THE UNIVERSITY OF CHICAGO

# ESSAYS IN ECONOMETRICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
CONNOR DOWD

CHICAGO, ILLINOIS
JUNE 2021

For my wife, and her unbounded patience

# CONTENTS

*

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I am honored to have an excellent group of advisors, Chris Hansen, Max Farrell, Panos Toulis, and Constantine Yannelis. I have benefited enormously from the generous support, advice, and encouragement they've offered. I'm incredibly lucky to have learned from them. Each of them has had a defining influence on my graduate school experience.

I have also benefited immensely from thoughtful discussions with seminar participants in numerous settings, as well as other faculty at this and other Universities. Seth Zimmerman, Eric Zwick, Michael Greenstone, Neale Mahoney, Nick Polson, Canice Prendergast, Ruey Tsay, and many others all helped me find my feet as a researcher. The UChicago Econometrics lunch group, the Tuesday working group, and other sources of feedback have been invaluable. The support of the UChicago Booth PhD program office has been tremendous – Malaina Brown, Cynthia Hillman, Kim Mayer, Amity James, and Ethan Simmonds all deserve medals for their hard work.

Without my fellow UChicago graduate students, there is no chance I would be completing this degree. My office mates provided incredible support – Peter Chen, Robbie Sanders, Xiao Zhang, and others – I appreciate it. Olivia Natan, Sam Hirshman, Uyen Tran, Vera Chau, Jessica Lopez, Jianfei Cao, Sondre Skarsten, Maxim Babush, Yewon Kim, Tony Ditta, and many more all helped in numerous ways, and I will never forget it.

I particularly thank my family – my parents, my brother, and numerous others. Their support during the last few years has been immense, but of course their support long before I was a PhD student was in many ways much more critical.

Finally, I deeply appreciate my lovely wife, Zoë. It is her love, sacrifice, and unconditional support in all things that made this dissertation possible.

Thank you all.

# ABSTRACT

This dissertation consists of three essays. The first essay focuses on regression discontinuity with a donut, the second essay looks at spillovers in synthetic controls, and the third essay examines a new two-sample test.

Regression discontinuity (RD) designs use policy thresholds to identify the causal effects of policy. RD Donut designs allow identification in situations with some manipulation, but they require extrapolation, typically projecting a polynomial, to identify treatment effects. Chapter 1 extrapolates into a donut by leveraging high-level smoothness conditions similar to those used to find optimal bandwidths. I start using known derivative bounds before using data-determined bounds.

The synthetic control method (SCM) is often used in treatment effect estimation with panel data where only a few units are treated and a small number of post-treatment periods are available. Current estimation and inference procedures for SCM do not allow for the existence of spillover effects. In a related paper [Cao and Dowd, 2021], we consider estimation and inference for SCM, allowing for spillover effects. In Chapter 2 we show simulations and empirical examples of this method.

Empirical cumulative distribution functions are used to test the hypothesis that two samples come from the same distribution. Chapter 3 describes a statistic that is usable under the same conditions as the Kolmogorov-Smirnov test, but provides more power than other extant tests in that vein. I prove the validity of the procedure, provide code, and show several simulations demonstrating substantial power.

# INTRODUCTION

Modern econometrics works to study quantitative methods by which we can perform social science. An important vein of work in this field struggles to assess the causal consequences of some action or policy using only observational data. There are many challenges in using observational data for determining causality.

In this essay, I detail two different common empirical approaches, identify fundamental challenges to assessing causality with those approaches, and provide guidance for resolving those challenges, allowing us to move forwards with the study of our fellow humans and ourselves. These techniques frequently rely on detecting a change in average outcomes, rather than some other distributional statistic. The third part of this essay focuses on computationally quick, and statistically powerful methods for detecting other changes to the distribution than purely mean shifts.

The first, and longest, chapter of this essay focuses on the regression discontinuity donut. Broadly, regression discontinuity leverages a threshold in some policy, allowing us to compare individuals who fall just on either side of the threshold, in order to understand the effects of the policy. However, this approach is susceptible to substantial bias when individuals who are aware of the policy threshold engage in manipulation to control their outcomes. A common response to this identification threat is to use a "Donut", dropping all observations in some region of the threshold. This has the undesirable property that we are removing the very observations that were most useful for identifying the parameter of interest. Chapter 1 works to

identify solutions for the implicit extrapolation problem that results. At a high level, the solution approach taken relies on some notion of smoothness in underlying mean functions.

The second chapter of this essay, coauthored with Jianfei Cao, addresses a problem arising with Synthetic Controls. Synthetic controls take advantage of a situation where there are many observations and time periods, and only some units receive policy treatment in some periods. The 'synthetic control' is built by finding the best weights for predicting the treated units in the pre-treatment periods. Relative to other naïve estimators which use equal weights, this can result in improved efficiency and a reduced burden of assumptions. However, if the policy of interest has effects that extend beyond the units that recieve treatment, we can wind up with a large asymptotic bias which doesn't shrink. An earlier paper has theoretical results giving an estimator which can adapt to this problem. Chapter 2 shows simulation results and empirical examples using that estimator, providing anecdotal evidence both of ease of use and of the actual improvements possible, even in datasets where original authors felt they had dealt with any potential spillover problems.

The final chapter addresses a problem I found myself with frequently. There are a number of ways to compare data from two different groups, but statistically and computationally efficient comparisons which can detect any difference between two distributions are rare. Chapter 3 provides details and intuition for a two-sample test statistic which is publicly available on CRAN. That test statistic builds on two well known test statistics – the Anderson-Darling statistic and the Wasserstein statistic,

by combining the best insights of each. Chapter 3 also provides proofs of consistency and validity, as well as showing simulations demonstrating substantial power relative to well known alternative test statistics in the same class.

# CHAPTER 1

# REGRESSION DISCONTINUITY: DONUTS AND EXTRAPOLATION

## 1.1 Abstract

Regression discontinuity (RD) designs use policy thresholds to identify the causal effects of policy. Under standard conditions, these effects are identified solely at the policy threshold. In many settings, treatment effects at other points may also be of interest. Conversely, RD donut designs require extrapolation, typically projecting a polynomial, to identify treatment effects anywhere. This paper extrapolates across the threshold or into a donut by leveraging high-level smoothness conditions similar to those used to find optimal bandwidths. The paper starts by using known derivative bounds before moving onto to learning those bounds from the data. By constraining the derivatives, we bound the Taylor expansion, which lets us identify a set containing the treatment effect and conduct inference. Shape restrictions derived from underlying economic logic, such as monotonicity or convexity, are excellent candidates for pre-specified derivative bounds. Under the assumption that we observe the maximum curvature, we can do away with the a priori bounds to provide more data-driven results. Leveraging recent results in Cattaneo et al. [2018], we conduct conservative inference for the maximum of a derivative, and subsequently for the treatment effects. This routine requires much weaker assumptions about the mean

functions relative to standard assumptions in the literature for Donut designs.

## 1.2 Introduction

We study extrapolation of treatment effects in regression discontinuity designs. The standard sharp regression discontinuity (RD) design can nonparametrically identify treatment effects for units local to the treatment threshold. However, policy makers may care about treatment effects for units away from the policy threshold. Moreover, there are applications, like "donut" designs, which drop all units local to the threshold, where nonparametric identification is not possible even at the threshold. This paper leverages global smoothness conditions to provide identification and inferential guarantees in these situations.

The distinguishing feature of RD designs is that weak smoothness conditions identify the treatment effect at the cutoff. Under the standard sharp RD design, our focus in this paper, the researcher observes an outcome of interest, $Y$, and a running variable, $X$. Units for which $X$ exceeds some known threshold, $c$, receive treatment, and those with $X < c$ do not recieve treatment. If the conditional expectations of $Y$ given $X$ are continuous on both sides of the threshold, the average effect of the policy on the outcome, for units at the threshold, may be causally identified.

Identifying causal effects at other values of $X$ requires additional assumptions. Our approach is to make the weakest possible assumptions consistent with performing

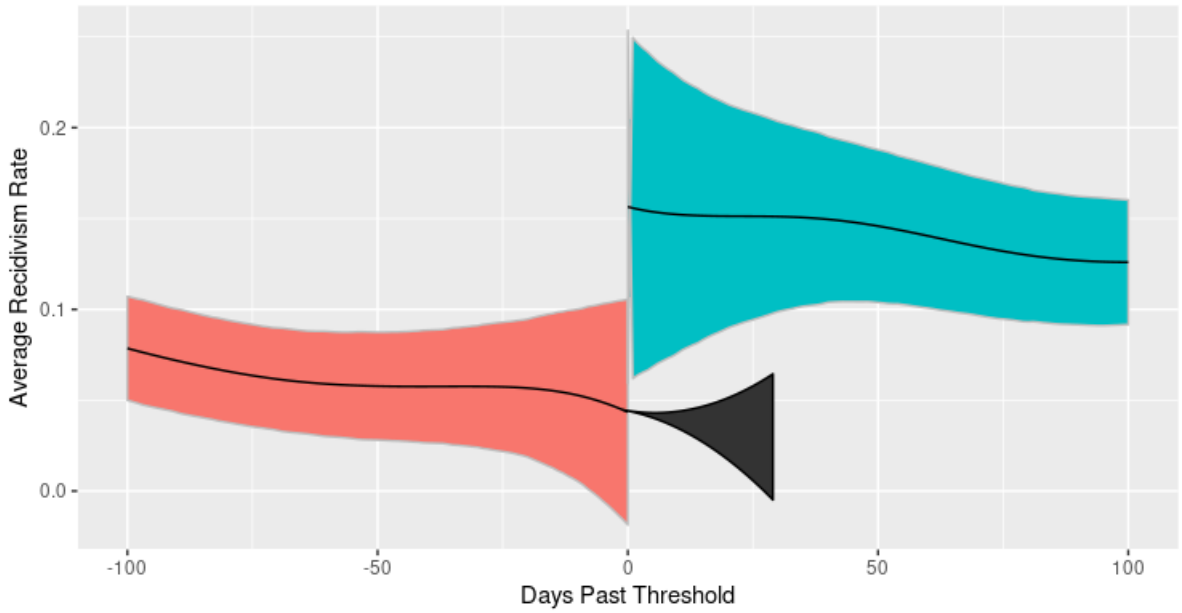


Figure 1.1: Recidivism Rate based on time of first offense relative to threshold. Local polynomials are used to estimate the confidence regions and point estimates. The dark region contains $\mu_0(x) = E[Y|X, T = 0]$ when that function has second derivatives between -0.0004 and 0.0006, which are numbers estimated from the data. Building a confidence region on top of that requires accommodating substantial uncertainty in both those estimates and the lower derivatives of the function.

estimation and inference – which will lead to partial identification of the parameter of interest. By using bounds on a given higher-order derivative of the conditional means, which are known a priori or estimated, we can bound the error terms in a Taylor expansion, allowing us to construct identified sets. Figure 1.1 shows an example of how bounding the second derivative leads to a range of possible values for the $E[Y|X, T = 0]$, and a corresponding range of possible treatment effects. In order to produce point estimates for the treatment effects, we would need to make much stronger assumptions. For instance, we could assume that the second derivative is 0 between the threshold and our point of interest – which would correspond to

making a linear projection from the threshold to any point we are interested in. This paper then shows how to construct those identified sets, as well as how to conduct inference on the sets containing the parameter of interest. We also show how the identified set responds to the strength of the assumptions the researcher imposes.

To build an identified set, we use regions of the support where we can obtain point identification of the conditional mean functions and their derivatives, as seen in figure 1.1. We extrapolate from those points to the desired parameter by bounding the derivative terms in a Taylor expansion of $E[Y|X, T = 0]$. We consider several approaches for bounding those derivatives, which are derivatives of the conditional expectation functions. We allow for bounding an arbitrary derivative $k$, and show how any given bound produces an identified set. The simplest bounds involve prior information available to the researcher – for instance about the convexity or concavity of the underlying function, or known limits on the curvature. We also consider estimating bounds for the derivatives under a global smoothness condition. Leveraging recent results of Cattaneo, Farrell, and Feng [2018], we can use the data to estimate bounds on this derivative using data away from the cutoff, allowing us to identify the treatment effects in a fully data-driven manner. The general procedure is described below:

*Outline of Estimation Procedure*

1. Set a confidence level $\alpha$ and a point of interest, $x_0$
2. Find a set $\mathbb{C}$ that contains the $k$th derivative of $\mu_1$ with probability at least $1 - \kappa > 1 - \alpha$

3. Estimate the first $k-1$ derivatives for $\mu_1$ at the threshold $c$.
4. Estimate the value of $\mu_0$ at the point of interest, $x_0$.
5. Estimate $\mu_1$ at $x_0$, using its first $k-1$ derivatives and a Taylor projection.
6. Estimate $\tau(x_0) = \mu_1(x_0) - \mu_0(x_0)$ and build a $1-\alpha+\kappa$ CI for this projection.

   This is a valid CI for the treatment effect if $\mu_1^{(k)}(x) = 0 \ \ \forall x \in (x_0, c)$.
7. Use the extreme values of $\mathbb{C}$ to find the maximal errors in the Taylor projection above.
8. Add the maximal errors to the $1-\alpha+\kappa$ CI for $\tau$.

   This new range is a conservative $1-\alpha$ CI for a region containing the treatment effect.

If we had precise estimates for the entire infinite sequence of derivatives,[1] we could make precise projections of the mean functions across the threshold to any point, identifying the treatment effect at any point. In the absence of that detailed information, we can make projections based on $k-1$ derivatives, but these projections will have large errors. By bounding the value of the $k$th derivative, we can bound the size of those projection errors by assuming that the $k$th derivative is immediately and forever at its bounds. Any point estimate outside the bounds created by that assumption would require that the $k$th derivative be outside its bounds at least briefly. This turns a problem of projection into a problem of bounding the derivatives of a mean function – which is also known as a smoothness condition. This paper

1. Along with a few technical conditions like the function being analytic.

will require that there are some bounds on a derivative, which are either known, or estimable somehow. Our primary technique will be to assume that we observe a maximal value for the $k$th derivative of the mean function, somewhere in the range of the data.

Because this approach relies solely on information which is identified in the support of the data, to obtain information outside the support of the data, we can generalize the approach. This allows us to study the "donut" design. Donut designs are a modification of sharp RD which attempt to deal with selection issues by dropping observations in some radius of the threshold. The radius around the threshold is dictated by the ability of units to manipulate their observed $X$ value, and thus decide on their treatment status. In an extension of the approach described above, we use the extrapolation techniques to identify the treatment effect at the threshold, after dropping all observations nearby – as is done in donut designs.

To fix ideas, we will revisit the recent application of Tuttle [2019]. Here the running variable is the days after new SNAP policy is implemented, while the outcome of interest is the recidivism behavior of convicted drug dealers. Broadly, convictions for drug dealing after day 0[2] lead to a lifetime ban for SNAP benefits in Florida. We are interested in the effect of the lifetime SNAP ban on recidivism for individuals who are caught after the threshold. The original paper estimates the effect for individuals on day 0, but policy makers may care substantially about the effect people beyond the threshold. In Figure 1.1, we see the mean functions before and after treatment.

2. The drug dealing must have been after day 0, not the conviction.

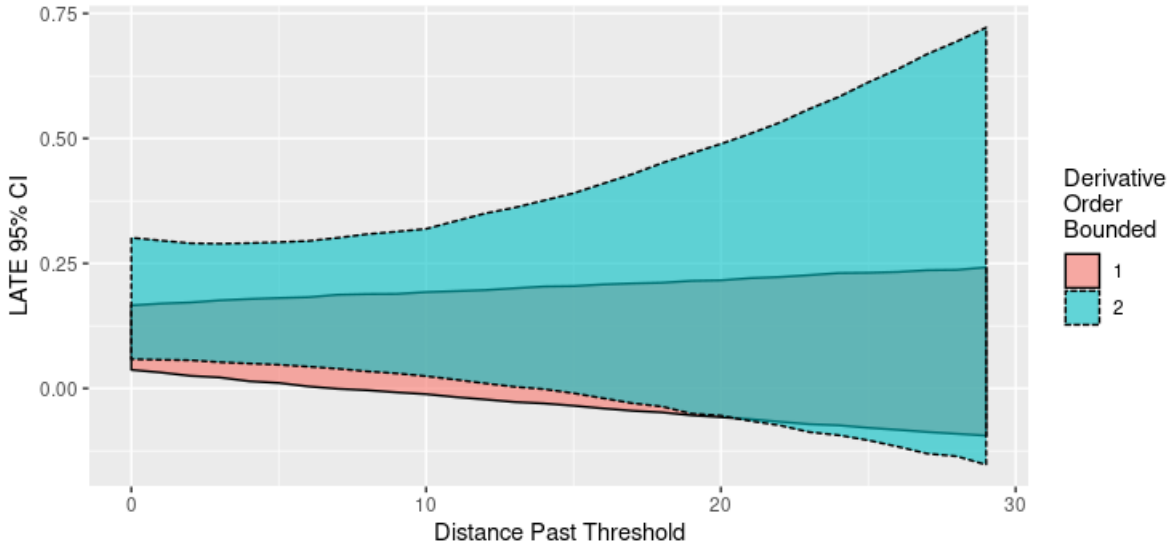


Figure 1.2: Local Average Treatment Effects of lifetime SNAP bans on Recidivism as we move away from the treatment threshold. Shown here are two different assumptions about structure of underlying mean functions, either that the first or second derivatives are globally bounded.

That same figure also shows an example taylor projection of one mean function for the control group using a second derivative bound. The area encompassed by the dark region is the identified set – not the confidence interval. The difference between that set and the point estimate for the treated group above it gives our estimate of the LATE for these regions beyond the threshold.

See Figure 1.2 to see the estimated 95% confidence intervals arising from the assumption of global bounds on either the first or second derivatives. This is the result of actually calculating the difference between the two sets and incorporating our uncertainty around them. Note that neither interval is a subset of the other – a clear demonstration that neither assumption is strictly weaker or stronger than the other.

Depending on the DGP, and point of interest, either set can encompass the other.

There is some prior work extrapolating RD treatment effects away from the threshold, differing substantially in the type and strength of assumptions used for idenfication. Angrist and Rokkanen [2012] assume that the researcher has access to additional variables, and that potential outcomes are mean independent of the running variable, given these additional covariates. Modelling the conditional expectation of outcomes given those covariates, they are able to identify causal effects for other values of the running variable. Our assumptions are much weaker, and we thus obtain weaker identification results, while encompassing more general situations. Dong and Lewbel [2015] use estimated derivatives, as we do, to consider causal effects of small changes to the position of the threshold, which is not our focus.

We work within the standard RD framework which uses nonparametric smoothness assumptions to identify the causal effects of interest [Hahn, Todd, and van der Klaauw, 2001]. The RD literature is large and still expanding: for a recent review, and numerous other references, see Cattaneo, Idrobo, and Titiunik [2019a]. For a comparison of the continuity-based and other approaches to RD identification and estimation, see Cattaneo, Titiunik, and Vazquez-Bare [2017].

This paper is organized as follows: Section 2 provides notation and necessary conditions. Section 3 details the main results. Section 4 looks at an example in the world of recidivism and SNAP benefits to examine the relative strengths of assumptions bounding different derivative orders. Section 5 discusses a natural extension to the

world of RD donuts. Section 6 concludes.

## 1.3 Framework

For each individual, the econometrician observes a variable $X$, known as the running variable or forcing variable, which has a compact domain $\chi \subset \mathbb{R}$. There is also a known threshold, $c \in \chi$, such that treatment status $T = \mathbb{1}[x \geq c]$. Without loss of generality, we assume that $c = 0$. We also observe another variable, $Y$, referred to as the outcome. We will focus on the standard heteroscedastic non-parametric framework.

$$Y = \mu_T(X) + \epsilon \qquad \mathbb{E}[\epsilon] = 0 \qquad Var(\epsilon) = \sigma_T^2(X) \tag{1.1}$$

Where $\mu_t$ and $\sigma_t$ are defined as:

$$\mu_t(x) = \mathbb{E}[Y|X = x, T = t] \qquad \sigma_t^2(x) = Var(Y|X = x, T = t)$$

The parameter of interest is the local average treatment effect (LATE) on the outcome variable at some known point $x_0$. Thus, our parameter of interest is:

$$\tau = \tau(x_0) = \mu_1(x_0) - \mu_0(x_0) = \mathbb{E}[Y|X = x_0, T = 1] - \mathbb{E}[Y|X = x_0, T = 0] \tag{1.2}$$

Most RD papers require that the point at which we evaluate the LATE, $x_0$, coincide with the treatment threshold, $c$. The restriction that $x_0 = c$ simplifies the problem

faced by the econometrician substantially, as in principle, $\mu_1(c)$ and $\mu_0(c)$ are both nonparametrically identified by the data. Notably, it is rare that the point of interest for policymakers is actually $c$.

