Covariate screening in high dimensional data: applications to text data and forecasting*

Adeline Lo

September 15, 2018

Abstract

High dimensional (HD) data, where the number of covariates and/or meaningful covariate interactions might exceed the number of observations, is increasing used in prediction in the social sciences. An important question for the researcher is how to select the most predictive covariates among all the available covariates. Common covariate selection approaches use ad hoc rules to remove noise covariates, or select covariates through the criterion of statistical significance or by using machine learning techniques. These can suffer from lack of objectivity, choosing some but not all predictive covariates, and failing reasonable standards of consistency that are expected to hold in most high-dimensional social science data. The literature is scarce in statistics that can be used to directly evaluate covariate predictivity. We address these issues by proposing a variable screening step prior to traditional statistical modeling, in which we screen covariates for their predictivity. We propose the influence (I) statistic to evaluate covariates in the screening stage, showing that the statistic is directly related to predictivity and can help screen out noisy covariates and discover meaningful covariate interactions. We illustrate how our screening approach can removing noisy phrases from U.S. Congressional speeches and rank important ones to measure partisanship. We also show improvements to out-of-sample forecasting in a state failure application. Our approach is applicable via an open-source software package.

1 Introduction

Prediction has gained prominence in the social sciences as scholars have become increasingly interested in forecasting events such as conflict (Colaresi and Mahmood, 2017; Beck, King and Zeng, 2000), state failure (Goldsmith et al., 2013; Brandt, Freeman and Schrodt, 2014; King and Zeng, 2001; Esty et al., 1999), elections (Lewis-Beck, Michael S; Tien, 2012; Lockerbie, 2008), and recessions (Dominguez and Shapiro, 2016; Stock and Watson, 2002). Many forecasting efforts now use big data. This is exemplified in the growing popularity of analyses of text data (Lucas et al., 2015), social media data (Wilson, Gosling and Graham, 2012), and larger cross country observational studies (e.g. Demographic and Health Surveys, Afrobarometer). Big datasets are often high dimensional (HD):

*I thank Kosuke Imai and the Imai research group for their thoughtful feedback and comments throughout the development of this research agenda. I would like to also thank Jesse Shapiro, who generously provided feedback and code related to his work on U.S. Congressional text records. Additionally, I am grateful to the following people for their insightful comments and suggestions for this article: Ted Enamorado, Ben Fifield, Shanthi Manian, Héctor Pifarré i Arolas, Soichiro Yamauchi, Yang-Yang Zhou **. The methods described in this paper can be implemented via the open-source statistical software, xx, available at `https://github.com/adelinelo/iscreen`.

they feature more covariates than observations.[1] This makes finding important covariates, covariate interactions, and accounting for their possible nonlinearities difficult to do in a principled manner.

Often social scientists can draw on a body of qualitative and quantitative research that explains the outcome phenomenon, and so one common approach to handle HD data is to select covariates theorized to influence the outcome variable, possibly informed by qualitative and quantitative literature surrounding the outcome. However, this approach may not select the most predictive variables. We may not reliably produce theories to cover the expanse of possible covariates that influence the outcome (see Ward, Greenhill and Bakke (2010) for a treatment of this phenomenon in forecasting conflict). A related approach is to select covariates based on associations between covariates and the outcome, or other correlation-based statistics. We demonstrate in other work, however, that covariates that are highly predictive for an outcome may not necessarily be statistically significant, and may evade the researcher relying on such association-based statistics (Lo et al., 2015).

Another broad category of approaches that is quickly growing in popularity is rather than selecting covariates using theories about the outcome, to rely on machine learning (ML) techniques to data mine for relevant covariates. Most of these approaches, such as random forests and lasso, were designed to handle large numbers of covariates. Rather than hand-picking a (possibly still large) list of covariates believed to be marginally or jointly related to the outcome in likely nonlinear ways, we rely on machine learning algorithms to do so for us. This has led to exciting advances in political forecasting, with works harnessing neural networks (inter alia, King and Zeng (2001)), random forests (Muchlinski, Siroky and Kocher, 2015), and lasso-type models. However, machine learning techniques are not directly applicable in every context. If the dimensions are sufficiently large compared to the number of observations, even ML techniques can be difficult to apply without dimension reduction. As dimensionality (the number of covariates compared to number of observations) grows, the curse of dimensionality can render dimension reduction prior to applying the preferred ML approach directly more attractive (see inter alia (Fan and Lv, 2008)). This problem is further complicated if there are many covariate interactions to consider — also referred to as 'higher order interactions' throughout this article — as these may become more relevant in certain applications where it is likely that many factors influence the outcome; accounting for interactions grows dimensionality mechanically.

