Matching Methods for High-Dimensional Data with Applications to Text*

Margaret E. Roberts, Brandon M. Stewart, and Richard Nielsen

This draft: July 16, 2015

*We thank the following for helpful comments and suggestions on this work: David Blei, James Fowler, Seth Hill, Gary King, Adeline Lo, David Mimno, Jennifer Pan, Caroline Tolbert, and audiences at the Princeton Text Analysis Workshop and the Visions in Methodology conference. We especially thank Dustin Tingley for numerous insightful conversations on the connections between STM and causal inference. Dan Maliniak, Ryan Powers, and Barbara Walter graciously supplied data and replication code for the gender and citations study.

Abstract

Matching is a popular technique for preprocessing observational data to facilitate causal inference and reduce model dependence by ensuring that treated and control units are balanced along pre-treatment covariates. While most applications of matching balance on a small number of covariates, we identify situations where matching with thousands of covariates may be desirable, such as causal inference where confounders are measured with text. With high-dimensional covariates, traditional matching methods are less effective and may be difficult or impossible to implement. We characterize the problem of matching in a high-dimensional context as a tradeoff between dimension reduction and imbalance bounding. We develop a new method called Topical Inverse Regression Matching (TIRM) that optimizes this tradeoff by including both a low-dimensional projection of covariates and information about the probability of treatment. We illustrate our approach by estimating the effect of censorship on the writing of Chinese bloggers, the effects of gender on citation counts in international relations, and the effects of targeted killings and capture by counterterrorists on the popularity of jihadist writings.

1 Introduction

Matching is a well-developed technique for finding appropriate counterfactuals for treated units within observational data (Rubin, 2006). Matching methods have been shown to improve balance along pre-treatment covariates and eliminate extreme counterfactuals, reducing model dependence (Ho et al., 2007; Morgan and Winship, 2014). Methods for matching have also been developed for cases with relatively few treated units, where synthetic matches, weighted combinations of control units, provide the counterfactuals (Abadie, Diamond and Hainmueller, 2010).

While the matching literature is large and well-developed, current methods for matching are predominately developed for cases with a relatively small number of pre-treatment covariates. Popular approaches such as propensity score matching (Rosenbaum and Rubin, 1983) and coarsened exact matching (CEM) (Iacus, King and Porro, 2011) either explicitly or implicitly assume that the dimension of the pre-treatment confounders is far smaller than the number of observations in the data set. For example, Rubin and Thomas (1996, 249) note that in "typical examples" of matching, "N_t [the number of treated observations] is between 25 and 250, the ratio $R = N_c/N_t$ [the ratio of control observations to treated observations] is between 2:1 and 20:1, and the number of matching variables, p, is between 5 and 50, although in some examples, N_t may be 1000 or more, R a hundred or more, and p may be a hundred or more." In computational social science the dimensionalty of rich data sources available to researchers is much larger than these figures, and is quickly outpacing the matching techniques available to condition on that information (King, 2009; Lazer et al., 2009). In this paper we address a particular type of high dimensional matching, where the pre-treatment confounder is captured in written text.

Matching on high dimensional confounders poses three distinct challenges that make the use of existing methods infeasible. The first challenge is that as the dimension of the pre-treatment covariates increases there will tend to be no observations that are nearly the same along all observed dimensions. Thus methods such as exact matching that require observations to match across all covariates will typically produce no matches, effectively

pruning the entire data set (Rubin and Thomas, 1996, 250). Methods such as coarsened exact matching (Iacus, King and Porro, 2011) weaken the restrictions of exact matching by matching within coarsened strata along each dimension, however such methods do not allow coarsening to be applied across variables, only within variables.

The second challenge is that high-dimensional data makes it difficult to estimate predictive models of treatment. Propensity score matching and related methods rely on the ability to build a predictive model of treatment (Rosenbaum and Rubin, 1983). To reduce bias, we only need to condition on the subset of variables that are related to treated and control. However in high-dimensional data sets, the analyst usually has too many variables to use standard regression techniques; moreover, many of these variables are not relevant to predicting treatment status. A combination of inverse regression and automated variable selection can be used to estimate the probability of treatment for any individual observation. These dimensionality reduction techniques are imperative to increasing the efficiency of matches, as matching on variables that are orthogonal to treatment will randomly prune observations from the matched dataset.

The third challenge is the relative lack of appropriate balance metrics for high-dimensional data in general and text data in particular. When the analyst has a well-defined balance metric that accurately measures the properties of the data that make observations suitable counterfactuals this balance metric can be directly optimized (Hainmueller, 2011; Diamond and Sekhon, 2013; Imai and Ratkovic, 2014). With a suitable balance metric the central decision for matching is identifying the appropriate trade-off between balance and sample size (King, Lucas and Nielsen, 2015). However, with text data, for example, the best measure of imbalance is often a human reading of matched pairs to evaluate if these pairs are sufficiently equivalent. As such, human balance checking is imperative.

In this paper, we set out to address each of these challenges with a particular focus on applications to text data. We develop two approximate analogs to existing matching methods. First, we show that a propensity score for text data can be estimated efficiently using multinomial inverse regression (Taddy, 2013*b*). Next, we show that matching on

topics in text is similar to coarsened exact matching where coarsening occurs across sets of variables, in this case indicators for words. We demonstrate that while each of these methods has desirable properties, both have specific weaknesses that limit their applicability in all real-world settings. We develop an algorithm called Topical Inverse Regression Matching (TIRM), which combines the two previous methods in a way that retains their desirable properties while ameliorating their weaknesses.

The paper proceeds as follows. First, we explain possible use cases of text matching in a social science context and introduce the examples we use in the paper. Next, we set up the problem and introduce some notation. We then adapt CEM and propensity score matching to textual data and discuss their respective strengths and drawbacks. We introduce our approach for text matching that alleviates these drawbacks. Last, we apply our methods to three social science examples: understanding the effects of being censored on the reactions of bloggers, the effect of author gender on academic article citations, and the effects of killing jihadist clerics on their subsequent popularity.

2 Use Cases of High-Dimensional Matching

There are a number of contexts where social scientists may wish to use matching with high-dimensional covariate data. In the case of text, whenever similarity between treated and control observations could be measured in terms of writing, text matching could be used to find appropriate counterfactuals. Since politics produces vast amounts of writing, our method is widely applicable. For example, if a political scientist in American politics were interested in the effect of veto threats on repositioning in Congress, she may want to control for the content of a bill, which might confound vetos and repositioning. The effect of proximity of two countries on trade between them may be confounded by the type of trade agreement between two countries; our method allows analysts to control for the agreement directly. Studies of bias in college admissions or employment — the observational equivalents of experimental audit studies (Neumark, Bank and Nort, 1996) — could use our approach to control for the content of applicants' letters of recommendation or CVs.

While the methods described here are focused on matching on textual data, other types of high-dimensional data are becoming increasingly available that could use similar methods to those described here. An analyst estimating the effects of height on politicians' popularity may want to control for facial expressions of politicians captured in image data. High-dimensional biological data are now frequently collected on participants in lab experiments and scholars may want to use these data as controls. Other extremely fine-grained time-series data streaming from anything from cell phones to MRIs could also be used for matching purposes.

In this paper, we focus on three examples of high-dimensional matching in the case of text data to assist with social science inference. First, we examine the question of how censorship affects bloggers in China. Do bloggers avoid sensitive topics after they experience censorship? Or does censorship backfire, causing bloggers to write on more sensitive topics? Matching on the text of the blog posts, we measure self-censorship by comparing bloggers who have been censored to those who have not when the content of the blog post is identical or nearly so. We find that bloggers react negatively to censorship, writing on more sensitive topics than those who were not censored.

Second, we examine how the gender of journal article authors affects the rate at which articles are cited by others. Maliniak, Powers and Walter (2013) find that women get many fewer citations than men in international relations while controlling for article content using covariates coded from article text by humans. We replicate this study using automated matching on the text data in place of human coding. We find even stronger gendered citation results than Maliniak, Powers and Walter (2013) after using automated matching.

Last, we examine how targeted killings of jihadist clerics influence the popularity of their writings among jihadist readers on the internet. Focusing on the death of Usama Bin Laden, we test whether documents authored by Bin Laden became more popular after his death by matching them to similar texts by other authors. We find that Bin Laden's death increased the popularity of his writings for at least six months after his death and perhaps longer.

In the next sections, we introduce the matching methods and evaluate their properties. To help the reader gain intuition, we use the case of censorship in China as a running example. We go into more detail of the performance and results of each of the models at the end of the paper.

3 The Setup of the Problem

We begin by introducing the notation for matching in the context of text. We start with a data set of n observations. We assume that for each observation i there is a binary treatment T_i, which takes a value of 1 for treated and 0 for control. We adopt the potential outcome framework so that the outcome variable Y_i takes on the value $Y_i(1)$ when unit i is treated and $Y_i(0)$ when unit i is not treated. In the censorship case, censorship is the treatment T_i and the outcome Y_i is the subsequent reaction of the blogger.

Because there is no random assignment, treated and control groups may not be similar before treatment. In the censorship case, censored bloggers write about very different topics than uncensored bloggers and this could explain both their treatment status and the outcome: their subsequent writing. To approximate random assignment, we control for pre-treatment covariates by matching on k pre-treatment covariates $\boldsymbol{X} = (X_1, X_2, \ldots X_k)$. In order to estimate the population average treatment effect on the treated, conditional on $\boldsymbol{X}$, treatment must be independent of the potential outcomes: $T_i \perp\!\!\!\perp Y_i(1), Y_i(1)|\boldsymbol{X}$.

In the cases we consider in this paper, some confounding covariates could be measured with text data, for example, the content of the blog post. In these cases of high-dimensional matching, k is at best large and at worst undefined. For text, it isn't immediately clear how the features of each blog post should be represented. Text includes not only individual words, but hierarchies of words (titles and section headings) as well as word order. It could be that overlapping sets of five words within each text should make up the feature space, in which case k would include each unique consecutive five word set within the corpus. Or, if only section headings are important confounders, then the words within the section headings should be included in k.

For simplicity, we will represent each document as a vector of counts of each word it

contains. This is the "bag of words" assumption (Grimmer and Stewart, 2013). We will describe all text within our examples in dimension V, which we define as the number of unique words within the corpus. Thus all the documents are collected within a sparse count matrix $\boldsymbol{X}$ whose typical element X_{ij}, contains the number of times the jth word appears within the text associated with observation i. In this case the $\boldsymbol{X}$ matrix has dimension n, the number of observations, by $k = V + r$, where V is the number of unique words in the corpus and r are other covariates to be matched on in addition to the text. The methods described below also apply to relaxations to the bag of words assumptions such as n-grams and with some modifications to methods which operate on sub-string sequences (Spirling, 2012).

The dimensionality problem arises because the number of unique words, V, can be very large relative to the number of documents, n. The example corpora we use within this paper have V on the order of $10,000 - 30,000$ and in two out of three of the cases $V > n$. These examples are not atypical for text data.

Except for documents which are exact copies of each other, exact matching can not be conducted. Nor is exact matching necessarily desirable – many of the examples described above do not require the texts to be identical in order for causal inference to be conducted, texts only have to be sufficiently similar so as not to confound their relationship between treatment T and the outcome Y. Thus variable selection or coarsening across variables (in this case words) must first be completed to simplify $\boldsymbol{X}$ so that matching is tractable. Importantly, in order to still be able to make causal inferences, $\boldsymbol{X}$ must be selected so that $T_i \perp\!\!\!\perp Y_i(1), Y_i(1) | \boldsymbol{X}$ still holds.

4 Analogs to Current Methods

Simplifying $\boldsymbol{X}$ so that matching is tractable is not new to the matching literature: most matching methods are different ways of simplifying the information in $\boldsymbol{X}$ so that matching can be completed when exact matching is impossible. In this section, we adapt two separate matching techniques from the current literature to the text case which address this problem in distinct ways: propensity score matching (PSM) which estimates the

probability that a unit receives treatment and then matches on this one-dimensional projection, and coarsened exact matching (CEM) which coarsens along each dimension of the covariates until exact matching is tractable.

