How to Make Causal Inferences Using Texts*

Naoki Egami[†] Christian J. Fong[‡] Justin Grimmer[§]
Margaret E. Roberts[¶] Brandon M. Stewart[‖]

July 3, 2017

Abstract

Texts are increasingly used to make causal inferences: either with the document serving as the treatment or the outcome. We introduce a new conceptual framework to understand text-based causal inferences, demonstrate fundamental problems that arise when using manual or computational approaches applied to text for causal inference, and provide solutions to the problems we raise. Our work introduces a methodology that connects traditional survey experiment methodology from the social sciences with A/B tests more common in industry and machine learning. Using this framework, we then show that the standard application of text methods leads to an Analyst Induced SUTVA Violation and we show how to resolve the problem using a training and test split. Taken together, our work provides a more rigorous foundation to build upon for applying text-based methods to causal inference.

*We thank Edo Airoldi, Gary King, Rich Nielsen, Thomas Richardson, Hanna Wallach and numerous seminar participants for useful discussions about making causal inference with texts. We also thank Dustin Tingley for early conversations about potential SUTVA concerns with respect to STM and sequential experiments as a possible way to combat it.

[†]Ph.D. Candidate, Department of Politics, Princeton University, negami@princeton.edu

[‡]Ph.D. Candidate, Graduate School of Business, Stanford University, cjfong@stanford.edu

[§]Associate Professor, Department of Political Science, University of Chicago, JustinGrimmer.org, grimmer@uchicago.edu.

[¶]Assistant Professor, Department of Political Science, University of California San Diego, meroberts@ucsd.edu

[‖]Assistant Professor, Department of Sociology, Princeton University, brandonstewart.org, bms4@princeton.edu

1 Introduction

Text data appears in many areas of social science and frequently within causal processes. Judicial opinions, political propaganda, and campaigning are all examples of texts written to persuade or have an effect on a subject or event. Social media posts, politicians' responses, and lawmaking are all affected by the political context, and can sometimes be direct responses to political phenomena or stimuli. Understanding how the text influences or is influenced are natural questions for social science researchers, and social science researchers have long incorporated text data into their own experiments and research designs (Lasswell, 1938; Laver, Benoit and Garry, 2003; Pennebaker, Mehl and Niederhoffer, 2003; Quinn et al., 2010).

While text data and causal inference may be a natural combination, text analysis poses many problems for causal inference because the data are high-dimensional – text can express many different concepts and can express the same concept in many different ways. As a result, the mapping between the text and the concepts expressed in the text that are of interest to the researcher is not always known before the experiment. When we collect texts in an experimental setting we may have no idea ex ante what they will contain and what concepts they will represent, and we may need to read or analyze a selection of them to understand how they map to the quantities that we are interested in measuring.

The need to discover and iteratively define measures and concepts from data is a fundamental component of social science research (Tukey, 1980), yet it also poses problems for causal inference. Because high-dimensional data like text represents many concepts and therefore can be sliced and categorized in many different ways, this discovery process could lead researchers down what Gelman and Loken (2014) have called the garden of forking paths, where we may reach false conclusions because the measures used to test the relationship between treatment and outcome are a function of the data themselves. But without the discovery process, the researcher may be left with imprecise measures of concepts of interest or may be missing important concepts all together. Thus, the researcher struggles to have it both ways: to find and discover new and appropriate measures from the data and to avoid well-known pitfalls of causal inference.

In this paper, we formalize the tradeoff between discovery and testing in causal inference with text as treatment and outcome and provide a simple solution. We define what we call the "analyst-induced SUTVA violation" (AISV) that explains why developing the mapping between the text and categories of interest in the same set of data that is used for measuring a treatment effect violates the SUTVA assumption commonly invoked in causal inference. We propose solving this problem with a simple split of the data that allows the researcher to freely define measures of interest through exploratory analysis of one dataset, then subsequently test them on the held out data to confirm that the relationships are not an artifact of exploration. This train-test split, common to machine learning and A/B tests in industry, approximates a sequential notion of science where first a finding is discovered and in a separate experiment confirmed. We apply this process of sequential science to two applications, one where text is the outcome of a survey experiment, and one where text is the treatment affecting a response.

2 Text For Causal Inference

In this section we introduce a framework for thinking about the role of text in causal inference. Our discussion here will be general, with our experimental framework incorporating several research designs—including open-ended text responses coded into categories with hand coding, conjoint experiments, and treatments with text-based appeals. To anchor our intuition, though, we focus on our two example applications of texts and causal inference: applying the Structural Topic Model (STM) to measure the dependent variable in texts (Roberts et al., 2014), and an experiment that uses texts to understand their causal effects (Fong and Grimmer, 2016). Throughout, our discussion will focus on instances where only either the dependent variable or intervention is text based—we defer to future work situations where both the dependent and independent variable are based on texts. Roberts, Stewart and Nielsen (2017) addresses the case of text as confounder.

Across our example designs, we will denote our dependent variable for each unit i ($i = 1, 2, \ldots, N$) with $\boldsymbol{Y}_i$, the treatment condition for unit i will be $\boldsymbol{T}_{d[i]}$ (with $\boldsymbol{T} = (\boldsymbol{T}_{d[1]}, \boldsymbol{T}_{d[2]}, \ldots, \boldsymbol{T}_{d[N]})$ and any covariates for a unit will be $\boldsymbol{X}_i$, where the index $d[i]$ refers to the index of the document that observation i has observed.[1] When dealing with text, we will use bold font for our dependent variable $\boldsymbol{Y}_i$ and treatment $\boldsymbol{T}_{d[i]}$. Note that for our framework it is possible that either the dependent or independent variable is a text-based treatment.

To motivate our approach to causal inference and text we first focus on a randomized experiment. First, we assume that the response depends only on the assigned treatment, often called the Stable Unit Treatment Value Assumption (SUTVA). Specifically:

Assumption 1 (SUTVA)**.** *For all individuals i, $\boldsymbol{Y}_i(\boldsymbol{T}) = \boldsymbol{Y}_i(\boldsymbol{T}_{d[i]})$.*

Second, we will assume that our treatment is randomly assigned:

Assumption 2 (Ignorability/Positivity)**.** *$Y_i(\boldsymbol{t}) \perp\!\!\!\perp \boldsymbol{T}_{d[i]}$ and $Pr(\boldsymbol{T}_{d[i]} = t) > 0$ for $\boldsymbol{t} \in \mathcal{T}$.*

For each observation we observe only a single potential outcome corresponding to the realized treatment.

The usual causal quantity of interest is the Average Treatment Effect (Imbens and Rubin, 2015). However, with text and other high-dimensional data it is often unclear how to determine the intervention or response of interest. This is true, in part, for conceptual reasons. We might follow a large text as data literature and treat each document as a bag of words and analyze how a particular intervention affects the use of particular words. This is well defined, but it is rarely the case that a single word (or a group of words considered independently) are the quantity of interest. Similarly, it is rarely the case that we are interested in how the presence or absence of a single word affects a response. For example, suppose we are interested in how a prior career as a lawyer affects support in an election (Bonica, Chilton and Sen, 2015). One might think that we could approximate this effect by measuring how

[1] This more flexible notation enables us to explicitly allow different observations to observe the same document

much the presence or absence of the word "lawyer" affects support. But there might be many other words or phrases that convey that someone has a legal background, such as "JD", "attorney", and "law school." Rather than individual words, social scientists are often interested in some emergent property of the text—such as the topic that is discussed, the sentiment expressed, or the ideological position taken.[2]

There are also statistical reasons to avoid working with texts directly. Text is high-dimensional and complicated, implying that distinct blocks of text that convey a particular topic, sentiment, or phrase are rarely replicated exactly across documents. The sparse nature of text data implies, then, that it would be difficult (if not impossible) to make a reliable inference about how interventions affect what is said or understand how various messages affect what is written because there will be insufficient data to make an inference, even in the largest data set. Rather than work directly with text, social scientists work with a low-dimensional representation of the text, though this low-dimensional representation is often only implicit (Laver, Benoit and Garry, 2003; Grimmer, Messing and Westwood, 2012; Spirling, 2012; Catalinac, 2016).

We unify the wide array of research designs with a framework that makes the move from the explicit text to a lower-dimensional representation. To make this representation, we will suppose that there is a function g that maps from the text to a set of numbers. The function g represents our coding rule and is used in both machine learning settings and when hand coding. For example, if assessing the effect of an intervention on open ended responses, the process of hand coding the open ended responses into categories represents the function g. And when designing text based treatments, the function g represents the *researcher's* coding of how the text corresponds to particular treatment. In general, we will suppose that $g : \mathcal{Y} \rightarrow \mathcal{Z}$ when the response is text and $g : \mathcal{T} \rightarrow \mathcal{Z}$ when the treatment is text. The set $\mathcal{Z}$ is a lower-dimensional representation of the text and can take on a variety of values depending upon the study of interest. For example, if we are hand coding our documents into two categories, then $\mathcal{Z}$ is $\{0, 1\}$. If we are using STM to measure the prevalence of K topics as our dependent variable, then $\mathcal{Z}$ is a $K-1$ dimensional simplex. And if we are using texts as a treatment, we might suppose that $\mathcal{Z}$ is the set of binary feature vectors, representing the presence or absence of an underlying treatment.

Using the function g we are able to define a causal quantity of interest. Consider first using hand coding to code text responses. For each unit i we will write $g(\boldsymbol{Y}_i(T_i))$ to be the coded text for unit i under treatment condition T_i. We can then define the average treatment effect to be:

$$\text{ATE} \quad = \quad E[g(\boldsymbol{Y}_i(1)) - g(\boldsymbol{Y}_i(0))] \tag{2.1}$$

While seemingly straightforward, we should note that this definition supposes that we already have a g function available. In practice, of course, we need to

[2]Of course there are counterexamples where scholars are interested in single words/phrases only, e.g. Gill and Hall (2015). This might be particularly true in the framing literature where a single word might substantially alter the content or meaning. Other experiments might make slight alterations that vary the magnitude of what is discussed.

estimate or otherwise develop a function g. We will now refer to the estimated function g. With a g in hand we can rewrite the ATE for a particular g as ATE $= E[g(\boldsymbol{Y}_i(1)) - g(\boldsymbol{Y}_i(0))]$.

The discovery of a g function represents an important, though often underappreciated, component of the experimental process. While the g function will define our dependent variable, many experimental designs implicitly assume it already exists. Rather than ignore the function or view it as a secondary issue, it is essential to place it at the center of the experimental design as one of the most important features of a valid experiment. By explicitly acknowledging the process of discovery we can develop procedures that support it and provide methodological guidance.

Obtaining the g function introduces new potential problems in the estimation of causal effects. First, estimating g creates the possibility of an *Analyst Induced SUTVA Violation* or (AISV). One component of SUTVA is that each observation's potential outcome depends on only its own treatment status and we have assumed that this is true for the texts. Once we use the texts to obtain a new g, we can violate SUTVA, creating a dependence across observations because the particular randomization may affect the g we estimate. This violation occurs because the specific treatment vector allocated affects the g that we obtain, causing dependence across *all* observations. This violation is induced by the analyst, which is why we call it an *Analyst* induced violation.

To see how this violation can emerge, suppose that we are developing a set of hand coding rules. To develop those rules, we might follow best practices and write down a set of coding rules, classify a set of statements, analyze intercoder agreement, and then iterate until we have reached a satisfactory codebook. We might make several passes over the entire data set if the set of responses are sufficiently small or if the existing codebook is insufficient for a large share of documents.