This article contributes to the high dimensional prediction literature in social sciences through the introduction of a simple, data-driven covariate screening step that evaluates covariate sets based on their estimated levels of predictivity. We propose incorporating this covariate screening step prior to the model selection and model fitting step to reduce the number of covariates, thereby reducing the high dimensional problem to a low dimensional one. Specifically, we suggest applying a single influence statistic to evaluate covariates for their predictivity. For researchers interested in applying their preferred machine learning approach, this would entail simply including a screening step of all covariates and covariate interactions, prior to their usual ML modeling step. The workhorse

[1] By high dimensionality, we follow Fan and Lv (2008) to mean that dimensionality grows exponentially in the sample size, $\log p = O(n^{\eta})$ for some $\eta \in (0, 1/2)$.

of our proposed covariate screening step is an influence statistic ("I-statistic") that is a nonparametric, lower-bound estimate of covariate predictivity. This statistic can evaluate predictivity both for covariates marginally related to the outcome of interest and for interacting covariates with no marginal relationship to the outcome. It also does not suffer from the problem of indiscriminate growth with the inclusion of additional covariates. We illustrate with simulations in Section 4 how the statistic can a) distinguish between noisy covariates and important covariates and b) recover covariates with joint but no marginal relationships with the outcome variable. We provide a simple workflow diagram for the applied researcher interested in dimension reduction for her high dimensional prediction problem.

The article is organized as follows. We first discuss how our proposed approach contributes to the literature on data-driven covariate screening. We review the common approaches to covariate screening with high dimensional data: theory-guided selection of covariates and machine learning approaches, and highlight our novel use of a screening statistic that is directly related to how predictive a covariate or set of covariates is. We lay the groundwork for understanding what predictivity of a covariate means precisely, by defining predictivity as a parameter of interest. We note that a naive sample estimator of the parameter cannot serve as a helpful criterion to evaluate covariates because the estimator grows indiscriminately with the number of covariates. We then introduce the I statistic as an appropriate alternative with accompanying simulations.

We then illustrate the possible advantages of adding a covariate screening step and utilizing the influence statistic for high dimensional prediction with two applications. The first application screens a large corpus of text drawn from U.S. Congressional speeches from 1873-2016 to determine whether screening results in removal of noise phrases and retention of partisan phrases reflective of important debates throughout history. Specifically, we follow Gentzkow, Shapiro and Taddy (2016) work on measuring partisanship using speeches on the congressional floor. Text data such as these are becoming more prevalent in political science research as online access to documents and different sources of media combined with developments in computational resources has grown. Like Gentzkow, Shapiro and Taddy (2016), we define measurement of partisanship based on how likely it is to be able to predict the party label of a speaker given a fixed amount of speech. Being able to predict the party label well reflects a more partisan Congressional session. To analyze such a large corpus of data, we require restricting the phrases we consider in each session to nonpartisan, nonprocedural phrases, eliminating any 'noise' phrases. We find that data-driven screening of noise phrases results in 1) sensible capture of phrases pertinent to major debates and 2) captures patterns of higher partisanship throughout history.