These two approaches use distinct approaches to handling dimensionality reduction of $\boldsymbol{X}$ and it is this difference which gives rise to their individual properties. We describe this difference with more mathematical precision below, but for pedagogical purposes we start with a simple example. Imagine that we want to estimate the causal effect of a job training program on income in observational data. To keep matters simple, assume we have only two covariates about each individual: age and education. The need for more elaborate matching methods arise because it is presumably nearly impossible to find a treated and control unit who have identical ages and levels of education, especially if both are measured continuously.

Propensity scores model the choice of each individual to enroll in the job training program as a function of age and education, and then matches along this estimated propensity. This means that two individuals with different ages and different levels of education can be matched together and treated as stochastically equivalent because both have a common propensity to enroll in the job training program. Put another way, the difference in age between a matched pair can compensate for the difference in education. Thus we address the difficulty of finding matches by relaxing the need for observations to match along all dimensions.

Coarsened exact matching adopts a different approach. First, each variable is separately coarsened; education, for instance, might be binned into categories such as *no high-school degree*, *high-school degree*, *college degree* and *post-graduate degree*. Second, the observations are exactly matched along the coarsened variable. For well chosen bins, all observations within a given stratum are stochastically equivalent, and observations are only matched when they share a stratum along every dimension. This means that while two matched observations may not have identical years of schooling, they fall into the same category and are thus fundamentally comparable. No difference in age can compensate for a difference in education as in the propensity score model. Indeed coarsened exact

matching uses no model of the probability of treatment, which means that observations must match along every included covariate.

PSM and CEM strategies yield a different set of costs and benefits (King and Nielsen, 2015). PSM requires treatment and controls units to match only along a single scalar and approximates a completely randomized experiment. CEM requires units to match across all covariate dimensions, but in return approximates a more powerful fully blocked design.

With these differences in mind, in the next section we develop analogs of these two methods for the text case which introduce information about the distribution of the high-dimensional confounders through the use of a generative model. We then show why each of these analogs is unsatisfying on its own. Indeed high-dimensional data seems to amplify the weaknesses of each of these methods, though we emphasize that this is not an indictment because these methods were not necessarily designed for such data. PSM and CEM do provide a useful starting point for the framework we present in Section 5.[1]

4.1 Matching on probabilities: the Propensity Score and MNIR

Propensity score matching is a common approach to use when exact matching fails (Rubin, 2006, 178, 264, 283). The basic idea is to simplify $\boldsymbol{X}$ by estimating the probability of treatment conditional on $\boldsymbol{X}$, or:

$$\hat{\pi}_i = p(T_i = 1|X_i) \tag{1}$$

[1]We have chosen to develop analogs of two popular matching methods but there are numerous others we might have chosen. We briefly comment on two alternatives: Mahalanobis distance matching (MDM) and reweighting methods such as Entropy Balancing (Hainmueller, 2011). MDM uses a distance metric which normalizes the Euclidean distance by the sample covariance of the confounders. We return to this approach in Section 6.3 but simply note here that the high-dimensional setting often makes it difficult to accurately estimate the covariance matrix, leading to an inefficient estimator. Entropy Balancing reweights observations to match the moments of the treated and control distributions (Hainmueller, 2011). Reweighting approaches are incredibly powerful when the correct balance metric is known and easily quantified. However, the nature of the weights on the observations makes it difficult to identify individual pairs of units which serve as counterfactuals and thus it is more difficult to evaluate the quality of those matches after the fact. This is unnecessary in a case where we have full faith in a particular balance metric, but for high-dimensional data — particularly text — it is helpful to be able to qualitatively examine matches.

In typical practice this involves estimating a logistic regression where the treatment is the outcome. Then, instead of matching on the full $\boldsymbol{X}$, the estimated probabilities $\hat{\pi}_i$, or the linear predictor, are used to match.

The challenge for a direct application of this methodology to text data is that $\boldsymbol{X}$ contains a very high-dimensional representation of the texts, typically the word count matrix. The standard advice for variable selection with propensity score matching is that a variable should always be included "unless there is consensus that it is unrelated to the outcome variables or not a proper covariate" (Rubin, 2006, 269). However, high-dimensional data, the estimation of the conditional distribution $p(T_i|X_i)$ will not be tractable or efficient unless the number of observations n scales well with the dimension of the pre-treatment covariates k. We can obtain an estimate of the conditional distribution using regularization but it will necessarily be model-dependent and noisy.

4.1.1 Inverse Regression

To address this problem of estimating the conditional distribution efficiently, we adapt propensity score matching to the text case using inverse regression (Cook and Ni, 2005; Cook, 2007). The central idea is to posit a parametric model for the inverse problem, $p(\boldsymbol{X}|T)$, which allows us to obtain a sufficient reduction of the information in $\boldsymbol{X}$ about the conditional distribution $p(T|\boldsymbol{X})$. When the feature space consists of word counts, a natural approach is to assume that the word counts arise from a multinomial distribution which leads to the Multinomial Inverse Regression (MNIR) framework developed in Taddy (2013*b*). This leads to the model for a given document

$$X_i \sim \text{Multinomial}(\vec{q}_i, m_i) \tag{2}$$

$$q_{i,v} = \frac{\exp(\alpha_v + \psi_v' T_i)}{\sum_{v=1}^{V} \exp(\alpha_v + \psi_v' T_i} \tag{3}$$

where T is an ℓ-length containing a categorical encoding of the treatment variable. The coefficients ψ are often given a sparsity-inducing regularizing prior (a point which we return to below).

Mechanically this amounts to estimating a multinomial logistic regression with the

words as outcomes and the treatment as the predictor. After estimating the model we can calculate a sufficient reduction score:

$$z_i = \psi'(x_i/m_i) \rightarrow T_i \amalg x_i, m_i | z_i \tag{4}$$

where the latter part of the equations comes from Propositions 3.1 and 3.2 of Taddy (2013*b*) which establish the classical sufficiency properties of the projection. This implies that given the generative model in Equation 2 we can condition on z_i and discard the higher dimensional data x_i.

Introducing this information about the generative process of the predictor $\boldsymbol{X}$ results in a gain in the efficiency of the estimator.[2] Under the standard propensity score model the variance in the MLE of the coefficients for the propensity score model descreases in the number of documents. However with MNIR the variance decreases with the number of total words (Taddy, 2013*c*, See Proposition 1.1). In high-dimensional data this is hugely advantageous. We defer to Taddy (2013*b*) and Taddy (2013*c*) for a more complete description of the technical properties of MNIR and Cook (2007) for inverse regression more generally.

To complete the analogy, we can estimate the propensity score using the forward regression $\hat{\pi}_i = \frac{1}{1+\exp(-z_i\beta)}$. This step would be necessary in cases where the propensity score itself was directly of interest, for example if it was used in a weighting scheme (Glynn and Quinn, 2010). For the purposes of matching the forward regression provides no new information and we can match directly on the sufficient reduction.

4.1.2 Inference for Multinomial Inverse Regression

The multinomial inverse regression model involves estimating a large number of coefficients. The coefficient matrix ψ has one row per level of the treatment (so typically 2 in this setting) and one column per word in the vocabulary. We don't however expect that

[2]Efficiency is important in its own right but as pointed out in Robins and Morgenstern (1987) and King and Nielsen (2015) high variance estimators can lead to bias in practice. The logic is that even well-intentioned researchers will run multiple models and pick the best result. Thus when the estimator is inefficient it raises the possibility that these models will yield radically different answers which can produce an effective bias in the published work. See King and Nielsen (2015) for more on this point in the context of propensity score matching in particular.

there will be treatment effects on every word in the vocabulary. To simultaneously provide variable selection and estimation we follow Taddy (2013*b*) in estimating the coefficients with a regularizing prior.

Taddy (2013*b*) develops a particular penalization scheme called the Gamma-Lasso. This is a sparsity-inducing concave penalization method has the attractive property that it is asymptotically unbiased for large coefficients. It is motivated as maximum a-posteriori (MAP) estimation under the Bayesian prior $\psi_v \sim \text{Laplace}(0, \tau_v)$, $\tau_v \sim \text{Gamma}(s, r)$ for some fixed hyperparameters s and r. This prior essentially zeroes out coefficient for words where the ratio of the use under treatment to use control is neither too large or too small.

A direct implementation of the above model would be prohibitively computationally expensive. However, because the sum of multinomial random variables is itself multinomial, we can collapse the word counts by treatment status. This radically simplifies computation. Leveraging later work in Taddy (2015*a*) we also distribute computation across words in the vocab using the connection between the multinomial and poisson. This estimation framework including techniques for computation are further developed in Taddy (2013*b*, 2015*c*,*a*).

4.1.3 Using MNIR to Estimate Propensity Scores

In standard regressions, we condition on our covariates and thus don't need to specify a parametric generative model. The idea of inverse regression is to use the assumed generative model to improve the efficiency of our estimates. We first estimate z, a sufficient reduction of $\boldsymbol{X}$, and then the propensity score can be estimated using a forward regression using only the low-dimensional z variable as a predictor. Due to the use of the sparsity-promoting prior, we are effectively only considering a subset of words which the model estimates have substantially different rates of use in the treated group in comparison to the control group.

The advantage of this framework is that it provides a fairly straightforward connection to propensity scores. The literature on propensity scores is quite developed in areas beyond matching and other forms of conditioning such as inverse propensity score weighting could also be used with these methods. Computation is straightforward even with

extremely large datasets. Importantly, MNIR with regularization excels at selecting a small subset of variables that are related to treatment. When strong dimensionality reduction is required to create matches, identifying variables that are related to treatment and ensuring that these variables are included increases the efficiency of the framework.

The assumptions required to estimate the MNIR propensity scores are quite strong. In addition to the usual assumptions for propensity scores, we also introduce assumptions about the suitability of the generative model. Under the model described here, ψ is estimating the population-average effect of the treatment on the word count vector and does not include, for example, the topic-specific types of generative models that we consider next.

Furthermore, matching on the propensity score may not result in matched texts that seem similar to human readers because propensity score matching only provides balance in distribution and does not necessarily recover (nearly) exactly matching pairs (King and Nielsen, 2015). As noted in the examples below, we find that matching on the MNIR projection sometimes matches very dissimilar documents. For example, in the censorship case a blog post about a protest may have a similar propensity to be censored as a blog post that contains pornography. While these posts and bloggers maybe similar in the probability that they were censored, they are otherwise completely dissimilar. If the model is correct, this will still result in the dataset being balanced in expectation, but in practice it dramatically complicates the ability to make manual, reading-based, assessments of balance.

4.2 Coarsening High-Dimensional Data: Topically Coarsened Exact Matching

In this section we provide a brief review of coarsened exact matching and then explain why a naive application of CEM to text data will inevitably fail. We then show how applying CEM to topics provides a more tractable form of coarsening.

4.2.1 Coarsened Exact Matching

Coarsened Exact Matching simplifies $\boldsymbol{X}$ by coarsening each variable into "substantively indistinguishable" bins and then performing exact matching. CEM creates strata for each variable in $\boldsymbol{X}$ so that exact matching can be performed at the level of these strata. Above we noted that if exact matches could not be found when matching on years of education, X_j education could be coarsened from years of education into bins: *no high-school degree*, *high-school degree*, *college degree* and *post-graduate degree*. Treated and control units in 9th and 10th grades, respectively, would be counted as "close enough."

CEM has desirable properties for matching because it is a monotonic imbalance bounding (MIB) method, meaning that by choosing strata, the researcher bounds the differences between treated and control to the extrema of the strata. If we use exact matching on the education variable described above, we know that we will never match a treated unit in high school to a control unit in middle school. Unlike propensity score matching which approximates a fully randomized experiment, CEM approximates the more efficient fully blocked randomized experimental design (Imai, King and Stuart, 2008; King and Nielsen, 2015).