The codebook that is created, however, depends on the specific randomization. A different randomization of the treatment would result in different responses from individuals (if the treatment has any effect at all). And if this is the case, we would obtain a different function g. This implies that the lower-dimensional representation of a unit's response depends on each unit's treatment status. And therefore, we have a SUTVA violation.

There are two ways to deal with this sort of violation. First, we might assume the problem away, but this leaves the researcher exposed to the practical problem that the same data used to estimate the g function is also used to estimate the causal effects, introducing an opportunity for fishing. A sufficient assumption that eliminates the possibility of an AISV is that g is stable across randomizations. This might be an attractive assumption, particularly when we have a large sample of texts or when we expect that there are only a few easy to discover categories. Further, this assumption is trivially satisfied in any instance when g is determined before the randomization, such as when scholars design text treatments themselves. But even when we assume it is stable, learning g from the data creates a second problem. If researchers are free to modify g as they analyze the data, it becomes possible that changes in the code book are made to create larger effects. The ability to modify the code categories of either the dependent or independent variable creates an opportunity for a much more serious fishing problem than standard experimental

analyses, though it is related to problems of recoding the dependent variable or ignoring some conditions in an experiment. Even if the fishing is done subconsciously (or with the best of intentions), it creates substantial bias in the estimates, implies that the reported confidence intervals fail to have their reported coverage rates, and leads to overfitting. Overfitting occurs when scholars tinker too much with the g function then they will be modeling noise in their data, rather than the systematic categories that matter for either the dependent or independent variable.

So even if we assume the technical problem of the AISV away, we continue to have the practical problem that we are learning the mapping from text to categories on the same data we will use to measure the casual effects. The second approach, and our preferred procedure, is to explicitly separate the creation of g and the estimation of treatment effects. This avoids both the AISV and eliminates the opportunity for fishing. To do this, we randomly divide our data into a training set and a test set. Specifically, we create a set of units in a training set S_{train}, $|S_{\text{train}}| = N_{\text{train}}$ and test set S_{test}, $|S_{\text{test}}| = N_{\text{test}}$. We use only the training set to estimate the g function and then discard it. We then use the test set exclusively to estimate the causal effect.

This division between the training and test set avoids the AISV and ensures that we are not overfitting our data. It avoids the AISV in the test set because the function g does not depend on the randomization in the test set, so that each test set unit's response depends only on its assigned treatment status. It avoids the problem of overfitting and fishing because of the division between the training and test set. Any fishing or overfitting in the training set will, by definition, not provide the same effects in the test set. By adopting this procedure we explicitly allow and account for the process of discovery. By contrast the most prominent proposed solution to fishing, pre-analysis plans, implicitly assumes that discovery has already taken place, focusing instead on public signaling (Humphreys, Sanchez de la Sierra and Van der Windt, 2013).

The use of a training and test set aligns with several ways researchers already create codebooks for their analyses. For example, scholars often run pre-tests on a subset of data in order to develop code books or to refine particular text-based treatments. Other times scholars doing handcoding in experiments will develop a code book on a subset of data, they then use the code book to classify only the remaining documents. Our approach extends the logic of these well established practices to instances where the function g is learned using machine learning methods.

The particular g that we obtain in our training set defines the categories that we will use for our analysis. Therefore, our goal is to obtain a consistent (or unbiased) estimator for the ATE (or other causal quantities of interest) assuming a particular g. By randomization of the treatment, a consistent estimator will be:

$$\begin{aligned} \hat{ATE} &= E[g(\boldsymbol{Y}_i(1))|T=1] - E[g(\boldsymbol{Y}_i(0))|T=0] \\ &= \sum_{i \in \mathcal{S}_{\text{test}}} \frac{I(T_i=1)g(\boldsymbol{Y}_i(1))}{\sum_{i \in \mathcal{S}_{\text{test}}} I(T_i=1)} - \sum_{i \in \mathcal{S}_{\text{test}}} \frac{I(T_i=0)g(\boldsymbol{Y}_i(0))}{\sum_{i \in \mathcal{S}_{\text{test}}} I(T_i=0)} \end{aligned}$$

Further, if we suppose that g is fixed, we can obtain variance estimates using the usual variance estimators.

We now discuss how to estimate a g function in the case of text as outcome (Section 3) and text as treatment (Section 5), providing applications of each (Sections 4 and 6 respectively). We conclude with a discussion about how we know we have learned something new in causal inference with texts (Section 7). An appendix with additional discussion and simulation follows the main paper (Appendix A).

3 Text as Outcome

In the text as outcome setting we are interested in measuring the effect of a manipulated non-text treatment on a function of the existing texts such that $g : \mathcal{Y} \rightarrow \mathcal{Z}$. Once the g is defined, the process is straightforward and the standard procedures of causal inference apply. However learning the g function, either statistically or manually, must be done carefully in order to avoid an Analyst Induced SUTVA violation. In this section we discuss the properties that we would like in a g function (3.1), how manual content analysts develop g functions (3.2), how we can learn the g function statistically from the data (3.3), how the train-test split avoids the AISV violation (3.4) and the tradeoffs of train-test splits (3.5). We conclude the section with implications for past and future work (3.6).

3.1 Desirable properties of the g function

Before describing the different ways that we can choose a g function it is useful to describe the stakes involved and the properties that we might want to see. The choice of g defines the property of the text along which we are able to measure the effect of the intervention $\mathbf{T}$. In many ways it has the same importance as choosing the outcome to measure in an experiment. We identify four properties (in decreasing order of importance) that we see as a starting point for choosing g: interpretability, theoretical interest, label fidelity, and tractability.

First and foremost, we should choose the g function to be *interpretable*. If we cannot communicate what we are measuring about the text we cannot claim that it is of theoretical interest or define label fidelity; at that point, tractability becomes irrelevant. What interpretability means is specific to a given set of texts and a particular research question, but our articles must communicate to the reader what our measure is capturing. The statistically learned g functions will generally try to learn bundles of words that share semantic meaning.

The g function should also capture something of *theoretical interest*. It would be simple and straightforward to evaluate the causal effect of a treatment on the number of words used in the document, but this will (generally) not help advance an interesting argument. Theoretical interest will obviously be application-specific, but often it is theoretical interest that draws us towards considering bundles of words. Many of the problems we raise with the g function could be resolved by simply comparing the counts of a pre-selected word, but it can be difficult to connect to a broader theory with only a single word. Instead, we typically will claim that the g function is capturing some property of the text, such as its general subject matter, that has a particular relevance to our research program.

We also want to choose our function such that there is high *fidelity* in the correspondence between the label we give to what the g function is capturing and what it actually encodes. It is easy to forget that when we develop a codebook that evaluates whether a given document is about the "economy" we have not actually measured that thing, but rather a surrogate function that will form the basis of our argument. This is a common exercise in the social sciences; there is always an implicit mapping between the labels we use for our variables and the reality of what our labels measure. Text analysis, and particularly unsupervised methods, clearly expose this problem by highlighting the importance of the label. While there is no objective definition of label fidelity, we think about it as minimizing the surprise that a reader would have in going from the label to reading the text.

Finally, we want the development and deployment of the g function to be *tractable*. In the context of manual content analysis this could mean that the codebook can be applied accurately and reliably[3] by human coders and that the number of documents to be coded is feasible for the resources available. In the case of learning the g function statistically, this would imply that we have a model which is sufficiently computationally tractable that we can estimate it and that it is able to learn a useful representation with the number of documents we possess.

An important component of all of these properties is the form of the latent representation $\mathbf{Z}$. Manual content analysis often chooses either a categorical representation (mutually exclusive and exhaustive categories), count of an event (a non-negative integer) or check-all that apply (a number of binary features). Statistical models use these forms as well as scales (a value in $\mathcal{R}^D$ or number in a constrained space) or mixed membership (a value on the simplex). Other representations are possible but these have historically provided the best compromise of intepretability and tractability.

There is an inherent tension between the four properties particularly between theoretical interest and label fidelity. It is often tempting to assign a label to a concept that is substantially more general than what the g function actually measures. However, allowing fidelity to decline in order to bolster theoretical interest is a way to lower the odds that the finding is replicable. We view any scientific finding as established by a sequence of studies in which a given result is replicated through multiple experiments, prefarably across research labs. Under this view we are incentivised to have a g function which tightly coincides with our chosen label because it increases the chances that our result can be replicated by a different research group who may measure the same concept with a slightly different mapping function.

3.2 Developing a g function in handcoding

Manual content analysts have been creating g functions for years. Content analysis handbooks such as Krippendorff (2004) or Neuendorf (2016) advocate the use of codebooks which are essentially a set of instructions for humans to map from texts to a latent representation; in the language of this paper, a g function. It is beyond the scope of this paper to describe how to develop a good coding scheme for manual content analysis, but our chosen properties are generally in line with the objectives.

[3]In a statistical sense we mean they have low bias and variance.

A manual content analysis scheme has many benefits. For example, there are many features of text which humans are quite adept at detecting which computers find more challenging (e.g. humor and sarcasm) and when these are the quantity of interest a manual content analysis scheme might be preferable. Because the investigator has more freedom to choose the component of text that interests them, they arguable have greater freedom to choose features of theoretical interest (although they are still bound by the capacity of human coders).[4]

Yet the manual content scheme has to come from *somewhere* and we rarely talk about the process of discovery. In practice, these coding schemes are developed through iteration between coding rules and the documents to be coded. In many ways this is a reasonable practice: looking at the documents let's us assess the 'model fit' and helps to ensure that the coding rules are completely specified. If we are developing a coding scheme for an open-ended survey question about 'why did you vote for Trump?' it would be inefficient at best and impossible at worst to enumerate every possible reason without reference to the texts. Yet engaging in this iterative process and then estimating the effect on the given set of texts results in an Analyst Induced SUTVA violation because the coding scheme is now dependent on a particular randomization (more on this in the next section).

3.3 Learning the g function

In the case of manual content analysis, the g function is learned through a generally unspoken process of discovery. We can also learn the g function statistically from the data. There are two strategies for this: supervised methods which ultimately produces a system close to manual content analysis and unsupervised learning which enables the inductive learning of the g function.

In supervised learning we devise a manual content analysis scheme that is applied to a subset of documents in the training set. The algorithm then uses this subset of hand-coded documents in order to learn the g function and predict the outcomes for the test set (see Grimmer and Stewart (2013) for more details). The threat of an Analyst Induced SUTVA violation requires that we do not use the test set in the development of our coding scheme or in the coding of the training documents that will be used to learn the g function.[5]

We can also use unsupervised learning methods to learn the g function, statistically operationalizing the process of discovery (Grimmer and King, 2011). Unsupervised methods search for a low-dimensional representation of the data that is able to accurately reconstruct the higher dimensional representation.[6] These low-dimensional representations capture covariance in the data (such as by grouping words that commonly co-occur into a single 'topic'). The most consequential choice

[4]Indeed the aspirations of investigators exceeding the capacity of human coders has led to trenchant critiques of the content analysis enterprise by those concerned that the coding is not able to capture the nuance of the underlying text and thus inadvertently 'invents facts' (Biernacki, 2012).

[5]This includes, for example, feature selection and representation. See the discussion of cross-validation in Grimmer and Stewart (2013) or Hastie, Tibshirani and Friedman (2013).