For notational simplicity, I will assume throughout that the point of interest, $x_0 < c$. This will allow me to denote the mean function being projected as $\mu_1$. This condition is by no means necessary, and will be relaxed in the section on donuts.

In order to learn about the LATE at points less than the threshold, we need to predict $Y$ in the counterfactual where those individuals were treated. Specifically, we need the ability to conduct inference for $\widehat{\mu_1(x_0)}$, which under standard RD assumptions is not possible. Then we can use our information about $\widehat{\mu_1(x_0)}$ to estimate the treatment effect using $\hat{\mu}_0(x_0)$. Learning $\widehat{\mu_1(x_0)}$ is not simple, so this paper finds weak assumptions that will bound the possible values of $\widehat{\mu_1(x_0)}$.

A simple fix to this would be to assume that the function $\mu_1$ is an order $k$ polynomial or some other known parametric function. Under that assumption, the projection becomes quite simple. We can use standard regression confidence intervals, projected over to the point $x_0$. However, this is not typically a reasonable assumption. Over the past 15 years a large literature developed examining the behavior of nonparametric estimators for the LATE in RD settings. This literature is the direct result of overly strong parametric RD estimates which frequently relied on polynomial regressions.

Non-parametric RD does not make strong enough assumptions to identify $\widehat{\mu_1(x_0)}$ when $x_0 \neq c$, but that literature frequently makes strong assumptions in order to

select the optimal bandwidth. In nonparametric estimators like local polynomial regressions, bandwidths balance a trade-off between the curvature of the underlying mean functions and the variance of the errors. As the curvature increases and the mean function becomes less smooth, relatively distant observations become less informative, and so dropping them by shrinking the bandwidth is sensible. At the same time, as the error variance increases, nearby observations become noisier and less informative, and so increasing our effective sample size by expanding the bandwidth becomes attractive.

Data-driven bandwidth choice requires making some assumption about the maximal extent of the curvature and the maximum variance, thus bounding the worst case outcome. Those conditions have a tendency to look like placing an upper bound on the $k$th derivative of the mean function. This paper will strengthen and extend those conditions. Broadly, I will require an assumption of the form:

**Condition 1** (Bounded $k$th Derivative)**.** For each $t = \{0, 1\}$, for some $k > 1$,

$$\partial_{L,t}^{(k)} \;\leq\; \mu_t^{(k)}(x) \;\leq\; \partial_{U,t}^{(k)} \quad \forall X \in \chi \tag{1.3}$$

where $\mu_t^{(k)}$ indicates the $k$th derivative of the function $\mu_t$, $\chi$ continues to represent the domain of the running variable, and $\partial_{L,t}^{(k)}$ & $\partial_{U,t}^{(k)}$ represent upper and lower bounds. In some settings, researchers may be able to use domain knowledge to state reasonable bounds $\partial_{L,t}$. Results in this paper will show the validity of that approach. Perhaps more frequently, researchers can make the additional assumption that $\mu_t^{(k)}(x)$ attains its extreme values in the region where we observe $T = t$. Under that condition, this

paper will show results for data driven methods to estimate the treatment effect. Bounds of the form in equation 1.3 are strictly weaker claims than requiring that $\mu_t$ be a polynomial of degree $k$, as that would imply that for some constant $C$, $\mu_t^{(k)}(x) = C \quad \forall x \in \chi$.

In order to make use of the bounds above, we need to recall the Taylor projection for a function.

$$\mathsf{P}_\infty(\mu_t, x) = \sum_{j=0}^{\infty} \frac{\partial^j \mu_t}{\partial x^j} \frac{(x-c)^j}{j!}$$

Notably, once we have finite bounds on the $k$th derivative and we know $c = 0$ we can simplify this somewhat to the Taylor projection below:

$$\mathsf{P}(\mu_t, x_0, \partial_{L,t}^{(k)}, \partial_{U,t}^{(k)}) = \sum_{j=0}^{k-1} \frac{\partial^j \mu_t}{\partial x^j} \frac{x_0^j}{j!} + \begin{pmatrix} \partial_{U,t}^{(k)} \\ \partial_{L,t}^{(k)} \end{pmatrix} \frac{x_0^k}{k!} = \begin{pmatrix} \mu_t(x_0)_U \\ \mu_t(x_0)_L \end{pmatrix} = \Phi_t \tag{1.4}$$

The vector created by that projection defines an interval, which Lemma 1 shows will contain the true $\mu_t(x_0)$. That interval is the identified set, which I will refer to as $\Phi$. In order to make the projection feasible however, we will have to estimate or know the first $k-1$ derivatives, as well as the two bounds on the $k$th derivative. The rest of this section will examine the conditions under which we show that estimating those derivatives to build an interval works.

**Condition 2** (Regularity Conditions)**.** Technical conditions on the DGP in order for the results below.

(i) (X,Y,T) are i.i.d. observations from a d.g.p. satisfying Eq (1.1)

(ii) $\mu_t(\cdot)$ has $k+2$ continuous derivatives

(iii) The density of the running variable, $f_x$ is absolutely continuous and bounded away from 0 over $\chi$.

(iv) The kernel function $K(x) = 0.5\mathbb{1}\left[|x| < 1\right]$.

(v) $\sigma_t(\cdot)$ is positive, bounded above, bounded away from 0, and has two continuous derivatives.

(vi) $\underset{x\in\chi}{sup}\ \mathbb{E}\left[|\epsilon_i|^3 exp(|\epsilon_i|)|x_i\right] = x < \infty$ which implies $\mathbb{E}\left[|\epsilon_i|^3 exp(|\epsilon_i|)\right] < \infty$.

(vii) There is no other treatment policy with a discontinuity in $\chi$ which affects $Y$.

Condition 1 and Condition 2(i,ii) are sufficient for us to establish that the taylor projection described in equation 1.4 is valid and will contain the true value of $\mu_1(x_0)$. Condition 2 parts (iii), (iv), and (v) are closely related to standard conditions for asymptotic normality of local polynomial estimates.[Fan et al., 1995] This makes them sufficient (with some mild rate conditions) for us to be able to take a known set of bounds on the $k$th derivative from equation 1.3 and make the projections feasible. At this point we could build a confidence region for $\mu_1(x_0)$ which is asymptotically valid for the true region. At the same time these conditions allow us to perform inference for $\mu_0(x_0)$. As the these two procedures use independent pieces of data, it is simple to construct a valid confidence region for $\tau$.

**Condition 3** (Rate Conditions)**.** For local polynomial estimates of derivatives to be asymptotically normal I will require $h_p(n) \to 0$ such that as $n \to \infty$:

(i) $nh_p^3 \to \infty$

(ii) $nh_p^{2k+3} \to 0$

Further, if we would like to estimate a global bound on the derivatives using b-splines I need the following conditions on a potentially different bandwidth, $h_b$:

(iii) $\frac{log(n)^{3/2}}{\sqrt{nh_b}} = o_{\mathbb{P}}(1/log(n))$

(iv) $\frac{log(n)^4}{nh_b} = o(1/log(n))$

(v) $nh_b^{1+2k} = o(1/log(n))$

The top conditions are sufficient for asymptotic normality of local polynomial regressions.Fan et al. [1995] The bottom half of these rate conditions will be necessary for us to get a valid uniform confidence band on the $k$th derivative. For that to be useful, we need the following condition to hold.

**Condition 4** (Derivative bounds are observed)**.** Recall that $c = 0$ and $T = \mathbb{1}[x \geq 0]$. Define $C_1$,...,$C_4$ as follows

$$\sup_{x>0 \in \chi} \frac{\partial^k \mu_1(x)}{\partial x^k} = C_1 \qquad \inf_{x>0 \in \chi} \frac{\partial^k \mu_1(x)}{\partial x^k} = C_2$$

$$\sup_{x<0 \in \chi} \frac{\partial^k \mu_0(x)}{\partial x^k} = C_3 \qquad \inf_{x<0 \in \chi} \frac{\partial^k \mu_0(x)}{\partial x^k} = C_4$$

Then, for known, continuous, weakly monotonic functions $f_1, ..., f_4$

$$\partial_{L,1}^{(k)} = f_1(C_1, C_2) \qquad \partial_{U,1}^{(k)} = f_2(C_1, C_2)$$
$$\partial_{L,0}^{(k)} = f_3(C_3, C_4) \qquad \partial_{U,0}^{(k)} = f_4(C_3, C_4)$$

Broadly, condition 4 says that we observe values which are known functions of the bounds in condition (1). In practice we will usually take these functions to be the identities. The generality allows for the inf and sup to be absolute values of the biggest observed derivative, as well as allowing for other situations – e.g. we know some maximal bound, but may wish to use the data-driven results below to tighten the bounds if possible.

## 1.4 Main Results

### *1.4.1 $\Phi$ Contains $\mu_t(x_0)$*

To prove that the set $\Phi$ contains the true value of the mean function, we will show that the approximation error of a Taylor projection using $k-1$ derivatives, when the $k$th derivative is bounded, are at most the projection of the bounds of the $k$th derivative.

**Lemma 1** ($\Phi$ contains the true value of $\mu_t(x_0)$)**.** *The projection error of a* $k-1$

*derivative Taylor projection is bounded by the Taylor projection of the kth derivative's bounds.*

*Proof.* $\mathsf{P}_\infty(\mu_t,x)-\mathsf{P}_{k-1}(\mu_t,x) = \sum_{j=0}^{\infty}\frac{\partial^j\mu_t}{\partial x^j}\frac{(x-c)^j}{j!}-\sum_{j=0}^{k-1}\frac{\partial^j\mu_t}{\partial x^j}\frac{(x-c)^j}{j!} = \sum_{j=k}^{\infty}\frac{\partial^j\mu_t}{\partial x^j}\frac{(x-c)^j}{j!}$
If $\mu_t^{(k)}(x) \leq \partial_{U,t}^{(k)} \forall X \in \chi$, then $\sum_{j=k}^{\infty}\frac{\partial^j\mu_t}{\partial x^j}\frac{(x-c)^j}{j!} \leq \partial_{U,t}^{(k)}\frac{(x-c)^k}{k!}$.
If $\mu_t^{(k)}(x) \geq \partial_{L,t}^{(k)} \forall X \in \chi$, then $\sum_{j=k}^{\infty}\frac{\partial^j\mu_t}{\partial x^j}\frac{(x-c)^j}{j!} \geq \partial_{L,t}^{(k)}\frac{(x-c)^k}{k!}$.
Thus $\partial_{L,t}^{(k)}\frac{(x-c)^k}{k!} \leq \mathsf{P}_\infty(\mu_t,x) - \mathsf{P}_{k-1}(\mu_t,x) \leq \partial_{U,t}^{(k)}\frac{(x-c)^k}{k!}$ □

### *1.4.2 Results for* $\Phi$

In order to actually estimate the values $C_1, ..., C_4$, much less perform inference on functions of them, we will need to rely on the results in Cattaneo et al. [2018]. If we use b-splines with equally sized partitions to estimate the $k$th derivative, that paper tells us that we can construct uniform confidence intervals for that derivative. Specifically we can find a $q(\alpha)$ such that we can build asymptotically valid uniform $(1-\alpha)$ CIs which are:

$$\left[\hat{\mu_t}^{(k)}(x) \pm q(\alpha)\sqrt{\hat{\Omega}_t(x)/n} : \ x \in \chi\right] \tag{1.5}$$

This implies that I can make statements like:

$$lim\ \mathbb{P}\left[\sup_{x\in\chi}\ \mu_t(x) \geq C\right] \leq \alpha/2 \tag{1.6}$$

Where $C = \underset{x \in \chi}{max} \left[ \hat{\mu}_j(x) + q_j(\alpha) \sqrt{\hat{\Omega}_j(x)/n} \right]$, i.e. $C$ is the upper bound.
The reverse is also true, and so we can make statements about the *sup* and $inf$ of the $k$th derivative over compact domains.

These statements are extremely conservative. This is a function of the confidence band construction which relies on fixed critical values to obtain uniformity. For the purposes of inference on extrema, this means that the bounds obtained will not achieve nominal size, even in the limit. Nevertheless, obtaining any valid probability statement for the sup of an unobserved function is a difficult problem. Chernozhukov et al. [2013] provide a direct approach to this problem, however, their bounds are conservative in the opposite direction, and so cannot be used in this paper.

With the ability to conduct inference for the extrema of derivatives, we can turn to conducting inference for the identified set. Recall that the set $\Phi_t(x_0)$ is the set identified by the taylor projection which contains the mean function at $x_0$. Asymptotically, without stronger assumptions on the DGP, it is impossible to identify as smaller set. Therefore, we will attempt to contain that region with given size.

**Theorem 1** (Containing $\Phi_t(x_0)$)**.** *Under conditions 1-4, using local polynomials to learn the 0,...,k-1 derivatives at 0 and using b-splines to learn the sup and $inf$ of the $k$th derivative, we can build a $1-\alpha$ confidence region $CR_g$ such that:*

$$lim\mathbb{P}\left[\Phi_t(x_0) \subset CR_g\right] \geq 1 - \alpha$$

The result in theorem 1 builds somewhat naturally on well known results about local polynomials, proofs are in the supplemental appendix. The projection $P$ is linear in the estimated derivatives and extrema, which makes for easy projections once we can make statements like the one in 1.6. Combining the results of the extrema estimation routine and the local polynomial is more difficult. For now, this paper relies on a union bound.

Namely, given two statements of the form $\mathbb{P}[X_i > q_i] \leq \alpha/2$, we can also state that

$$\mathbb{P}[X_1 + X_2 > q_1 + q_2] \leq \alpha$$

As each estimation routine can return a straightforward confidence region, we can combine those regions upper and lower bounds as above.

### *1.4.3 Inference for $\tau$ in Standard RD*

The focus of this section so far has been inference of the region $\Phi_t$. The results above give a region which asymptotically contains $\Phi$ with at least given size. In the same way that union bounds let us move from just the CR for extrema to a region for $\Phi$, we can extend to a region around $\tau$. But first we should discuss the identified set.

Once again, the assumptions above are not adequate to identify a point estimate.

Rather, the nature of the derivative bounds in 1.3 is that they allow us to identify a set which will contain the value of interest. In this case, we can identify the following set:

$$\mathsf{T}(x_0) = \Phi_1(x_0) - \mu_0(x_0) = \begin{pmatrix} \mu_1(x_0)_U - \mu_0(x_0) \\ \mu_1(x_0)_L - \mu_0(x_0) \end{pmatrix} \tag{1.7}$$

Recall that for simplicity, we are relying on $x_0 < 0 \in \chi$. As a result, we know that we can identify the parameter $\mu_0(x_0)$ using standard results for local polynomials. Theorem 1 gives us a region containing $\Phi$, and so we can combine the two for $\hat{\mathsf{T}}$.

Applying the union bound again leads to the following lemma regarding the estimand of interest, $\tau$.

**Lemma 2** (CR for $\tau$)**.** *Under all the conditions of theorem 1, we can build a* $1-\alpha$ *confidence region* $CR_\tau$ *such that:*

$$lim\mathbb{P}\left[\mathsf{T}(x_0) \subset CR_\tau\right] \geq 1-\alpha$$

This result follows naturally from theorem 1, but in many ways this is the real meat of the paper. Given a point, we can take an RD design and some higher order derivative bounds, and with them we can partially identify treatment effects at that point – even when it doesn't overlap with the threshold. In section 5 the paper I will discuss applications of this idea to the closely related setup that is an RD donut design.

## 1.5 Example: Snap Benefits and Recidivism

This example is from the paper by Tuttle [2019]. That paper looks at recidivism as affected by a food assistance program. The treatment effect is identified by leveraging a discontinuity in policy which imposed a lifetime ban on SNAP benefits for individuals who engage in drug trafficking after August 23rd, 1996. The high level finding of that paper is that individuals who received a lifetime ban were about 10% more likely to commit more crimes in the future, with the effects predictably concentrated among crimes with financial benefits.

This is an excellent paper. I merely use the setting to demonstrate the extrapolations discussed here, and certainly not because of concerns about the results. A common question around these extrapolations is what derivative order makes the most sense. The notion that higher order derivative bounds are weaker conditions seems quite intuitive to many people. One critical takeaway from this example is those comparisons are not as straightforward as they may seem. Broadly speaking, as we change the derivative order being bounded, the other assumptions we make are also changing, which may make the overall procedure more conservative or not. Moreover, the location of the LATE to be estimated also can affect the relative strength of these assumptions.

To see this, recall that a second derivative bound will grow at $O(x^2)$, while a third derivative bound will grow at $O(x^3)$. For $x$ near the threshold, the third derivative may well be a stronger assumption, while far away, the second derivative can be more restrictive. I will compare the use of several different derivative bounds for

extrapolating the treatment effect.

In context, the assumption of bounds on a derivative corresponds to a bound on the changes in probability of recidivism for treatment and control groups. For a second derivative bound, this suggests that the acceleration of the control group's recidivism is restricted. Perhaps more importantly, the assumption that we observe the extrema of the derivative implies that there are not other structural changes on August 23rd which would cause the control group function to change drastically.

Figure 1.2 shows the CI for a set containing the LATE across a number of different derivative restrictions. As we can see, the first and second derivative bounds each are fairly comparable for the LATE at the threshold. Nevertheless, the second derivative bound starts substantially wider than the first derivative bound. As time goes on, the second derivative bound also grows faster, eventually containing the entire first derivative set.

The difference in starting positions and variances comes down to the additional information and variance associated with estimating more parameters in the local polynomial regression at the threshold. The more rapid growth is the natural result of allowing the first derivative to grow without limit.

The important point here is that these are different assumptions. One is not necessarily weaker or stronger, but rather different.

## 1.6 Special Case: Donut Designs

A special case of extrapolation in RD settings is that of a Donut. Donut designs are used when we have fairly standard RD settings – that is some sort of policy threshold – but we are worried that individuals have control over where they fall relative to the threshold. If individuals can shift their $x$ position by a bounded amount, then there may be selection across the threshold. Some individuals may choose to cross it, while others do not. In order to retrieve the LATE for individuals who were exogenously at the threshold, we need to eliminate the selection effect. Donut designs do this by dropping all individuals within some distance $d$ of the threshold. In essence, this corresponds to saying that we do not trust those observations.

However, having thrown out the observations near the threshold, we have gotten rid of the very observations that identify the LATE at the threshold under standard assumptions. Currently donut designs deal with this by implicitly projecting polynomials across the region of the donut.[3] This paper presents an alternative – estimate the extrema of some derivative $k$, and use that to project an identified set across the region of the donut.

In order for this to be useful, we need to make an additional assumption.

**Condition 5** (Donuts)**.**

(i) *Donut Exclusion.* There is a known region, $\mathbb{D} = (d_-, d_+)$, with $d_- < d_+$, hereafter referred to as the donut, such that all manipulation ($M = 0$ in the

3. This is what dropping those observations and re-running your local polynomial RD estimator does.

absence of manipulation) is contained to the donut.

$$\forall i \;\; s.t. \;\; \mathsf{M}_i \neq 0, \;\; x_i', x_i \in \mathbb{D}$$

(ii) *Unique Threshold.* There is one, and only one, policy relevant to the outcome of interest, which has a threshold inside the region defined as the donut and its boundaries, $[d_-, d_+]$.

Condition 5(i) ensures that all manipulation is contained to the interior of the donut. Nobody from outside that region was induced to change their behavior by the presence of the policy. This is critical – without this assumption, we retain the selection problems which we had before we decided to use a donut.
Condition 5(ii) replaces condition 2(iii) in the donut setting. This assumption looks much more like a natural extension of the standard RD assumption that there is no other co-located policy threshold.

**Lemma 3** (Donut)**.** *Under Conditions 1-5, we can find a* $1-\alpha$ *confidence region* $CR_d$ *such that*

$$lim\mathbb{P}\left[\mathsf{T}\left(0\right) \subset CR_d\right] \geq 1-\alpha$$

This is the result that we need in order to perform inference for the LATE at the threshold under a donut design. In the absence of this, or some other extrapolation result, the LATE is not asymptotically identified in donut designs. The problem is

that in most situations, the ability to manipulate the running variable is unrelated to sample size. Thus asymptotics based on observing data arbitrarily close to the threshold don't work without these extrapolation results.