An implicit assumption of CEM is that the set of conditioning variables is not too large relative to the total number of observations. To see why this is the case consider a case with a single variable Education which is coarsened to have 4 categories. This results in 4 strata that must be populated with treated and control units. Now add a second variable Age which also has 4 categories. Now we have $4 \times 4 = 16$ strata. Thus the growth is exponential in the number of variables. Consider now a text corpus with a very small number of unique words: 100. Applying the maximal amount of coarsening we coarsen each dimension to a binary variable indicating whether or not the document contains the word at all. This still results in 2^{100} strata which is a number so incredibly large that we cannot possibly expect to see many matches unless all combinations of words are almost perfectly correlated with each other. When applying CEM to individual word occurrences, any unique word in a text will eliminate all possible matches.

However, the logic of CEM still applies if variables (in this case, indicators for words)

can be grouped into similar "bins." This procedure is already familiar to students of statistical text analysis because word stemming is a type of coarsening across words. When analysts encounter a corpus containing, say, the words "censor", "censoring", and "censored," they often conclude that these words are similar enough to group into one variable "censor" by a stemming algorithm (reducing the dimension of $\boldsymbol{X}$ by two). Applying CEM to stemmed text data maintains the MIB property because any matched documents must have equal counts of the stem "censor" (though they no longer have necessarily equal counts of the unstemmed words "censor", "censoring", and "censored"), so semantic distance between texts remains bounded.

Stemming is not enough, however. Even if we stemmed all words within our text, k would still not be small enough to be tractable. Besides, stemming algorithms are not well-developed for languages like Chinese (where verbs are not conjugated) or Arabic (where words are modified by infixing), so stemming is not currently a viable general-purpose solution for dimension reduction in text matching applications.

4.2.2 Topically Coarsened Exact Matching

An alternative strategy, which we call topically coarsened exact matching, is to estimate a topic model and then apply traditional matching methods to the resulting topics. This maintains the monotonic imbalance bounding property in the topical space but it also can be seen as an analog to coarsened exact matching on words where coarsening is allowed to happen across bundles of related terms.

Recall that in the generative process for a topic model each individual word in the document has a topic assignment. Under this model, two words with the same topic assignment are stochastically equivalent. Thus applying coarsened exact matching to the topic proportions on each document assures that each matched pair of documents has approximately the same proportion of words assigned to (for example) the "censorship" topic, but the model is indifferent to which censorship words are used.[3]

[3]We note that this is related to but distinct from a type of data-driven stemming. In stemming all variants of a word ("censor", "censored", "censoring") map to the same stem. In a topic model a word is mapped to a topic based on the other words in the document. For example, the word "block" might be alternatively mapped to a topic about web censorship, childrens' toys, or karate moves depending on the other words in the document.

We interpret this approach in two ways. If the topics themselves are semantically meaningful then the imbalance bounding properties are a meaningful and desirable aspect of the method. It is essentially a statement that the important thing that makes two texts good counterfactuals for each other is not the exact words they use but the subject matter they discuss. Alternatively we can see the topic model as an estimate of the joint density of the confounders. On this view, matching on the density estimate is a way of reducing variance at the risk of introducing a small amount of bias (compared to CEM on the full set of confounders).[4] While we are not aware of any work which matches on a density estimate, this perspective does suggest connections to prior work in statistical genetics[5] and optimal design for manual coding.[6] The density estimation view is premised on the idea that two observations with a common density estimate are stochastically equivalent and deviations between them are essentially random noise.

Intuitively both of these interpretations are connected to the idea that if two words commonly co-occur in the corpus as a whole they are essentially interchangeable for the purposes of identifying counterfactuals.[7] The experimental analog of this might be thought of as a kind of partial or hierarchical blocking in which balance is enforced across the collections of words but not within each collection.

[4]CEM bounds the sample imbalance which is directly related to bias in the causal effect estimate (Imai, King and Stuart, 2008; Iacus, King and Porro, 2011). The use of the density estimate allows for the possibility that balance is achieved in the density estimate but imbalance remains in the space of the original confounders.

[5]In an influential and highly cited article Price et al. (2006) suggest a procedure which amounts to an eigen decomposition of genotype data followed by a regression based adjustment using that decomposition. Although they are using regression they explicitly invoke the idea that this creates "a virtual set of matched cases and controls" (Price et al., 2006, pg. 904). The connection our proposed procedure is clear by consider the basic Latent Dirichlet Allocation topic model as a form of model-based discrete principal components analysis (Buntine and Jakulin, 2004)

[6]Taddy (2013*a*) addresses the problem of how to select documents for manual coding in supervised learning. The crux of the problem is that you want to choose documents to code somewhat optimally which suggests a space filling design on the text. Unfortunately this is impractical because in high dimensions every document is very far away from every other document. His solution is to estimate a topic model and then use a D-optimal space filling design in the lower-dimensional topic space. He dubs the strategy "factor-optimal design." Although neither the applications nor the set up are framed in these terms, the findings here fairly clearly suggest a framework for performing approximate blocking in treatment assignment of an experimental design. In this sense the work suggests an experimental analog of the strategy pursued here for observational data.

[7]This is a bit imprecise as the topic assignment allows for words to be context sensitive. Thus, the word "bat" can be in a topic with small mammals or sports depending on the other words in the document. These two uses of "bat" are not interchangeable from the point of the view of the model.

As with all parametric topic models it is necessary to choose the number of topics. Thankfully in this setting the choice is less fraught than in cases where semantic interpretation is the primary concern (Grimmer and Stewart, 2013). In general, more topics will result in closer matches. Redundant topics will not cause bias but will drive down the efficiency of the estimator (as it will begin to approach the efficiency of simply applying CEM to the raw word counts). The primary goal is to set the number of topics high enough that matched documents are good counterfactuals, as determined by the demands of the research design. However the risks of choosing too few topics are much greater than choosing too many.

4.2.3 Properties

Topically coarsened exact matching has a set of desirable properties that make it useful for the text case. Similar to CEM, topic matching bounds the differences between words in two matched documents by ensuring that groups of words are treated similarly. This ensures that two documents that have completely different topical content could not be matched. Topic matching in this respect does not have the same pitfalls at MNIR discussed in the previous section where two completely different documents could be matched if they had the same probability of treatment.

However, topic matching does not include treatment in selecting variables to use for matching and therefore can fail to pick up on sets of words that are important for predicting treatment. Consider the censorship example. Two blog posts were matched both talking about a particular city, based on a topic with words about that city. However, one post may be about an ongoing protest in the city, something that would be typically censored, and one post may be about a new construction project within the city. If this were the case, the topics might be too coarse to distinguish important characteristics related to treatment and the assumption $T_i \perp\!\!\!\perp Y_i(1), Y_i(1)|\boldsymbol{X}$ would not hold.

To understand why this happens, it is important to realize that the topics need to explain all the words in the corpus. Typically this means that topics will generally capture the subject matter of the document rather than the sentiment, even though sentiment may be a strong predictor of treatment assignment. Thus ideally we want a method that can

combine the imbalance bounding properties on the topics with the directed assessment of words which affect the probability of treatment.[8]

5 Topical Inverse Regression Matching

In this section we combine elements of MNIR propensity score matching and topically coarsened exact matching to develop a method called topical inverse regression matching (TIRM). TIRM allows us to estimate both the topics to match on and also within-topic propensities for treatment. This type of matching forces matched documents to be topically similar, while also increasing weights on words that are related to treatment within a topic, thus incorporating the within-topic perspective of the treated group. This disallows two types of matches. First, it ensures that a document related to, say, pornography will not be matched with a document related to protest because they are topically dissimilar (even though they have a similar propensity to be censored). At the same time, this model also prevents a document about construction in a city from being matched to a document about the protest within the city (because they have very different propensities to be censored). This allows us to prioritize variables which are related to treatment assignment while approximating a blocked design on the full set of confounders (similar in spirit to the combination of PSM and nearest neighbor matching on "prognostic" covariates proposed in Rubin and Thomas (2000)).

The TIRM method also estimates topic-specific probabilities of treatment which makes it more appropriate to many text applications than simply matching on the overall probability of treatment. In the censorship case, the word "delete" may only be related to censorship if the topic has to do with censorship itself[9] and not if the topic has to do with coding software. TIRM allows words to be related to treatment in one topical context and not in another. This provides an additional value-added to matching simply on MNIR-estimated probabilities of treatment, which apply across all topics. TIRM-estimated probabilities of treatment also do as well as or better than MNIR-estimated

[8]We note that this problem arises due to the necessity of combining multiple dimensions with topically coarsened exact matching. Under the standard CEM method for low-dimensional data, matches are enforced across all dimensions of the confounders.

[9]Criticism of the censors is often censored, as described in King, Pan and Roberts (2013).

values of treatment out-of-sample, see a further discussion of this in the Appendix.

We estimate the components of TIRM using the Structural Topic Model (STM). In this section, we first review the set up of the STM model. We focus in particular on the content covariate, which allows us to estimate the topic-specific probability of treatment. We then explain how to match with the estimators from the STM model before moving to the examples.

5.1 Review of STM model

The Structural Topic Model is an extension of the popular Latent Dirichlet Allocation model (Blei, Ng and Jordan, 2003; Blei, 2012) which is designed for use with covariates (Roberts et al., 2014; Roberts, Stewart and Airoldi, 2015). Covariates in the STM can be included to affect topic prevalence (the frequency with which a topic is discussed) and topical content (the words used to discuss a topic). In particular, we show that using STM with the treatment indicator as a topical content covariate effectively combines the MNIR and topic modeling framework.

In the simplest version without covariates the data generating process of the Structural Topic Model can be given as:

$$\vec{\gamma}_k \sim \text{Normal}_P(0, \sigma_k^2 I_P), \qquad \text{for } k = 1 \dots K-1, \tag{5}$$

$$\vec{\theta}_d \sim \text{LogisticNormal}_{K-1}(\mathbf{\Gamma}'\vec{x}_d', \mathbf{\Sigma}), \tag{6}$$

$$\vec{z}_{d,n} \sim \text{Multinomial}_K(\vec{\theta}_d), \qquad \text{for } n = 1 \dots N_d, \tag{7}$$

$$\vec{w}_{d,n} \sim \text{Multinomial}_V(\mathbf{B}\,\vec{z}_{d,n}), \qquad \text{for } n = 1 \dots N_d, \tag{8}$$

$$\beta_{d,k,v} = \frac{\exp(m_v + \kappa_{k,v}^{(\text{t})} + \kappa_{y_d,v}^{(\text{c})} + \kappa_{y_d,k,v}^{(\text{i})})}{\sum_v \exp(m_v + \kappa_{k,v}^{(\text{t})} + \kappa_{y_d,v}^{(\text{c})} + \kappa_{y_d,k,v}^{(\text{i})})}, \qquad \text{for } v = 1 \dots V \text{ and } k = 1 \dots K, \tag{9}$$

Note in particular that the form of the model in Equation 9 mirrors the MNIR model in Equation 2 but with the addition of topic-specific effects and (optionally) topic-covariate interactions. Thus we can equivalently see this form of the STM as embedding a multinomial inverse regression into a topic model or embedding topic-specific random effects

inside the MNIR model.[10]

5.2 Matching Quantities

Using the same framework as Taddy (2013*b*) we can derive a sufficient reduction of the information contained in the word counts about treatment. In this case though the projection represents the information about the treatment not carried in the topics. When we omit topic-covariate interactions, the sufficient reduction takes the simple form $(\kappa^{(c)})'(x_i/m_i)$. This projection was explored in work a related work by Rabinovich and Blei (2014) who are focused primarily on using the sufficient reduction as a way to improve prediction.

When we include the interaction of topics and the content covariate the projection becomes more complex due to the coupling of the estimated topics and the treatment. The complication arises because we need to reweight the interaction term by the topic use in the given document. Thus the form is: $(\kappa^{(c)})'(x_i/m_i) + \frac{1}{m_i}\sum_v x_{i,v}\left(\left(\kappa_v^{(\text{int})}\right)'\theta_i\right)$.

Once we have computed the analog of the MNIR sufficient reduction we can match on both that reduction and the estimated topics, ensuring matches are both topically similar and have similar within-topic probabilities of treatment. In order to ensure that topics comparable irrespective of the treatment/control effects we make an additional pass through all control documents re-estimating their topic proportions as though they were observed as treated. We then match on these new topic proportion vectors and the projection using CEM to provide the imbalance bounding guarantees. In Section 6.3 we give an example of where it is appealing to match in a different way in order to maintain the estimand of interest.