[6]In practice this high-dimensional representation is itself already a reduction of the original text such as a document-term matrix.

to make in this setting is the statistical model to be used which dictates the form of the latent representation. One popular representation of texts is provided by topic models where the latent representation, $\mathcal{Z}$, is a point on the $K-1$ dimensional simplex (Blei, Ng and Jordan, 2003; Blei, 2012). The representation of each document has an interpretation as proportional membership in a set of K topics (Airoldi et al., 2014). Our causal estimand is specified in terms of counterfactual differences in the proportion of the document devoted to a particular topic k. For this paper we use the Structural Topic Model (STM) which is a particular type of topic model (Roberts, Stewart and Airoldi, 2016).[7] We could also chose non-topic model representations which would give rise to other forms for $\mathcal{Z}$ including, for example, continuous variables, single categories and binary features.

While it might be tempting to make a sharp distinction between causal inference employing supervised and unsupervised inference, both are instances of our general framework. Whether doing hand coding, supervised learning, or unsupervised methods scholars are still required to specify some function g that moves from the high-dimensional text to some lower-dimensional representation of the text. Certainly how we learn that g function will matter—it could introduce measurement error or bias (as we detail below). But the procedure used to discover the function does not alter the same basic framework.

Because we are now explicitly representing the process of discovery, it means at the outset we do not initially know what the outcome of our study will be. We must find the representation of the text that we are interested in, connect it to our theoretical interest and estimate the appropriate g function for mapping documents to that construct. This creates the risk for the Analyst Induced SUTVA violation because the particular randomization of our data may affect the discovery we make and thus our potential outcomes.[8] This connects the Analyst Induced SUTVA problem to overfitting or what Gelman and Loken (2014) call 'the garden of forking paths.' Because we are engaged in a process of discovery we want to avoid places stringent limits on the capacity of the analyst to explore other possibilities.

Once we have found a particular g function that has the desirable properties we discussed above, we can apply it to a new set of documents. We emphasize that conceptually this does not imply that we have to believe there is a "true" underlying representation or that we have to assume the model, rather the causal effect in terms of the g function simply becomes the estimand of interest (in the same way that choosing a given outcome measure in an experiment defines how we measure the effect of treatment). If we could collect new data, we could apply the g function to that data and it would be statistically no different than if we were measuring any other standard outcome.

[7] See the appendix for a summary of the model and an explanation of how to extract the g function.

[8] We note that the Analyst Induced SUTVA problem could be resolved if we wanted to make a technical assumption that the g function was stable across all randomizations. This is a fairly heroic assumption from the point of view of the model, but also requires a commitment to a particular model choice before looking at the data, which is overly restrictive from a data analysis standpoint.

3.4 Train-Test Splits

In order to promote discovery while also avoiding an Analyst induced SUTVA violation we advocate the use of a train-test split, a technique that has an increasing number of advocates in political science (Cranmer and Desmarais, 2017; Grimmer, Messing and Westwood, 2017; Ward, Greenhill and Bakke, 2010) as well as scholars in causal inference (Wager and Athey, 2017; Chernozhukov et al., 2017). By confining our exploration to the training set and estimating only in the test set, we avoid the Analyst Induced SUTVA concerns and align researcher incentives. The analyst can "fish" and explore as much as they like in the train set, but are incentivized to find a robust underlying pattern. This reduces concerns about overfitting and the garden of forking paths while also allowing the analyst to develop theoretically important and precise measures.

Procedurally we start by creating a random partition of our document set into two portions (generally halves although not necessarily). It is important that we don't do any preprocessing first which requires the complete dataset. For example, text scholars often remove words that are either too rare or too frequent (Grimmer and Stewart, 2013), however doing so prior to the train-test split uses data from the test set.[9] Using only the training data we are free to specify as many models as we like in our search for the outcome of interest. Only after we have made a final decision should we move to the test set.

Validation is an important part of the text analysis process and researchers should apply the normal process of validation to establish that the model has a strong correspondence between the chosen label and the actual measurement instrument (label fidelity). These validations are often application-specific and draw on close reading of the texts.[10] These validations should be completed in the training set as part of the process of discovering and labeling the g function.

The train-test split requires an additional validation procedure because we need to ensure that the model fits nearly as well on the test set as it did on the training set. When both the training and test sets are random draws from the same population this will generally be true, but if we have succumbed to some level of overfitting or we got a bad randomization, it may not be. The same types of techniques used to validate the original model can be used in the test set as well as common measures of model fit such as log likelihood. Unlike the validation in the training set, once we validate in the test set we cannot return to make changes to our model. Nevertheless, validation in the test set helps us to understand the substantive meaning of what we have estimated and provides guidance for future experiments. When the test set is truly a random draw from the same population as the test set, we should expect these validations to be successful.

Applying the g function in the test set is relatively straightforward and is essentially the process of making a prediction for a new text. In the appendix we provide a review of the Structural Topic Model and discuss the mechanics of applying the g function including how to deal with the covariate-based prior distribution.

[9] See Denny and Spirling (2017) for a discussion and an approach to assess the impact of these text processing decisions.

[10] See Grimmer and Stewart (2013) for more detail on types of validation and the `stm` package (Roberts, Stewart and Tingley, 2017) for tools designed to assist with validation.

Once we have our quantities $g(\mathbf{Y})$ we can use standard estimators appropriate to our estimand (such as the difference of means to estimate the average treatment effect).[11]

3.5 Tradeoffs in Train-Test Split

While the train-test split addresses many of our concerns, it is not without cost. The most notable concern is the efficiency loss in splitting the data. In a 50/50 train-test split we have lost half our data for both estimating the g function and estimating the treatment effect. It is difficult to assess how much data we need for either the training or the test set. The challenge in the test set is that we don't yet know what the outcome will be when we have to make the choice. However, while we in the discovery phase we will be able to make a reasonable assessment of whether the sample size we have allotted for the test set will be able to estimate an effect the size we are interested (Gelman and Carlin, 2014).

The problem is even more complicated for setting the size of the training set because fundamentally *we don't know the power we need for discovery.* As a technical matter the results on posterior contraction rates for topic models are few and far between and generally assume the data generating process as well as consider relatively small numbers of topics (Tang et al., 2014; Nguyen, 2015). Even if we trust our intuition about how many documents we need, it is difficult to know how many documents we will need to discover and estimate the most interesting outcome. In practice, we have to simply make a choice.

A concern we might have about this process is the stability that the entire procedure has over different train-test splits. This will change what we discover given any train-test split but it will not invalidate our results. Still there is an understandable preference to have a sufficiently large train-test split that the particular randomization of the train-test is irrelevant. We empirically explore the stability of STM under different train-test splits in the appendix. Our results suggest that a productive area of future research interest is to develop initialization procedures for the model which are more stable under different splits of the data (even perhaps at some expense of the quality of the initialization).

The train-test split offers many advantages; most notably peace of mind for both

[11] Once we have applied the g function to our test data we can calculate confidence intervals using usual variance estimators that capture uncertainty about our estimate given a limited sample size conditional on the function g. In prior work we have explicitly taken the view of $g(\mathbf{Y})$ as a latent variable about which there is some additional measurement uncertainty and advocated approaches to incorporate this additional uncertainty into our confidence intervals (Roberts et al., 2014; Fong and Grimmer, 2016). For example, Roberts et al. (2014) advocates a simulation approach to integrate over the variational approximation to the posterior distribution which conditions on the learned topic-word distribution, but accounts for the fact that the document-topic proportion θ cannot be known with certainty for a particular document because it has a finite length. Fong and Grimmer (2016) use a bootstrap approach which captures measurement uncertainty both in the topic-word parameters and the document-topic representation. While this approach is intuitively appealing, it complicates the definition of g as a function because we run the risk of the same text mapping to two different values of the latent variable (failing the vertical line test). In the interest of simplicity we do not include this form of measurement error in this article and leave to future work the incorporation of this uncertainty into the causal framework.

the analyst and the reader that the discovered finding holds in a held-out sample. For the analyst this allows the process of discovery to be more unconstrained. It also allows the study to make rigorous claims of having estimated a causal effect. However, the methodology we outline significantly decreases the amount of data available to us for both discovery of the g function and estimating the causal effects. Whether or not the benefits are worth the costs is a function of the particular researcher's objective.

3.6 Prior Work, Implications and Open Questions

There has been comparatively little work on causal inference with latent variables. Lanza, Coffman and Xu (2013) consider causal inference for latent class models but do not give a formal statement of identifying assumptions or acknowledge the set of concerns we identify as an analyst induced SUTVA violation. There is also a line of work on treatment effects for ordinal outcome variables. Volfovsky, Airoldi and Rubin (2015) present a variety of estimands and estimation strategies for causal effects where the dependent variable is ordinal. They provide approaches based both on the observed data as well as latent continuous outcomes. Volfovsky, Airoldi and Rubin (2015) express caution about the latent variable formulation due to identification concerns and the subsequent literature (e.g., Lu, Ding and Dasgupta, 2015) has moved away from it. Unfortunately many of the strategies based directly on the observed outcomes are unavailable in the much higher dimensional setting of text analysis.

There is also a burgeoning literature on causal inference using machine learning (Van der Laan and Rose, 2011; Athey, 2015; Athey and Imbens, 2016; Bloniarz et al., 2016; Hartford et al., 2016; Chernozhukov et al., 2017; Wager and Athey, 2017; Ratkovic and Tingley, 2017). Much of this work focuses on estimating causal parameters on observed data and addressing a common set of concerns such as estimation and inference in high-dimensional settings, regularization bias and overfitting. Our work complements this literature by exploring the use of latent treatments and outcomes. Many pieces in this area call for sample splits or cross-validation for estimation and inference, providing additional justification for our preferred approach (see e.g. Chernozhukov et al., 2017).

Although text is an obvious area of concern for Analyst Induced SUTVA violation, the points we make are applicable in any case where there is a latent construct/mapping. This could include latent measures common in political science such as measures of democracy (e.g. Polity), voting behavior (e.g. ideal points) and forms of manual content analysis. Any time a process of discovery is necessary, we should be concerned if the discovery is completed on the same units where the effect is estimated. In certain circumstances this process will be unavoidable- Polity scores were developed by looking at the full population of world democracies so there is no test set we can access, but in the development of future methods, we argue that there is substantial value in acknowledging and accounting for the discovery phase.

What then does this mean for applied work on open-ended survey experiments? The AISV concern clarifies a subtle way that one of the standard assumptions in causal inference could be violated, but it does not mean any work not employing a

train-test split is invalid. However, as estimands based on latent constructs become more common in the social sciences, we hope to see an increased use of the train-test split and possibly the development of new methodologies to explicitly address the process of discovery.

4 Application: Using the Structural Topic Model for an Immigration Experiment

In this section, we illustrate our experimental process for text as outcome by applying the Structural Topic Model to an experiment about immigration and crime in the United States. The research question of interest is how a known history of violent crime for a person who has illegally entered the United States affects what how respondents believe the United States criminal justice system to respond to the knowledge that the person has illegally entered the U.S.

Despite the relevance of this question to current events within the United States, as a starting point we analyze data from an existing survey experiment conducted and made available by Cohen, Rust and Steen (2004). The survey experiment was administered in the context of a larger study of public perceptions of the criminal justice system. The survey was conducted in 2000 by telephone random-digit dial and includes 1,300 respondents.[12]

In the survey experiment of interest, respondents were given two scenarios of a criminal offense. In both conditions, the crime was illegal entry into the United States. In the treatment condition, respondents were told that the person had previously committed a violent crime and had previously been deported. In the control condition, respondents were told that the person had never been imprisoned before.[13]

In the treatment condition, respondents were told:

> "A 28-year-old single man, a citizen of another country, was convicted of illegally entering the United States. Prior to this offense, he had served two previous prison sentences each more than a year. One of these previous sentences was for a violent crime and he had been deported back to his home country."