### *1.6.1 An Example Donut*

In their paper, Lindo, Sanders, and Oreopoulos [2010] assess the effect of academic probation using a treatment threshold at a GPA of 1.5. They look at a variety of outcomes split on multiple dimensions, but the focus is about whether students' GPAs rise in subsequent semesters. They acknowledge the risk of manipulation – specifically that students may be "convincing teachers to give them a higher grade". After testing for a discontinuity and finding nothing, as well as checking that a number of covariates are smooth across the threshold, the authors move on.[4]

Assuming that students are able to convince 1/3 professors to raise their grade one partial letter (a max of 0.6 on the GPA scale), that implies that students have the ability to move up to 0.2 units of GPA in a semester (and year). This creates the circumstances in which a donut is reasonable.

In order to make progress, we will rely on the assumptions in section 2.[5] We set the

4. Thanks to Cattaneo et al. [2019a] the data and code needed to replicate these results are widely available.

5. Continuity of the density of the RV is somewhat questionable, the RV takes 160 unique values in the region containing the donut and bandwidths, while there are 16,000 observations in that region.

donut $\mathbb{D} = (-0.25, 0.25)$. We will use k=2 – thus it is the second derivative which is bounded. A full plot of the outcome variable against the RV shows that $\mu_0$ and $\mu_1$ may be exactly linear, so assumptions about the second derivative are reasonable. We use the bandwidth the authors selected of 0.6. Together, these assumptions give us the results seen in table 1.1.

In the original paper, the authors find a treatment effect of 0.233 GPA (95% CI: [0.18,0.285]) points gained by a person on the threshold. Breakdowns across subgroups give results of a similar magnitude. Those results are consistent with the outcomes from a donut.

| | Estimate | CI Lower | CI Upper |
|---|---|---|---|
| Bias-Corrected | 0.213 | 0.136 | 0.291 |
| Robust | 0.213 | 0.122 | 0.304 |
| Derivative Bounds: $\hat{\tau}$ | $[0.275, 0.407]$ | 0.034 | 0.727 |

Table 1.1: Comparison of Estimates from rdrobust and Donut routines

Overall, this example lets us conclude that the treatment effect of academic probation is inside the region $[0.07, 0.68]$ with confidence. This is consistent with the results in both the original paper and the replication by Cattaneo et al. [2019a]. The treatment effect on future GPA is not the only outcome of probation which should be considered for policymakes, but if there was no effect, the justification for such a policy would be thin.

## 1.7 Conclusion

This paper provides a simple approach to extending the LATE regime of regression discontinuity away from the threshold. I provide asymptotic size control for the partially identified set. An application of this work to the world of RD donuts was discussed. I hope to demonstrate the utility of this work in the future by looking at other example settings, as well as looking into the possibility of estimating the ATE using this design.

A task closely related to this paper's goals involves attempting to detect *where* the donut should be. In general this is infeasible, however, under similar (but stronger) smoothness conditions on the mean functions, we may be able to determine where relevant manipulation exists, and place the boundaries of our donut there. This is a rich vein for future work.

## 1.8 Proofs

### *1.8.1 Theorem 1*

Condition 2(ii) implies that a taylor projection is a valid technique for approximating the functions $\mu_t$. Condition 1 is somewhat unusual in combination with taylor projections – which usually are infinite sums – but in this case, but putting bounds on the extreme values of the derivative, we can say with certainty that $\mu_t(x_0) \in \Phi_t(x_0)$.

The issues here arise from the feasibility of estimating $\Phi_t$, and worse, conducting inference.

Conditions 1-4 are substantially stronger than the needed conditions for asymptotic normality of point estimates and derivatives using local polynomial estimators. They are sufficient for the conditions in Section 5.4 of Fan et al. [1995]. This will allow us to conduct inference on the vector $\theta_t(\cdot) = (\mu_t^{(0)}(\cdot), ..., \mu_t^{(k)}(\cdot))^T$ at x=0 for each of t=0,1. This will also allow inference on the point $\mu_t(x_0)$ for whichever treatment status is observed at $x_0$. This is critical for Lemma 2.

Conditions 1-4 also imply the necessary conditions for Lemma SA-5.1 and Theorems SA-5.1, SA-5.3, and SA-5.7 in Cattaneo et al. [2018]. Many of the rate restrictions and technical conditions come directly from that paper. That paper provides us with the ability to construct a confidence band that contains the entire function $\mu_t^{(k)}(\cdot)$ with given probability. By finding the extreme values of that band, and using the mappings defined in Condition 4, we can learn about the distribution of $\partial_{U,t}^{(k)}$ and $\partial_{L,t}^{(k)}$.

As the projection $P$ is linear in the derivatives, we are simply taking the parameters we have now built confidence regions for, scaling them as the projection requires, and adding them. The scaling does not affect our size control.

We have several options to add the parameters together and retain a valid confidence region. If we had a full distribution for the extrema, we could think about the joint distribution and the optimal adding of the two. However, in pushing a supremum through the results in Cattaneo et al. [2018], the outcome statement is substantially conservative, and does not correspond to a proper distribution for the

true value. Thus we will use union bounds. This means we can take any $\alpha_1$ and $\alpha_2$ such that $\alpha_1 + \alpha_2 = \alpha$, and where we know that $\mathbb{P}\left[\partial_{U,t} > C\right] \leq \alpha_1$ (with $C$ as defined in equation 1.6) and $\mathbb{P}\left[\sum_{j=0}^{k-1} \frac{\partial^j \mu_t}{\partial x^j} \frac{x_0^j}{j!} > C_5\right] \leq \alpha_2$, and conclude that $\mathbb{P}\left[\partial_{U,t} \frac{x_0^k}{k!} + \sum_{j=0}^{k-1} \frac{\partial^j \mu_t}{\partial x^j} \frac{x_0^j}{j!} > C_5 + C\frac{x_0^k}{k!}\right] \leq \alpha$. As the first part of that probability defines the upper bound of our identified set $\Phi_t$, and the statement is true for the lower bound as well, we can contain the set $\Phi_t$, with whatever probability given. See Imbens and Manski [2004] for details about construction of such a set.

### *1.8.2 Lemma 2*

With a set containing $\Phi_1(x_0)$ with some probability, and an asymptotically normal estimate of $\mu_0(x_0)$ from Fan et al. [1995], we can again apply the union bounds to build a set $\mathsf{T}$ which contains the value $\tau$ with given probability. Because there is no other policy which affects $Y$ with a threshold in $\chi$, the difference here is the LATE. This is not the most efficient construction of the LATE however. There are substantial power gains to be had from constructing the LATE equation, which can be decomposed into the projection of the extrema, the projection of a normal, and a normal. By combining the normal distributions then using the needed union bound, we manage to limit the power loss associated with union bounds.

### *1.8.3 Lemma 3*

Donuts are an interesting application of Theorem 1. In order to use them properly, we need to recenter our projection on the boundaries of the donuts. Condition 5 tells us that the donut has successfully gotten rid of all selection issues. Theorem 1 tells us that projections from those boundaries to the threshold will give us something meaningful. Taking the difference between the two set-identified parameters projected from the edges of the donut uses the same union bound procedure as above.

# CHAPTER 2

# SCM WITH SPILLOVERS: EXAMPLES AND SIMULATIONS

This paper is coauthored with Jianfei Cao.

## 2.1 Abstract

The synthetic control method is often used in treatment effect estimation with panel data where only a few units are treated and a small number of post-treatment periods are available. Current estimation and inference procedures for synthetic control methods do not allow for the existence of spillover effects, which are plausible in many applications. In a related paper [Cao and Dowd, 2021], we consider estimation and inference for synthetic control methods, allowing for spillover effects. We propose estimators for both direct treatment effects and spillover effects and show they are asymptotically unbiased. In addition, we propose an inferential procedure and show it is asymptotically unbiased. Our estimation and inference procedure applies to cases with multiple treated units and/or multiple post-treatment periods, and to ones where the underlying factor model is either stationary or cointegrated. In this paper we show simulations and empirical examples. With them, we confirm that the presence of spillovers renders current methods biased and have distorted sizes, whereas our methods yield properly sized tests and retain reasonable power.

We apply our method to a classic empirical example that investigates the effect of California's tobacco control program as in Abadie et al. [2010] and find evidence of spillovers.

## 2.2 Introduction

The synthetic control method is often used in treatment effect estimation with panel data where only a few units are treated and a small number of post-treatment periods are available. Current estimation and inference procedures for synthetic control methods do not allow for the existence of spillover effects, which are plausible in many applications. This paper alleviates these concerns by showing that given some knowledge about the spillover effects, it is possible to provide asymptotically unbiased estimators and inference in the presence of spillovers. The theoretical results for this reside in a related paper [Cao and Dowd, 2021], while this paper is our practical demonstration of the technique. Our results extend to scenarios with multiple treated units and periods, and cases with stationary or cointegrated factor models.

The synthetic control method (SCM) has gained popularity in empirical studies since its introduction in Abadie and Gardeazabal [2003]. When we observe panel data with only a few treated units and post-treatment periods, the SCM can estimate treatment effects. This setting is common in program evaluation, where we often consider state-level policies and have state-level aggregate data. The SCM models the relationship between the treated and untreated units using pre-treatment data.

Then the SCM uses the post-treatment data from untreated units to predict the counter-factual values of the treated unit. This process gives us the synthetic control, while the difference between the outcome and predicted counter-factual outcome is the treatment effect estimate. The SCM exploits the pre-treatment data to form better counter-factual values, and so in comparative case studies it is often favored over other program evaluation methods such as difference-in-differences. See Abadie and Cattaneo [2018] for review and comparison of econometric methods used in program evaluation.

However, the SCM and its variants assume explicitly or implicitly that untreated units are not affected by the treatment. That is, they rely on the Stable Unit Treatment Value Assumption (SUTVA). This dependence is natural since the SCM uses post-treatment control units to predict the counter-factual values of the treated units, which, however, is not always realistic. In our empirical example in Section 2.5, when California imposes a cigarette tax, SUTVA implies (among other things) that nobody decides to shift their cigarette purchases to Nevada.

Under SUTVA and a few other regularity conditions within a factor model, treatment effect estimators using a demeaned version of SCM are shown to be inconsistent but asymptotically unbiased by Ferman and Pinto [2019], even when the pre-treatment fit is imperfect. Unfortunately, in the presence of a spillover effect, this estimator can be severely biased. Intuitively, the reason is that post-treatment controls are contaminated by the spillover effect, resulting in a biased estimator of the counter-factual value of the treated unit in post-treatment periods, which implies a biased

treatment effect estimate. Contamination inducing bias is a standard problem in program evaluation, even within difference-in-differences and RCTs. This problem is worse for the SCM. If by chance the spillover is concentrated in control units that the synthetic control method puts significant weight on, the bias will be substantially worse than in difference-in-differences. Moreover, it is possible the spillovers propagate along the same channels as the underlying factor model, which would mean that the SCM may actively select for units which will induce bias. In our simulation section, we will explore this bias in more depth.

It is worth noting that the problem caused by spillover effects cannot fully solved by naïve methods such as not including contaminated units in estimation. This is because the contaminated units are often the most important control units that can be useful in forming the synthetic control. Simply not including them in estimation can potentially cause efficiency loss. Moreover, there are cases where most or even all control units are affected by the spillover, which cannot be solved by throwing away affected control units. This is also true for synthetic control methods that are modified to estimate treatment effects with multiple treated units, since the current methods in the literature use only the units that are not affected by the treatment in order to form the synthetic control. For examples of multivariate synthetic control methods, see Cavallo et al. [2013], Firpo and Possebom [2018], Kreif et al. [2016], Robbins et al. [2017], and Xu [2017].

The goal of this paper is to demonstrate our relaxation of the SUTVA condition and to perform estimation and testing. Particularly, we look at the cases where

there are spillover effects, which are defined by a Rubin model as the difference between the actual outcomes and the counterfactual ones. To facilitate estimation, we assume some knowledge about the spillover effects is known. More specifically, the treatment effect and the spillover effects are linear in some unknown parameters. We give examples where this assumption is plausible. For each unit of observation, we estimate a model between it and all the other units, using the SCM with pre-treatment data. Thanks to the known spillover structure, we obtain asymptotically unbiased estimators for the treatment and spillover effects. We also characterize the asymptotic distribution of the estimator. Unlike the current methods, our method uses information from all control units in estimation.

In addition, we proposed an inferential procedure based on Andrews [2003]'s end-of-sample instability test, or $P$-test. We first generalized the $P$-test to the synthetic control method without spillover effects and then generalize it further to incorporate cases with spillover effects. Similar to the $P$-test, our testing procedures use the idea of approximating the null distribution of the statistic using pre-treatment data.

In our related paper, we gave high-level conditions under which our methods are valid. Specifically, our conditions adapt to factor models with either stationary or cointegrated common factors, which are often used to justify the usage of synthetic control methods. Furthermore, we consider extensions where treatment applies to multiple units or periods, and where there are extra covariates.

In this paper, we examine an empirical example from Abadie et al. [2010]. In 1989

California implemented a cigarette tax. Abadie et al. [2010] gather data from 38 states starting in 1970 for comparison. They dismiss 12 states for potentially being affected by spillovers or later treatment. Despite this precaution, we find evidence of spillover effects in every year after 1990. Moreover, those spillovers appear to have a substantial impact on the treatment effect estimate in 4 of the first 5 years after treatment.

This paper mainly contributes to three developing literatures. First, it complements the fast-developing literature on synthetic control inference by relaxing SUTVA. Due to its popularity among empirical researchers, many formal results have been developed for statistical inference in similar settings. For example, Conley and Taber [2011] consider hypothesis testing in a similar data structure where only a few units are treated and both pre- and post-treatment periods are short. They consider difference-in-differences, and use control units to form the null distribution of the statistic. In this particular setting with only a few treated units, difference-in-difference estimator can be treated as a speical case of the SCM with equal weights. In Ferman and Pinto [2017] and Hahn and Shi [2017], similar ideas are used to conduct placebo tests which permute across observed units. Among all, Chernozhukov et al. [2017] is the most related to our work, since they also use outcomes across periods rather than across units like the above citations. Li [2019] proposes a testing procedure that is based on the idea of projection onto convex sets and results in Fang and Santos [2018]. However, none of the papers mentioned above allows for the existence of spillover effects. Our methods provide formal statistical results in this setting, without assuming SUTVA. Furthermore, our estimation and testing proce-

dure applies to factor models with cointegrated common factors, which is of special interest even in cases without spillover effects.

We also contribute to the literature on spillover effects. This fast-growing literature looks into both estimation of treatment effects in the presence of spillover effects, as well as estimation of spillover effects themselves. For example, Vazquez-Bare [2017] consider a framework where observations are grouped into clusters, and spillover effects are allowed within a cluster, but not across clusters. It discusses estimation of heterogeneous treatment effects as a function of the number of treated units within the same cluster, and spillover effects as a function of whether the unit is treated, and number of treated units within the same cluster. Basse et al. [2017] and Rosenbaum [2007] use randomization test for inference in the presence of spillover effects. Also see Basse et al. [2017] and Vazquez-Bare [2017] for a literature review on spillover effects. However, this literature seldom looks at the panel data setting with only a few treated units and short post-treatment periods. This limitation is in part because we usually do not have enough information about the spillover effects in this particular setting. We overcome this problem by requiring a potentially weak assumption that the spillover structures be pre-specified and follow a pattern that is linear in some underlying parameters. With that specification, we can estimate the spillover effects and perform statistical tests on the spillovers.

Third, our results extend the literature on Andrews [2003]'s end-of-sample instability tests. Andrews [2003] uses data across time periods to approximate the null distribution of the test statistic, and apply this idea to OLS, IV, and GMM.

Chernozhukov et al. [2017] propose a permutation test that is more general, but similar in cases where serial correlation matters. We extend this idea to the the SCM case, and further to more complicated cases with spillover effects. As Andrews and Kim [2006] extends Andrews [2003]'s results to the cointegrated cases, we also show that our method is still valid for a cointegrated factor model.

The remainder of this paper is organized as follows. Section 2.3 introduces a model with spillover effects, proposes an estimator of the spillover effects and discusses its asymptotic distribution. In the companion paper [Cao and Dowd, 2021], we consider the $P$-test introduced by Andrews [2003] and explains how it can be applied in our settings, with proofs in the Appendix Section. In that paper, we also extend our methods to cases with multiple treated units and/or multiple post-treatment periods, and briefly discusses cases with extra covariates. In this paper, we present Monte Carlo simulation results in Section 2.4 and in Section 2.5 we present an empirical example of our method. Section 2.6 concludes.

## 2.3 Model and Estimation

### *2.3.1 A Rubin Model with Spillover Effects*

We consider Rubin's potential outcome model. In Rubin's model with violation of SUTVA, the potential outcomes are functions of treatment assignments on all units.

$$
\begin{array}{cccc|c}
y_{1,1}(0,\ldots,0) & \ldots & y_{1,T}(0,\ldots,0) & y_{1,T+1}(1,0,\ldots,0) & \} \text{ treated unit} \\
\hline
y_{2,1}(0,\ldots,0) & \ldots & y_{2,T}(0,\ldots,0) & y_{2,T+1}(1,0,\ldots,0) & \\
\vdots & \ddots & \vdots & \vdots & \} \text{ control units} \\
y_{N,1}(0,\ldots,0) & \ldots & y_{N,T}(0,\ldots,0) & y_{N,T+1}(1,0,\ldots,0) & \\
 & & \uparrow \text{treatment} & &
\end{array}
$$

Figure 2.1: Example Synthetic Controls Data Structure

Namely, the outcome of unit $i$ at time $t$ is

$$y_{i,t} = y_{i,t}(d_t),$$

where $d_t = (d_{1,t}, \ldots, d_{N,t})'$ and $d_{i,t} = 1$ if unit $i$ has been treated at time $t$.

We consider a standard synthetic control setting where only one unit is treated and only one period is available after the treatment is implemented. We consider cases with multiple treated units and multiple post-treatment periods in the related paper. Let unit 1 be treated between time $T$ and $T+1$, and there be another $N-1$ units that are not directly treated by the policy. Thus, we observe an $N \times (T+1)$ panel as shown in Figure 2.1.

Note that we only observe outcomes with $d_{T+1} = (0, \ldots, 0)'$ or $d_{T+1} = (1, 0, \ldots, 0)'$. This is the fundamental limitation of the dataset we are currently studying. Unless other homogeneity conditions are assumed, we cannot say anything about $y_{i,T+1}(d_{T+1})$ for $d_{T+1} \notin \{(0, \ldots, 0)', (1, 0, \ldots, 0)\}$ because only a few units are treated and only a

few post-treatment periods are available. For notation simplicity, let

$$\begin{cases} y_{i,t}(0) = y_{i,t}(0,\ldots,0) \\ y_{i,t}(1) = y_{i,t}(1,0,\ldots,0) \end{cases}$$

for each $(i,t)$. Let $\alpha_i = y_{i,T+1}(1) - y_{i,T+1}(0)$ be the potential deviation from unit $i$'s counterfactual outcome $y_{i,T+1}(0)$ where no unit is treated at time $T+1$. That is, $\alpha_1$ is the direct treatment effect on unit 1, while $\alpha_i$ with $i \neq 1$ is the indirect effect or spillover effect. Throughout, we consider the case where $N$ is fixed and $T$ goes to infinity.

In case studies, we are often interested in estimating the treatment effect $\alpha_1$. For example, Abadie et al. [2010] consider the direct treatment effect on California of the tobacco control policy implemented in the state. A common choice is the synthetic control estimator. Namely, we obtain the synthetic control weights by solving the optimization problem

$$\begin{bmatrix} \widehat{a}_1 \\ \widehat{b}_1 \end{bmatrix} = \underset{\tilde{a}\in\mathbb{R},\tilde{b}\in W^{(1)}}{\arg\min} \sum_{t=1}^{T} (y_{i,t} - \tilde{a} - Y_t'\tilde{b})^2, \tag{2.1}$$

where $Y_t = (y_{1,t},\ldots,y_{N,t})'$ and $W^{(1)} = \{(w_1,\ldots,w_N)' \in \mathbb{R}_+^N : w_1 = 0, \sum_{j=2}^N w_j = 1\}$. An estimator of the treatment effect $\alpha_1$ is given by

$$\widehat{\alpha}_1 = y_{1,T+1} - (\widehat{a} + Y_{T+1}'\widehat{b}),$$

i.e., the counter-factual value $y_{1,T+1}(0)$ is approximated by $\widehat{a} + Y'_{T+1}\widehat{b}$. For this paper we use an constraint set as in the demeaned synthetic control method [Ferman and Pinto, 2019]. That is, we do not restrict the intercept but require the other coefficients to be positive and sum up to one. [1]

### *2.3.2 Assumptions*

#### Spillover Structure

Throughout the paper, we assume some knowledge about the spillover effects is known. Namely, assume that the full effect vector $\alpha$ is a linear transformation of some unknown parameter $\gamma \in \mathbb{R}^k$, i.e. $\alpha = A\gamma$. Typically, $\gamma$ has less dimensions than $\alpha$ does. Here are some examples that fit in this framework.