5.3 Limitations of TIRM

Matching on the STM topics and projections inherits the attractive properties of the MNIR propensity score and topically coarsened exact matching procedures. Indeed a significant benefit is that by jointly estimating both the propensity score and the topics,

[10]We note a minor difference arises in the prior distribution used for κ. Taddy (2013*b*) uses the Gamma-Laplace scale mixture prior whereas we use the more basic Laplace prior. This doesn't appear to matter an enormous amount in practice.

we ensure that our matching procedure aligns cases which have similar probabilities of treatment but also are broadly similar in the collections of words they use.

However, a key limitation of course is that the matches do rely heavily on parametric models of a complex data generating process.[11] The topic model in particular can be interpreted as a density estimate of the data and the resulting matches will only be useful if the underlying topic model provides an accurate representation of the texts. Thankfully the topics themselves provide some indication of how useful they will be for matching. An analyst is able to substantively interpret the topics and determine if documents which shared a similar topic profile would serve as reasonable counterfactuals for the analysis in question.

An important area for future work and a limitation of the current state of the method is the lack of theoretical development. Our interest in matching on texts and other high-dimensional data is born out of a practical necessity. As we argued above and show through examples in subsequent sections, there are a variety of research problems where documents themselves are clearly the relevant pre-treatment confounders. We believe a particular fruitful avenue for future work is the theoretical properties of using density estimates in matching including the implications that such strategies have for how a unit's density estimate is connected to other units within the sample.[12] Nevertheless, we do not believe the lack of such development is an impediment to practical work in the short term.

TIRM also inherits some limitations that are common to its predecessors. In particular, we need to carefully consider the risks of interference between units (Bowers, Fredrickson and Panagopoulos, 2013; Aronow and Samii, 2013). Essentially all matching methods require that Stable Unit Treatment Value Assumption (SUTVA) holds (Rubin, 1980); however, when a text in a corpus was written to influence the writers of other

[11] We do however note that nothing about our framework requires this particular data generating process. The mixed-membership multinomial model we propose here could easily be replaced by a different density estimator. For example, Taddy (2015*b*) recently proposed an inverse regression technique based on distributed representations from deep learning. We prefer the topic model representation as it allows for context sensitivity for words, but as other techniques evolve there will almost certainly be better options available. Our framework is sufficiently general to handle these changes.

[12] We note that this problem isn't exactly new to our method. Well-studied approaches such as propensity score matching and Mahalanobis distance matching have a dependence on the sample through the estimation of the propensity score model and sample covariance respectively.

texts a corpus (as in academic articles, for example), this can be a difficult assumption to justify. While this concern is significantly more general than the method we propose here, we highlight it in order to emphasize that these issues should be considered on a case by case basis. We also caution that as with other matching methods, when units are dropped from the sample we must be careful to correctly characterize the group to which the estimated effect applies (King, Lucas and Nielsen, 2015; Rubin, 2006, 221-230).

5.4 Balance Checking

Balance checking is an important part of any matching procedure, but it can be difficult to clearly specify an appropriate balance metric for text matching applications.[13] Even our numeric representation of the text as a count matrix is a substantial simplification of the true text. For this reason, we have emphasized a series of methods where it is easy to show which observations are matched to each other so those matches can be evaluated in both a quantitative and qualitative way.

Without a single, unifying balance metric, we find it useful to check balance across a variety of quantitative metrics in addition to more thorough, but labor-intensive, qualitative reading. In the first and simplest balance check, we identify words that in the full sample were very associated with treatment. We then compare average word appearance of these words in the matched sample, verifying that the treatment and control uses of the word are similar in the matched set. Second, we compare the distribution of topics in the matched sample between treated and control. If we used a variant of topically coarsened exact matching, these will necessarily be close. However, it can be a useful mechanism of investigating differences in the matched sample even when using a different approach.

Our third balance test is the most demanding quantitative metric. We compare similarity of matched texts using string kernels, which measures similarities in sequences of characters (Lodhi et al., 2002; Spirling, 2012). String kernels allow us to include a metric of similarity that includes word order information. Recall that we have discarded word order information in all of the approaches we have previously presented in the paper, thus

[13] One could argue that it is a natural extension of the "No Free Lunch" theorem that there cannot be a single balance metric appropriate for all situations (Wolpert and Macready, 1997).

we find it useful to have a measure that incorporates that information for assessing our matched sample. Unfortunately string kernels are extremely computationally intensive and so they are only practical for relatively small samples.

As a final and most important test of balance, we conduct qualitative readings of matched documents. Here the assessment of what constitutes good balance is particularly subjective, but it does play an important role in communicating to the analyst the nature of the counterfactuals we are using to estimate our effect of interest.

6 Examples

In this section, we use text matching on each of the examples described in Section 2. In the first example, studying the effects of censorship on bloggers' writings, we show how TIRM matching and topic matching can recover close to identical matched texts while MNIR performs poorly, matching texts that are not as similar. In the second example, we estimate the effect of author gender on citations to academic articles in the field of International Relations. We show the benefits of matching using TIRM in comparison to using topic-only matching and we compare TIRM matching to matching on human codes. In the last example, we estimate the effect of bin Laden's death on the popularity of his writings. This example shows the benefits of matching on topics in conjunction with additional covariates.

6.1 How does censorship in China affect bloggers?

In this example, we study the effect of censorship on bloggers' writing in China. While many have maintained that a self-censorship is one of the primary mechanisms through which censorship functions on the Chinese Internet (Wacker, 2003), scholars have yet to empirically estimate self-censorship behavior among typical bloggers in China. In a perfect experimental setting, we could estimate how censorship affects bloggers by randomly assigning censorship to bloggers and measuring the bloggers' reaction to censorship: did they discontinue writing? Write on less politically sensitive topics? Or continue writing on similar topics?

Of course, censorship cannot be randomly assigned to bloggers, both for technical

and ethical reasons. However, we can identify counterfactuals to censored bloggers by identifying censors' "mistakes". While China's censorship program is very thorough and employs thousands of people (King, Pan and Roberts, 2013, 2014), there are instances where two almost identical blogposts have different censorship statuses; the censors find and censor one blogpost, but miss the other. In a companion paper, Roberts (2015) collects a panel dataset of 6 months of posting from 593 bloggers. Roberts (2015) collects the censorship statuses on each post by collecting the post before it could be censored and returning to it later to see if it had been removed. This dataset, containing thousands of unique posts provides an opportunity to search for censor mistakes and record how censored bloggers' subsequent writings were different than those who had written a nearly identical post, but were not censored.

The text matching challenge in this example is to find posts with different censorship statuses that are for all practical purposes identical. In other words, we hope to match the text so closely that at most a few words are different between matching posts. For the most part, this means we want to match very closely on a large number of topics; posts that have nearly identical topical estimates will have nearly identical text. Topical matching will also allow for small differences in posts, which is important because even reposts are not always identical, as they might have different dates or authors. However, if there are small differences in matched posts, we want to make sure that these small differences are not those that are related to censorship in the larger dataset. TIRM therefore becomes extremely important as the projection will reflect whether the non-overlapping sentences between the two posts have words that are differentially related to censorship.

In this example, we compare the unmatched sample to 1) matching only on the projection from a Multinomial Inverse Regression (propensity score matching) to 2) matching on 200 topics only without a projection (topically coarsened exact matching) to 3) TIRM, matching on 200 topics in addition to the topic-specific projection. In each case, we also match on other covariates that we think would be important to censorship, including the date of the post, the previous censorship rate of each of the bloggers, the previous post rate of the bloggers, and the average estimated sensitivity of the bloggers' previous

posts. We find that even exact matching on the MNIR projection does poorly at finding nearly identical posts, as it matches posts with similar propensities of being censored, but with very little in common otherwise. Topical matching does much better at finding close to identical matches, but will include matches where non-overlapping sentences may be related to censorship. TIRM finds nearly identical matches, outperforming the other two.[14]

We show this result with two balance metrics. First, in Figure 1 we compare the average difference in the 10 topics estimated to be most and least associated with censorship in the unmatched dataset. We find that while TIRM and topic matching do quite well on this metric, MNIR matching reduces topical differences to a lesser degree than the other matching methods.

Second, we use string kernel similarities between matched blog posts to compare each matched dataset and the unmatched dataset. String kernels examine consecutive sequences of text, in this case groups of five consecutive words, and estimate the percentage of these kernels that are shared between two documents. Figure 2 shows the density of string kernel similarities between matched blog posts in the MNIR matched dataset and topic matched dataset. We see that topic matching produces much more similar blog posts than MNIR matching, which only in a few case produces significantly more similar posts than the unmatched dataset.

Given clear evidence that TIRM found the most similar blog posts than matching on probabilities, Roberts (2015) uses this method to identify censors' mistakes, producing identical or nearly identical matches. Using nearly identical topical matching and matching on date and blogger history, Roberts (2015) finds that censorship causes bloggers to react negatively against the state; experience with censorship seems to backfire against the states intent to stop the spread of information on a particular topic. Experience with censorship induces bloggers to continue to write about the sensitive topic, write about

[14]There is nothing inherently wrong with matching posts which are dissimilar other than propensity to be censored and would be expected in the fully randomized experiment that PSM approximates. However, such radical differences as matching pornography with protests makes it difficult to impossible to make human-oriented assessments of the balance between treated and control texts. It is also a less efficient design than one where matched units are similar on all pre-treatment dimensions.

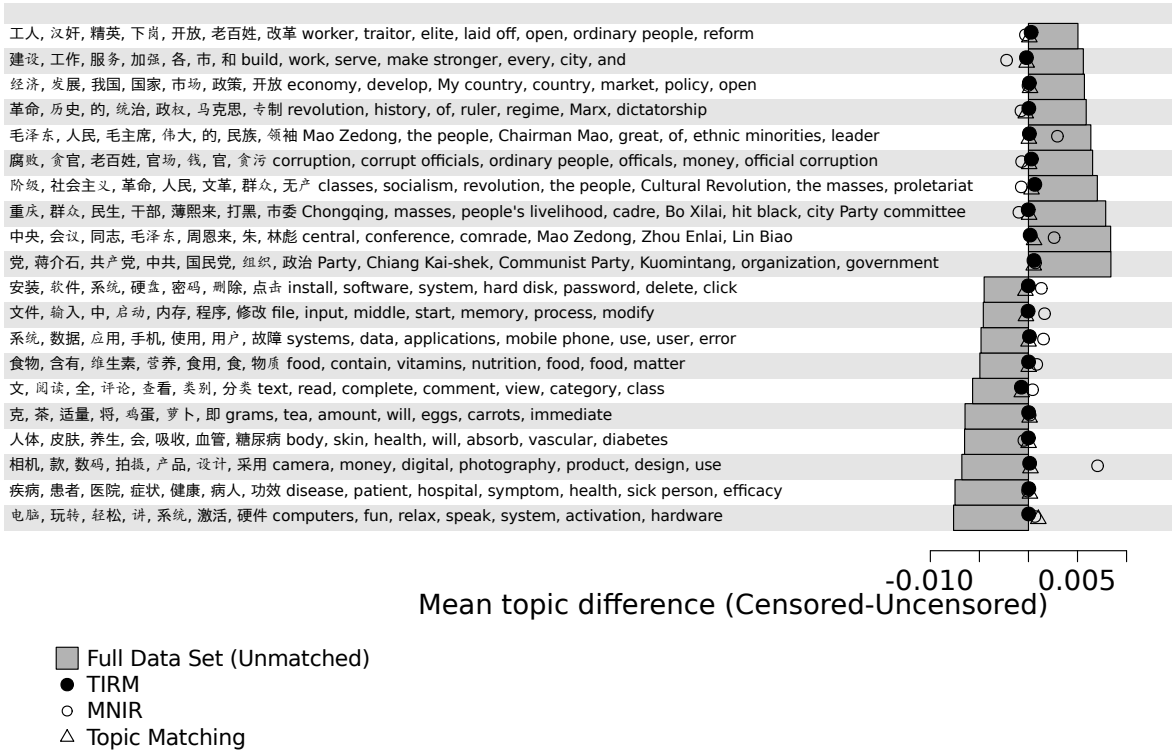


Figure 1: Topic balance comparison between unmatched, topic matched, MNIR matched, and TIRM datasets.