In the control condition, respondents were told:

[12] More details about the survey are available in Cohen, Rust and Steen (2002).

[13] The treatment and control conditions were placed in a slightly different order in the survey and we are aware that there may be order effects. To address some of the features of the experiment that we would design differently, we are currently re-running the survey experiment in the U.S. on Mechanical Turk.

"A 28-year-old single man, a citizen of another country, was convicted of illegally entering the United States. Prior to this offense, he had never been imprisoned before."

Respondents were then asked a close-ended question about whether or not the person should go to jail. If they responded that the person should not go to jail, they were asked to respond to an open-ended question, "Why?" Based on the subsequent questions in the survey, the researchers who designed the survey experiment were interested in estimating whether or not the respondent believed that the person should not go to jail because he should be deported, or whether the respondent believed that the person should not go to jail for some other reason.[14]

The challenge for analyzing the open-ended survey experiment like this one is to define a mapping $g()$ from the text to the categories of interest. The $g()$ mapping is challenging to define before seeing the open-ended responses. The researcher may know some of the categories of interest – like deportion – before the fact, but may not know the function that maps between the text and these categories. For example, in this survey, many respondents explained that they thought the person in the scenario should be deported, but many did not use the word 'deport'; saying instead, "He should be sent home." The researcher would need to understand the many different ways that a person could express the same concept before defining the the g function.

In addition, the researcher may not know some of the categories before collecting the open-ended responses, and the categories themselves need to be known before the mapping $g()$ can be defined. For example, many people within the survey responded that they did not think that the person in the vignette had committed a crime. Because both scenarios explain that the person has been "convicted" of illegally entering the country, the researcher may not have anticipated that the respondents would disagree with the court conviction entirely.

Because the researcher may not know the mapping from the text to the categories of interest or even the categories themselves, the discovery phase of the research process allows the researcher to explore and define the $g()$ function before testing the treatment effect. In this example, we split the data into two equally sized training and test sets. Using the training set, we fit many structural topic models with treatment as a prevalence covariate using various preprocessing techniques and numbers of topics. This allows us to explore a variety of mappings and decide on one that best measures the concepts of interest and reflects the range of ideas expressed within the data. We settled on an STM specification of ten topics that we believed best measured the concepts within the open-ended responses. The ten topics estimated within the training set as well as representative documents from each topic are reported in Tables 1 and 2.

To create a final estimate, we cluster the topics into two main categories – topics

[14]Immediately following the experiment, the survey enumerator was told to follow up with a few other questions if the respondent used the word "deport" in their open ended response.

	Label	Highest Probability Words
Topic 1	He wants a better life	didnt, want, pay, better, life, probabl, isnt
Topic 2	Send him back	back, countri, send, home, well, charg, troubl
Topic 3	Small punishment	offens, reason, like, chanc, first, can, citizen
Topic 4	Depends on circumstances	come, depend, doesnt, free, feel, law, shouldnt
Topic 5	Crime was not violent	crime, commit, violent, immigr, wasnt, look, never
Topic 6	Deport	deport, that, give, counti, peopl, look, guilti
Topic 7	Prison is too strict	enter, anyth, right, live, realli, illeg, anybodi
Topic 8	Right to freedom	just, tri, get, hes, came, freedom, put
Topic 9	Deport bc overcrowded	sent, prison, think, alreadi, anoth, done, hasnt
Topic 10	Deport bc expense	dont, think, know, time, need, serv, crimin

Table 1: Words most representative of topics.

	Label	Representative Document
Topic 1	He wants a better life	we're the land of opportunity everybody wants a better life
Topic 2	Send him back	send him back to his country
Topic 3	Small punishment	"it was his first offense, didn't hurt anybody, maybe a fine though, probation or something. that's nice serious like murder or robbery"
Topic 4	Depends on circumstances	it depends on reaason why he is coming into state if he was coming to beter himself its ok if he has a record he should be disbarred or deported
Topic 5	Crime was not violent	because he didnt commit a crime that was effecting someone else's individual liberties
Topic 6	Deport	he should be deported
Topic 7	Prison is too strict	because he didnt do anything except illegally enter
Topic 8	Right to freedom	Because he's just trying to get his freedom. Maybe he's trying to away from a tough situation/that country-maybe it's not good for him.
Topic 9	Deport bc overcrowded	he should be sent to prison in another country our prisons are over crowded already
Topic 10	Deport bc expense	because i think he shold be deported-p-i don't think he should be supported in our prison system and i don't think he should be allowed to immigrate here

Table 2: Representative documents of each topic.

related to deportion (Topics 2, 6, 9, and 10) and topics that indicate the respondent would neither deport nor jail the person in the vignette (Topics 1, 3, 4, 5, 7, and 8). This allows us to estimate the relationship between treatment and more general

topical categories in the training set on the combination of the closed and open ended responses, reported in Table 3. Within the training set, the treated group (where the respondents were told that the person had previous criminal history) is more likely to report that they think the hypothetical person should go to jail, second most likely to report that the hypothetical person should be deported, and unlikely to report that the hypothetical person should be neither deported or sent to jail. Within the control group, fewer people thought the person should be sent to jail and instead thought the person should either be deported or another course of action should be considered.

	Jail	Deport	Neither
Treatment	0.68	0.18	0.14
Control	0.23	0.38	0.39

Table 3: Impact of previous crime treatment on perceptions of just outcomes, training set.

	Jail	Deport	Neither
Treatment	0.66	0.20	0.14
Control	0.23	0.38	0.39

Table 4: Impact of previous crime treatment on perceptions of just outcomes, test set.

To ensure that this relationship is not simply an artifact of the data within the training set, we apply the fitted STM model we decided on within the training set onto the data in the test set that was not used to estimate the mapping $g()$. We compare the relationship between treatment and topics in Figure 1. For the most part, the relationships between treatment and topics are similar in the training and test set.[15] The aggregated topic results are also very similar, as shown in Table 4. Not only do the training and test sets have similar results, these results match up almost exactly with the hand coding of the open ended responses from the authors' original paper (Cohen, Rust and Steen, 2002, pg 30). The evidence from this full analysis suggests that the treatment condition increases the probability that a respondent will think that the person in the vignette should be deported or put in jail, and that the control condition makes it more likely that a respondent will think that the person should be neither put in jail nor deported.

5 Text-Based Treatments

We now consider the case where text is the intervention in an experiment and the dependent variable is a uni-dimensional variable. For example, say that individuals are all randomly assigned different text-based descriptions of the same candidate for a political office. After reading the description they are then asked to rate their evaluation on a feeling thermometer scale from 1 to 100. In this case the treatment is the description of the candidate assigned to the respondents $T_d[i]$. The potential outcomes $Y_i(T_d[i])$ describes respondent i's rating of the candidate under the $d[i]^{\text{th}}$ version of the treatment.

[15] To be clear in a true application we would advocate simply showing the effect in the test set. Here we show the train/test comparison solely to demonstrate that the effects are relatively stable.

Figure 1: Comparison of the relationship between treatment and topics in the training and test sets.

5.1 Comparing Text Treatments Without a Latent Representation: A/B Tests

A/B tests, a widely used methodology in industry applications, rely on testing and comparing different versions of text based treatments without a latent-feature representation. In these cases, the average treatment effect comparing two treatments is straightforward: $E[Y_i(T_{d[i]} = 1) - Y_i(T_{d[i]} = 0)]$. Randomly assigning the texts identifies the average treatment effect of comparing one treatment to another. More generally, we can compare any pair of treatments through similar random assignment of texts to individuals.

This methodology is a workhorse tool across industry. A variety of industry methodologies attempt to determine the optimal treatment to deliver, or the optimal treatment based on a set of categories. To do this, a set of texts (or other high-dimensional data, such as a website design) are provided and then a response is evaluated.

A/B tests are useful for applications where a best text is needed to use for an application; they are less useful for social science research. Social scientists are rarely interested in obtaining the "best" version of a text or in comparing two messages without reference to some underlying feature.

5.2 Binary Latent Treatments

Rather than A/B tests to estimate the response to an entire block of text, social scientists are often interested in how some underlying feature of a document affects responses—that is the researcher is interested in estimating on how an *aspect* or *latent* value of the text influences the outcome. For example, the researcher might be interested in whether including military service in the description has an impact on the respondents' ratings of the candidate. Military service is a latent variable – there are many ways that the text could describe military service that all would count as the inclusion of military service and many ways that the text could omit military service that all would count as the absence of the latent variable. The researcher might assign 100 different candidate descriptions, some which mention the candidate's military service and some which do not. In this case, the treatment of interest is $Z_i = g(T_d[i])$ which maps the treatment text to an indicator variable that indicates whether or not the text contains a description of the candidate's military service.

The latent representations of the text-based treatment presents complications to causal inference and not just if machine learning methods are used to infer the g function.

5.2.1 Analyst-induced SUTVA violation and overfitting

First, the latent-representation of text creates the possibility of an analyst-induced SUTVA violation (AISV). As we described above, if the researcher decides on the mapping g between text based intervention $\boldsymbol{T}_{d[i]}$ and binary treatments Z_i based on the full text of all $\boldsymbol{T}$s and responses, then the inferred treatments for each unit, Z_i, will depend on the texts allocated to all other units. Further, if researchers are

able to search for different g functions then they will be able to fish for significant treatments across a much broader search than standard experiments.

Therefore, when a researcher decides to represent treatment as a latent variable that is a function of the text – whether through hand coding or statistical modeling – they should take precautions to avoid analyst-induced SUTVA assumptions and overfitting by ensuring that g is not estimated based on the full corpus of treatment texts or on the full outcome variables Y. Either the researcher should define g from the outset or should split the data into a training and test set and estimate g only from the training set, as discussed above. This addresses both the problem of analyst-induced SUTVA violations and of overfitting.

5.2.2 Sufficiency of Latent Treatments

Focusing on inferring the effect of latent treatments, rather than entire texts, enables social scientific inferences from text-based experiments. The tradeoff, however, is is that randomization is often done at the text level. In order to make an inference about the underlying latent treatments after a text-based randomization, we have to assume that our g function captures all the information relevant to the response in T is contained in Z. We present two versions of this assumption that is necessary to identify the causal effect. (For proofs see Fong and Grimmer (2016) and Fong et al. (2017)). The stronger assumption is the most intuitive and that is that $Y_i(\boldsymbol{T}_{d[i]}) = Y_i(g(\boldsymbol{T}_{d[i]}))$ for all documents and for all respondents. In words, this assumption requires that the potential outcome response to the text be identical to the potential outcome response to all documents with the same latent feature representation. This assumption is strong because it requires that there is no other information contained in the text that matters for the response beyond what is contained in the latent feature representation. In our running example about military service, this would mean that the inclusion or exclusion of military service is the only aspect relevant to the effect of the document on the individual's rating. Particularly for text, we could imagine that this assumption could easily be violated. If both versions of the treatment contain "The candidate served in the military", but one also adds "The candidate was dishonorably discharged" we might expect that this additional text added in addition Z may be relevant to the responses.