*Example* 1. Assume a subset of control units, but not all of them, are equally affected

1. Other choices of constraint set for $(\widehat{a}_1, \widehat{b}'_1)'$ include $\{0\} \times \{0\} \times \Delta_{N-1}$ as in the original synthetic control method of Abadie and Gardeazabal [2003] and Abadie et al. [2010], and $\mathbb{R} \times \{0\} \times \mathbb{R}_+^{N-1}$ as in the modified synthetic control of Li [2019], where $\Delta_{N-1} = \{w \in \mathbb{R}^{N-1} : w_i \geq 0 \text{ for each } i, \sum_{i=1}^{N-1} w_i = 1\}$ is a $(N-1)$-dimensional simplex. See Doudchenko and Imbens [2016a] for a discussion of other restriction sets.

by the spillover effects, i.e.

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}, \ \gamma = \begin{bmatrix} \alpha_1 \\ b \end{bmatrix}.$$

*Example* 2. Assume the spillover effect shrinks as the geometric distance goes up. For $i = 2, \ldots, N$, $\alpha_i = b \exp(-d_i)$ where $d_i$ is the distance between unit 1 and unit $i$ and $b$ is some unknown parameter of interest. Then, we have

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \exp(-d_2) \\ \vdots & \vdots \\ 0 & \exp(-d_N) \end{bmatrix}, \ \gamma = \begin{bmatrix} \alpha_1 \\ b \end{bmatrix}.$$

*Example* 3. Assume the spillover effect is likely to take place at some known locations, but not at other locations, while the sizes of spillover effects are allowed to vary across those units. For example, assume there are potential spillovers at locations whose distance to unit 1 is less than $\bar{d}$. Then, the treatment and spillover effect vector can also be represented by $A\gamma$. WLOG order the units by increasing distance from unit

1, and let $p$ the number of units experiencing spillovers. Then

$$A = \begin{bmatrix} 1 & 0_{1\times p} \\ 0_{p\times 1} & I_p \\ 0_{(N-p-1)\times 1} & 0_{(N-p-1)\times p} \end{bmatrix}, \ \gamma = \begin{bmatrix} \alpha_1 \\ \alpha_{k_1} \\ \vdots \\ \alpha_{k_p} \end{bmatrix}.$$

Thus the units indexed $2, ..., (p+1)$ each experience their own size spillover effect.

The assumptions in Example 3 are often plausible. We give an empirical example in Section 2.5. If mis-specification of the spillover structure is a concern, one can always choose an $A$ matrix that incorporates more potential spillovers, i.e., a bigger $p$.

## Invertibility Assumption

In order to back out the spillover effects, we proceed as follows. We first define the individual synthetic control weights and their limits. Namely, let

$$\begin{bmatrix} \widehat{a}_i \\ \widehat{b}_i \end{bmatrix} = \underset{\tilde{a}\in\mathbb{R},\tilde{b}\in W^{(i)}}{\arg\min} \sum_{t=1}^{T} (y_{i,t} - \tilde{a} - Y_t\tilde{b}')^2, \tag{2.2}$$

where $W^{(i)} = \{(w_1, \ldots, w_N)' \in \mathbb{R}_+^N : w_i = 0, \sum_{j=1}^N w_j = 1\}$. Then, let

$$a_i = \text{plim } \widehat{a}_i, \ b_i = \text{plim } \widehat{b}_i,$$

and we only consider cases where they are well-defined. We show in the related paper that $a_i$ and $b_i$ exist for each $i$ in factor models with stationary or cointegrated common factors. In general, $a_i$ and $b_i$ do not coincide with the weights that reconstruct the factor loadings [Ferman and Pinto, 2019].

For each $(i,t)$, define the specification error by

$$u_{i,t} = y_{i,t}(0) - (a_i + Y_t(0)'b_i). \tag{2.3}$$

Note that the $i$-th entry of $b_i$ is zero. Define $a = (a_1, \ldots, a_N)'$, $B = (b_1, \ldots, b_N)'$, and $M = (I - B)'(I - B)$. Stacking Equation (2.3) for all $i$'s gives

$$u_t = Y_t(0) - (a + BY_t(0)),$$

where and $u_t = (u_{1,t}, \ldots, u_{N,t})'$. For $t = T + 1$, this becomes

$$u_{T+1} = (I - B)(Y_{T+1} - \alpha) - a, \tag{2.4}$$

where $Y_{T+1} = (y_{1,T+1}, \ldots, y_{N,T+1})'$. We will use this equation to estimate the spillover effect.

Defining $M = (I - B)'(I - B)$, we introduce the following invertibility assumption:

*Condition IN.* $A'MA$ is non-singular.

First note Condition IN is testable in principle. We can consistently estimate $B$ so

the data informs us of the validity of this assumption. To understand this assumption better, we replace $\alpha$ by $A\gamma$ in Equation (2.4) and have

$$(I - B)A\gamma = (I - B)Y_{T+1} - a - u_{T+1}. \tag{2.5}$$

Equation (2.5) is the key to learning $\alpha$. Under mild regularity conditions, $a$ and $B$ are identified from the model and learned by the synthetic control method. We do not observe $u_{T+1}$, but the distribution of $u_{T+1}$ can be learned using pre-treatment data under stationarity of $\{u_t\}_{t\geq 1}$. Therefore, if $A'MA$ is non-singular, or equivalently, $(I - B)A$ has full rank, we can form an estimator of $\gamma$ whose limiting distribution is identified by multiplying both sides of Equation (2.5) by $(A'MA)^{-1}A'(I - B)'$. Note that we do not identify $\gamma$ or $\alpha$. This is because we have only one observation of the outcome in post-treatment periods.

We illustrate Condition IN in the following toy example.

*Example* 4. Assume there are 3 units in total, where unit 1 is treated. Let the synthetic control weight matrix $B$ be

$$B = \begin{bmatrix} 0 & w_1 & 1 - w_1 \\ w_2 & 0 & 1 - w_2 \\ w_3 & 1 - w_3 & 0 \end{bmatrix}.$$

Suppose the researcher first assumes unit 2 and 3 are equally exposed to the spillover

effects. That is, they assume

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \text{ and } \alpha = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_2 \end{bmatrix}.$$

Then, Condition IN does not hold, because

$$(I - B)A_1 = \begin{bmatrix} 1 & -1 \\ -w_2 & w_2 \\ -w_3 & w_3 \end{bmatrix}.$$

If they instead assumes only one of the controls is exposed to the spillover effects, Condition IN is satisfied in general. In this case,

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \text{ and } \alpha = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ 0 \end{bmatrix},$$

and

$$(I - B)A_2 = \begin{bmatrix} 1 & -w_1 \\ -w_2 & 1 \\ -w_3 & w_3 - 1 \end{bmatrix}.$$

It can be shown that $(I - B)A_2$ always has full rank for $(w_1, w_2, w_3) \in [0, 1]^3$.

This applies to more general settings. That is, if all controls are equally hit by the

spillover effects, then $(I - B)A$ does not have full rank and we lose Condition IN. Allowing a few units to be exempt from the spillover effects makes $(I - B)A$ have full rank in general.

A more interesting case is Example 3, where we only restrict the range of spillover effects and allow the levels to vary. In this case, $(I - B)A$ can be obtained by eliminating columns that correspond to units that are neither treated nor exposed to spillover effects. Again, as long as at least one control is not exposed to the spillover effects, $(I - B)A$ has full rank in general. This assumption is more convincing if a moderate number of columns are eliminated from $(I - B)$, i.e. only a few units are exposed to the spillover effects.

### *2.3.3 Estimation*

We form estimators for $(a, B)$ using synthetic control methods as in (2.2). We do that for each $i = 1, \ldots, N$, as if each $i$ is the treated unit and other units are controls. Then, the estimators for $a$ and $B$ are $\widehat{a} = (\widehat{a}_1, \ldots, \widehat{a}_N)'$ and $\widehat{B} = (\widehat{b}_1, \ldots, \widehat{b}_N)'$ respectively. Let $\widehat{M} = (I - \widehat{B})'(I - \widehat{B})$ be an estimator for $M$. Let an estimator of $\gamma$ be such that

$$\begin{aligned}\widehat{\gamma} &= \arg\min_{g \in \mathbb{R}^k} \|(I - \widehat{B})(Y_{T+1} - Ag) - \widehat{a}\| \\ &= (A'\widehat{M}A)^{-1}A'(I - \widehat{B})'((I - \widehat{B})Y_{T+1} - \widehat{a}). \end{aligned} \tag{2.6}$$

Note that the FOC implies

$$A'(I-B)'u_{T+1}=0,$$

i.e. it requires that some weighted sum of the residuals to be zero. Under that condition, the treatment and spillover effect vector $\alpha$ can be estimated by $\widehat{\alpha}=A\widehat{\gamma}$.

*Assumption* 1. *(a)* $\{u_t\}_{t\geq 1}$ is stationary, and has mean zero.

*(b)* $\|\widehat{a}-a\|=o_p(1)$, $\|\widehat{B}-B\|=o_p(1)$

*(c)* $\|(\widehat{B}-B)Y_{T+1}(0)\|=o_p(1)$.

*(d)* $A'MA$ is non-singular.

Note that Part (c) excludes polynomial time trends.

*Theorem* 1. Suppose Assumption 1 holds. Then, $\widehat{\alpha}-(\alpha+Gu_{T+1})\to_p 0$ as $T\to\infty$, where $G=A(A'MA)^{-1}A'(I-B)'$. Moreover, $E[Gu_{T+1}]=0$.

The structure of the limiting distribution is similar to the case as in Ferman and Pinto [2019], as it is inconsistent but asymptotically unbiased (i.e. that the difference between the estimator and the true value has zero mean). Note that consistent estimators are impossible because only one post-treatment period is available.

Moreover, we can form an estimator of $\alpha$ with possibly lower variance. For some positive definite matrix $W\in\mathbb{R}^N$, we minimize $\|W^{1/2}\epsilon_{T+1}\|$ instead of $\|\epsilon_{T+1}\|$. The

resulting estimator is

$$\begin{aligned}\widehat{\gamma}_W &= \arg\min_{g\in\mathbb{R}^k} \|W^{1/2}((I-\widehat{B})(Y_{T+1}-Ag)-\widehat{a})\| \\ &= (A'\widehat{M}_W A)^{-1}A'(I-\widehat{B})'W((I-\widehat{B})Y_{T+1}-\widehat{a}),\end{aligned}$$

where $\widehat{M}_W = (I-\widehat{B})'W(I-\widehat{B})$. The corresponding estimator for $\alpha$ is $\widehat{\alpha}_W = A\widehat{\gamma}_W$. In the spirit of GMM with an efficient weighting matrix, let $\Omega = Cov[u_1]$ and $W_T^e$ be a consistent estimator of $\Omega^{-1}$. Then an estimator of $\alpha$ with lower variance can be achieved by $\widehat{\alpha}^e = \widehat{\alpha}_{W_T^e}$.

Let $M_W = (I-B)'W(I-B)$, $G_W = A(A'M_W A)^{-1}A'(I-B)'W$ for some weighting matrix $W$, $W^e = \Omega^{-1}$, $M^e = M_{W^e}$, and $G^e = G_{W^e}$. Then, we have the following results.

*Proposition* 1. Suppose Assumption 1 holds, $W_T$ is a consistent estimator for $W$, and $W_T^e$ is a consistent estimator for $W^e$. Then, $\widehat{\alpha}_{W_T} - (\alpha + G_W u_{T+1}) \to_p 0$, and specifically, $\widehat{\alpha}^e - (\alpha + G^e u_{T+1}) \to_p 0$, as $T \to \infty$. Moreover, $(Cov[G_W u_{T+1}] - Cov[G^e u_{T+1}])$ is positive semi-definite.

In practice, we need to estimate $\Omega$, and for that we would need relatively large sample size (large $T$) to have a good approximation.

## 2.4 Simulation

We present Monte Carlo simulation results in this section. For each case considered, we use 1000 simulation repetitions.

### *2.4.1 Estimation with Spillover Effects*

In this subsection we examine the finite sample performance of our estimation procedure proposed in Section 2.2. The model considered here is similar to Li [2019], where $y_{i,t}(0)$ follows a factor model structure. We show both stationary and $\mathcal{I}(1)$ case.

#### Stationary Case

The underlying factor model is

$$y_{i,t}(0) = \eta_t + \lambda_t' \mu_i + \epsilon_{i,t},$$

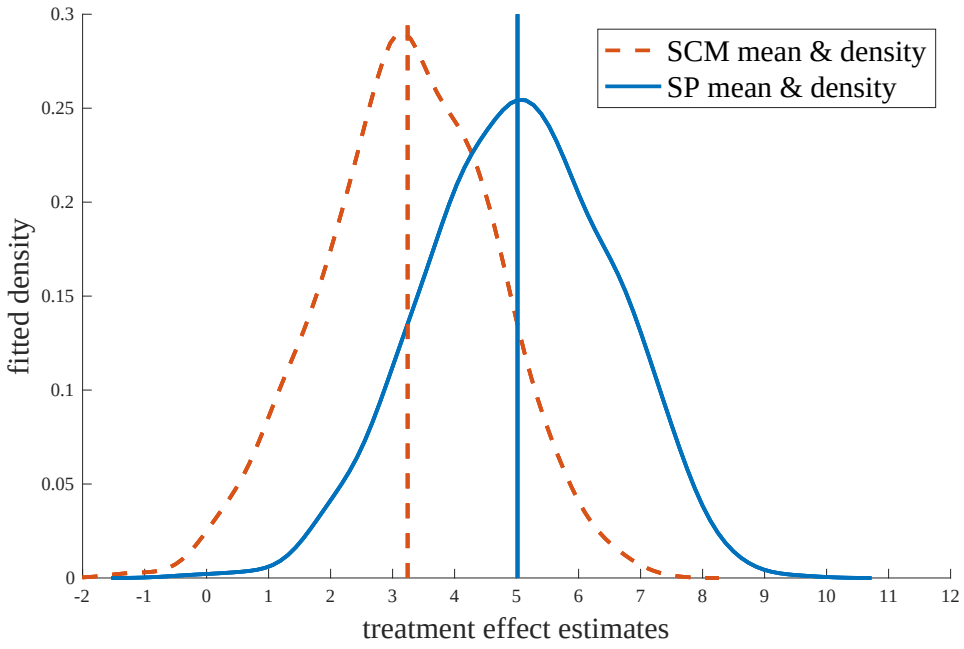


Figure 2.2: Distribution of treatment effect estimates. The true treatment effect is 5. SCM is using the standard synthetic control method assuming no spillover effects. SP is the estimation procedure proposed in this paper that takes spillover effects into account. Estimates are fitted using kernel density.

where $\lambda_t = (\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t})'$,

$$\eta_t = 1 + 0.5\delta_{t-1} + \nu_{0,t},$$

$$\lambda_{1,t} = 0.5\lambda_{1,t-1} + \nu_{1,t},$$

$$\lambda_{2,t} = 1 + \nu_{2,t} + 0.5\nu_{2,t-1},$$

$$\lambda_{3,t} = 0.5\lambda_{3,t-1} + \nu_{3,t} + 0.5\nu_{3,t-1},$$

and $\epsilon_{i,t}$ and $\nu_{j,s}$ is i.i.d. $N(0,1)$ for each $(i,j,s,t)$. Each entry of $\mu_i$ is drawn from an independent uniform distribution on $[0,1]$ and fixed for each repetition. At $t = T+1$, the observed outcome is $y_{i,T+1} = y_{i,T+1}(0) + \alpha_i$, where $\alpha_i$ is either treatment effect or spillover effect and is specified below. The treatment effect is set to 5 and the spillover effect is 3.

The empirical bias and variance (in parentheses) of the treatment effect estimator using two methods are shown in Table 2.1. We consider three spillover patterns. *No spillover effects* is the case where unit 1 receives a treatment effect of 5 at $t = T+1$ and other units are not affected. *Concentrated spillover effects* is the case where 1/3 of the control units receive a spillover effect of 3. *Spreadout spillover effects* is the case where 2/3 of the control units receive a spillover effect of 3. SCM is the original synthetic control method, and SP is the corrected synthetic control method proposed in Section 2.3.3. Throughout the simulations we assume the coverage of spillover effect is known, but not other information, so $A$ is constructed as in Example 3. For *No spillover effects*, we are being conservative in our use of the SP estimator and run it as if 1/3 of the control units are exposed to spillover effects.

To better compare results, we fit the simulation results using kernel density for the $(N,T) = (10,50)$ case with concentrated spillover effects and plot it in Figure 2.2.

## $\mathcal{I}(1)$ Case

For the $\mathcal{I}(1)$ case, the underlying factor model follows

$$y_{i,t}(0) = \lambda_t' \mu_i + \epsilon_{i,t},$$

where $\lambda_t = (\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t})'$,

$$\begin{aligned}
\lambda_{1,t} &= \lambda_{1,t-1} + 0.5\nu_{1,t}, \\
\lambda_{2,t} &= \lambda_{2,t-1} + 0.5\nu_{2,t}, \\
\lambda_{3,t} &= 0.5\lambda_{3,t-1} + \nu_{3,t},
\end{aligned}$$

and $\epsilon_{i,t}$ and $\nu_{j,s}$ follows i.i.d. $N(0,1)$ for each $(i,j,s,t)$. The factor loadings are constructed such that condition CO is satisfied. Namely, we let $\mu_1 = (1,0,0)'$, $\mu_2 = (0,1,0)'$, $\mu_3 = (1,0,0)'$, $\mu_4 = (0,1,0)'$, and for $\mu_j$ with $j = 5, \ldots, N$, we draw independent uniform distribution on $[0,1]$ for each entry and then normalize each loading vector such that three entries of each $\mu_j$ sum up to one. The constructed factor loadings are fixed for each repetition while other settings are same as the stationary case. The results are shown in Table 2.2.

### *2.4.2 Test for Treatment Effects*

In this section we compare test procedures against the null hypothesis $H_0 : \alpha_1 = 0$, i.e. the treatment effect is zero. The results are shown in Table 2.3 and Table 2.4. The DGP is exactly the same as in Section 2.4.1 (the stationary case), except that $\alpha_1 = 0$ (the null) for Table 2.3 and $\alpha_1 = 5$ (the alternative) for Table 2.4. Placebo test is as in Abadie and Gardeazabal [2003] and Hahn and Shi [2017]. Andrews' test is as in Andrews [2003]. SP is the spillover-adjust test proposed in Cao and Dowd [2021].

Among the three testing procedures, SP test has correct sizes and outperforms the other two methods in power. Placebo test has correct sizes in some cases but has lower power, and Andrews' test over-rejects under null. The reasons are discussed in Cao and Dowd [2021].

### *2.4.3 Test for Existence of Spillover Effects*

In this section we examine the power of the proposed test against the null hypothesis that there are no spillover effects. We also look into its behavior when the range of the spillover effect is not correctly specified. In this set of experiments, the level of spillover effects varies from 0 to 2, corresponding to the strength of alternative hypotheses. We set $(N, T) = (20, 50)$ and $\alpha_1 = 5$. There are 9 units that are affected by spillover effects. Other settings follow exactly as in Section 2.4.1 (the stationary

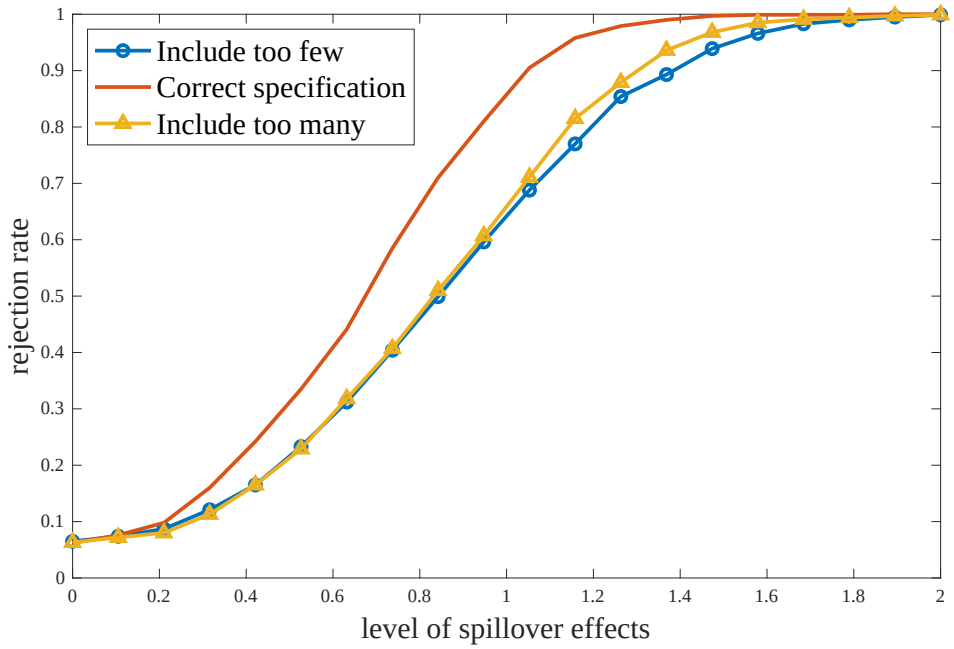


Figure 2.3: Empirical rejection rate of testing for existence of spillover effects. There are 20 units in total and half of them are affected by the treatment. *Include too few* is assuming only 5 of them are affected by the treatment. *Correct specification* assumes the researcher knows exactly which set of units are affected. *Include too many* assumes 15 units are affected, 5 of which are in fact not affected.

case). The model for the range of spillover is as in Example 3.