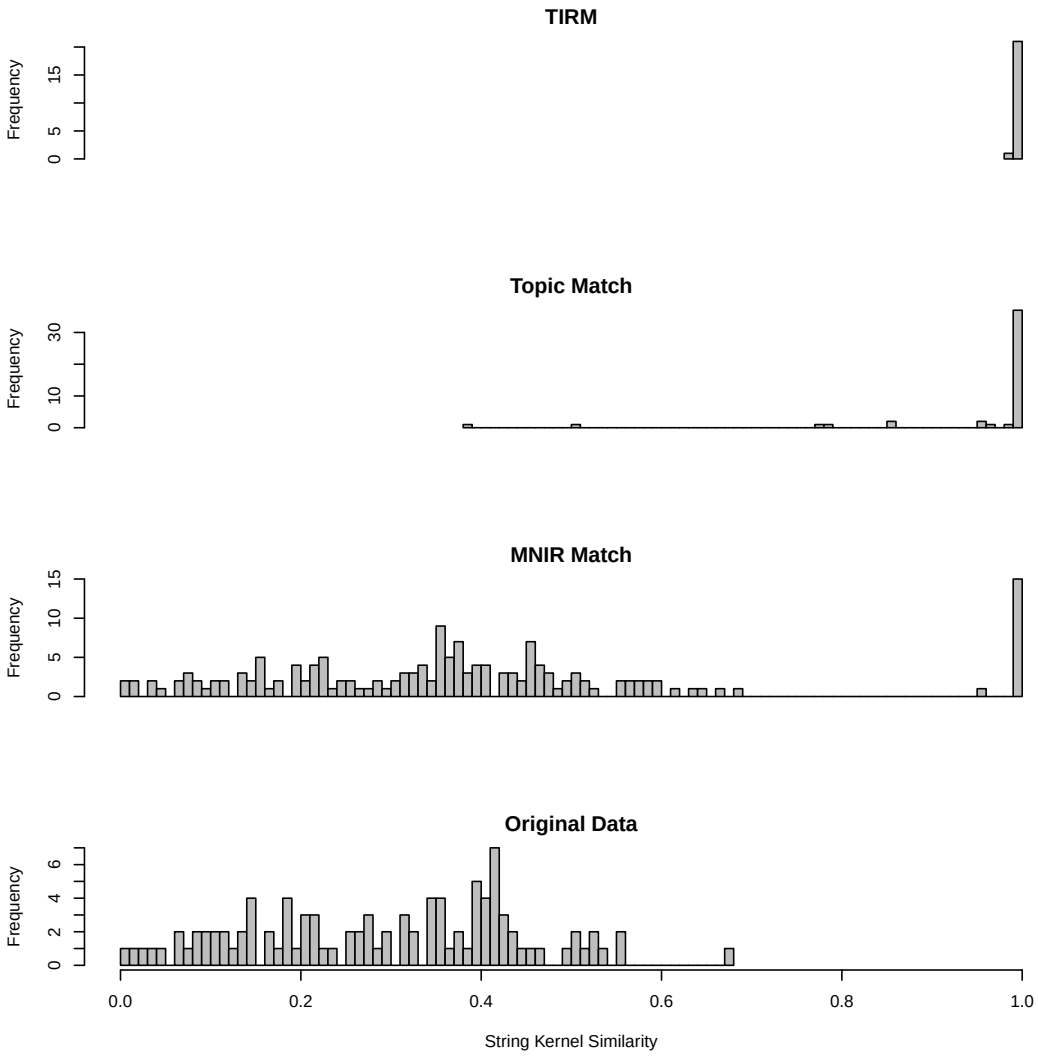


Figure 2: String kernel similarities,from top to bottom: TIRM, topic matched, MNIR matched, and unmatched.

more sensitive topics, and these bloggers are also more likely to be subsequently censored.

6.2 How does author gender affect article citations?

In this example, we study the effect of gender on citation counts of articles in the field of International Relations in political science. In a widely-discussed study, Maliniak, Powers and Walter (2013) estimate the effect of gender on citation counts and find that papers with female authors have lower citation counts even after using using a parametric model to control for potentially confounding variables such as article age, venue, status of the author, issue area, and methodology. They further analyze citation networks between authors and conclude that self-citations and the tendency for men to cite other men may account for the differences in citations between women and men.

Maliniak, Powers and Walter (2013) note that textual features of articles could explain why men and women may be cited differently: they write on different topics within international relations, use different methods, and come from different epistomological schools. Because subfields in political science are sometimes highly gender imbalanced, eliminating treated or control units without a close match may also help reduce extreme counterfactuals and model dependence (King and Zeng, 2006; Ho et al., 2007). To address these concerns Maliniak, Powers and Walter (2013) use extensive human coding provided by the TRIP Journal Article Database (Peterson and Tierney, 2009) to measure article sub-fields, methodologies, paradigms, and epistemologies. This allows us to compare TIRM to matches based on human-coded categories, providing a strict test of our method.

We obtain the text of each article in the Maliniak, Powers and Walter (2013) data set.[15] Following Maliniak et al, we consider articles with no male authors to be treated, and articles with at least one male author to be controls. In total, we have 3,201 articles, 333 of which are authored by women.[16] We conduct TIRM matching on the data. First, we estimate a 15 topic Structural Topic Model on the data, including the treatment as

[15]The data was generously provided by JSTOR's Data for Research Program.

[16]Our dataset is larger than that used by Maliniak, Powers and Walter (2013) because originally TRIP coding was only applied to articles from issue numbers 1 and 3. Since Maliniak, Powers and Walter (2013), TRIP coding has been applied to all issue numbers, which we include in our analysis. We also include data from all years. We did however lose some articles for which we were not able to obtain the raw texts as they were not in JSTOR's dataset.

a content covariate. We obtain estimates of how much each article talks about a given topic. We also obtain topic-specific projections that indicate which words within-topics are more likely to be associated with all female authors than all male or co-ed authors. For comparison, we estimate the Multinomial Inverse Regression to obtain estimates of non-topic specific probabilities that the article is associated with all female authorship. We obtain three types of automated matched data sets for comparison: MNIR match, topic-only match, and TIRM. We also conduct exact matching on the human-coded categories to create a human analog to text matching.

There are multiple reasons why this example in particular is conducive to TIRM matching. We are not looking for exact matches; in fact, we hope there are not identical articles by different authors, as this would indicate plagiarism or reprinting. Instead, we want to compare articles on similar issues, and we would like to control for clusters of words that women tend to use more frequently within these topics.

For example, Figure 3 shows words within two topics that are more and less likely to occur within all-female articles. The topic on the right has to do with political theory. We see that within political theory, women tend to focus on gender issues, using words like "women", "gender", and "children". The topic on the left has to do with political economy. We see that within this topic men use words associated with quantitative methodology more than women, like "model", "estimate", and "data". Ideally, we would find matching articles that were similar both in topic (political economy for example) *and* were not different in methods or focus within that topic in ways that might be systematically related to gender.

To compare the various matching methods in this case, first we calculate the mutual information as it pertains to the all female variable for each of the words within the original corpus vocabulary. Mutual information measures the extent to which a word contains information about whether the document was written by all female authors. In Figure 4 we show how the difference in average appearance of the word between treated and control documents is related to mutual information. Words with high mutual information are either very related to female documents such as "interview", "feminist",

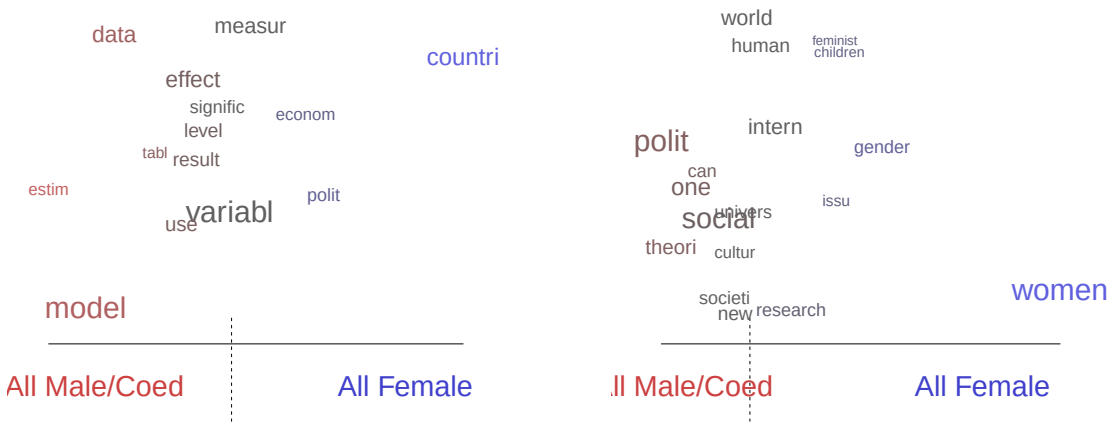
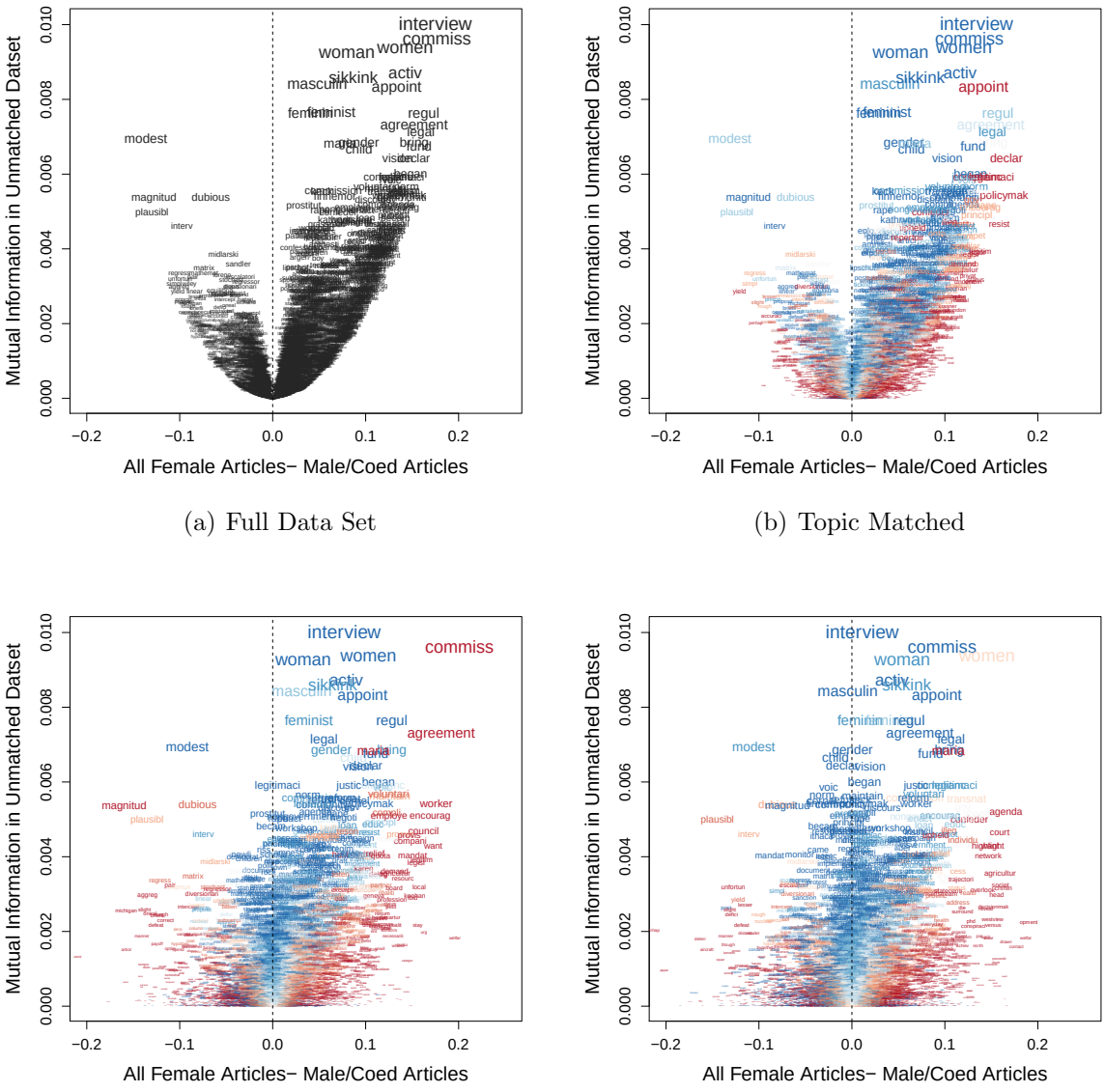


Figure 3: Topic-specific words that are associated with gender. Words that are farther to the right are more likely to occur within an all female article.