We can weaken this strong assumption to allow for some features of the text not contained in the latent feature representation to matter for the potential outcome. Rather than assume all responses are equal, Fong et al. (2017) shows we can assume that for any two treatments $T, T^{'} \in \mathcal{T}$, with the same feature representation ($g(T) = g(T^{'})$), then $E_i[Y_i(T)] = E_i[Y_i(T^{'})]$. In words, this assumption states that for any individual the response to two documents with the same latent feature representation might differ, but on average over individuals the responses are the same. Fong et al. (2017) shows that this assumption is equivalent to supposing that the components of the document that affect the response and are not included in the latent feature representation are orthogonal to the latent feature representation. Technically, we can write $\epsilon_i(T) = Y_i(T) - Y_i(g(T))$. Then this more general assumption is equivalent to assuming that $E_i[\epsilon_i(T)] = 0$ for all T.

5.2.3 No Hidden Values of Treatment

Even if we avoid AISVs and the latent treatments are sufficient, the latent variable representation of treatment also requires that Z does not contain hidden values of the treatment. This can emerge because a description of a particular treatment might obscure some underlying variation in the treatment. For example, suppose we create a dictionary based coding rule that considers a candidate's biography to refer to military experience if it mentions the word `army` or `military`. Then this will include individuals who have served in the military, but also to candidates whose parents served in the military, or even candidates who worked for the Salvation Army.

In this case we can still obtain an unbiased estimate the treatment effect, though the estimated effect will differ from the effect that we think we are estimating. This is problematic for interpretation and for applications to other experiments that attempt to build on the first round. With careful validation of the text-based treatment we can limit the risk of hidden versions of the treatment in confounding our analysis.

5.3 Multidimensional Latent Treatments

With complicated interventions, such as text, it is likely that there are more than just one latent treatment. The presence of multiple interventions requires defining new causal quantities of interest and suggests new potential questions we can ask about how the features of a text affect a response. With multidimensional latent treatments we can go beyond asking what effects including or excluding military service has on perceptions of a candidate, we can also learn how military service interacts with other components of the text to influence responses—such as a candidates career background, education prior to military service, or life after the military. We might think of texts as combinations of multiple treatments – a candidate description could include military service, occupation description, or a description of their family. In this case, we map the treatment text $\boldsymbol{T}$ to a vector of K binary treatments $\boldsymbol{Z}_j \in \mathcal{Z}$ where $\mathcal{Z}$ represents all 2^K possible combinations of the treatments.

Scholars often create the multidimensional treatments explicitly. For example, Grimmer, Messing and Westwood (2017) examine how claiming credit for funding in a district with a press release affects the way constituetents allocate credit to elected officials. Grimmer, Messing and Westwood (2017) design an experiment that varies many aspects of legislators' messages to their participants. In the treatment condition of the experiment, the authors are able to vary many different aspects of credit claiming. Treatment T_j is the combination of all of the texts. This maps directly into a binary treatment vector Z_j which describes the components of the text to which the participant was assigned.

By using a multidimensional treatment, researchers are able to identify the aspects of the text that are most relevant to the outcome variable. The researchers can estimate the Average Marginal Component Effect (AMCE) for each individual component k, using methods developed for conjoint analysis (Hainmueller, Hopkins and Yamamoto, 2013). The AMCE is the average effect of the component k, averaged over all other possible combinations:

$$AMCE_k = \int_{Z_{-k}} E[Y(Z_k = 1, Z_{-k}) - Y(Z_k = 0, Z_{-k})]m(Z_{-k})dZ_{-k}$$

The $AMCE_k$ describes the average effect of component k, summed over all other values of k, weighted by $m(Z_{-k})$, or the distribution of Z_{-k}. The AMCE can be thought of as an estimate of the effect of component k, given the distribution of other components in the population—therefore providing a sense of how an intervention will matter given other characteristics.

The analyst might also be interested in estimating the effect of an interaction between two components k and l. For example, the researcher might be interested if including military service into a candidate profile has a different effect on candidate ratings if the profile also includes that the candidate is female. This could be estimated as the Average Component Interaction Effect (ACIE) (Hainmueller, Hopkins and Yamamoto, 2013) :

$$\begin{aligned} ACIE_{k,l} &= \int_{Z_{-k,-l}} E[(Y(Z_k = 1, Z_l = 1, Z_{-k,-l}) - Y(Z_k = 1, Z_l = 0, Z_{-k,-l})) \\ &- (Y(Z_k = 0, Z_l = 1, Z_{-k,-l}) - Y(Z_k = 0, Z_l = 0, Z_{-k,-l})]m(Z_{-k,-m})dZ_{-k,-m} \end{aligned}$$

The ACIE will be the difference between the AMCE for military service for a candidate description that includes information that the candidate is female and the AMCE for military service for a candidate description that does not include this information.

Note that the three complications from the last section also pertain to the case of multidimensional treatments. If the mapping $g()$ between $\boldsymbol{T}$ and $\boldsymbol{Z}$ is not known before defining and reading the treatment texts or the outcome is used in the estimation of these mapping, then an AISV will occur. Even when using hand coding, researchers should either use a pre-test to determine their coding scheme or use a training/test split to first learn a coding scheme using the responses and then separately estimate the treatment effects.

5.4 Unsupervised Discovery of Latent Treatments

In the cases described above, the researchers came up with their own representation of g either through their substantive knowledge, a test of a theoretical argument, or through careful reading of documents in the training set. They then constructed the text treatments to reflect the treatments they wanted to test. In many cases, researchers will know that they want to estimate the effect of including military service or the effect of including the amount of funding raised on the outcome. However, in other cases, researchers may not have a clear sense of the treatments they would like to test, or would like some guidance from the data about what the interesting treatments to examine might be. This is equivalent to saying that researchers are not sure what $\mathcal{Z}$ should be, or if they have texts of interest in hand, they are not sure of the mapping from the set of texts $\mathcal{T}$ to the set of treatments $\mathcal{Z}$.

One approach to estimating the function g would be to estimate features that explain the texts wells, such as the principal components of a document-term matrix. Using these estimates for $\boldsymbol{Z}$, the impact of $\boldsymbol{Z}$ on $\boldsymbol{Y}$ can be estimated directly. The use of principal components to estimate causal quantities of interest leads to several issues. First, the set of features that are discovered best explain the texts, but do not condition on the responses. This can lead to the discovery of features that explain the content of texts but do not explain the response to those texts. Second, the use of principal components leads to a continuous treatment. This might be reasonable, however, estimating the response to a continuous treatment requires strong functional form assumptions and often a great deal of data to estimate more than a linear response. Alternative approaches, such as an Indian Buffet Process (Griffiths and Ghahramani, 2011), yield a binary feature vector about the treatments that are present or absent in a text, but fail to include information about the responses.

Certainly focusing on the texts alone is a reasonable approach, but including information about the response can improve the latent features that the g function maps texts to. Fong and Grimmer (2016) create an unsupervised method for estimating treatments from text data. They develop a supervised Indian Buffet Process (sIBP) that discovers the topics within the documents that are related to the outcome. The authors assume that the proportion of documents in each latent feature k is π_k, where π_k is generated by a stick-breaking algorithm (Doshi et al., 2009). Each document can be summarized by treatment vector Z_j where $z_{j,k} \sim \text{Bernoulli}(\pi_k)$. Note that because each individual $z_{j,k}$ is drawn from a Bernoulli that a treatment document can have more than one latent feature, allowing for multi-dimensional treatments.

The authors assume a mapping from Z_i to the standardized term-document matrix X_i through the D-dimensional vector A_k, where $X_i \sim \text{MvtNormal}(Z_i A, \sigma_n^2 I_D)$. The latent feature vector Z_i also affects the response Y_i through the normal, $Y_i \sim \text{Normal}(Z_i\beta, \tau^-1)$ where $\tau \sim \text{Gamma}(a, b)$. Thus with the model the authors both want to discover the latent treatments Z_i *and* estimate their influence on the outcome by estimating β. The authors use variational approximation to estimate these parameters.

Fong and Grimmer (2016) apply the sIBP to the training data in order to learn a g function. In the test set Fong and Grimmer (2016) use the g function to infer the treatments that are present in a particular text, but alter the inference to avoid conditioning on the dependent variable. They do this because otherwise the inferred treatments present in the test set will depend upon the observation's response to that text, which creates obvious problems for causal inference.

Once the latent treatments are inferred in the test set documents, their effect can be estimated using any procedure that might be used to analyze an experiment. Fong and Grimmer (2016) use a simple linear regression with each of the latent features as the regressors to estimate the effects of the treatments. More complicated models could be used to estimate interactions or to extrapolate effects to a different population of documents.

6 Application: Consumer Financial Protection Bureau

To illustrate the use of text-based treatments, we analyze the determinants of a timely resolution to complaints filed at the Consumer Financial Protection Bureau (CFPB). The CFPB is a product of Dodd-Frank legislation and is (in part) charged with offering protections to consumers. The CFPB solicits complaints from consumers across a variety of financial products and then addresses those complaints. It also has the power to secure payments for consumers from companies, impose fines on firms found to have acted illegally, or both.

The CFPB is particularly compelling for our analysis because it provides a massive database on the text of the complaint from the consumer and how the company responded. If the person filing the complaint consents, the CFPB posts the text of the complaint in their database, along with a variety of other data about the nature of the complaint. For example, one person filed a complaint stating that "the service representative was harsh and not listening to my questions. Attempting to collect on a debt I thought was in a grace period ...They were aggressive and unwilling to hear it" and asked for remedy. The CFPB also records whether a business offers a timely response once the CFPB raises the complaint to the business. In total, we use a collection of 113,424 total complaints downloaded from the CFPB's public website.

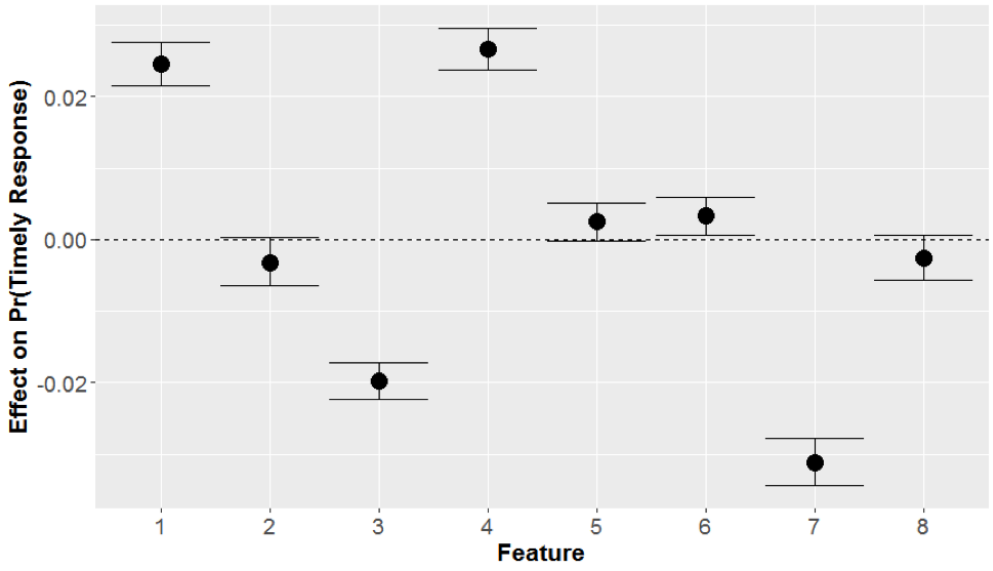
While the texts are not randomly assigned to the CFPB, the way complaints are recorded at the CFPB makes it plausible that we have all the *observable* features necessary to understand the determinants. The bureaucrats at the CFPB have little information other than what is in the complaints and have little capacity to collect additional information. Given these limitations, the assumption that the texts provide all the information for the outcome seems reasonable. It could be violated if there are other non-textual factors that correlate with the text content. For example, if working with the CFPB directly to resolve the complaint were important and individuals who submitted certain kinds of complaints were less well equipped to assist the CFPB, then we would have concern the selection on observables assumption holds.