The empirical rejection rates against various levels of spillover effects using our method proposed in Cao and Dowd [2021] are plotted in Figure 2.3. Here *Include too few* misses half of the units that are actually affected by the treatment (assuming that unit 1 as well as four other units are affected), *Correct specification* assumes we know exactly which units are affected, and *Include too many* assumes 15 units are affected in estimation, 5 of which are actually not affected by spillover effects.

The simulation results show that the proposed test is quite robust to model misspecification. Among the three cases, *Include too many* is still a correct specification but is supposed to be more conservative, so it has less power than *Correct specification* does. The range of spillover effects is misspecified in *Include too few*, but the test is still correctly sized under the null[2] and has reasonable power under alternatives.

## 2.5 Empirical Example

To demonstrate our method, we use it on the classic SCM example from Abadie et al. [2010] (ADH), which looks at the effect of Proposition 99 on California cigarette consumption. In this section, we will walk through the results from our method, with interruptions to point out key features and issues.

Proposition 99 intended to disincentivize smoking, which was primarily achieved by introducing a $0.25 tax on each pack of cigarettes. By measuring sales in California, ADH and others have attempted to determine the effect of the policy on smoking rates. However, traditional SCM is not guaranteed to produce an unbiased treatment effect estimator in the presence of spillover effects. In this tobacco control program example, we are concerned about two kinds of spillover effects. The first spillover is based on concerns about "leakage". A common problem with cigarette taxes is that measured local consumption might fall as people move their purchasing behavior across legal boundaries. In order to accommodate this, we allowed for a spillover

2. The model is always correctly specified under null.

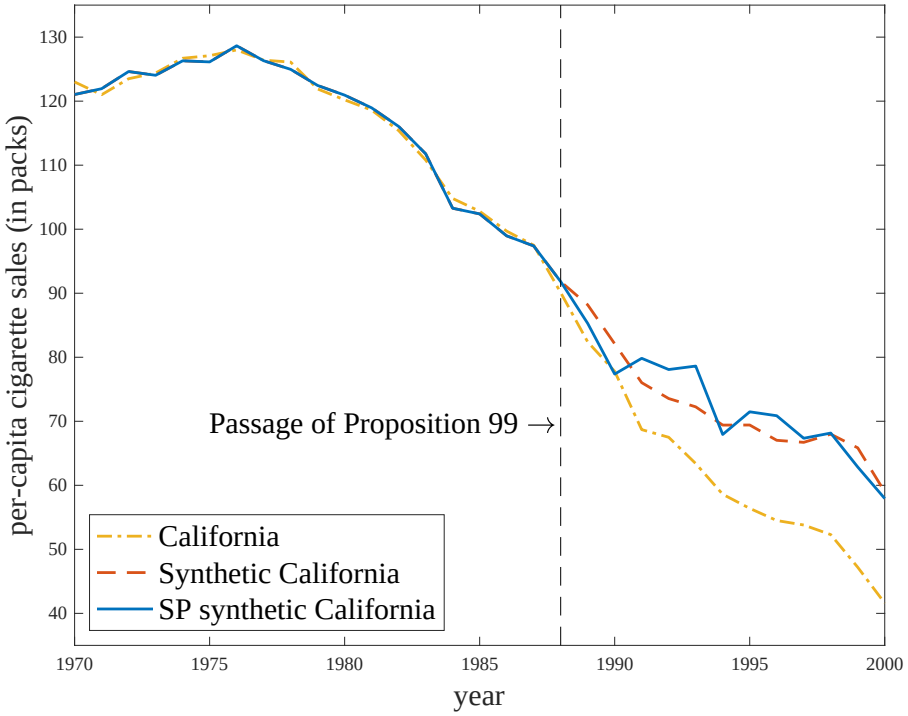


Figure 2.4: Trends in per-capita cigarette sales: California, synthetic California, and spillover-adjusted synthetic California. SP synthetic California is using our estimation procedure, which accounts for spillover effects. The vertical line indicates the start of treatment.

affecting states neighboring California and a spillover affecting states which a state away from California. The second spillover type we considered was a cultural change. If tobacco is discouraged in California, it might reduce the cultural appeal of smoking. Reasoning that the northeast is culturally close to the west coast, we allowed for the northeastern states to experience this cultural spillover.

One might also think that there could be a policy contamination whereby culturally close states also enact policies with similar targets. Our method can allow for this

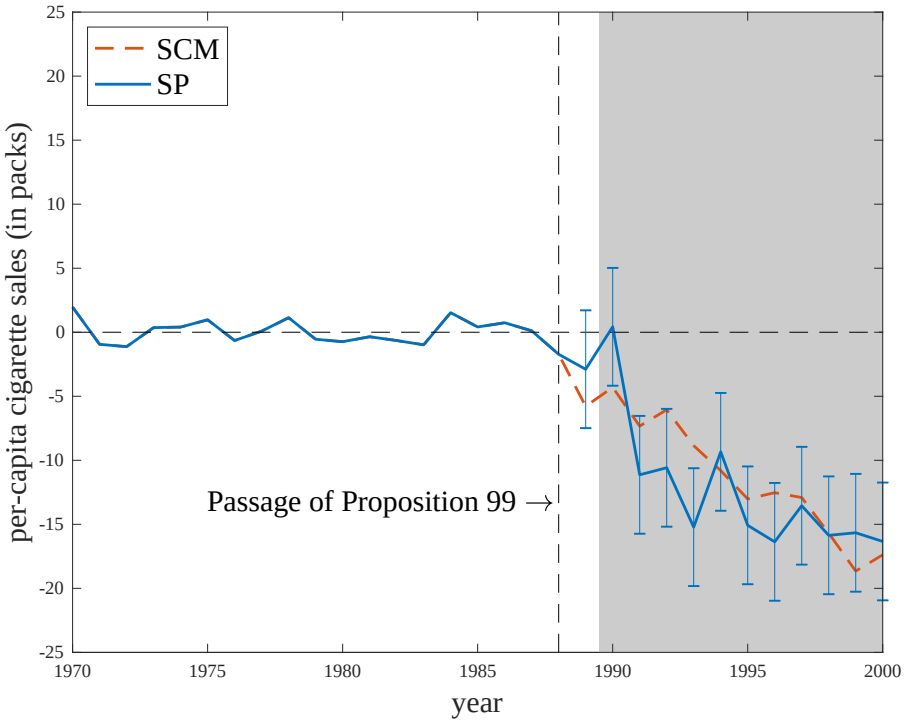


Figure 2.5: Per-capita cigarette sales gap between California and (spillover-adjusted) synthetic California (with 90% confidence interval). The lines to the right of passage of Proposition 99 are treatment effect estimates. SCM is obtained by using standard synthetic control method. SP is using our estimation procedure, which accounts for spillover effects. Shaded area denotes our test rejects there is no spillover effects in those years.

kind of spillover in our estimation. However, the initial paper took that type of problem into account, and 12 states which experienced legislative changes in the ensuing years were removed in that paper (and thus in our data).

The data used is per capita cigarette consumption in 38 of the 50 states running from 1970 to 2000. Twelve states were removed from the data because of concerns that they were either contaminated, or received treatment later on. In 1989 California enacted Proposition 99, so all periods from 1989 onwards are considered post-treatment

periods. We replicate this program evaluation using the method introduced in previous sections, allowing for possible spillover effects. We use the spillover structure as in Example 3. That is, we allow for arbitrary spillover effects in those geographically close and culturally similar states as described in the last paragraph, but not the others. We also perform hypothesis testing on both treatment effects and spillover effects.

The results are shown in Figure 2.4 and Figure 2.5. The method in Abadie et al. [2010] is indexed by SCM and our method is SP. Figure 2.4 shows the "synthetic California" and Figure 2.5 elaborates on this by specifically looking at the estimated treatment effects. The error bars are built using the methods described in this paper, at a significance level of 90%. We do not use a 95% significance level because there are only 19 pre-treatment periods.

As Figure 2.4 shows, our estimated consumption in the "synthetic California" does not differ qualitatively from what a standard SCM would predict. Quantitatively, Figure 2.5 shows that our results are more consistent with an addiction story, that tobacco consumption is addictive and should not fall immediately after the policy. From the tests of spillover effects (shaded area of Figure 2.5), we see that likely there were substantial spillover effects, which in some periods lead to statistically significant changes in the treatment effect estimates. For example, the SCM estimate of year 1990 lies outside our confidence interval, which potentially results from the over-estimation of scale of the treatment effects in the presence of spillover effects.

## 2.6 Conclusion

The synthetic control method is a powerful tool in treatment effect estimation in the panel data settings, but it does not work in the presence of spillover effects. In a related paper, we relax this assumption and propose an estimation and testing procedure that is robust to the presence of spillover effects. Our method requires specification of the spillover structure, which can be weak (Example 3). We derive a set of conditions under which our estimators are asymptotically unbiased. We develop a testing procedure based on Andrews [2003]'s end-of-sample instability tests and show that it is asymptotically unbiased under a set of conditions. We show our conditions are satisfied by the commonly used factor models, with either stationary or cointegrated common factors. Our methods can be extended to cases with multiple treated units and multiple post-treatment periods, and with extra covariates. In this paper, we provide simulation results certifying the validity of our estimation and testing procedure in the presence of spillover effects. The simulations also indicate that our testing procedure is relatively robust to misspecification of the spillover structure. Finally, we illustrate our method by applying it to Abadie et al. [2010]'s California tobacco control program data.

Table 2.1: Treatment effect estimation with stationary common factors.

| | $N = 10$ | | $N = 30$ | | $N = 50$ | |
|---|---|---|---|---|---|---|
| | $T = 50$ | 200 | 50 | 200 | 50 | 200 |
| *No spillover effects* | | | | | | |
| SCM | 0.011 | -0.003 | 0.114 | 0.016 | -0.041 | -0.033 |
| | (1.249) | (1.586) | (1.244) | (1.273) | (1.290) | (1.182) |
| SP | 0.013 | 0.018 | -0.012 | 0.010 | -0.031 | -0.040 |
| | (1.417) | (1.710) | (1.362) | (1.486) | (1.516) | (1.270) |
| *Concentrated spillover effects* | | | | | | |
| SCM | -0.986 | -1.333 | -0.880 | -1.543 | -1.070 | -0.796 |
| | (1.451) | (2.065) | (1.654) | (1.392) | (1.638) | (1.461) |
| SP | 0.025 | 0.140 | 0.038 | 0.025 | -0.055 | 0.110 |
| | (1.425) | (1.756) | (1.435) | (1.250) | (1.581 ) | (1.408) |
| *Spreadout spillover effects* | | | | | | |
| SCM | -1.910 | -2.114 | -1.859 | -2.398 | -2.112 | -2.154 |
| | (1.470) | (1.696) | (1.472) | (1.369) | (1.538) | (1.313) |
| SP | 0.007 | 0.029 | -0.025 | 0.018 | -0.048 | -0.028 |
| | (1.438) | (2.061) | (1.296) | (1.602) | (1.450) | (1.290) |

Notes: The numbers without parentheses are empirical bias in simulation. The ones with parentheses are empirical variance. SCM is the standard synthetic control method assuming no spillover effects. SP is the estimation procedure proposed in this paper that takes spillover effects into account. *No spillover effects* stands for the cases where the true DGP has no spillover effects. *Concentrated spillover effects* is the case where 1/3 of the control units receive a spillover effect. *Spreadout spillover effects* is the case where 2/3 of the control units receive a spillover effect of the same level.

Table 2.2: Treatment effect estimation with $\mathcal{I}(1)$ common factors.

| | $N = 10$ | | $N = 30$ | | $N = 50$ | |
|---|---|---|---|---|---|---|
| | $T = 50$ | 200 | 50 | 200 | 50 | 200 |
| *No spillover effects* | | | | | | |
| SCM | -0.018 | -0.043 | -0.088 | -0.031 | 0.038 | -0.038 |
| | (1.642) | (1.772) | (1.539) | (1.900) | (1.810) | (1.866) |
| SP | -0.057 | -0.017 | -0.053 | -0.044 | 0.013 | -0.017 |
| | (2.249) | (4.523) | (2.121) | (2.184) | (1.849) | (1.952) |
| *Concentrated spillover effects* | | | | | | |
| SCM | -1.400 | -2.234 | -2.026 | -1.954 | -1.408 | -2.325 |
| | (1.854) | (1.856) | (1.921) | (2.079) | (2.043) | (1.976) |
| SP | -0.057 | -0.017 | -0.053 | -0.044 | 0.013 | -0.017 |
| | (2.249) | (4.523) | (2.121) | (2.184) | (1.849) | (1.952) |
| *Spreadout spillover effects* | | | | | | |
| SCM | -2.599 | -2.885 | -2.536 | -2.465 | -2.402 | -2.889 |
| | (1.779) | (1.795) | (1.759) | (2.037) | (1.921) | (1.900) |
| SP | 0.027 | -0.022 | -0.008 | 0.010 | 0.006 | -0.045 |
| | (3.447) | (7.367) | (2.412) | (2.740) | (2.279) | (2.712) |

Notes: The numbers without parentheses are empirical bias in simulation. The ones with parentheses are empirical variance. SCM is the standard synthetic control method assuming no spillover effects. SP is the estimation procedure proposed in this paper that takes spillover effects into account. *No spillover effects* stands for the cases where the true DGP has no spillover effects. *Concentrated spillover effects* is the case where 1/3 of the control units receive a spillover effect. *Spreadout spillover effects* is the case where 2/3 of the control units receive a spillover effect of the same level.

Table 2.3: Empirical rejection rate of testing for treatment effects under null.

| | $N = 10$ | | | $N = 30$ | | | $N = 50$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $T = 15$ | 50 | 200 | 15 | 50 | 200 | 15 | 50 | 200 |
| *No spillover effects* | | | | | | | | | |
| Placebo | 0.000 | 0.000 | 0.000 | 0.072 | 0.053 | 0.062 | 0.034 | 0.031 | 0.040 |
| Andrews | 0.076 | 0.061 | 0.060 | 0.108 | 0.082 | 0.065 | 0.141 | 0.078 | 0.072 |
| SP | 0.048 | 0.049 | 0.058 | 0.055 | 0.064 | 0.052 | 0.066 | 0.046 | 0.059 |
| *Concentrated spillover effects* | | | | | | | | | |
| Placebo | 0.000 | 0.000 | 0.000 | 0.066 | 0.046 | 0.116 | 0.035 | 0.029 | 0.026 |
| Andrews | 0.411 | 0.207 | 0.224 | 0.417 | 0.279 | 0.346 | 0.519 | 0.346 | 0.184 |
| SP | 0.065 | 0.050 | 0.043 | 0.111 | 0.069 | 0.061 | 0.109 | 0.092 | 0.054 |
| *Spreadout spillover effects* | | | | | | | | | |
| Placebo | 0.000 | 0.000 | 0.000 | 0.129 | 0.063 | 0.147 | 0.060 | 0.059 | 0.072 |
| Andrews | 0.576 | 0.478 | 0.399 | 0.685 | 0.563 | 0.616 | 0.741 | 0.621 | 0.544 |
| SP | 0.036 | 0.035 | 0.042 | 0.034 | 0.042 | 0.046 | 0.030 | 0.042 | 0.044 |

Notes: SP is the estimation procedure proposed in this paper that takes spillover effects into account. *No spillover effects* stands for the cases where the true DGP has no spillover effects. *Concentrated spillover effects* is the case where 1/3 of the control units receive a spillover effect. *Spreadout spillover effects* is the case where 2/3 of the control units receive a spillover effect of the same level.

Table 2.4: Empirical rejection rate of testing for treatment effects under alternative.

| | $N = 10$ | | | $N = 30$ | | | $N = 50$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $T = 15$ | 50 | 200 | 15 | 50 | 200 | 15 | 50 | 200 |
| *No spillover effects* | | | | | | | | | |
| Placebo | 0.000 | 0.000 | 0.000 | 0.908 | 0.939 | 0.966 | 0.922 | 0.936 | 0.931 |
| Andrews | 0.797 | 0.948 | 0.926 | 0.785 | 0.901 | 0.983 | 0.797 | 0.972 | 0.827 |
| SP | 0.835 | 0.956 | 0.923 | 0.823 | 0.937 | 0.965 | 0.839 | 0.964 | 0.993 |
| *Concentrated spillover effects* | | | | | | | | | |
| Placebo | 0.000 | 0.000 | 0.000 | 0.461 | 0.502 | 0.448 | 0.465 | 0.434 | 0.464 |
| Andrews | 0.651 | 0.765 | 0.329 | 0.704 | 0.754 | 0.542 | 0.680 | 0.746 | 0.737 |
| SP | 0.860 | 0.932 | 0.991 | 0.957 | 0.918 | 0.967 | 0.834 | 0.816 | 0.853 |
| *Spreadout spillover effects* | | | | | | | | | |
| Placebo | 0.000 | 0.000 | 0.000 | 0.348 | 0.378 | 0.331 | 0.305 | 0.255 | 0.294 |
| Andrews | 0.337 | 0.403 | 0.277 | 0.563 | 0.414 | 0.278 | 0.406 | 0.309 | 0.343 |
| SP | 0.866 | 0.978 | 0.981 | 0.969 | 0.950 | 0.991 | 0.909 | 0.985 | 0.974 |

Notes: SP is the estimation procedure proposed in this paper that takes spillover effects into account. *No spillover effects* stands for the cases where the true DGP has no spillover effects. *Concentrated spillover effects* is the case where 1/3 of the control units receive a spillover effect. *Spreadout spillover effects* is the case where 2/3 of the control units receive a spillover effect of the same level.

# CHAPTER 3

# TWO-SAMPLE TEST

## 3.1 Abstract

Empirical cumulative distribution functions (ECDFs) have been used to test the hypothesis that two samples come from the same distribution since the seminal contribution by Kolmogorov and Smirnov. This paper describes a statistic which is usable under the same conditions as Kolmogorov-Smirnov, but provides more power than other extant tests in that vein. I demonstrate a valid (conservative) procedure for producing finite-sample p-values. I outline the close relationship between this statistic and its two main predecessors. I also provide a public R package (CRAN: `twosamples`[1]) implementing the testing procedure in $O(N \log(N))$ time with $O(N)$ memory. Using the package's functions, I perform several simulation studies showing the power improvements.

## 3.2 Introduction

Determining whether two samples came from the same distribution is an old problem with constant relevance. Particularly when two distributions may have the same

1. Package was published in 2018.

mean, but differ in other important ways, testing their similarity can be both difficult, and critical. In this paper, I characterize a new testing procedure for this situation, which builds on the literature started by Kolmogorov and Smirnov.

Consider the following situation: there are two independent samples: $A$ and $B$, of sizes $n_a$ and $n_b$. Within each sample, all observations are independently drawn from the same distribution: $a \overset{iid}{\sim} E$ and $b \overset{iid}{\sim} F$. Our null hypothesis is that the two (cumulative) distributions are the same, $H_0 : E = F$. Without making any further assumptions, we would like a valid (and ideally consistent/powerful) test of this hypothesis.