"woman", "women", or "legal", or are very related to male documents such as "modest", "magnitude", and "plausible".

For matching to be successful, we want to reduce discrepancies between men and women on the dimensions on which they systematically differ. Therefore, successful matching should reduce differences between words that have high mutual information between the treated and control corpus. In Figure 4, we plot the average difference in appearance of a word between treated and control for each of the matching methods by the mutual information in the unmatched corpus. Panel 1 show the difference in word occurrences in the unmatched corpus. In panel 2, we see that topic matching reduces the average difference between high mutual information words in the corpus slightly. Panel 3 shows the match based on human coding, which reduces the differences on words with high mutual information more than topic matching. Panel 4 shows TIRM, which explicitly incorporates information about words that are highly associated with all female authors. TIRM reduces the difference in high mutual information words between the all female and male/coed corpora the most out of all of the matched datasets.

Second, we compare various matching methods on the estimated differences in topics



(a) Full Data Set

(b) Topic Matched

(c) Matched on human codes

(d) TIRM

Figure 4: Relationship between mutual information and difference in word appearance (all female - male/coed) a) Full Data Set b) Topic Matched c) Matched on human codes d) TIRM. In panels b, c, and d, words for which matching decreased the absolute value of the difference in word appearance are in blue and words for which matching increased the absolute value of the difference in word appearance are in red.

between the treated and control groups. In Figure 5, we show that matches based on MNIR may not have similar topics. In this case, the matches based on MNIR are more dissimilar on topic than the original unmatched sample. This is because matching on a

uni-dimensional projection will often match documents that are estimated to have similar propensities to have all female authorship, but are not similar in any other way. TIRM and topic matching do much better than the original data on this metric, while matching based on human coding is also better than the original data for the most part, but exacerbates imbalance on some topics. This is likely because the topics estimated by the model are not always similar to those identified *a priori* by the human coders.

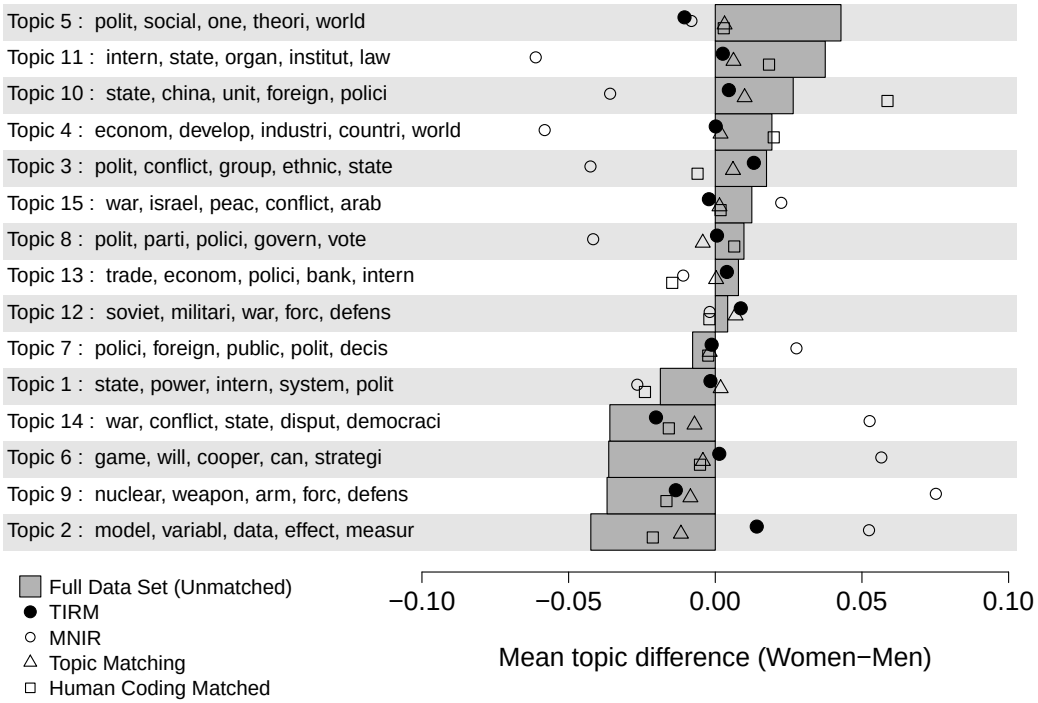


Figure 5: Matching Comparison for Topics

Next, we compare matching methods across the human coded categories – showing the average difference between treated and control within the human coded categories for each matched dataset. Again, we see that MNIR matching does quite poorly, in many cases worse than the original data when the human coded categories are used as a metric of comparison. Topic matching and TIRM decrease imbalance between human coded categories more consistently than MNIR. TIRM matching in particular performs best at reducing differences between categories that are most related to gender, in this case the Qualitative and Quantitative methods category.

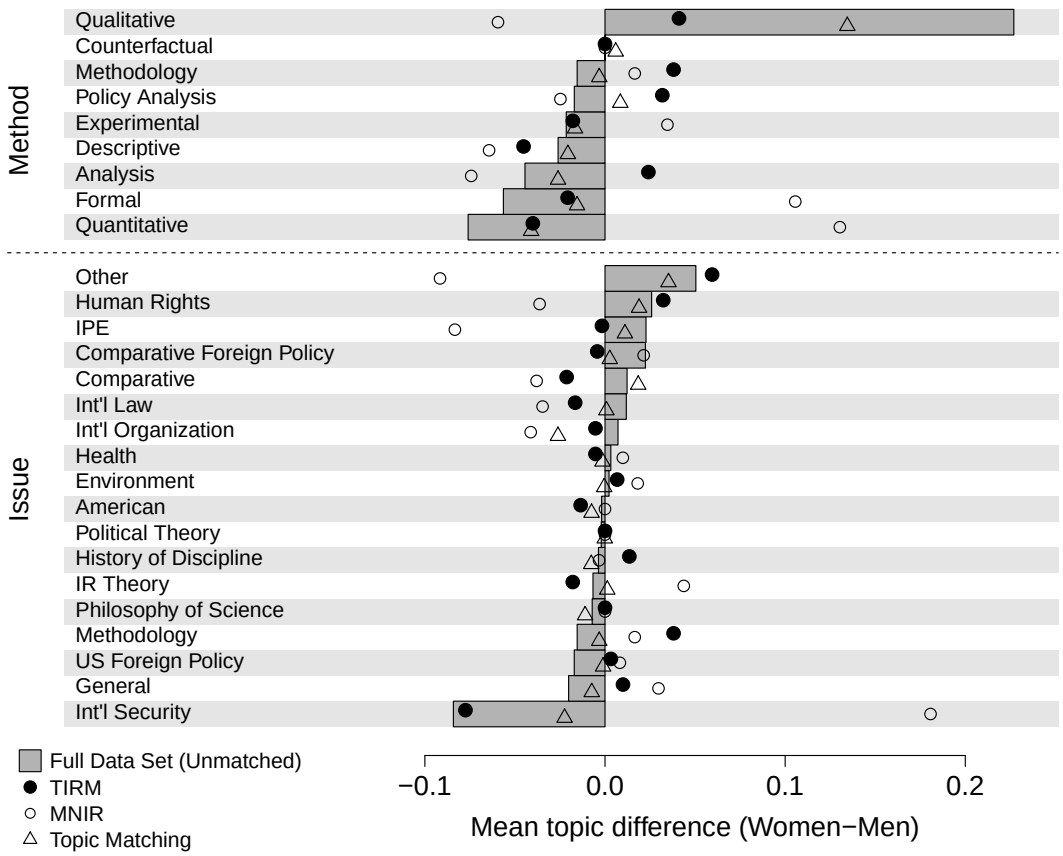


Figure 6: Automated Matching Comparison and Human Categories

Last, we compare the matching based on human coding to TIRM using a string kernel similarity metric. Figure 7 shows the similarity between matched documents in the corpus matched using TIRM and corpus matched exactly on human codes. Overall, TIRM performs as well to the human-coding matching in producing semantically similar documents when measured with string kernel similarity.

Using TIRM as well as matching on covariates such as journal and article age, we re-estimate the results produced by Maliniak, Powers and Walter (2013). We find that after automated text matching, gender differences in citations are even more pronounced than those reported in Maliniak, Powers and Walter (2013), with an average 16 fewer citations of all female articles than of their co-authored or male-only authored counterparts. We report negative binomial models similar to those Maliniak, Powers and Walter (2013)

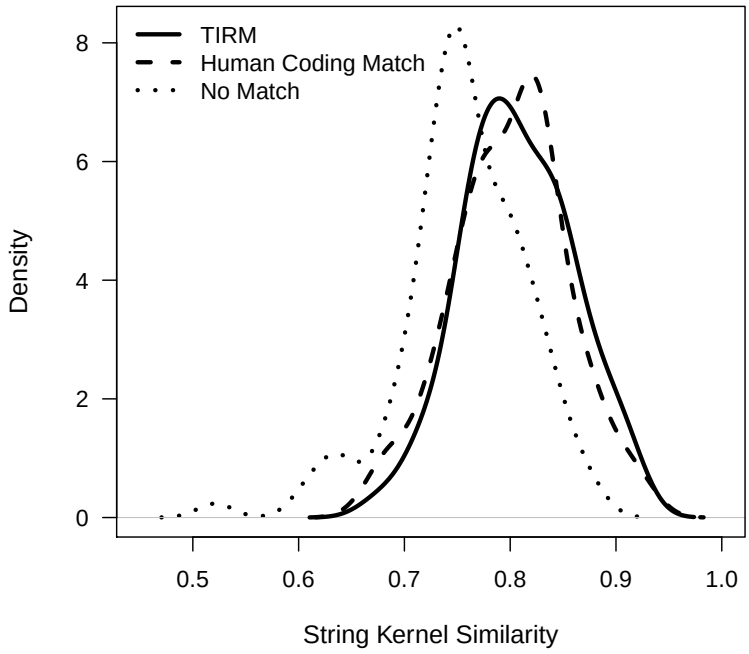


Figure 7: String Kernel Similarity Comparison

estimated in the Appendix.

6.3 The Effect of Bin Laden's Death on the Popularity of His Writings

In this example, we use data from a Jihadist website to explore the question of the targeted killing of Usama Bin Laden in May 2011 made his writings more popular with other jihadists. Sunni Muslim Jihadists have been targeted in a variety of ways during the "War on Terror," but there is little systematic evidence about the impact of this targeting. Scholars disagree about whether targeting the leaders of terrorist organizations helps or hurts counterterrorism efforts, (Jordan, 2009; Johnston, 2012) and the death of Bin Laden raised concerns that his killing would lionize him and make his ideas more popular (Cronin, 2006, 40).

We use TIRM to estimate whether Bin Laden's death re-popularized his writings. Nielsen (2015) has collected data from a large Jihadist online library. This web-library contains 6,115 documents (and counting) authored by 492 individuals. Virtually all of

these documents are written by Arab-speaking Jihadists to promote some aspect of jihadist ideology, and although we cannot track website visitors, evidence suggests that most are jihadist sympathizers.