The second application is a classic example of forecasting state failure, using the well-known Political Instability Task Force dataset (PITF). We adopt the usual goal of such exercises and attempt to forecast country-specific state failure two years into the future. The PITF data exemplifies a common type of conflict data, with observations collected at the country-year level and numbering in the thousands, with a rich set of covariates numbering in the

hundreds. [2] As it is well-known that covariates predicting state failures and conflicts are likely to be both nonlinear and interactive (King and Zeng (2001)), what appears to be a dataset that is not too high dimensional quickly becomes so with the necessary inclusion of interactions of covariates in possibly nonlinear forms. For instance, simply accounting for two-and-three-way interactions in the Political Instability Task Force data requires exploring $p = 310,902,591$ covariates with $n = 8,580$ observations.[3]

In this setting, we show that our covariate screening step improves predictive performance relative to both theory-guided covariate selection alone, and machine learning techniques alone. Selecting covariates through guidance by theory alone leads us to miss some important predictors. We compare the predictive performance of models with covariates selected via theory with models with covariates selected via screening for predictive variables and find that, in general, the latter performs better than the former. We also implement common ML approaches with and without screening and compare out of sample prediction rates. We find that screening improves the out of sample prediction rate across a variety of models.

Finally, we discuss the implications of applying covariate screening to prediction exercises and possible extensions to inference-based work.

2 Related literature

A common approach for the applied researcher interested in forecasting some phenomenon is to rely on a preferred model of choice with which to forecast. Often the evaluation of the forecasting effort is based on testing data, or data set aside from the model training stage, to properly calculate error rates without the risk of overfitting. A wide range of possible models are available to the researcher, whose final choice may result from qualitative and quantitative expertise of the factors that are most important to the outcome. A related, but no less important, choice the researcher may make is in selecting covariates that enter into the chosen model.

An intuitively attractive and popular approach is to rely on expert knowledge on the substantive topic and select variables measuring factors known to (causally) influence the outcome. For instance, a number of works in the state failure literature utilize Bates (2008) work theorizing the importance of the interplay of the size of public revenues, magnitude of rewards the leadership can gain from plundering of citizen wealth, and the leader's discount rate in affecting the likelihood of state failure. Likewise, others build on work by Fearon and Laitin (2003) to predict conflict, adding covariates that capture variation in recorded human rights repression (Rost, Schneider and Kleibl, 2009).

The literature on elections forecasting features several prominent examples of theoretically-motivated covariate-choice as well. Lewis-Beck, Michael S; Tien (2012) propose jobs and proxy models that rely on covariates related to presidential popularity, economic growth, incumbent status, jobs creation, measures of national business condi-

[2] Accounting for all covariates, including intermediary ones used to develop other variables, there are over one thousand.

[3] More generally, this is $\sum_{k=1}^{3} \frac{n!}{k!(n-k)!}$ where n is the number of covariates and k is the highest order of interactions we are considering.

tions. Lockerbie (2008) takes a similar approach to identifying the two important variables, citizen evaluations of their own economic health and the length of service of an incumbent party in the White House, in his forecasting model for the presidential election.

Similarly, in the choice of financial variables as predictors for recessions, researchers on focused on a hypothesized relationship with domestic spread Bernard and Gerlach (1998), interest rate differential between considered countries Nyberg (2010), and forward looking variables such as stock market return Estrella and Mishkin (2016) with likelihood of economic recession. Welch and Goyal (2008) reviews the literature on stock market returns and finds that variables that are suggested to be good predictors lead to poor in and out-of-sample prediction performance.

Theory-guided covariate choice for forecasting models is not limited to the social sciences either. See Jakobsdottir et al. (2009) for a survey of the many works in biological sciences and genetics that rely on similar methods to handpick important covariates to feature in a variety of prediction models.

However, there is evidence that covariates that are theoretically significant in explaining an outcome are not necessarily good predictors of the outcome. A common decision rule is that empirical evidence supports a particular theory if it passes some test of statistical significance, often in a form of a regression. As Ward, Greenhill and Bakke (2010) aptly warns us though, covariates that are statistically significant are not necessarily predictive. While Ward, Greenhill and Bakke (2010)'s caution was directed towards conflict forecasting, we find similar warnings in economics (Welch and Goyal, 2008) and even genetics (Jakobsdottir et al., 2009; Gransbo et al., 2013).