Validity in the testing setting refers to a testing procedure which has the correct rejection rate when the null is true. That is to say, when we set a critical value of 5%, the test should only reject 5% of the time if the null hypothesis is true. Consistency refers to the ability of the test to detect small differences in the limit. For a consistent test, there is a sample size beyond which it rejects the null with high probability, when the null is false. I prove validity for this test statistic, and I will outline a proof that the test is consistent – able to eventually detect any differences between two CDFs.[2] But two tests which detect a difference asymptotically may still have massively difference performance. Power is how we discuss performance differences in pre-asymptotic samples. In some situations, there are already tests

2. This is weaker than being consistent for any difference between two distributions. The PDF, not the CDF, uniquely identifies a distribution. By the same token, there are different PDFs which do not cause differences in the PDF. No test based solely on the ECDF will be able to detect the difference between such distributions.

which have been proven to be maximally powerful, and I will compare directly to those tests. The primary benefit of this new test is greater power across a wide range of possible differences between distributions.

There are several non-standard use cases which readers may be interested in. I discuss using weighted observations, parallelizing, and comparisons to a known null distribution in Appendix A. In Appendix B I discuss code runtime and memory usage, as well as showing some real world runtime data.

The rest of the introduction will introduce the extant testing procedures in the literature, and compare their methods in a single example simulation. That simulation consists of one sample from a standard normal, and another sample from a $N(0.5, 4)$.

### *3.2.1 Test Statistics*

Kolmogorov and Smirnov were the first two to study this problem Kolmogorov [1933], Smirnov [1948]. Their statistic calculates the two empirical cumulative distribution functions, takes the difference, and finds the maximal absolute value of the resulting function.

$$KS = \max_{x \in \mathbb{R}} |\hat{F}(x) - \hat{E}(x)|$$

They also used innovative techniques to find the resulting asymptotic distribution, and generate p-values. Figure 3.1 shows a black line, the height of which is the KS test statistic in that simulation. Other versions of the KS statistic are one-sided in

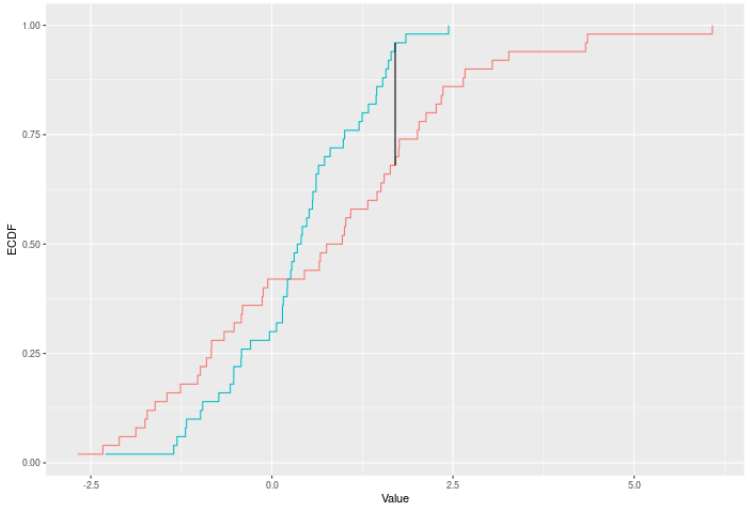


Figure 3.1: A demonstration of the Kolmogorov-Smirnov statistic – the height of the black line is the KS stat. The other two lines represent the ECDFs of two independent samples.

the sense that they focused solely on $max(\hat{F}(x) - \hat{E}(x))$ or $min(\hat{F}(x) - \hat{E}(x))$ – thus focusing their power on mean shifts in either direction.

Later researchers realized that by focusing solely on the maximum value, power against other hypothesis was being lost. The Kuiper test sums the max and min values of that difference.Kuiper [1960]

$$Kuiper = \max_{x \in \mathbb{R}}(\hat{F}(x) - \hat{E}(x)) + \min_{x \in \mathbb{R}}(\hat{F}(x) - \hat{E}(x))$$

This provides more power against possible variance changes – which produce the situation in figure 3.2. The Kuiper stat would be the sum of the heights of the two black lines.

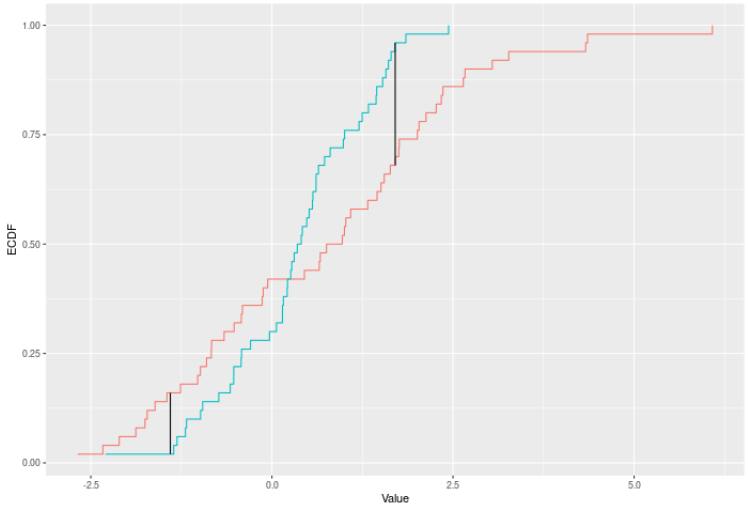


Figure 3.2: A demonstration of the Kuiper statistic – the sum of the heights of the black lines is the Kuiper stat. The other two lines represent the ECDFs of two independent samples.

Cramer and Von Mises developed this further.Cramér [1928], von Mises [1928] Their test statistic takes the sum of all the observed differences between the two ECDFs. Denoting the combined sample $X$, it can be written as follows.

$$CVM = \sum_{x \in X} |\hat{F}(x) - \hat{E}(x)|$$

Figure 3.3 shows an example of this. The CVM test statistic would be the sum of each black line.

At this point, Anderson and Darling [1952] noticed a central issue. Under the null, the variance of $\hat{F}(x) - \hat{E}(x)$ is not remotely stable across $x$. At any given point $x_0$, $n_a \hat{E}(x_0) \sim Binomial(n_a, E(x_0))$. That is to say, if the true CDF makes it such that $P[x < x_0] = 0.5$ for any one observation, then the ECDF, which is the fraction which

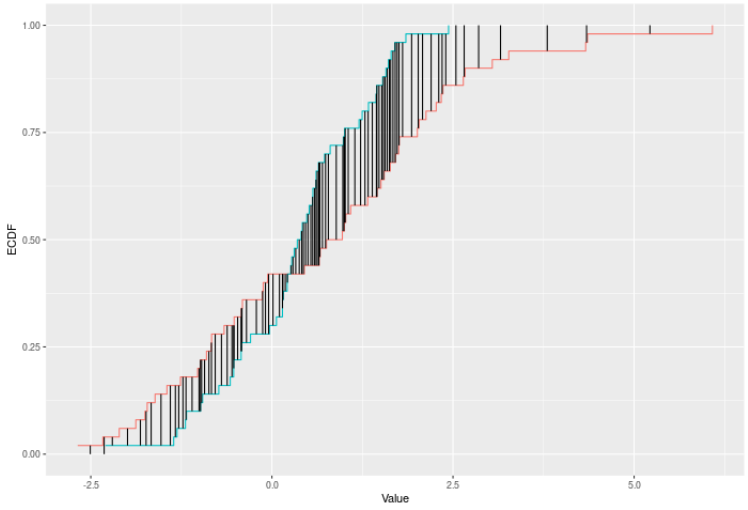


Figure 3.3: A demonstration of the Cramer-Von Mises statistic – the sum of the heights of all the black lines is the CVM stat. The other two lines represent the ECDFs of two independent samples.

are less than a given point, is distributed like a binomial with that same probability. And we know that binomial distributions have variance $np(1-p)$, which is maximized when $p = 0.5$. We also know that in general, putting less weight on high variance observations, and more on low-variance observations will improve the power of a procedure. This leads naturally to the question – how do we estimate the variance to adjust for in a two-sample version of the test? The answer is to compensate for the predicted variance of the combined sample's ECDF (denoted $\hat{D}$ below) at each point. It turns out, that under the null that the two samples come from the same distribution, this is the best estimate of the variance we can get at each point, and in the limit, it is a constant multiplier away from the true variance of $\hat{F}(x) - \hat{E}(x)$. Their estimator is:

$$AD = \sum_{x \in X} \frac{|\hat{F}(x) - \hat{E}(x)|}{\hat{D}(x)(1 - \hat{D}(x))}$$

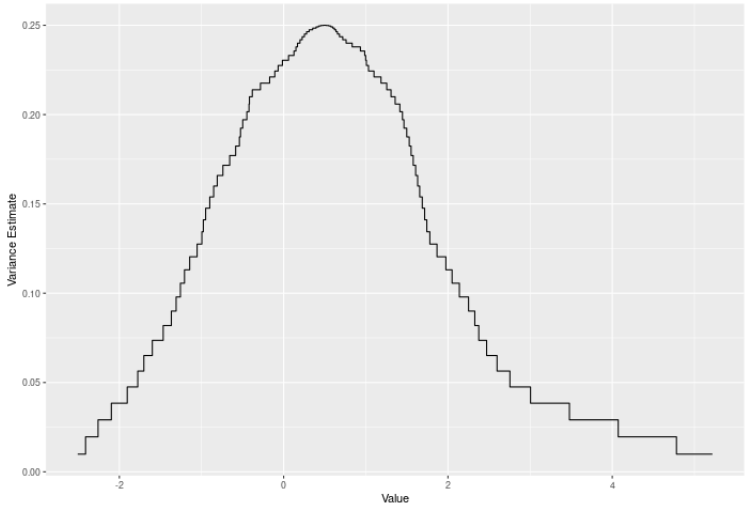


Figure 3.4: A plot showing how the variance of $\hat{F}(x) - \hat{E}(x)$ varies over $x$

Figure 3.4 plots the predicted variances of the combined sample we've been examining. Figure 3.5 shows a plot very similar to the CVM plot above, but where the bars are shaded by the weight they will recieve in the estimation routine.

In a different area of statistics, Wasserstein developed a metric closely related to optimal transport problems.Vaserstein [1969] Broadly, the question "how little can I have to move to get from this position to that?" turns out to be related to the question of distance metrics between pdfs. The answer is not merely the sum of the distances between the ECDFs at each point, but the integral of the distance between ECDFs. Intuitively, this is a statement that the distance between observations is important and should be considered in this framework. The estimator is:

$$wass = \int_{-\infty}^{\infty} |\hat{F}(x) - \hat{E}(x)| dx$$

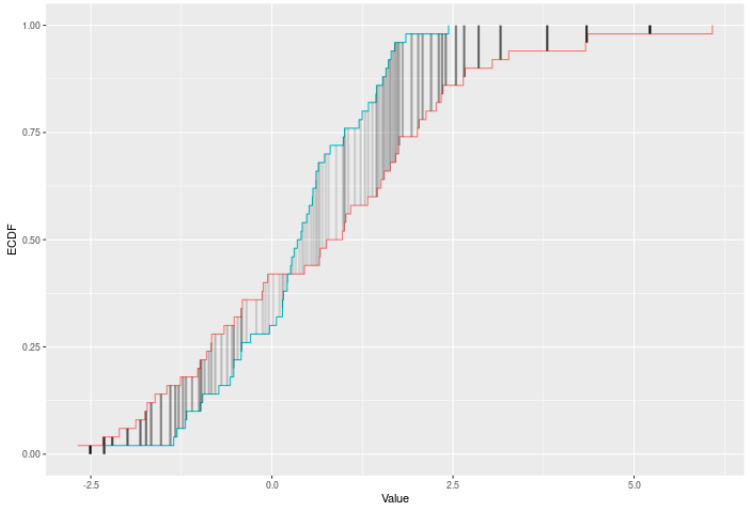


Figure 3.5: A demonstration of the Anderson-Darling statistic – the weighted sum of the dark vertical lines is the AD stat. The color of those lines represents the weight each line will get. The other two lines represent the ECDFs of two independent samples.

We can see what this looks like in figure 3.6.

In building up to the last two estimators, substantial power to detect differences between two distributions has been wrung out of the simple ECDF framework. However, each of the last two contributions has moved in a different direction – each incorporating a different important bit of information. The test statistic I provide here is a synthesis of these two distinct strands in the literature. It combines the Wasserstein notion of distance as important with the Anderson-Darling realization that the variance of the estimator is changing rapidly.

$$DTS = \int_{-\infty}^{\infty} \frac{|\hat{F}(x) - \hat{E}(x)|}{\hat{D}(x)(1 - \hat{D}(x))} dx$$

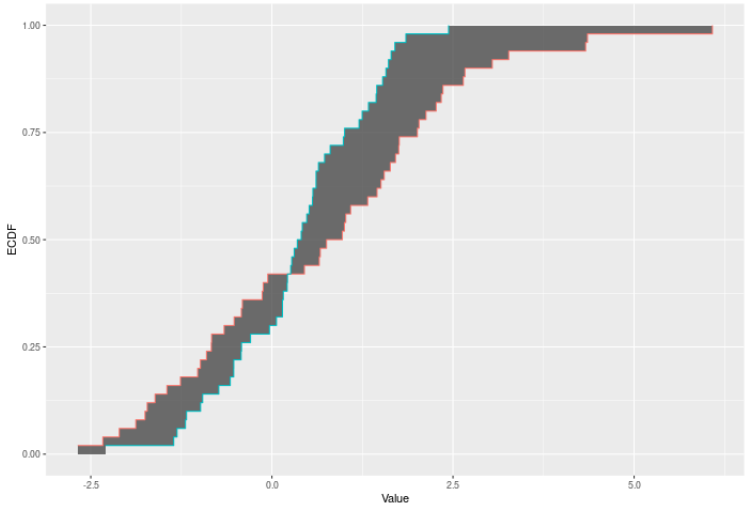


Figure 3.6: A demonstration of the Wasserstein statistic – the two colored lines represent the ECDFs of two independent samples, and the Wasserstein statistic is the area between them.

Broadly, this serves to offer substantial power improvements in many situations over both the Wasserstein test and the AD test, as I'll show in simulations below. Figure 3.7 shows what the test statistic looks like in an example, where the weights are being represented by the shading of the area being integrated.

## 3.3 Theory

Before diving into the proper theoretical results a few notes about the applicability of the test are in order. Broadly, this test will work with any two samples that are made up of ordered data. Unlike the Chi-squared test, it cannot test categorical data. However, for data which is purely ordered, with no meaningful distances

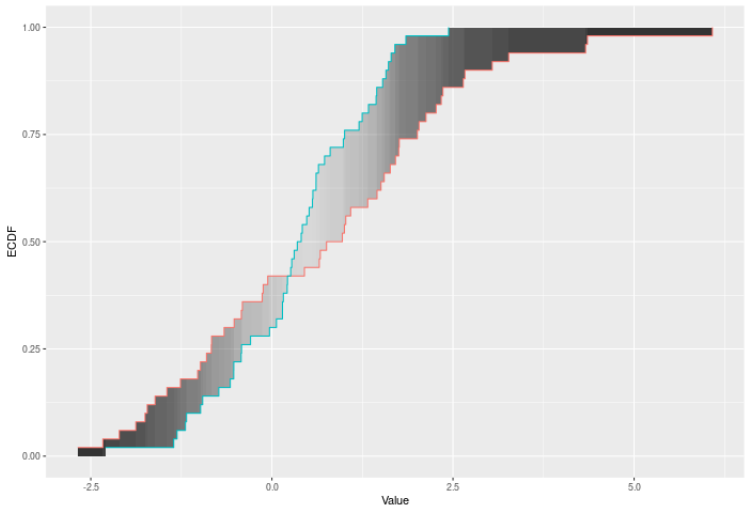


Figure 3.7: A demonstration of the DTS statistic – the two colored lines represent the ECDFs of two independent samples, and the DTS statistic is the weighted integral of the their difference. The color of each region represents the weight it receives.

between the ranks, the Anderson-Darling test is the best bet. Nevertheless, if the implied distances between ranks are all the same (e.g. if we give the test the ranks of the observations in a sample), it can be shown that the AD test and the DTS test will give the same result.

As for distributional features which make this test more or less appropriate, I am not aware of any features that would break this test. Bounded domains, non-continuous PDFs, and more all lead to the same valid testing procedure.[3] There are of course some assumptions that can be made allowing a more powerful testing regime. Assuming that the two samples have smooth PDFs may allow for entirely different constructions. I do not compare to such tests in simulation or the rest of this paper,

3. Even though, as discussed below, some discontinuities in a PDF will be undetectable, the test will remain valid.

as the comparison is not apples-to-apples.

### *3.3.1 DTS Test Validity*

Broadly, to get from the test statistics to a full testing procedure requires finding the sampling distribution of the test statistic and then comparing our observation to that distribution. Fortunately, in our setting, this will be quite straightforward. As all the observations are independent from eachother, and under the null both distributions are the same, we get a nice exchangeability condition. Specifically, under the null, our observed sample labels are just as likely as any other permutation of labels from the joint sample.

Thus any permutation of labels is another draw from the same sampling distribution. We can generate a p-value by permuting labels a large number of times, and then finding the quantile in that distribution that the observed test statistic represents. Because "smaller distance between observed samples" is not indicative of anything we care about, this will be a one-sided test statistic – i.e. we only need to compare to the right tail.

Further, because each of the permutations generated in this way is equally likely under the null, the rank order of our observed test statistic in the set of possible permuted test statistics is uniformly distributed over the possible rank orders. The test will only reject when there are sufficient unique test statistic values to generate a 1-in-20 event,[4] and we observe that event.

4. Or whatever level you set your critical threshold to.

Thus, under the null, the observed test statistic is drawn from a distribution of equally likely test statistics, which we can ennumerate. In order to reject the null with confidence level $\alpha$, the test statistic must be larger than $100 - 100\alpha\%$ of the ennumerated values, which will only happen $100\alpha\%$ of the time. Therefore the test is valid under the null.

### *3.3.2 DTS Test Consistency*

Proving test consistency is a bit trickier than proving test validity, so I will only outline such a proof here. Broadly there are two questions: what is the distribution of the test statistic, and what is the distribution the test statistic is compared to. Under the null, both of these distributions are the same, which substantially eases the proof of validity. However, to show consistency, we must assess these two separately.

As I discussed in the introduction, this test will only be consistent for distributions which have different CDFs. This is not the same as the distributions being different, as two different PDFs can have the same CDFs – and be totally undetectable to testing regimes relying on the ECDF. Also importantly, we can only detect distinctions between CDFs that are not measure 0. A 0-width difference between two CDFs will be indetectable even in the limit – something that is true for all other ECDF based test statistics.

