S}}}^{\text{eval}} \widehat{\mu}(\widehat{g}) - \rho^{\text{eval}}\, \mu(\widehat{g})^\top \Sigma_{w,\mathcal{S}_0} \mu(\widehat{g}) \xrightarrow{p} 0$$

Combining these two results establishes $\widehat{V}_{\widehat{g}} - V_{\widehat{g}} \xrightarrow{p} 0$, as desired. □

F Proofs for Section 5

F.1 Proof of Proposition 7

Proof. By the triangle inequality, we can then write:

$$|\widehat{\tau}_g - \tau_g| \le |\widehat{\tau}_g - \widetilde{\tau}_g| + |\widetilde{\tau}_g - \tau_g|$$

Furthermore, by Theorem 1, we have $\widehat{w}^{\text{ts}} \xrightarrow{p} w^0$. By the proof of consistency in Theorem 2, we then have $|\widehat{\tau}_g - \widetilde{\tau}_g| = o_p(1)$. By definition, we can write:

$$\widetilde{\tau}_g - \tau_g = \int_{\mathcal{X}} g(x)\, d\Big\{P_0 - \widetilde{P}_0\Big\}(x)$$

Adding and subtracting $h(\mathcal{E}(x))$ and applying the triangle inequality yields:

$$|\widetilde{\tau}_g - \tau_g| \leq \left|\int_{\mathcal{X}} h(\mathcal{E}(x))\, d\Big\{P_0 - \widetilde{P}_0\Big\}(x)\right| + \left|\int_{\mathcal{X}} (g - h\circ\mathcal{E})(x)\, dP_0(x)\right| + \left|\int_{\mathcal{X}} (g - h\circ\mathcal{E})(x)\, d\widetilde{P}_0(x)\right|$$

By assumption, the last two terms are each bounded by δ. The reproducing property and Cauchy-Schwarz applied in the RKHS for the first term give:

$$\begin{aligned}\left|\Big\langle h,\, \mu_{P_0} - \mu_{\widetilde{P}_0}\Big\rangle_{\mathcal{H}_k}\right| &\leq \|h\|_{\mathcal{H}_k}\left\|\mu_{P_0} - \mu_{\widetilde{P}_0}\right\|_{\mathcal{H}_k} \\ &= \|h\|_{\mathcal{H}_k}\text{MMD}_k\Big(\widetilde{P}_0, P_0\Big)\end{aligned}$$

Collecting these results yields:

$$|\widehat{\tau}_g - \tau_g| \leq \|h\|_{\mathcal{H}_k}\text{MMD}_k\Big(\widetilde{P}_0, P_0\Big) + 2\delta + o_p(1)$$

This establishes the result. □

F.2 Proof of Proposition 8

Proof. Conditional on $\mathcal{J}_K$, $\widehat{g}$ and h are fixed. The following population inequalities therefore hold almost surely, and all $o_p(1)$ statements are understood conditionally on $\mathcal{J}_K$. By the proof of Theorem 3, we have $|\widehat{\tau}^{\text{eval}}_{\widehat{g}} - \widetilde{\tau}_{\widehat{g}}| = o_p(1)$. By definition, we can write

$$\widetilde{\tau}_{\widehat{g}} - \tau_{\widehat{g}} = \int_{\mathcal{X}} \widehat{g}(x)\, d\Big\{P_0 - \widetilde{P}_0\Big\}(x)$$

Adding and subtracting $h\circ\mathcal{E}$ and applying the triangle inequality to the previous integral expression gives:

$$|\widetilde{\tau}_{\widehat{g}} - \tau_{\widehat{g}}| \leq \left|\int_{\mathcal{X}} h(\mathcal{E}(x))\, d\Big\{P_0 - \widetilde{P}_0\Big\}(x)\right| + \left|\int_{\mathcal{X}} (\widehat{g} - h\circ\mathcal{E})(x)\, dP_0(x)\right| + \left|\int_{\mathcal{X}} (\widehat{g} - h\circ\mathcal{E})(x)\, d\widetilde{P}_0(x)\right|$$

Since both P_0 and $\widetilde{P}_0$ are probability measures and $\|\widehat{g} - h\circ\mathcal{E}\|_\infty \leq \delta$ almost surely, the final two terms are each bounded by δ almost surely. Meanwhile, the first term is bounded by the RKHS

inequality in the proof of Proposition 7. Together, this yields:

$$\left|\widetilde{\tau}_{\widehat{g}} - \tau_{\widehat{g}}\right| \leq \|h\|_{\mathcal{H}_k} \mathrm{MMD}_k(\widetilde{P}_0, P_0) + 2\delta$$

By the triangle inequality, we can then write:

$$\begin{aligned} |\widehat{\tau}^{\mathrm{eval}}_{\widehat{g}} - \tau_{\widehat{g}}| &\leq |\widehat{\tau}^{\mathrm{eval}}_{\widehat{g}} - \widetilde{\tau}_{\widehat{g}}| + |\widetilde{\tau}_{\widehat{g}} - \tau_{\widehat{g}}| \\ &\leq o_p(1) + \|h\|_{\mathcal{H}_k} \mathrm{MMD}_k(\widetilde{P}_0, P_0) + 2\delta \end{aligned}$$

This establishes the result. □