Lo et al. (2015) explain that the main difference between what makes a covariate good for classification and good for statistical significance depend on different properties of the underlying distributions – while the test of significance is a test of the null hypothesis that the distributions of X under the two outcome states are the same, prediction of a class involves testing whether X belongs in one state or another. Even if we were to avoid selecting covariates based on statistical tests that point to theoretical importance, and were to find other means with which to find support for theoretically relevant covariates, we may fall short in theorizing and finding *all* applicable factors. When researchers don't know the true pattern of interactions that generates an outcome, they risk selecting models with only a subset of the relevant interactions when using statistical significance as a decision rule for inclusion. If selection through theory can suffer from overpruning and/or misselection of variables, can machine learning techniques offer a data-driven panacea then?

We are certainly not the first nor the only proponents of covariate screening as a way to dimension reduce. We contribute to a rich computer science literature on covariate screening (also known as feature selection), whereby a subset of an original group of variables is selected for later use in the construction of a model. Screening is often relied upon when the data consists of a modest sample with many covariates, only a few of which may be related to the outcome of interest. Most screening approaches focus on searching for the covariates that satisfy

a specific performance measure, e.g. root mean squared error (RMSE). An approach that bears some similarity to the previously mentioned approach of using theory-guided, statistically significant covariates, is to screen *all* covariates for statistical significance and to conduct statistical models on significant covariates. As such, current approaches do not return a direct assessment of covariates predictivity.

With some minor exceptions, there exist three broad categories of covariate screening in the literature. These are filter, wrapper and embedded methods of screening. Filter methods tend to rely on specific statistics to rank covariates (correlation coefficients, entropy, chi square etc.). Our proposed approach falls in the category of filter methods, as we suggest the usage of an influence statistic to rank covariates (and sets of covariates) to select appropriate covariates for the prediction modeling stage. Perhaps the most relevant comparison approach is sure independence screening (SIS) developed by Fan and Lv (2008), which relies on pairwise correlations to rank covariates. The technical conditions for SIS fail however, if a covariate is marginally unrelated but jointly related to the response (Saldana and Feng, 2018). For instance, if rugged terrain must be jointly present with a large young population for the right circumstances for rebellion to occur, SIS screening may have difficulties identifying variables measuring the terrain and population as important. Our influence statistic differs from SIS in that the influence statistic is nonparametric and can provide information on the predictivity of covariates for the outcome directly. It is also designed to evaluate joint effects of covariates even when the covariates are marginally unrelated to the response. We compare screening performance between the two filtering approaches in simulations in Section 4.

Wrapper methods select covariates based on the error rates of trained models with different subsets of covariates on held out data; they tend to be computationally intensive and are by nature, model-dependent. A popular example of a wrapper method is a stepwise regression. Stepwise (whether forward or backwards) regressions, while quite intuitive may identify different sets of covariates depending on the initial choice of model. An additional problem is that these approaches often have difficulties with correcting for (low) p-values (see Taylor and Tibshirani (2015), Flom and Cassell (2009)).

Embedded methods learn which covariates best contribute to predictive accuracy of the model while estimating the model. Regularization algorithms such as lasso, ridge regression and elastic net are popular examples of embedded approaches. Other commonly used embedded methods include neural nets.[4] If the number of covariates (or interactions between covariates) grow to be much larger than the number of observations, some of these approaches deteriorate in performance. Toy examples show that the lasso performance can deteriorate as the number of predictor variables increases; furthermore, while the lasso can model interactions (see Bien, Taylor and Tibshirani (2013)) it is susceptible to lower prediction performance. This deterioration problem pervades other regularization

[4]Many social scientists embracing machine learning techniques for forecasting political phenomena have utilized popular embedded methods (a burgeoning literature is evident especially in conflict (Beck, King and Zeng, 2000; Ward, Greenhill and Bakke, 2010; Muchlinski, Siroky and Kocher, 2015))

procedures as well (Flynn, Hurvich and Simonoff (2017)). In fact, many approaches are only selection consistent if the number of observations is larger than the predictor dimension (Wang (2009)).