We are interested in learning both the salient features that determine a salient response and then the effect of those features on the probability of a response. To do this, we first randomly divide the data, placing 10% in the training set and 90% of the data in the test set. We place more data in the test set because our large sample provides ample opportunity to discover the latent-treatments in the training set and to provide greater power when estimating effects in the test set. In the training set we apply the sIBP process to the text of the complaints and whether there was a timely response to infer the latent treatments. We use an extensive search to determine the number of features to include and the particular model run to use. The sIBP is a nonparametric Bayesian method, so it will estimate the number of features to include in the model, though the number estimated from a nonparametric method rarely corresponds to the optimal number for a particular application. To select a final model we then evaluate the candidate model fits utilizing a model fit statistic introduced in Fong and Grimmer (2016) that provides a quantitative

Table 5: Consumer Financial Protection Bureau Latent Treatments

No.	Automatic Keywords	Manual Keyword
1	payment, payments, amount, interest, balance, paid, month	loan
2	card, called, call, branch, money, deposit, credit_card, told	bank
3	debt, debt_collection, account, number, validation, dispute, collection	debt collection
4	xxxx, account, time xxxx_xxxx, request, copy, received, letter	detailed complaint
5	payment, payments, pay, told, amount, month, called	disputed payment
6	loan, mortgage, modification, house foreclosure, payments	mortgage
7	debt, debt_collection, collection, credit_reporting, proof, credit_report	threat
8	fcra, credit_report, credit_reporting, reporting, debt, violation, law	credit report

Figure 2: The Effect of Complaint Features on a Prompt Response



measure of model fit. And because we are using a training set, we can also refit the model several times choosing the model that provides the features that provide the best substantive insights.

Once we have fit the model in the training set, we use it to the infer the treatments in the test set. Table 5 provides the inferred latent treatments from the CFPB complaint data. The *Automatic Keywords* are the words with the largest values in the estimate latent factors for each treatment, and the manual keyword is a phrase that we assign to each category after assessing the categories. Using these features we can then infer their presence or absence in the treated documents and then estimate their effect. To do this we use the regression procedure from Fong and Grimmer (2016) and then use a boostrap to capture uncertainty from estimation.

Figure 2 shows the effects of each latent feature on the probability of a timely response. The black dots are point estimates and the line are 95-percent confidence intervals. Figure 2 reveals that when consumers offer more detailed feedback (Treatment 4) and when complaints are made about payments to repay a loan. In contrast, the CFPB is much less successful at obtaining prompt responses from debt collectors—either when those collectors are explicitly attempting to collect a debt (Treatment 3) or when the debt collectors are threatening credit reports (Treat-

ment 7). The inability to obtain a prompt response from debt collectors is perhaps not surprising—debt collection companies exist to successfully recoup funds and are likely less concerned with their perceived reputation with debtors. It also demonstrates that it can be harder to remedy consumer complaints in some areas than others, even if the FSPB is generally able to assist complaints.

7 How do we know we've learned something?

With data as complicated as high-dimensional as text and so many different possible mappings between text and the latent variables, we might be concerned about the generalizability of our inferences and with validating our conclusions in other data and domains. Do the findings hold only in this dataset? Did our randomly drawn training/test split result in a specific analysis that would not occur in other dataset? Might other researchers have chosen a different g function that would have provided different conclusions?

We argue that the only way to ensure generalizable findings is through a variety of replications. Each step in the research process offers some place where different decisions could have been made or different strategies could have been pursued. While this "garden of forking paths" is often used to criticize empirical research, recognizing the forked paths and varying as many of the decisions as possible will provide guidance on the robustness of our findings. We argue for replication because results achieved with a training/test split are internally valid, results that hold after replicating all stages of the research process are more general than those that do not withstand replication in new datasets or to new models. Here we introduce various levels of replication that serve as levels of validation of the generalizability of the findings for the types of experiments we discuss above.

Fit the model on new test data. The researcher might use the model they obtained through exploration in the training set to estimate the impact of treatment in a different test set than their original test set. This new test data could simply be another randomly split subset of the original data. Or it could be test data collected in a different sample or even in a different domain. If the results were similar in a number of varied test sets, we might be more certain that the findings generalize more broadly.

Estimation of the model – model selection The researcher might re-estimate the model in the original training set and provide a set of model estimates based off of a variety of models. This would show that the results were not specific to a particular coding scheme g, regardless of whether that g came from hand-coding or a topic model.

Training split The researcher might re-sample the training/test split. This would require re-estimating the model and validating the results in a new test set. If the results were robust to the training/test split it would show that the result could be

discovered under multiple training/test splits. Re-estimating by repeatedly drawing new training and test splits introduces the possibility for an AISV. But even when confronted with that violation, redrawing the split (and following some fixed procedure) enables an assessment of the stability of the algorithm.

Collect data The researcher might completely recollect the data. This would automatically generate a new training/test split because the data would all be completely different. This would show that the results are robust to variation in the implementation of data collection and to a different set of individuals.

Choose a domain/context The researcher might decide to ask the same question in a different domain or context, i.e. in a different population or time period. Replication in this context would show the most broad generalizability of the findings as the result would hold in broader populations of individuals.

7.1 Conclusion

Making causal inferences with textual data is difficult to do, even within an experimental context. While high-dimensional data like text requires the researcher to discover the mapping between the data and the quantities of interest, this process of discovery undermines the researcher's ability to make causal inferences. We clarify this issue by describing the Analyst-induced SUTVA violation and show that this type of SUTVA violation can be solved with a simple split of the dataset into a training set used for discovery and a test set used for testing. More broadly, we advocate for research designs that allow for sequential experiments that explicitly set aside research degrees of freedom for discovery of interesting measures, while rigorously testing relationships within experiments once these measures are defined explicitly.

While we set out the basic framework for doing causal inferences with text data here, there is much more work to be done to understand the optimal research design within this framework, for example we need a deeper understanding of the tradeoffs in train/test splits. Very little research has explicitly tackled methods for discovery of new measures, and we hope that future work will help spell out the types and amount of data best suited for discovery from data. Last, the analyst-induced SUTVA assumption does not only apply to text data, but to any latent representation of treatment and outcome variables. The framework of the train/test split should be considered in the context of other types of data that require a low-dimensional representation of high-dimensional data.

A Using the Structural Topic Model for Text as Outcome

This appendix provides a short review of the Structural Topic Model, explains how to obtain and use the g function and provides an analysis of stability of topic model results across train-test splits. In this appendix we deviate slightly from the notation in the main paper to use bold-faced text to indicate vectors or matrices.

A.1 A Brief Review of the Structural Topic Model

The Structural Topic Model is a mixed membership model of texts related to Latent Dirichlet Allocation (Blei, 2012) which was developed in Roberts et al. (2014); Roberts, Stewart and Airoldi (2016) and implemented in Roberts, Stewart and Tingley (2017). It allows for the analyst to incorporate observed document metadata which is able to affect either topical prevalence (the amount which a topic is discussed) and topical content (the way in which a topic is discussed). In this paper we consider the case in which a set of observed metadata which includes the treatment and pre-treatment covariates are allowed to affect topic prevalence and there are no topical content covariates. Denoting the pretreatment covariates for document d as $\boldsymbol{X}_d$ and the treatment as T_d, the generative process can be given as:

$$\begin{aligned}
\boldsymbol{\eta}_d &\sim \text{Normal}(\boldsymbol{X}_d\boldsymbol{\gamma}_X + T_d\gamma_\tau, \boldsymbol{\Sigma}) \\
\theta_{d,k} &= \frac{\exp(\eta_{d,k})}{\sum_{j=1}^{K}\exp(\eta_{d,j})} \\
z_{d,n} &\sim \text{Categorical}(\boldsymbol{\theta}_d) \\
w_{d,n} &\sim \text{Categorical}(\boldsymbol{\beta}_{z_{d,n}})
\end{aligned}$$

Where $\boldsymbol{\theta}_d$ is a K-dimensional vector on the simplex indicating the proportion of the document allocated to each topic formed by applying the softmax function to $\boldsymbol{\eta}_d$ a vector in $\mathcal{R}^{K-1}$ where the K-th element is fixed to zero. $z_{d,n}$ is a token level latent variable containing the assignment for token n of document d. $\boldsymbol{\beta}$ is a K by V dimensional matrix where each row contains the conditional probability of seeing word v given that is about topic k. The model differs from Latent Dirichlet Allocation in its use of a logistic normal prior distribution for the document-topic proportions and through the ability to have that prior centered at a document-specific location determined by the document metadata.

The model is estimated using partially-collapsed, non-conjugate, variational inference. $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ are given regularizing priors of the user's choice and β is point estimated. The model optimization problem is non-convex and so a careful initialization strategy is necessary (Roberts, Stewart and Tingley, 2016). Roberts, Stewart and Tingley (2016) advocate a deterministic initialization based on the spectral method of moments (Arora et al., 2013) which we refer to below as the 'Spectral Initialization.'

A.2 Obtaining and Using the g Function

In this section we walk through the procedure for obtaining and using the g function with the Structural Topic Model. The experimental protocol is summarized below. Note that any other method can be used for the discovery stage but here we demonstrate with the STM for simplicity.

- Create the train-test split
- In the training set (discovery)
  - explore the documents as desired using STM
  - choose an estimand (including assigning and validating a label)
  - Identify the mapping function g such that

$$\hat{\boldsymbol{\theta}}_i = g(\boldsymbol{Y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}, \hat{\boldsymbol{\Sigma}})$$

- In the test set (evaluation)
  - Using the g function, obtain our transformed outcome for each document. (see below for details)
  - Estimate treatment effects (using for example the difference of means)
  - Validate model fit and label fidelity in the test set.

Application of the g function in STM is equivalent to predicting $\boldsymbol{\theta_i}$ for a held-out document i. This can be accomplished with the recently added `fitNewDocuments` function in the `stm` package. In the STM model, the latent variable $\boldsymbol{\theta_i}$ is a function of a global prior $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the topic word parameters $\boldsymbol{\beta}$ and the observed words $\boldsymbol{W_d}$. The token-level latent variables $\boldsymbol{Z}$ are integrated out. We have estimated $\boldsymbol{\beta}$ in the train set and in many ways this communicates what the topics substantively contain. We must also decide how to set our priors $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The `stm` package offers three options: no prior, the covariate-specific prior and the average prior. The 'no prior' setting sets $\boldsymbol{\mu}$ to a vector of zeroes and $\boldsymbol{\Sigma}$ to be a diagonal matrix with very large diagonals. The covariate-specific prior uses the observed covariates in the new documents to construct the document-specific prior. The average prior averages over the values of $\boldsymbol{\mu}$ in the training set and provides a single average prior for all documents.[16]

If we have used only pre-treatment covariates in the STM model we can use any of these strategies. In our application we do include the treatment and so we cannot use the covariate-specific prior because then the same text would yield two different values of the outcome depending on the treatment assignment. For our application we use the average prior. In future work we plan to implement functionality in `stm` that will allow us to marginalize out a single covariate such as treatment so that we

[16] More specifically we take the column means of the D by $K-1$ matrix $\boldsymbol{\mu}$ in the training set which we call $\tilde{\boldsymbol{\mu}}$. We then recalculate $\boldsymbol{\Sigma}$ as though the update had been made using the new value of μ. The update is then $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \left(\sum_d (\boldsymbol{\eta_d} - \boldsymbol{\mu_d})(\boldsymbol{\eta_d} - \boldsymbol{\mu_d})^T\right) + \left(\sum_d (\boldsymbol{\eta_d} - \tilde{\boldsymbol{\mu}}_d)(\boldsymbol{\eta_d} - \tilde{\boldsymbol{\mu}}_d)^T\right)$.