For starters, we are given two samples, $\mathbf{e} \sim E$ (length $n_e$) and $\mathbf{f} \sim F$ (length $n_f$. Given those $n = n_e + n_f$ observations, we can estimate the CDF for each using the ECDF, defined as

$$\hat{E}_n(x) = \frac{1}{n_e} \sum_{i=1}^{n_e} \mathbb{1}\left[e_i \leq x\right]$$

and

$$\hat{F}_n(x) = \frac{1}{n_f} \sum_{i=1}^{n_f} \mathbb{1}\left[f_i \leq x\right]$$

We can also combine the two samples into one sample of length $n$, $\mathbf{d}$ which is drawn from a distribution $D$ which is a mixture of $F$ and $E$ dependent on the relative sample sizes $n_f$ and $n_e$. We can estimate the CDF $D$ the same way.

$$\hat{D}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[d_i \leq x\right]$$

Our test statistic is defined as:

$$DTS_{observed} = \int_{-\infty}^{\infty} \frac{|\hat{E}(x) - \hat{F}(x)|}{\hat{D}(x)(1 - \hat{D}(x))} dx$$

A minor problem occurs for values $x$ outside $range(\mathbf{d})$. Namely, for those values, $\hat{F}(x) = \hat{E}(x) = \hat{D}(x) \in \{0, 1\}$, which in turn means that $|\hat{E}(x) - \hat{F}(x)| = \hat{D}(x)(1 - \hat{D}(x)) = 0$. Thus, for values of $x$ outside the range of the data, the interior of the

integral above becomes $\frac{0}{0}$. In practice, defining $\frac{0}{0} = C < \infty$ for any constant $C$ will simply shift both our observed test statistic and all permutation statistics by the same amount – and so for simplicity, we take $C = 0$. Defining $\mathbb{H}_n = range(\mathbf{d})$ and its complement ${\mathbb{H}_n}^C$ we can rewrite our integral as:

$$\begin{aligned}
DTS_{observed} &= \int_{-\infty}^{\infty} \frac{|\hat{E}_n(x) - \hat{F}_n(x)|}{\hat{D}_n(x)(1 - \hat{D}_n(x))} dx \\
&= \int_{\mathbb{H}_n} \frac{|\hat{E}_n(x) - \hat{F}_n(x)|}{\hat{D}_n(x)(1 - \hat{D}_n(x))} dx + \int_{{\mathbb{H}_n}^C} \frac{|\hat{E}_n(x) - \hat{F}_n(x)|}{\hat{D}_n(x)(1 - \hat{D}_n(x))} dx \\
&= \int_{\mathbb{H}_n} \frac{|\hat{E}_n(x) - \hat{F}_n(x)|}{\hat{D}_n(x)(1 - \hat{D}_n(x))} dx + \int_{{\mathbb{H}_n}^C} \frac{0}{0} dx \\
&= \int_{\mathbb{H}_n} \frac{|\hat{E}_n(x) - \hat{F}_n(x)|}{\hat{D}_n(x)(1 - \hat{D}_n(x))} dx + \int_{{\mathbb{H}_n}^C} 0 dx \\
&= \int_{\mathbb{H}_n} \frac{|\hat{E}_n(x) - \hat{F}_n(x)|}{\hat{D}_n(x)(1 - \hat{D}_n(x))} dx + 0 \\
&= \int_{\mathbb{H}_n} \frac{|\hat{E}_n(x) - \hat{F}_n(x)|}{\hat{D}_n(x)(1 - \hat{D}_n(x))} dx
\end{aligned}$$

The distribution of the test statistic is the easier of the two notions to examine. We know that ECDFs converge almost surely to the CDF of the underlying distribution, so $\hat{E}(x) \to E(x)$, and likewise $\hat{F}(x) \to F(x)$. Holding the ratio of sample sizes constant asymptotically,[5] our joint ECDF also converges almost surely to the CDF of the mixture of the two distributions, $\hat{D}(x) \to D(x)$. Thus, in the limit, our test statistic $\hat{DTS} = \int |\hat{F}(x) - \hat{E}(x)|/(\hat{D}(x)(1 - \hat{D}(x)))$ will converge in probability to

5. This is critical for there to be a stable mixture for the joint ECDF to converge towards.

the true distance between the distributions $\gamma = \int |F(x) - E(x)|/(D(x)(1 - D(x)))$. This distance is strictly some positive number whenever $E \neq F$ over some measurable range. We can prove this because if $E \neq F$ over a measurable range, there must be some range of $x$ for which $|F(x) - E(x)|/(D(x)(1 - D(x)) \neq 0$. At the same time, by construction there can be no $x$ for which that same expression is negative. Thus the integral of that expression, $\gamma$, must be positive if $E \neq F$ over any measurable set.

We can now state that if $E \neq F$, $\hat{DTS} \to \gamma > 0$. However, to make claims about consistency, we need to know about the distribution to which $\hat{DTS}$ will be compared.

In the permutation procedure outlined above, we take our joint sample and permute the labels indicating from which sample each observation came, which keeps sample sizes constant. In the limit, with each permutation we are essentially taking two independent samples from the mixture distribution characterized by $D(x)$. As $n \to \infty$, each of those samples ECDFs will converge towards the ECDF of the mixture $\hat{D}_1(x), \hat{D}_0(x) \to D(x)$. Denoting the permuted observations $\mathbf{e}_{(j)}$ and $\mathbf{f}_{(j)}$, we can estimate their respective ECDFs $\hat{E}_{(j),n}$ and $\hat{F}_{(j),n}$. Finally, the joint sample stays the same, and so we continue to use the same denominator. Thus for each permutation $j$, we observe the permuted test statistic:

$$DTS_{(j),n} = \int |\hat{E}_{(j),n}(x) - \hat{F}_{(j),n}(x)|/(\hat{D}_n(x)(1 - \hat{D}_n(x)))$$

We know that $\hat{E}_{(j),n} \overset{a.s.}{\to} D$ and likewise $\hat{F}_{(j),n} \overset{a.s.}{\to} D$. Thus, in the limit, $|\hat{E}_{(j),n}(x) - \hat{F}_{(j),n}(x)| \overset{a.s.}{\to} 0$. As the numerator converges almost surely to 0, the permuted test statistics will converge to 0.

The final step is to link the two statements. We know that the each resampled test statistic, $\hat{DTS}_{(j),n}$ is converging in probability to 0, and we know that our observed test statistic, $\hat{DTS}_n$ is converging in probability $\gamma$, which is positive when $E \neq F$ over a measurable set. In order to complete the proof, we need to establish the relationship between $DTS_n$ and $DTS_{(j),n}$. Assuming that conditional on a sample and in the limit $DTS_n \perp DTS_{(j),n}$, we can say that $P[DTS_n > DTS_{(j),n}|\mathbf{d}] = P[DTS_n > 0|\mathbf{d}]P[DTS_{(j),n} = 0|\mathbf{d}] \to 1$. Extending that probability statement across possible samples $\mathbf{d}$ , and demonstrating the independence holds are areas for further work. The independence within a sample should derive from the randomness creating the permuted labels, but formally demonstrating this is incomplete.

## 3.4 Simulation Results

To demonstrate the strengths of each test, I've run several simulations, leveraging the speed of the `twosamples` package which implements these test statistics. In the first set of simulations, I compare all the tests described above[6] as we change the amount of information observed. Two samples are generated, then each test runs on those two samples, then the whole process is repeated several thousand times to get

6. As well as either the T-test or F-test depending on context.

estimates for the power of the testing procedures.

All of the test statistics above are calculated in a standard manner, however the p-values are calculated using the resampling logic described in the theory section above, so that they should all be perfectly sized under the null. For some, such as the Kolmogorov-Smirnov test, this gives substantial power improvements relative to the usual asymptotic p-values – which are known to be very conservative in smallish samples. It should be clear that without ensuring that all the tests had the same size control, the comparisons between tests wouldn't be fair. I do not make this adjustment for the T-test and F-test – as they are very standardized, and I want to compare to the standard implementation.

For each simulation exercise, I plot every test's power across the parameter that is changing. The color assigned to each test is ordered by the mean power that test had in the simulation – so that by looking at the legend, we can see which tests performed best.

The first four simulation exercises are simple comparisons of shifted means and inflated variances – situations which most tests should have been benchmarked on previously. Even here we see the DTS test showing performance improvements beyond its predecessors on the variance changes, while remaining competitive at detecting mean shifts. The later simulations involve both mean and variance changes, or mixtures of normals. In all of those simulations, the DTS test outperforms every other test run. At times the DTS test demonstrated a power 1.8 times as much as the best

of the other advanced ECDF tests.[7]

### *3.4.1 Simulations Across Parameter Values*

In this section I look at two simple simulations showing the rejection rate as the difference between two sample's distributions grows. In each example, the leftmost point is a no-difference condition, so the rejection rate there represents the size of the test. As we will see, that rate consistently equals 5% – the nominal test size – so all of these tests are properly calibrated. For each of these, both samples will have a sample size of 50 – representing a total sample of 100 observations every time a test is run.

#### Mean Shifting

In this example, the first distribution is a standard normal, and the second distribution is a $N(\mu, 1)$, where $\mu$ is on the x-axis. Because it is provably optimal in the circumstances, I also include a T-test. See Figure 3.8 for results. Broadly we can see that performance for all the tests except KS and Kuiper was very comparable.

7. I exclude the Kuiper and KS tests from the 'advanced' category for their abysmal performance on the first four simulations. 71% rejection/38% rejection = 1.87.

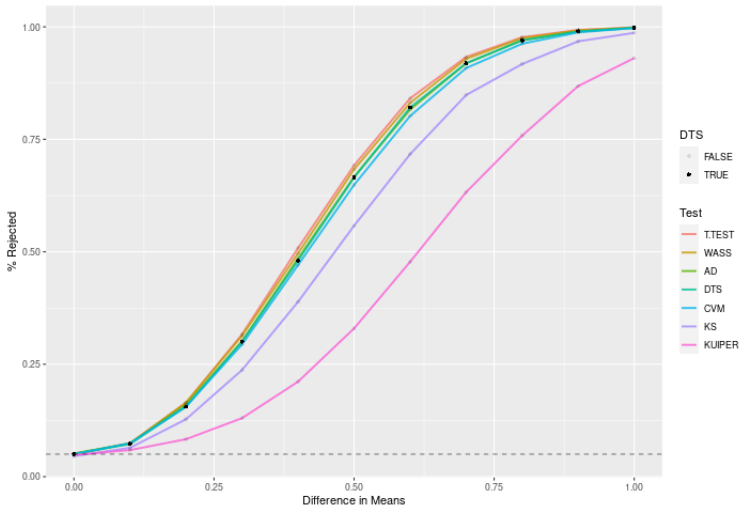


Figure 3.8: Rejection rates for different two-sample tests as the mean changes. $N = 50$ for both samples. When the difference in means is 0, this is the test size.

### Variance Inflation

In this simulation, one sample is from a standard normal, and the other is from a $N(0, \sigma^2)$, where $\sigma^2$ is on the x-axis. Because it is provably optimal in this circumstance, I also include an F-test.See Figure 3.9 for results. We see immediately that the F-test does very well. DTS follows at some distance, with Kuiper and Wasserstein a little behind it. The rest of the tests lag behind.

## *3.4.2 Simulations Across N*

In this section I look at five simulations, as the size of both samples grow.

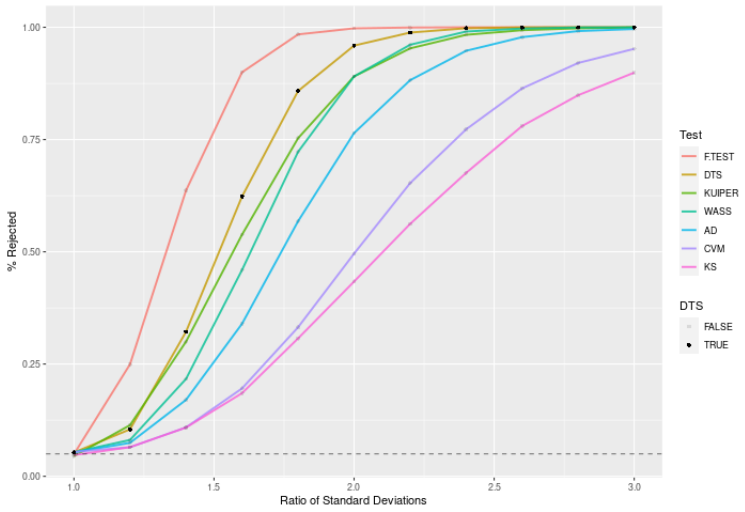


Figure 3.9: Rejection rates for different two-sample tests as the variance changes. $N = 50$ for both samples. When the ratio of variances is 1, this is the test size.

## Mean Shift

In this simulation, one sample is from a standard normal distribution, and the other is from a $N(1, 1)$, i.e. the same distribution with a mean shift. Because it is provably optimal in the circumstances, I also include a T-test for comparison. See Figure 3.10 for results. Much like the mean simulation above, we can see that the KS and Kuiper tests lag behind, while the rest of the tests are very comparable, and indeed near the optimal power of the T-test.

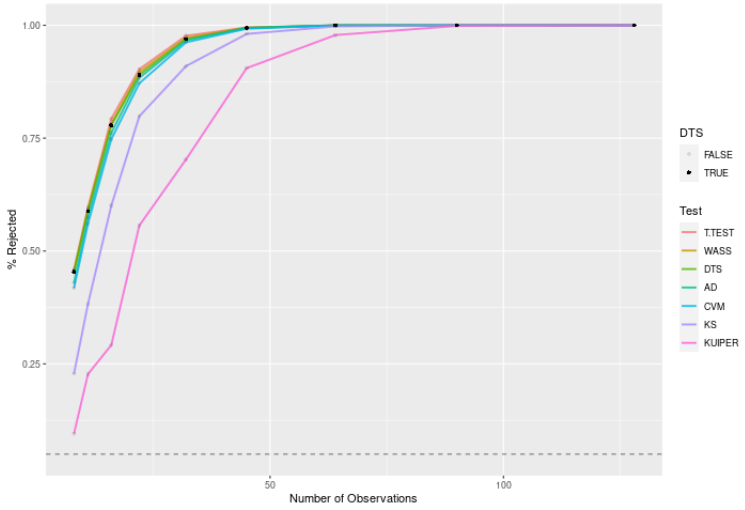


Figure 3.10: Rejection rates for different two-sample tests as the sample size changes. The difference in means between the samples is 1.

## Variance Inflated

In this DGP, one sample is from from a standard Normal, and the other is from a $N(0, 4)$, i.e. the same distribution but with larger variance. Because it is provably optimal under the circumstances, I also include an F-test for comparison. See Figure 3.11 for results. Much like the variance simulation above, we can see that the F-test is far and away the best test, with DTS lagging it and followed by Kuiper and Wasserstein tests.

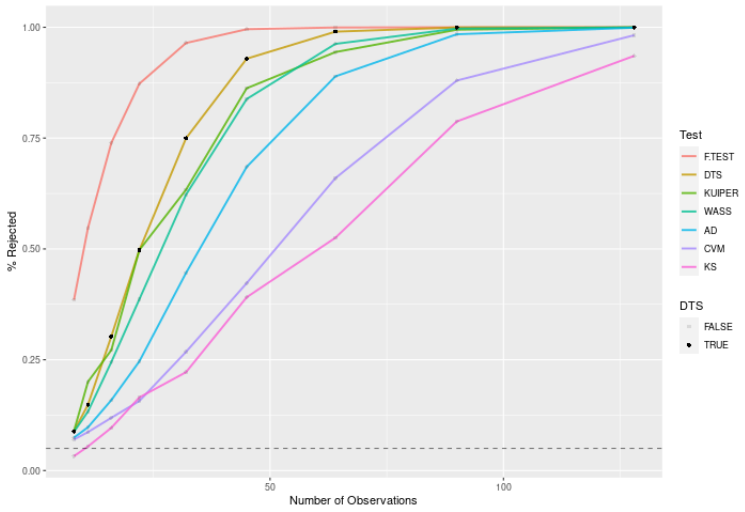


Figure 3.11: Rejection rates for different two-sample tests as the sample size changes. The ratio of variances between the two samples is 4.

## Mean and Variance Shift

In this DGP, one sample is from a standard Normal, and the other is from a $N(0.5, 2.25)$. I also include a $t_test$ for comparison, though strictly speaking it is not testing the same null hypothesis. See Figure 3.12 for results. In this test, we see the T-test fall away, with its performance decaying substantially relative to the competition. DTS takes a consistent lead, with Wasserstein and Anderson-Darling following.

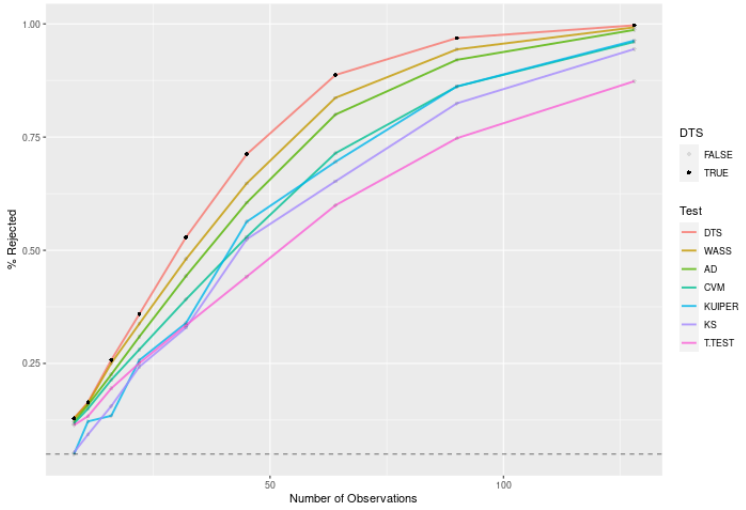


Figure 3.12: Rejection rates for different two-sample tests as the sample size changes. One sample is a standard normal and the other is a $N(0.5, 1.5)$.

## Mixtures

In this section I look at a much harder to detect class of distributions – mixtures. The first sample in each dgp will be a standard normal, while the second sample in each will be a mixture of two normals. However, in order to make the problem more difficult, the mixture itself is recentered and scaled so that it has a mean of 0 and variance of 1 – leaving only the higher moments to identify it as different from the standard normal. Because these are harder differences to detect, I've inflated the sample sizes.

The first mixture is a $N(0.8, 1)$ with probability 0.2, and a $N(-0.2, 1)$ with probability 0.8. See Figure 3.13 for results. On the whole, the DTS test has the best performance, followed by the Kuiper test. As expected, the t-test detects no differ-

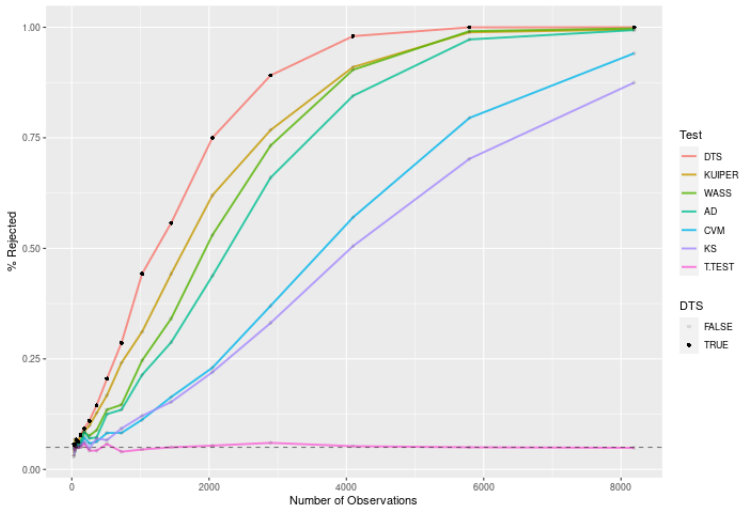


Figure 3.13: Rejection rates for different two-sample tests as the sample size changes. The second sample is a mixture of two normals with different means, centered so that the overall mixture has a mean of 0.

ence in the means of the two samples.

The second mixture is a $N(0, 0.625)$ with probability 0.2, and a $N(0, 2.5)$ with probability 0.8. See Figure 3.14 for results. Again we see the DTS test outperforming everything else, followed by the Kuiper test. By the time the DTS test had power to detect the difference 71% of the time, the Wasserstein test was only detecting a difference 38% of the time, and other advanced tests were even worse. Thus at that point, 1-in-3 simulations the DTS test could reject the null while other advanced tests couldn't.

The third mixture is a $N(0.8/1.7607^{0.5}, 4/1.7607)$ with probability 0.2 and a $N(-0.2/1.7607^{0.5}, 1/1.7607)$ with probability 0.8 – which holds the mean at 0 and variance at 1. See Figure 3.15 for results. Again, we see the DTS test as the most

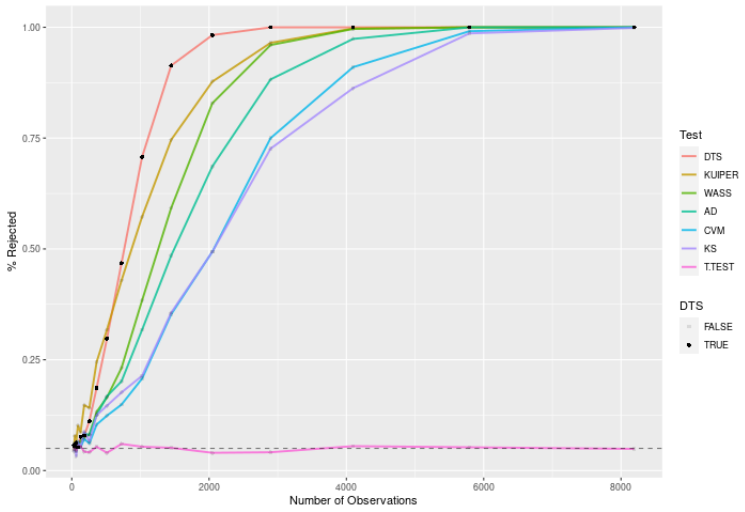


Figure 3.14: Rejection rates for different two-sample tests as the sample size changes. The second sample is a mixture of two normals with different variances, scaled so that the overall mixture has a variance of 1.

powerful, followed by the Kuiper test. Here we see that when the DTS test can reject 71% of the time, the Wasserstein test only rejects 47% of the time – so that in nearly 1-in-4 samples at that sample size, the DTS test rejected while other advanced tests couldn't.