The proposed approach is meant to address some of these issues by suggesting the inclusion of a straightforward covariate screening step prior to the usual preferred modeling techniques with 1) no modeling assumptions, 2) screening for covariates with joint relationships to the response, and 3) screening based on a criterion directly related to covariate predictivity. Furthermore, since the approach entails directly calculating statistics on (unmanipulated) covariates, the results from the covariate screening step are also highly interpretable, as the selected covariates do not undergo any further transformations or weighting, as might be the case for more 'black box' approaches.[5]

3 Method

3.1 Defining predictivity, θ_c

To construct a statistic that estimates predictivity, we first present the concept of predictivity itself. We focus on data where covariates are discrete or have been discretized and the outcome variable is binary. The following results are generalizable to outcome variables that take on a finite number of possible values.

For instance, in forecasting conflict or state failure, the outcome is whether a country-year experiences a conflict ($Y = 1$) or does not experience a conflict ($Y = 0$). In a traditional Bayesian binary classification setting we have a prior over a country experiencing conflict $\pi(Y = 1)$ and experiencing no conflict $\pi(Y = 0) = 1 - \pi(Y = 1)$. We can also define the joint distribution of a group of covariates $\mathbf{X}$ and Y as $P(\mathbf{x}, y) = \pi(y|\mathbf{x}) \cdot P(\mathbf{x}) = P(\mathbf{x}|y) \cdot \pi(y)$, where $\pi(y|\mathbf{x})$ is the posterior distribution and $\pi(y)$ is the prior. We proceed to simplify the setting for presentational purposes, by assuming that priors over conflict and no-conflict are equally likely ($\pi(Y = 1) = \pi(Y = 0) = \frac{1}{2}$) and the cost of incorrect prediction is the same for both conflict and no-conflict. We then relax the assumption over the priors. For generalization to different loss and cost functions, see Appendix **Section #**.

The best classification rule here can be derived using Bayes' decision rule for minimizing the posterior probability of error: predict $Y = 1$ if $\pi(Y = 1|\mathbf{x}) > \pi(Y = 0|\mathbf{x})$, otherwise predict $Y = 0$. We consider here a set of discrete covariates $\mathbf{X} = (X_1, X_2, ...X_m)$ each of which may take different levels. For example, if $\mathbf{X} = (X_1, X_2, X_3)$ and each covariate in the set has three levels, then we can say that $\mathbf{X}$ forms a partition, denoted by $\Pi_{\mathbf{X}}$ with $3^3 = 9$ cells. We may refer to particular cells in $\Pi_{\mathbf{X}}$ as j, where here $j = 1, ..., 9$.

The correct prediction rate θ_c on $\mathbf{X}$ using the full Bayes' decision rule can be calculated as:[6]

$$\theta_c(\mathbf{X}) = \frac{1}{2} \sum_{j \in \Pi_{\mathbf{X}}} \max\{\Pr(j|Y = 1), \Pr(j|Y = 0)\} \tag{1}$$

[5] A prominent example are neural networks (Tomandl and Schober, 2001); though see Olden and Jackson (2002) for developing specific statistical tools to understand variable contributions to neural networks).

[6] The correct prediction rate can be written also as one minus the error rate, $1 - \theta_e = \theta_c$.

We can rewrite θ_c into the following:

$$\begin{aligned}\theta_c(\mathbf{X}) - \frac{1}{2} &= \frac{1}{2}\sum_{j\in\Pi_\mathbf{X}} \max\{\Pr(j|Y=1), \Pr(j|Y=0)\} - \frac{1}{2} \\ &= \frac{1}{2}\sum_{j\in\Pi_\mathbf{X}} \left(\max\{\Pr(j|Y=1), \Pr(j|Y=0)\} - 1\right) \\ &= \frac{1}{2}\sum_{j\in\Pi_\mathbf{X}} \left(\max\{\Pr(j|Y=1), \Pr(j|Y=0)\} - \frac{1}{2}(\Pr(j|Y=1) + \Pr(j|Y=0))\right)\end{aligned}$$

The last equation is equivalent to half the difference between the maximum and the mean of two probabilities, which can be rewritten as half the absolute difference between the two probabilities:

$$\theta_c(\mathbf{X}) = \frac{1}{2} + \frac{1}{4}\sum_{j\in\Pi_\mathbf{X}} |\Pr(j|Y=1) - \Pr(j|Y=0)| \tag{2}$$