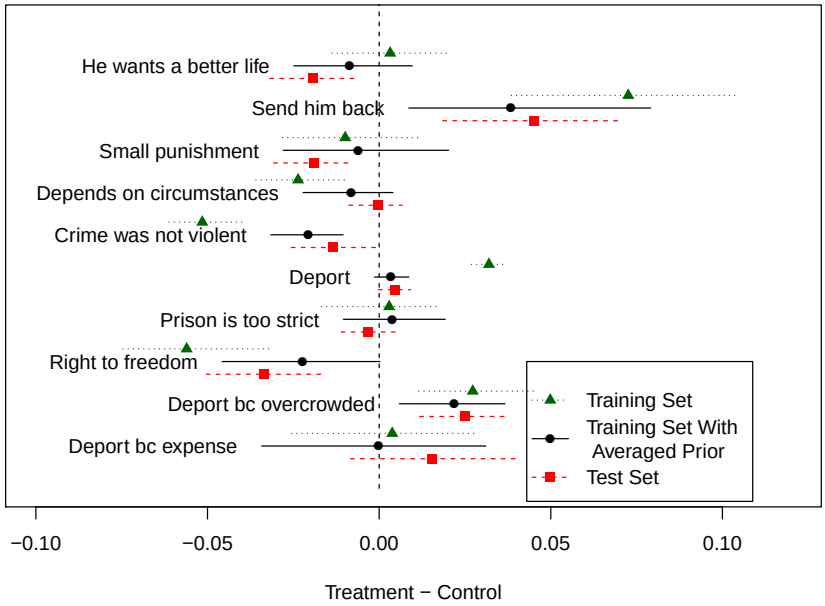


Figure 3: Train-Test set effect comparing g function using the model estimates (training set), the training set with averaged prior and the test set. Note that while the estimates are broadly similar, in general the training set with averaged prior is a closer approximation to what we end up seeing.

can still make use of pre-treatment covariates in the test set. This still requires us to make a choice about the what distribution of the treatment to marginalize over.

When using a version of the g function which is not the covariate-specific prior, we recommend that analysts assess effects in the training set using the same procedure as in the test set. While the effects will generally not be very different (particularly for long documents), maintaining the same procedure should provide a better expectation of test set behavior. For example, in our application Figure 3 compares our training set estimates using both the covariate-specific prior and the averaged prior and compares them to the test set (which uses the averaged prior). Using the average prior to make predictions in the training set before calculating effect estimates gives us a better indication of what we will eventually observe in the test set.

A.3 Stability Across Train-Test Splits

Our approach does not require stability of analysis across different train-test splits. Different train-test splits might lead to different discovery phases which in turn yield different estimands and test sets where we can evaluate that estimand. Nevertheless, we might be slightly uncomfortable with the idea that the particular randomization into the train-test spit yield quite different estimands (and papers) at the end of the

process. As such we wanted to evaluate the stability of the STM under different samples of a fixed population.

In a formal sense we are interested in studying the posterior contraction rates of the model, a problem taken up analytically in Tang et al. (2014) for the related Latent Dirichlet Allocation model. However, we are far more interested in understanding performance in practice and whether different train-test splits lead to substantively different topic-word distributions ($\boldsymbol{\beta}$), different document-topic proportions ($\boldsymbol{\theta}$) and different covariate effects. As the number of documents increase or the topics are more sharply defined stability will improve. For this demonstration we use the Poliblog data (Eisenstein and Xing, 2010), a collection of around 13,246 posts from six different political blogs in the runup to the 2008 American presidential election. We use this because it is readily available for use with the `stm` package and is roughly representative of the document lengths that we often see in `stm` applications overall. We would expect that the diversity of topics in political blogs would make the problem harder than the more focused open-ended response case, but the length of the documents would make it easier.

We started by estimating the model on the full set of documents with 20 topics using the spectral initialization. We consider this to be the "truth" because the unattainable stability ideal would be that the subsamples provide the same answer as the full set of documents. We then choose two prominent topics to be our "outcomes" a topic about Obama and a topic about War (particularly Iraq and Afghanistan). In each simulation we choose the topic that most closely approximates our two chosen outcomes, emphasizing that the labels 'Obama' and 'War' may well not be good approximations for the topic in the subsample.

Because of the multimodality problem in topic models, instability could arise from two sources: differences in the local mode discovered and differences in the data observed. We investigate this by considering three different initialization strategies:

1) Cold Spectral Start
   Using the spectral initialization on the subsample. This is reflective of current practice.

2) Warm Spectral Start
   Use the complete data to initialize the model. This would create an analyst-induced SUTVA violation as it shares information from the test set. However, it is suggestive of what might be achievable by providing more stable intializations.

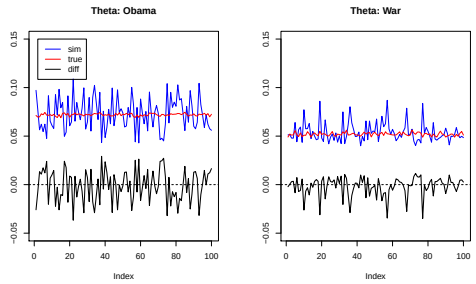
3) Warm Oracle Start
   Use the results of the *converged* model on the full sample to initialize each subsample. This is an infeasible estimator. The instability in this estimate cannot be reduced by a better initialization strategy.

In each case we run the model on the sample sizes 100 times and plot the results along with the 'truth' as defined by the full document set. Figure 4 shows the results for the average proportion of the topic use in the corpus.

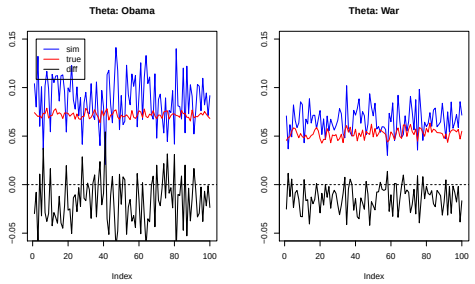
As we can see the results are reasonably stable at 5000 documents for a corpus of this complexity and less so with 1000. The warm spectral start shows considerable

Figure 4: Stability of θ in Simulations of Train-Test Splits on Real Data.

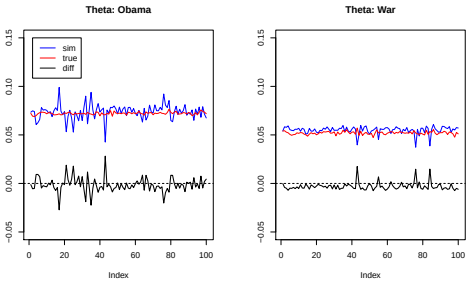
Sample=5000 with Cold Spectral start



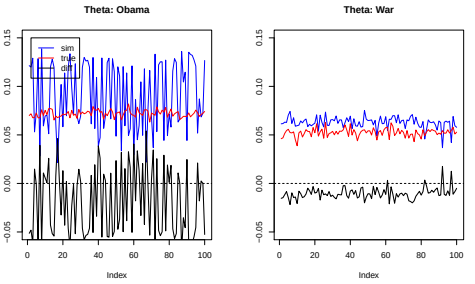
Sample=1000 with Cold Spectral Start



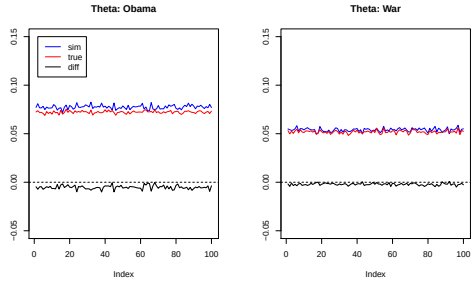
Sample=5000 with Warm Spectral Start



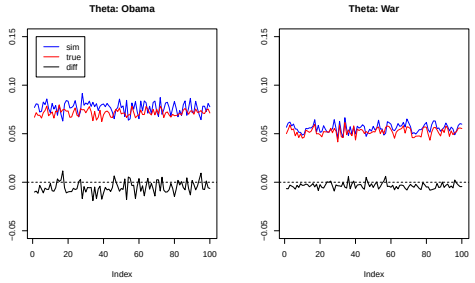
Sample=1000 with Warm Spectral Start



Sample=5000 with Warm Oracle Start



Sample=1000 with Warm Oracle Start



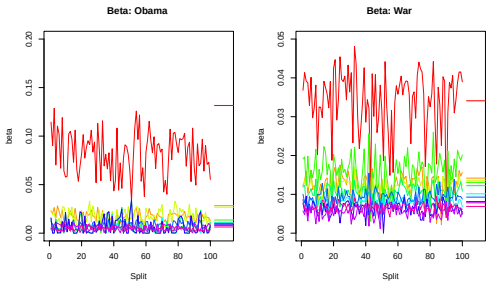
improvement for the 5000 document case suggesting that at least at this scale, we might see substantial gains from an initialization specificly designed to be more stable across splits. The near perfect stability of the warm oracle start for the 5000 document case suggests it is a matter of the initialization and not necessarily the data itself, where for 1000 documents there is evidence that some level of the instability is unavoidable given the model.

We can also examine the word-distributions themselves. Figure 5 shows the proportion of mass associated with each of the top ten words in the topic (as chosen by the full model). The horizontal tickmarks on the right show the estimate in the full data.

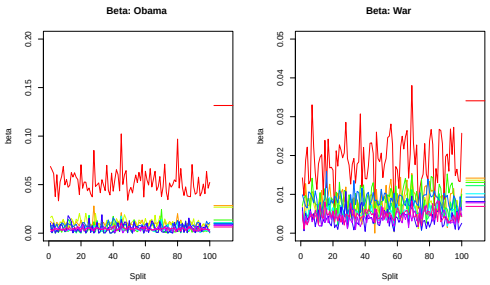
Generally speaking the models correctly preserve the rank ordering of the most prominent words in each topic, but the estimates can often be substantially incorrect. We do emphasize that there is relatively little information with which to

Figure 5: Stability of β in Simulations of Train-Test Splits on Real Data.