## 3.5 Conclusion

The DTS test is a powerful tool for detecting differences between two samples. Particularly when we don't know how what kind of difference exists, or how much power to detect that difference we will have, the DTS test shows promise, with uniformly large power against a wide variety of distributional test statistic. As shown in sim-

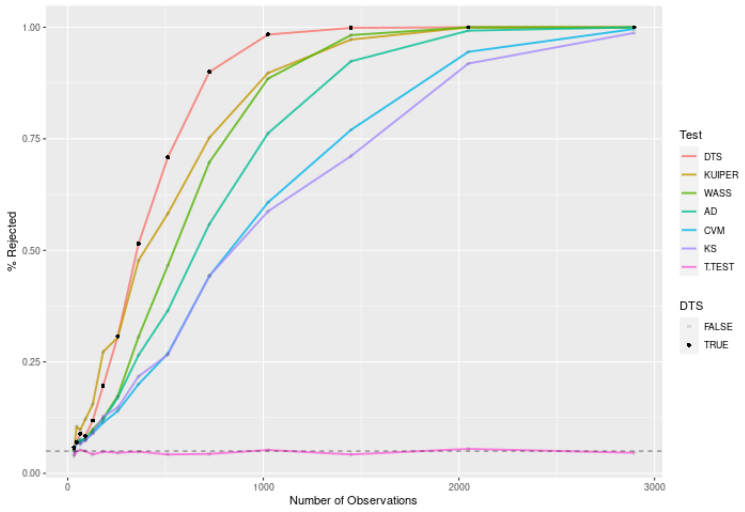


Figure 3.15: Rejection rates for different two-sample tests as the sample size changes. The second sample is a mixture of two normals with different means and variances, centered so that the overall mixture has a mean of 0, and scaled so that the overall mixture variance is 1.

ulations above, it offers power which exceeds that of other tests making comparably few assumptions about the data. Moreover, while other test statistics like the Kuiper test tended to do well in some simulations and poorly in others, the DTS test was consistently at the top, or very close to it. This suggests that while it is not uniformly more powerful than other tests, the tradeoff it manages sacrifices very little power in some situations, to gain substantial power in most situations. However, it is by no means the last word. While the DTS stat adjusts appropriately for the variance in $\hat{F}(x) - \hat{E}(x)$, it is not clear that it adjusts correctly for the variance in $\int \hat{F}(x) - \hat{E}(x)$.

A different perspective on the situation could be desribed as the following. The chi-squared test is proven optimal for this task when the values are sample from a

distribution which takes nominal values. The Anderson-Darling test, under the same assumptions, is optimal for samples from distributions with merely ordinal values. The DTS test, under the same assumptions, offers substantial improvements towards optimality for distributions with interval values. Work towards finding the truly optimal test in this setting may require making additional technical assumptions to optimally accommodate the variance in the width of each weighted cell that the DTS stat adds. Not to mention the substantial work I've left untouched simply proving things like consistency, and power improvements over other tests. But work could also progress to the next step in the measurement framework – looking for tests which make sense for a ratio scale. It seems likely that either taking the log of all the values before using the DTS stat, or taking the log of the widths between observations is a good step towards such a framework – but that is a question for someone else.

## 3.6 Alternate Uses

This section will address some nonstandard uses of the code in `twosamples`. Specifically, it will look at parallelization, testing against a known distribution, and weighting observations.

### *3.6.1 Parallelizing*

To keep things simple, `twosamples` does not internally support parallelizing the test statistics. However, the outputs from any of the `_test` functions make it easy to run your own parallelized code. The p-values the code outputs are a count of the number of more extreme values of the test statistic observed, divided by the number of resamplings performed. If you wish to run many resamplings across cores, you could run the same code (e.g. `dts_test(x,y)`) on each core. So long as you keep the number of resamples the same on each core, the average of your p-values will give you a new, valid p-value. If you wish to change the sample sizes, you merely need to find the appropriately weighted average of the p-values.

This works because when the number of resamples in each core is the same, the number of resamples automatically drops to the outside of the 'correct' sum procedure.

$$\frac{1}{n_{cores}} \sum_{i=1}^{n_{cores}} pval_i = \frac{1}{n_{cores}} \sum_{i=1}^{n_{cores}} \frac{extreme_i}{n_{resamples}} = \frac{1}{n_{cores} n_{resamples}} \sum_{i=1}^{n_{cores}} extreme_i$$

A small issue may arise when running the test code in parallel. When the observed test statistic is the most extreme value observed, without ties, the code defaults to a p-value of $1/(2n_{resamples})$, instead of 0. This prevents the test from rejecting the null inappropriately when either the number of observations or number of resamples is quite small. However, when performing the averaging procedure I outlined above,

it would make sense to take any p-values that are $1/(2n_{resamples})$, and convert them into the more accurate 0. Then, after averaging all the p-values together, if the new p-value is 0, you should move it back up to $1/(2n_{cores}n_{resamples})$.

Because each core will calculate the observed test statistic independently, there is some duplication of effort. However, so long as the cores are not doing anything else important, and $n_{cores} > 2$, this can still make sense even when $n_{resamples} = 1$ on each core.

### *3.6.2 Weighted Observations*

Sometimes the data being used is weighted. In principle, incorporating those weights into these functions should be possible. The resampling routine can resample with appropriate weights, and then the observations contributions to the ECDF's height merely need to be adjusted to accomodate the appropriate weights. Someday, I may update the code to incorporate this ability. In the meantime however, I suggest that users desiring an ECDF test for weighted observations, find some integer $k$ such that $\min_i(k * weights_i) = 1$, and all values of $k * weights_i$ are within 0.1 of an integer. Then, create new vectors $A$ and $B$, which contain the elements of $a$ and $b$, $k*weights_i$ times for each element. At that point you can run $dts\_test(A, B)$. This should give each observation the appropriate resampling probability and weight in the ECDF.

### *3.6.3 One Sample Tests*

This entire paper has examined the use of `dts_test` for the two sample problem. However, we should be able to use it to test against known distributions, like the normal. At the moment, to test whether sample A comes from a known distribution, I suggest taking $n_b = k * n_a$ draws from the known distribution to create sample B, with $k \in 10, 100$ and running `dts_test(A,B)`. In practice, when $k = 10$ this compares your sample A to a distribution which is a mixture of 91% the known distribution and 9% the distribution A came from.[8] While there is some power lost in this procedure, relative to comparing our observed test statistic against the true null sampling distribution[9], this is a valid (and consistent) procedure which is easy to implement. In practice, particularly with large $k$, the power loss is quite minimal.

8. The mixture probabilities become 99.1% and 0.9% when $k = 100$.

9. This is what a true one sample version of `dts_test` would look like.

## 3.7 Run Time and Memory Usage

The code in `twosamples` has three main components of relevance to run time. The test function consists of a loop running the test stat function $n_{resamples}$ times. Each iteration of the test stat function requires one sorting of $n_a + n_b$ observations, and then another loop calculating the test statistic using that sorted data, which has $n_a + n_b$ iterations. The sort uses `std::sort`, which operates $O(n \log(n))$ on average,[1011] when denoting $n = n_a + n_b$. The internals of the loop through $n$ points is not of obviously growing complexity, so this is $O(n)$ on average. Thus the test stat function operates $O(n \log(n))$ on average. Therefore the overall test function takes $O(n_{resamples} n \log(n))$ time. Of course, theory and reality rarely interact in the way we would like. Figure 3.16 shows real world execution times as sample sizes change, holding $n_{resamples}$ at 2000.

### *3.7.1 Memory Usage*

The memory usage of the `twosamples` package is approximately $O(N)$. The test function primarily performs a loop which operates sequentially, that loops' only side effect is to increment counters, so from iteration to iteration memory usage shouldn't change. Outside the loop the test function creates a duplicate vector of length $N$.

10. Because we run this many times, the average is the main thing that matters, but `std::sort` claims an $O(n \log(n))$ worst case as well.

11. Special thanks to Github user sircosine for spotting some bugs with speed.

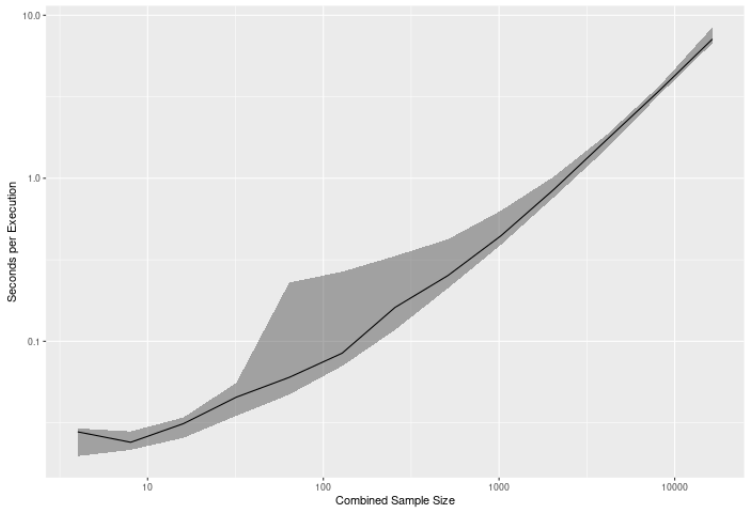


Figure 3.16: Execution time for `dts_test` as $n = n_a + n_b$ grows. The black line is the mean execution time, while the ribbon represents a 95% predictive interval for the execution time in one set of simulations.

Inside the loop, there are 4 vectors created with lengths totalling $2N$, half of which are passed on to the test stat function. That function creates 7 more vectors of length $N$, as well as about a dozen counters, for an anticipated memory use of $7N$. Thus the total memory usage from running the function once should be on the order of $10N$. Planned updates to this code should reduce the internal vector creation by $3N$, for a total memory usage of $O(7N)$. This is theoretical usage, not real-world measured memory, so plan accordingly.

# REFERENCES


Alberto Abadie and Matias D. Cattaneo. Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10(1):465–503, aug 2018.

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

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505, jun 2010.

Douglas Almond, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. Estimating marginal returns to medical care: Evidence from at-risk newborns. *The quarterly journal of economics*, 2010.

Douglas Almond, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. The role of hospital heterogeneity in measuring marginal returns to medical care: a reply to barreca, guldi, lindo, and waddell. *The quarterly journal of economics*, 2011.

Michael L Anderson. As the wind blows : The effects of long-term exposure to air pollution on mortality *, 2015.

T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, 23(2):193–212, 06 1952. doi: 10.1214/aoms/1177729437. URL `https://doi.org/10.1214/aoms/1177729437`.

Donald W. K. Andrews. End-of-Sample Instability Tests. *Econometrica*, 71(6): 1661–1694, 2003.

Donald W. K. Andrews and J. Kim. Tests for Cointegration Breakdown Over a Short Time Period. *Journal of Business & Economic Statistics*, 24(4):379–394, oct 2006.

Joshua Angrist and Miikka Rokkanen. Wanna get away? rd identification away from the cutoff, 2012.

Joshua D Angrist, Victor Lavy, Jetson Leder-Luis, and Adi Shany. Maimonides rule redux. *NBER Working Paper Series*, 2017.

Alan I. Barreca, Melanie Guldi, Jason M. Lindo, and Glenn R. Waddell. Saving babies? revisiting the effect of very low birth weight classification. *The quarterly journal of economics*, 2011.

Guillaume Basse, Avi Feller, and Panos Toulis. Conditional randomization tests of causal effects with interference between units. *arXiv preprint arXiv:1709.08036*, sep 2017. URL `http://arxiv.org/abs/1709.08036`.

Ylenia Brilli and Brandon J Restrepo. Birth weight , neonatal intensive care units , and infant mortality : Evidence from macrosomic babies, 2017.

Sebastian Calonico, Matias D. Cattaneo, and Roćıo Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 2014.

Sebastian Calonico, Matias D. Cattaneo, and Max H. Farrell. Optimal bandwidth choice for robust bias corrected inference in regression discontinuity designs, 2018.

Jianfei Cao and Connor Dowd. Synthetic controls with spillovers: Theory for estimation and inference. *Working Paper*, 2021.

Cattaneo, Idrobo, and Titiunik. A practical introduction to regression discontinuity designs: Foundations, 2019a.

Matias Cattaneo, Brigham Frandsen, and Rocío Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, 2014.

Matias D. Cattaneo, Rocío Titiunik, and Gonzalo Vazquez-Bare. Comparing inference approaches for rd designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*, 36(3):643–681, 2017. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.21985`.

Matias D. Cattaneo, Max H. Farrell, and Yingjie Feng. Large sample properties of partitioning-based series estimators, 2018.

Matias D. Cattaneo, Luke Keele, Rocio Titiunik, and Gonzalo Vazquez-Bare. Extrapolating treatment effects in multi-cutoff regression discontinuity designs, 2019b.

Eduardo Cavallo, Sebastian Galiani, Ilan Noy, and Juan Pantano. Catastrophic Natural Disasters and Economic Growth. *Review of Economics and Statistics*, 95 (5):1549–1561, dec 2013. ISSN 0034-6535. doi: 10.1162/REST_a_00413. URL `http://www.mitpressjournals.org/doi/10.1162/REST{_}a{_}00413`.

Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 2013.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *arXiv preprint arXiv:1712.09089*, dec 2017.

Timothy G. Conley and Christopher R. Taber. Inference with "Difference in Differences" with a Small Number of Policy Changes. *Review of Economics and Statistics*, 93(1):113–125, feb 2011.

Harald Cramér. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928. doi: 10.1080/03461238.1928.10416862. URL `https://doi.org/10.1080/03461238.1928.10416862`.

Gordon B. Dahl, Katrine Vellesen Løken, and Magne Mogstad. Peer effects in program participation. *AER*, 2014.

N. Meltem Daysal. Spillover effects of early-life medical interventions, 2015.

Thomas S. Dee and Emily K. Penner. The causal effects of cultural relevance : Evidence from an ethnic studies curriculum, 2016.

Juan Manuel Ospina Díaz, Nicolás Grau, Tatiana Reyes, and Jorge Rivera. The impact of grade retention on juvenile crime, 2016.

Yingying Dong and Arthur Lewbel. Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 2015.

Nikolay Doudchenko and Guido W. Imbens. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *arXiv preprint arXiv:1610.07748*, oct 2016a.

Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016b.

Arindrajit Dube, Laura Giuliano, Jonathan, and Léonard. Fairness and frictions : The impact of unequal raises on quit behavior, 2015.

Esther Duflo and Emmanuel Saez. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *Framed Field Experiments*, 2003.

Jianqing Fan, Nancy Heckman, and M.P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 1995.

Zheng Fang and Andres Santos. Inference on Directionally Differentiable Functions. *The Review of Economic Studies*, 86(1):377–412, sep 2018.

Bruno Ferman and Cristine Pinto. Placebo Tests for Synthetic Controls, apr 2017.

Bruno Ferman and Cristine Pinto. Inference in differences-in-differences with few treated groups and heteroskedasticity. *The Review of Economics and Statistics*, jul 2018.

Bruno Ferman and Cristine Pinto. Synthetic Controls with Imperfect Pre-Treatment Fit, sep 2019.

Sergio Firpo and Vitor Possebom. Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets. *Journal of Causal Inference*, 6(2), sep 2018. ISSN 2193-3685. doi: 10.1515/jci-2016-0026. URL `http://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2016-0026/jci-2016-0026.xml`.

Dirk Foremny and Albert Solé-Ollé. Who ' s coming to the rescue ? revenue-sharing slumps and implicit bailouts during the great recession, 2016.

Romain Gauriot. Winner effect in dynamic contests, 2014.

Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. Working Paper 20405, National Bureau of Economic Research, August 2014.

François Gerard, Miikka Rokkanen, and Christoph Rothe. Bounds on treatment effects in regression discontinuity designs with a manipulated running variable, 2019.

Ruud Gerards and Pierre M Theunissen. Becoming a mompreneur: Parental leave policies and mothers' propensity for self- employment, 2018.

Sarena Goodman, Adam Isen, and Constantine Yannelis. A day late and a dollar short: Liquidity and household formation among student borrowers, 2018.

Jinyong Hahn and Ruoyao Shi. Synthetic Control and Inference. *Econometrics*, 5 (4):52, nov 2017.

Jinyong Hahn, Petra Todd, and Wilbert H van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 2001.

James D. Hamilton. *Time series analysis*. Princeton University Press, 1994.

Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

Cheng Hsiao and Qiankun Zhou. Panel parametric, semiparametric, and nonparametric construction of counterfactuals. *Journal of Applied Econometrics*, 34 (4):463–481, jun 2019. ISSN 0883-7252. doi: 10.1002/jae.2702. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2702`.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 2004.

A Kolmogorov. Sulla determinazione empirica di una legge di distribuzione, 1933.

Noémi Kreif, Richard Grieve, Dominik Hangartner, Alex James Turner, Silviya Nikolova, and Matt Sutton. Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units. *Health Economics*, 25(12):1514–1528, dec 2016. ISSN 10579230. doi: 10.1002/hec.3258. URL `http://doi.wiley.com/10.1002/hec.3258`.

N. H. Kuiper. Tests concerning random points on a circle, 1960.

David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 2010.

Kathleen T. Li. Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods. *Journal of the American Statistical Association*, pages 1–40, oct 2019. ISSN 0162-1459. doi: 10.1080/01621459.2019.1686986. URL `https://www.tandfonline.com/doi/full/10.1080/01621459.2019.1686986`.

Jason M. Lindo, Nicholas J. Sanders, and Philip Oreopoulos. Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2010.

Jason M. Lindo, Peter M. Siminski, and Oleg Yerokhin. Breaking the link between legal access to alcohol and motor vehicle accidents: Evidence from new south wales. *Health economics*, 2016.

Charles F. Manski. Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, 60(3):531, jul 1993.

Benjamin Marx. The cost of requiring charities to report financial information, 2018.

Justin McCrary. Manipulation of the running variable in the regression discontinuity design : A density test. *The Journal of Econometrics*, 2008.

Michael George Mueller-Smith and Kevin T. Schnepel. Diversion in the criminal justice system : Regression discontinuity evidence on court deferrals, 2017.

Abdul Munasib and Dan S. Rickman. Regional economic impacts of the shale gas and tight oil boom: A synthetic control analysis. *Regional Science and Urban Economics*, 50:1–17, jan 2015.

Whitney K. Newey and Daniel McFadden. Chapter 36 Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, jan 1994.

Ben Ost, Weixiang Pan, and D. B. Webber. The returns to college persistence for marginal students : Regression discontinuity evidence from university dismissal policies, 2016.

Jack Porter. Estimation in the regression discontinuity model. *Monograph*, 2003.

Michael W. Robbins, Jessica Saunders, and Beau Kilmer. 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, jan 2017. ISSN 0162-1459. doi: 10.1080/01621459.2016.1213634. URL `https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1213634`.

Paul R Rosenbaum. Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, 102(477):191–200, mar 2007. ISSN 0162-1459. doi: 10.1198/016214506000001112. URL `http://www.tandfonline.com/doi/abs/10.1198/016214506000001112`.

Alexandra Roulet. Unemployment insurance and reservation wages : Evidence from administrative data, 2016.

Carlos Zamarrón Sanz. Direct democracy and government size : Evidence from spain, 2015.

Judith Scott-Clayton and Lauren Schudde. Performance standards in need-based student aid, 2016.

Firpo Sergio and Possebom Vitor. Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets. *Journal of Causal Inference*, 6(2):1–26, 2018. URL https://ideas.repec.org/a/bpj/causin/v6y2018i2p26n1.html.

Hitoshi Shigeoka. The effect of patient cost sharing on utilization , health , and risk protection. *AER*, 2014.

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19, 1948. doi: 10.1214/aoms/1177730256. URL https://doi.org/10.1214/aoms/1177730256.

M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974. ISSN 01621459. URL http://www.jstor.org/stable/2286009.

Jörg Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009. doi: https://doi.org/10.3982/ECTA7347.

Cody Tuttle. Snapping back: Food stamp bans and criminal recidivism. *American Economic Journal: Economic Policy*, 2019.

Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

Gonzalo Vazquez-Bare. Identification and Estimation of Spillover Effects in Randomized Experiments. *arXiv preprint arXiv:1711.02745*, nov 2017. URL http://arxiv.org/abs/1711.02745.

R. E. von Mises. Wahrscheinlichkeit, 1928.

Yiqing Xu. Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*, 25(01):57–76, jan 2017.

Ming Yu, Varun Gupta, and Mladen Kolar. Constrained High Dimensional Statistical Inference. *arXiv preprint arXiv:1911.07319*, nov 2019. URL http://arxiv.org/abs/1911.07319.

Eduardo H. Zarantonello. Projections on Convex Sets in Hilbert Space and Spectral Theory. In *Contributions to Nonlinear Functional Analysis*, pages 237–424. Elsevier, 1971.

Seth D. Zimmerman. The returns to four-year college for academically marginal students. *Journal of Labor Economics*, 2014.