A key feature of this website is that the number of page views accumulated by each document is updated in real time next to the title of the writing and the author's name. Nielsen (2015) collected this data daily, starting in February 2011 and ending in September 2014, resulting in extremely granular information about the popularity of jihadist writings.

In this example, we focus on the death of Usama Bin Laden because of the intrinsic importance of his writings and because his death occasioned substantial speculation that his popularity would experience a "martyrdom bump." By comparing the post-targeting popularity of Bin Laden's writings to the popularity of texts by non-targeted authors, we can precisely estimate whether the death of Bin Laden made his ideas more or less popular in the short- to medium-term. Usama Bin Laden is the author of 33 documents on the jihadist web library which had collectively been viewed 448,584 times on May 1, 2011, the day before his death. As of the end of the data collection on September 6, 2014, this cumulative count had increased to 673,182. Our task is to estimate how many of these 224,598 page views were the result of Bin Laden's death and how many would have accrued if he had remained alive.

If targeting were random, we might obtain unbiased estimates of the effect of Bin Laden's death by comparing the popularity of his writings to all other documents on the website. However, Bin Laden was not targeted randomly and the reasons for his targeting might be correlated with features of his writing that in turn drive page views of his online documents. We have two available sources of data to assist us in developing credible counterfactual statements about what might have happened to the popularity of treated documents if targeting had not occurred. First, we have the content of targeted documents, giving us the opportunity to control for aspects of each text that make it attractive to readers. Second, we have page view data for each document prior to targeting which is arguably the best predictor of subsequent page views. For the purpose of demonstrating ways of conditioning on text in the presence of other covariates, we try various ways of

conditioning on each of these sources of information.

To condition on the text of each of Bin Laden's documents, we estimate a structural topic model with an (arbitrarily chosen) 50 topics. After examining the topic model to ensure that the topics are generally meaningful, we identify control documents that have similar topic proportions to each of Bin Laden's 33 texts. Despite the benefits of CEM, it discards treatment units that do not have sufficiently similar matches which changes the estimand from the sample average treatment effect on the treated (SATT) to the *feasible* sample average treatment effect on the treated (FSATT) (King, Lucas and Nielsen, 2015). Here, our quantity of interest is the popularity of Bin Laden's writings *as a whole* so we prefer not to risk discarding any of them. We modify our TIRM procedure by replacing the CEM component with one-to-two nearest neighbor matching such that each of Bin Laden's 33 documents is paired with its two closest matches for a matched sample size of 99 texts (we use 1:2 matching rather than 1:1 matching for added statistical precision). We calculate nearest neighbors using a weighted combination of the STM results (topic proportions and the respective projections of the MNIR model) and the trend of pre-treatment page views. For illustration, we present results with matching only on topics, matching only on page view trends, and matching equally on both; in practice we find the latter most compelling.

To evaluate the quality of the matching, we use Figure 8 to compare the topics in Bin Laden's writings to the topics in the unmatched and matched control samples. In the unmatched sample, we see that Bin Laden's writings have much more text devoted to a few topics about fighting, the United States, and Afghanistan. Matching on topics from TIRM greatly reduces, though fails to completely eliminate, this imbalance. When we match on both topics and trends, the topic similarity between treated and control texts remains almost identical, with only slight movement on some topics. This suggests there is not too much of a trade-off between including the topic covariates and other covariates in this application. When we match only on the prior page view trend, topics are almost as imbalanced as in the raw data.

Manually validating the quality of matches in this corpus by comparing matched

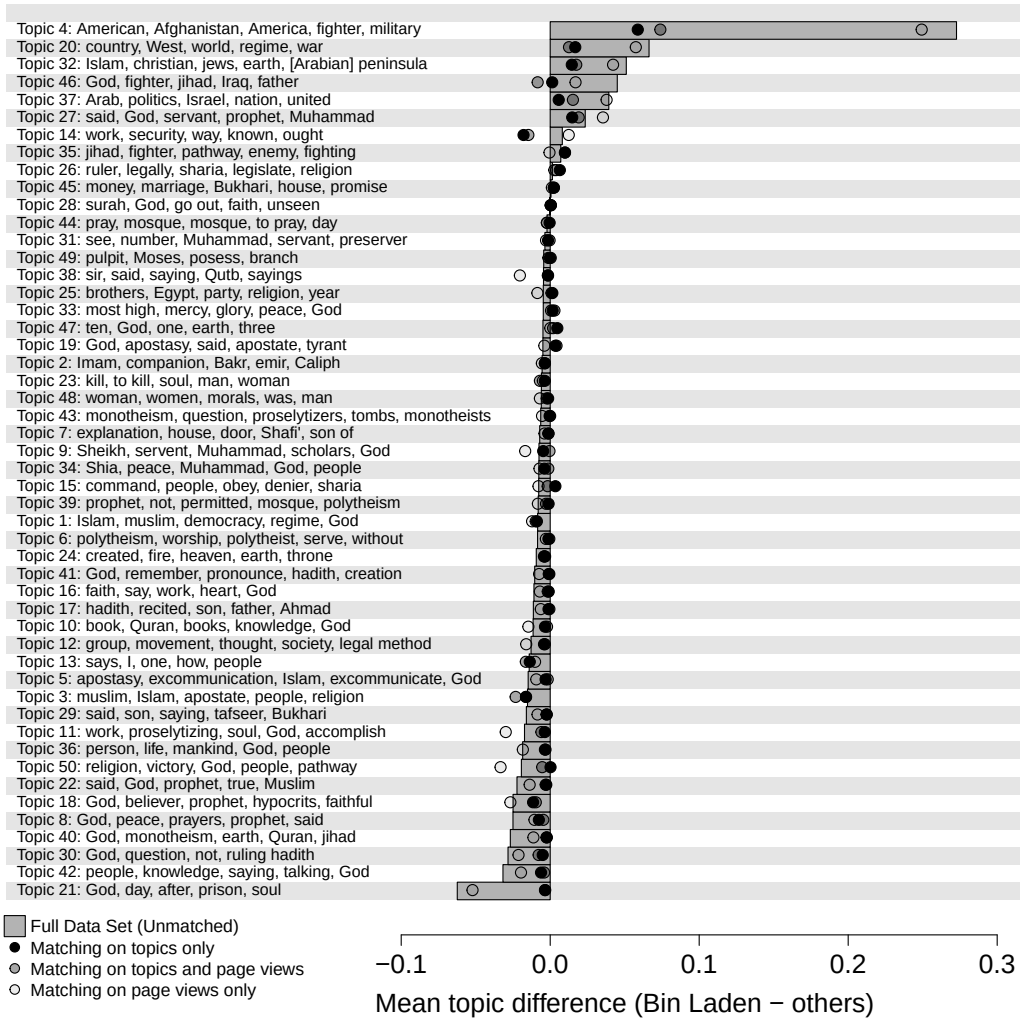


Figure 8: Matching Comparison for Topics in Jihadist Corpus

pairs of documents to randomly paired documents is infeasible because the documents are long (average wordcount is 5,403, maximum wordcount is 445,810) and the material is technical. However, the structure of the jihadist web library provides a unique way to validate our matching procedure; the website is organized in a series of subpages and documents on the same subpage are placed there because website administrators deem them similar in some way. We find that on average, documents matched by our method are 8 times more likely to co-occur on a subpage than randomly selected pairings; our method are correlates with website administrators' judgments of similarity.

To calculate treatment effects, we estimate a series of linear regressions on the matched data sets. We observe the post-treatment page views of every document on every day for the next three years. Rather than selecting a single day to serve as the evaluation point, we present the estimate treatment effect of Bin Laden's death on on all subsequent days for which we have data. We present these together so as not present a multiple testing problem, but note that isolating a single estimate after looking at all of them will result in an overstatement of statistical precision of the standard errors of that individual estimate are not corrected. In each regression, we include two pretreatment covariates: the number of page views of each document on May 1, 2011 (the day before Bin Laden's death) and the difference in page views between April 1, 2011 and May 1, 2011. Together, these control for the level and slope of each documents pretreatment trend which accounts for much of the variation in post-treatment page views and allows us to estimate more precise causal effects. This estimation setup is equivalent to a difference-in-difference design with matching to ensure that pretreatment trends are as similar as possible between treated and control units.

The results are shown in Figure 9. In each panel, the x-axis denotes calendar time and the y-axis displayes the estimated effect of Bin Laden's death on the popularity of his documents relative to matched control documents. These effects are expressed in page views, meaning that an estimate of 1000 means that the average Bin Laden document gained an additional 1000 page views by a given date as a result of treatment. The panels show alternative weightings of the topics and pre-treatment page view trends. In the left panel, only topic information is used for document matching. With this sampling strategy, we find that Bin Laden's death increased the popularity of his writings in the short-term, but that there is no long-term effect. The treatment effect climbs to just under 200 page views a week after Bin Laden's death but shortly after treatment, the effect is no longer statistically distinguishable from zero. The confidence intervals are wide because the matched documents have very different trends. This suggests that page view trends and topical content are weakly correlated and that more precise estimates of counterfactual page views for Bin Laden's documents can be obtained using matching on

past page views.

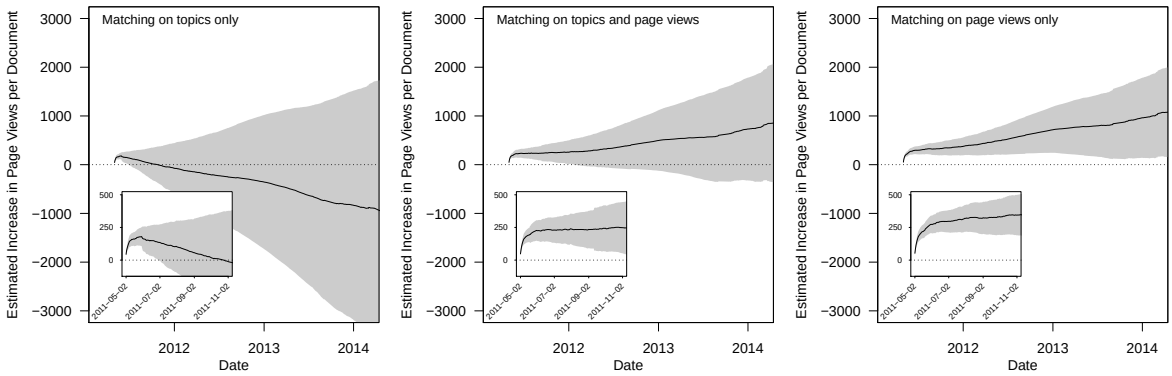


Figure 9: Estimated effects of Usama Bin Laden's death (on May 2, 2011) on subsequent page views of his documents on a large jihadist web-library.

The right panel presents the results when we match exclusively on pre-treatment page views. The estimated short-term effects are similar, but the estimated long-term effects are now positive and statistically significant at all dates after Bin Laden's death. By the end of data collection Semptember 6, 2014, the estimated treatment effect has increased to approximately 1000 additional page views for the average Bin Laden document.

The center panel shows the results from a matched sample where information about topics and pre-treatment page views are weighted equally. The results of these regressions are similar in magnitude to those in the right panel, but the estimates are less precisely estimated. We again find Bin Laden's death increases viewership of his documents by an average of 200 page views in the first week and by approximately 1000 page views three years on. However, the estimates become statistically indistinguishable from zero in mid-2012, approximately a year after Bin Laden's death.

We trust some of these results more than others. In particular, it is very clear that Bin Laden's death resulted in approximately 200 additional page views per document in the following week, or in cumulative terms, approximately 6,600 additional page views for Bin Laden's writing as a whole. This effect is very precisely estimated regardless of how heavily we weight the text and page view information in our matching procedure. Further out, the estimated effects are more varied which is understandable — constructing

credible counterfactuals becomes harder as the time since treatment grows larger. We find the estimates that combine information about page views and topical content to be most plausible, so we interpret our findings to mean that killing Bin Laden may have increased his popularity in the long-term but that this estimate is rather imprecise. If we take the estimated effect as of September 6, 2014 seriously, then our best guess is that the death of Bin Laden caused approximately 1000 additional page views for his average piece or writing. This means that of Bin Laden's 224,598 page views between May 2, 2011 and September 6, 2014, approximately 33,000 (15 percent) are attributable to his death.