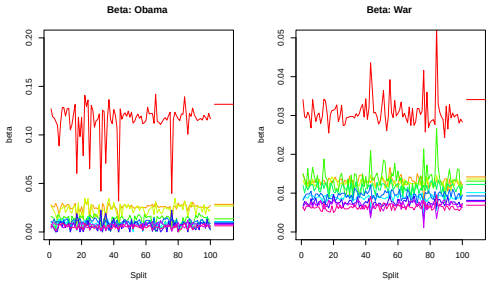
Sample=5000 with Cold Spectral Start



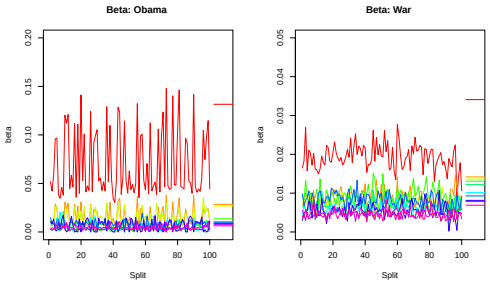
Sample=1000 with Cold Spectral Start



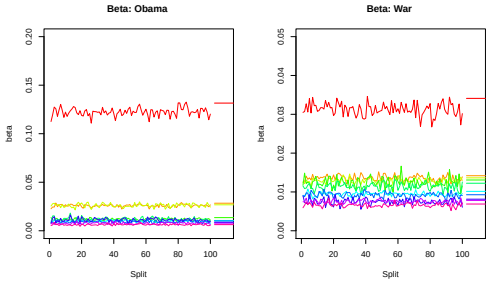
Sample=5000 with Warm Spectral Start



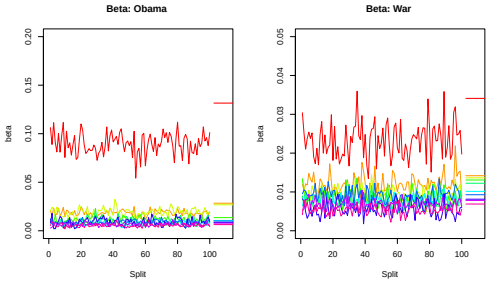
Sample=1000 with Warm Spectral Start



Sample=5000 with Warm Oracle Start



Sample=1000 with Warm Oracle Start



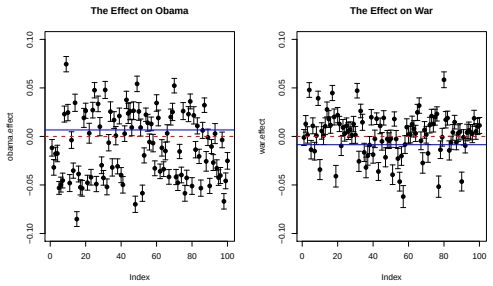
estimate these parameters and so we would expect to see more instability than in the simulations for θ.

Finally we present estimates of "treatment effects." Here we use the binary rating variable (indicating whether the blog is liberal or conservative) as a treatment. This is clearly not randomly assigned and we use it simply because it is a binary covariate we would expect to influence the outcome in some way. We plot the estimate with a 95% confidence interval in Figure 6 along with the estimate in the complete dataset shown in blue.

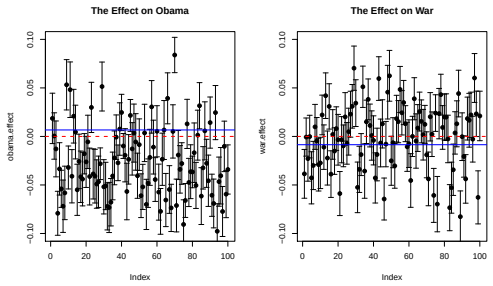
Once again we can see some substantial variability that appears to be reducible via a more stable initializations. We emphasize that we should not expect these confidence intervals to have proper coverage as in every case the estimand is different. Indeed the high variance on the Obama topic of the warm spectral start is a good indication that the estimand is changing substantially in each different split. What

Figure 6: Stability of Covariate Effect in Simulations of Train-Test Splits on Real Data.

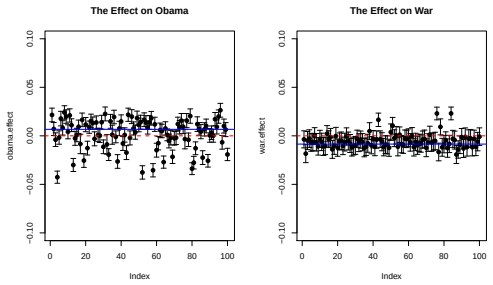
Sample=5000 with Cold Spectral Start



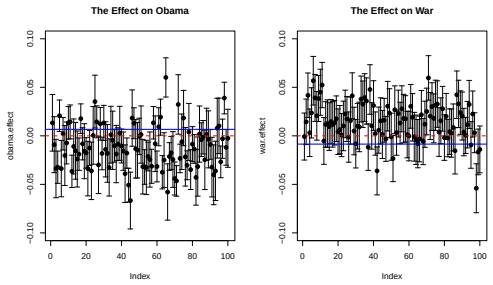
Sample=1000 with Cold Spectral Start



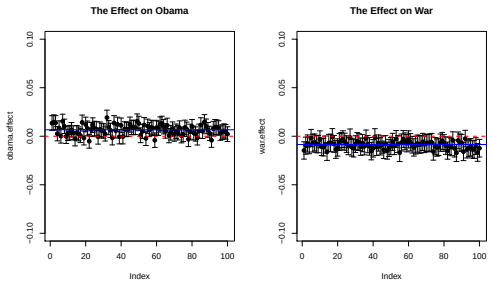
Sample=5000 with Warm Spectral Start



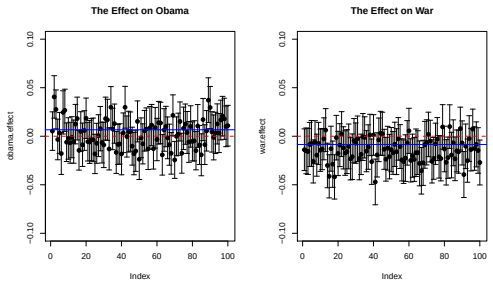
Sample=1000 with Warm Spectral Start



Sample=5000 with Warm Oracle Start



Sample=1000 with Warm Oracle Start



the relatively tight set of estimates in the two warm starts suggest is that this might be avoidable with a different initialization.

In summary, there is a significant degree of instability across splits. Again, this is not a problem in a technical sense as the g function applied to the test set will still provide a proper estimator of that specific estimand. What these simulations do suggest is that further research into more stable initialization strategies might substantially reduce the amount of instability across train-test splits.

There are several limitations to this simulation study: most notably that we neither know the actual truth nor can we be sure what the scope conditions are

for these results to apply to other datasets. We also cannot simulate the stability of the entire discovery process, only that a particular model is comparable across subsamples. Hoping for stability in discovery may be quixotic as the very idea of discovery itself may imply some level of instability.

References

Airoldi, Edoardo M, David Blei, Elena A Erosheva and Stephen E Fienberg. 2014. *Handbook of mixed membership models and their applications*. CRC Press.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*. pp. 280–288.

Athey, Susan. 2015. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM pp. 5–6.

Athey, Susan and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Biernacki, Richard. 2012. *Reinventing evidence in social inquiry: Decoding facts and variables*. Springer.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55(4):77–84.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3(Jan):993–1022.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon and Bin Yu. 2016. "Lasso adjustments of treatment effect estimates in randomized experiments." *Proceedings of the National Academy of Sciences* 113(27):7383–7390.

Bonica, Adam, Adam S Chilton and Maya Sen. 2015. "The Political Ideologies of American Lawyers." *Journal of Legal Analysis* 8(2):277–335.

Catalinac, Amy. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *The Journal of Politics* 78(1):1–18.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2017. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* pp. n/a–n/a.

Cohen, Mark A, Roland T Rust and Sara Steen. 2002. "Measuring public perceptions of appropriate prison sentences: Report to National Institute of Justice." *NCJ Report* (199365).

Cohen, Mark A, Roland T Rust and Sara Steen. 2004. "Measuring perceptions of appropriate prison sentences in the United States, 2000. ICPSR version. Nashville, TN: Vanderbilt University [producer], 2000." *Ann Arbor, MI: Inter-university Consortium for Political and Social Research.[distributor]* .

Cranmer, Skyler J. and Bruce A. Desmarais. 2017. "What Can We Learn from Predictive Modeling?" *Political Analysis* 25(2):145166.

Denny, Matthew J and Arthur Spirling. 2017. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it.".

Doshi, Finale, Kurt Miller, Jurgen V Gael and Yee W Teh. 2009. Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*. pp. 137–144.

Eisenstein, Jacob and Eric Xing. 2010. "The CMU 2008 Political Blog Corpus.".

Fong, Christian and Justin Grimmer. 2016. Discovery of Treatments from Text Corpora. In *Association of Computational Linguistics*.

Fong, Christian, Justin Grimmer, Brandon M. Stewart and Margaret E. Roberts. 2017. "Exploratory and Confirmatory Causal Inference for High Dimensional Interventions.".

Gelman, Andrew and Eric Loken. 2014. "The Statistical Crisis in Science Data-dependent analysisa garden of forking pathsexplains why many statistically significant comparisons don't hold up." *American Scientist* 102(6):460.

Gelman, Andrew and John Carlin. 2014. "Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors." *Perspectives on Psychological Science* 9(6):641–651.

Gill, Michael and Andrew B Hall. 2015. "How Judicial Identity Changes The Text Of Legal Rulings.".

Griffiths, Thomas L and Zoubin Ghahramani. 2011. "The indian buffet process: An introduction and review." *Journal of Machine Learning Research* 12(Apr):1185–1224.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

Grimmer, Justin and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2012. "How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation." *American Political Science Review* 106(4):703–719.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* .

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2013. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* 22(1):1–30.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown and Matt Taddy. 2016. "Counterfactual Prediction with Deep Instrumental Variables Networks." *arXiv preprint arXiv:1612.09596* .

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2013. *The elements of statistical learning*. Springer.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter Van der Windt. 2013. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration." *Political Analysis* 21(1):1–20.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. Sage.

Lanza, Stephanie T, Donna L Coffman and Shu Xu. 2013. "Causal inference in latent class analysis." *Structural equation modeling: a multidisciplinary journal* 20(3):361–383.

Lasswell, Harold Dwight. 1938. "Propaganda technique in the world war.".

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331.

Lu, Jiannan, Peng Ding and Tirthankar Dasgupta. 2015. "Sharp bounds of causal effects on ordinal outcomes." *arXiv preprint arXiv:1507.01542* .

Neuendorf, Kimberly A. 2016. *The content analysis guidebook*. Sage.

Nguyen, XuanLong. 2015. "Posterior contraction of the population polytope in finite admixture models." *Bernoulli* 21(1):618–646.

Pennebaker, James W, Matthias R Mehl and Kate G Niederhoffer. 2003. "Psychological aspects of natural language use: Our words, our selves." *Annual review of psychology* 54(1):547–577.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Ratkovic, Marc and Dustin Tingley. 2017. "Causal Inference through the Method of Direct Estimation." *arXiv preprint arXiv:1703.05849* .

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2016. Navigating the Local Modes of Big Data: The Case of Topic Models. In *Computational Social Science: Discovery and Prediction*, ed. R. Michael Alvarez. New York: Cambridge University Press chapter 2, pp. 51–97.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2017. *stm: R Package for Structural Topic Models*. R package version 1.2.3.

Roberts, Margaret E, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M. Stewart and Edoardo M Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Roberts, Margaret E, Brandon M. Stewart and Richard Nielsen. 2017. Matching methods for high-dimensional data with applications to text. Technical report Working paper.

Spirling, Arthur. 2012. "US treaty making with American Indians: Institutional change and relative power, 1784–1911." *American Journal of Political Science* 56(1):84–97.

Tang, Jian, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning*. pp. 190–198.

Tukey, John W. 1980. "We need both exploratory and confirmatory." *The American Statistician* 34(1):23–25.

Van der Laan, Mark J and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Volfovsky, Alexander, Edoardo M Airoldi and Donald B Rubin. 2015. "Causal inference for ordinal outcomes." *arXiv preprint arXiv:1501.01234* .

Wager, Stefan and Susan Athey. 2017. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* .

Ward, Michael D, Brian D Greenhill and Kristin M Bakke. 2010. "The perils of policy by p-value: Predicting civil conflicts." *Journal of Peace Research* 47(4):363–375.