Conclusion

In this paper, we set out to address the challenges of matching observations when confounders are measured by high-dimensional data. We developed analogs to two different existing matching methods: propensity score matching and coarsened exact matching, addressing the desirable properties and drawbacks of each. We created a new matching approach, Topical Inverse Regression Matching (TIRM) that address both drawbacks, while preserving these desirable properties. We applied MNIR propensity score matching, topically coarsened exact matching, and TIRM to three different datasets to compare their performance and answer social science questions.

Our research extends the current matching literature by allowing for matching on low-dimensional projections of high-dimensional data to improve inference in social science research when confounders are measured in text. In future work, we hope to extend the current focus on text to other types of high-dimensional data, including image data and biological data.

Appendix

7 Out-of Sample Prediction: TIRM and MNIR

In this section, we compare the out of sample performance of the TIRM projection and MNIR projection. We set aside 500 documents from the Maliniak, Powers and Walter (2013) dataset and train both the MNIR and TIRM models on the remainder of the documents. We then predict the gender of the authors in the article for the heldout documents and compare these predictions to the true author's gender. Figure 10 plots the Receiver-Operator curves for the TIRM and MNIR predictions. Both perform similarly well out of sample.

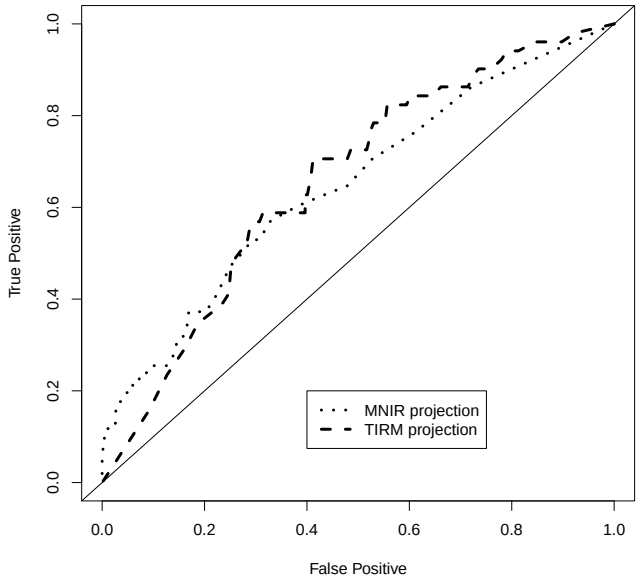


Figure 10: Receiver-operator curve for prediction on out of sample documents from MNIR and TIRM models

8 Gender Citations Results

Here we reproduce the models in Maliniak, Powers and Walter (2013) using the matched data set from automated text matching. Following Maliniak, Powers and Walter (2013), we use a negative binomial model to estimate the effects of gender on citations. For robustness checks, we also include the additional covariates provided by Maliniak, Powers and Walter (2013) in the model. In the last model, we include all covariates included in their "Kitchen Sink" model. Some of the covariates are not identified because no variation exists within the matched data sets and so we do not include them in the table below.

Table 1

	Dependent variable:			
	Citation Count			
	(1)	(2)	(3)	(4)
all_female	-0.694^{***}	-0.931^{***}	-0.968^{***}	-0.660^{***}
	(0.169)	(0.201)	(0.200)	(0.173)
article_age			0.107^{**}	-0.001
			(0.045)	(0.041)
article_age_sq			-0.003^{***}	0.0003
			(0.001)	(0.001)
tenured			-0.492^{***}	-0.548^{***}
			(0.186)	(0.168)
tenured_female		0.544^{*}	1.105^{***}	1.332^{***}
		(0.307)	(0.347)	(0.302)
gender_compAll Female				
gender_compCoed		-0.654^{**}	-0.410	0.023
		(0.327)	(0.327)	(0.288)
coauthored		0.043	0.007	0.037
		(0.187)	(0.186)	(0.169)
R1		-0.163	-0.026	
		(0.154)	(0.150)	
issue_american				
issue_cfp				-0.644
				(0.565)
issue_comparative				-2.752^{***}
				(0.858)
issue_env				1.241
				(0.805)
issue_general				0.318
				(0.961)
issue_health				1.996^{*}
				(1.096)
issue_hist_disc				1.190
				(1.254)
issue_hr				0.833
				(0.644)

Table 1 cont

	Dependent variable:			
	Citation Count			
	(1)	(2)	(3)	(4)
issue_ir				0.940^{*}
				(0.561)
issue_io				0.173
				(0.544)
issue_ipe				0.245
				(0.557)
issue_is				0.445
				(0.545)
issue_meth				0.871
				(0.623)
issue_other				-0.141
				(0.580)
issue_pos				
issue_political_theory				
issue_usfp				-0.366
				(0.609)
meth_qual				-0.151
				(0.254)
meth_quant				-0.050
				(0.255)
meth_exp				-1.074^{**}
				(0.528)
meth_formal				0.256
				(0.229)
meth_anal				0.109
				(0.330)
meth_policy				-0.700
				(0.464)
meth_desc				-0.392
				(0.376)
meth_count				
positivist				1.197^{***}
				(0.256)
material				0.792^{*}
				(0.442)
idea				0.565^{***}
				(0.150)

Table 1 cont

	Dependent variable:			
	Citation Count			
	(1)	(2)	(3)	(4)
AJPS				0.811**
				(0.347)
APSR				0.945**
				(0.421)
BJPS				0.733**
				(0.358)
EJIR				
IO				1.486***
				(0.214)
IS				0.654**
				(0.312)
ISQ				1.023***
				(0.241)
JCR				1.397***
				(0.247)
JOP				1.344**
				(0.522)
SS				
RIPE				
WP				1.522***
				(0.286)
Constant	3.475***	3.611***	2.895***	−0.245
	(0.084)	(0.144)	(0.423)	(0.905)
Observations	289	289	289	289
Note:				*p<0.1; **p<0.05; ***p<0.01

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American Statistical Association* 105(490).

Aronow, Peter M and Cyrus Samii. 2013. "Estimating average causal effects under interference between units." *arXiv preprint arXiv:1305.6156* .

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." *the Journal of machine Learning research* 3:993–1022.

Bowers, Jake, Mark M Fredrickson and Costas Panagopoulos. 2013. "Reasoning about interference between units: A general framework." *Political Analysis* 21(1):97–124.

Buntine, Wray and Aleks Jakulin. 2004. Applying discrete PCA in data analysis. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press pp. 59–66.

Cook, R Dennis. 2007. "Fisher lecture: Dimension reduction in regression." *Statistical Science* pp. 1–26.

Cook, R Dennis and Liqiang Ni. 2005. "Sufficient dimension reduction via inverse regression." *Journal of the American Statistical Association* 100(470).

Cronin, Audrey Kurth. 2006. "How al-Qaida Ends: The Decline and Demise of Terrorist Groups." *International Security* 31(1):7–48.

Diamond, Alexis and Jasjeet S Sekhon. 2013. "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies." *Review of Economics and Statistics* 95(3):932–945.

Glynn, Adam N and Kevin M Quinn. 2010. "An introduction to the augmented inverse propensity weighted estimator." *Political Analysis* 18(1):36–56.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* p. mps028.

Hainmueller, Jens. 2011. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* p. mpr025.

Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15(3):199–236.

Iacus, Stefano M, Gary King and Giuseppe Porro. 2011. "Multivariate matching methods that are monotonic imbalance bounding." *Journal of the American Statistical Association* 106(493):345–361.

Imai, Kosuke, Gary King and Elizabeth A Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the royal statistical society: series A (statistics in society)* 171(2):481–502.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Johnston, Patrick B. 2012. "Does Decapitation Work? Assessing the Effectiveness of Leadership Targeting in Counterinsurgency Campaigns." *International Security* 36(4):47–79.

Jordan, Jenna. 2009. "When heads roll: Assessing the effectiveness of leadership decapitation." *Security Studies* 18:719–755.

King, Gary. 2009. The Changing Evidence Base of Social Science Research. In *The Future of Political Science: 100 Perspectives*, ed. Gary King, Kay Schlozman and Norman Nie. New York: Routledge Press.

King, Gary, Christopher Lucas and Richard Nielsen. 2015. "The Balance-Sample Size Frontier in Matching Methods for Causal Inference.".

King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107:1–18. http://j.mp/LdVXqN.

King, Gary, Jennifer Pan and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722.

King, Gary and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131–159.

King, Gary and Richard Nielsen. 2015. "Why Propensity Scores Should Not Be Used for Matching.".

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann et al. 2009. "Life in the network: the coming age of computational social science." *Science (New York, NY)* 323(5915):721.

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini and Chris Watkins. 2002. "Text classification using string kernels." *The Journal of Machine Learning Research* 2:419–444.

Maliniak, Daniel, Ryan Powers and Barbara F Walter. 2013. "The gender citation gap in international relations." *International Organization* 67(04):889–922.

Morgan, Stephen L and Christopher Winship. 2014. *Counterfactuals and causal inference*. Cambridge University Press.

Neumark, David, Roy J. Bank and Kyle D. Van Nort. 1996. "Sex discrimination in restaurant hiring: an audit study." *Quarterly Journal of Economics* .

Nielsen, Richard A. 2015. "Can Ideas be "Killed?" Evidence from Counterterror Targeting of Jihadi Ideologues." Unpublished manuscript.

Peterson, Susan and Michael J Tierney. 2009. "Codebook and User's Guide for TRIP Journal Article Database." *Revised May* .

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick and David Reich. 2006. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics* 38(8):904–909.

Rabinovich, Maxim and David Blei. 2014. The inverse regression topic model. In *Proceedings of The 31st International Conference on Machine Learning*. pp. 199–207.

Roberts, Margaret E. 2015. "Experiencing Censorship Emboldens Internet Users and Decreases Government Support in China." Unpublished manuscript.

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 Airoldi. 2015. "A model of text for experimentation in the social sciences." *Unpublished manuscript* .

Robins, JM and H Morgenstern. 1987. "The foundations of confounding in epidemiology." *Computers & Mathematics with Applications* 14(9):869–916.

Rosenbaum, Paul R and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1):41–55.

Rubin, Donald B. 1980. "Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test"." *Journal of the American Statistical Association* 75:591–593.

Rubin, Donald B. 2006. *Matched sampling for causal effects*. New York: Cambridge University Press.

Rubin, Donald B. and Neal Thomas. 1996. "Matching using estimated propensity scores: Relating theory to practice." *Biometrics* 52:249–264.

Rubin, Donald B. and Neal Thomas. 2000. "Combining propensity score matching with additional adjustments for prognostic covariates." *Journal of the American Statistical Association* 95(450):573–585.

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.

Taddy, Matt. 2013*a*. "Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression." *Technometrics* 55(4):415–425.

Taddy, Matt. 2013*b*. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association* 108(503):755–770.

Taddy, Matt. 2013*c*. "Rejoinder: Efficiency and Structure in MNIR." *Journal of the American Statistical Association* 108(503):772–774.

Taddy, Matt. 2015*a*. "Distributed Multinomial Regression." *arXiv preprint arXiv:1311.6139* .

Taddy, Matt. 2015*b*. Document Classification by Inversion of Distributed Language Representations. In *Proceedings of The Association of Computational Linguistics*.

Taddy, Matt. 2015*c*. "One-step estimator paths for concave regularization." *arXiv preprint arXiv:1308.5623* .

Wacker, Gudrun. 2003. The Internet and censorship in China. In *China and the Internet: Politics of the digital leap forward*, ed. Christopher R Hughes and Gudrun Wacker. Routledge.

Wolpert, David H and William G Macready. 1997. "No free lunch theorems for optimization." *Evolutionary Computation, IEEE Transactions on* 1(1):67–82.