After having defined the predictivity of covariate set $\mathbf{X}$, $\theta_c(\mathbf{X})$, we turn to forming statistics for estimating this quantity of interest. It is easy to see in this theoretical setting that θ_c increases or remains the same value when adjoining additional covariates to $\mathbf{X}$ and thereby expanding the resulting partition $\Pi_\mathbf{X}$; a naïve sample estimate of $\hat{\theta}_c$, also known as the training rate, is non-decreasing with the addition of any covariate – be it a covariate with information or simply noise. Such an estimator would encourage the researcher to continually add covariates until the partition space is maximized and each cell has at most one observation (see Appendix for an in depth explanation of the phenomenon). Since the sample analog is not dependable in discerning between noise and important covariates, and is therefore not a helpful screening statistic, we require a different statistic that can differentiate between these types of covariates.

3.2 Influence statistic I

The influence statistic was originally introduced in Chernoff, Lo and Zheng (2009). It is defined as:

$$I(\mathbf{X}) = \frac{\sum_{j=1}^{J} n_j^2(\bar{Y}_j - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3}$$

where $\bar{Y}_j$ is average value of Y in cell j , n_j are the number of observations in cell j. The collection of all cells, each representing a different combination of levels for each covariate in $\mathbf{X}$, comprises the partition of $\mathbf{X}$, $\Pi(\mathbf{X})$. The curious reader may notice the statistic bears some resemblance to the Pearson chi-square statistic; an important difference between the two statistics lies in the n_j^2 term outside the squared difference of local mean value of Y, $\bar{Y}_j$, and global mean value of Y, $\bar{Y}$. This places greater weight on cells with more observations. Intuitively, we might expect that "emphasizing where the data is" might be helpful in a high dimensional setting of many cells and fewer observations.

Since the influence statistic is measured at the partition level, it essentially captures the influence of a *set* of covariates. Attaining high influence statistic values does not require lower order or marginal relationships with the outcome; this is a particularly attractive feature for which approaches such as the hierarchical lasso (Bien, Taylor and Tibshirani (2013)) and random forests (Breiman, 2001) have greater difficulties achieving due to the reliance on the existence of lower order signals in order to retrieve higher order interactions. Covariate sets with values equal to or less than 1 are denoted as noise; for details and proofs related to this cut off, see Chernoff, Lo and Zheng (2009).

3.3 Practical application

For ease of application, we provide a general workflow for the social scientist interested in incorporating screening in his prediction research in Figure (1). The idea is to simply insert screening before application of the scientist's preferred prediction model to the data.

In Step 1, the data requires some preprocessing; in particular, the current implementation of our approach requires discretization of continuous covariates. We recommend using a data-driven approach to discretization, such as applying k-means clustering, with a small size for k, such as 2 or 3, as large k can lead to overly sparse partitions. Our R package provides a function for such discretization. The researcher also makes a choice, based on his expert opinion on the topic, of the order of interaction,m, that covariates can achieve in jointly affecting the outcome. For instance, if the researcher believes that his application may involve three-way interactions of covariates, then he selects $m = 3$. The selection of m helps directly define the total screening space to search for important covariates.

In Step 2, we begin screening. If our computing resources permit, we conduct a full screening of all m-wise interactions, as well as all lower orders, for covariates and calculate I statistics for all sets. However, if the full number of covariate sets grow too large for a full search in a sensible amount of computing time, we suggest a random search. Here, the researcher randomly draws m covariates many times from the full set of covariates. After each draw (Steps 3 and 4):

1. Calculate $I(m)$, then randomly drop one covariate in the m set at hand, resulting in m' covariates.

2. Calculate $I(m')$ and compare $I(m')$ with $I(m)$. If the former is larger, then the dropped covariate was noise and we allow m' to become our new m. If the latter is larger, then the dropped covariate was important *when combined with the remaining set* and should be returned. Should we find that the latter is larger, and we return the dropped covariate, we proceed with removing *another* random covariate from m.

3. Repeat until the covariate set cannot be reduced – no random drops of covariates results in larger I values – and retain the resulting covariate set. We repeat this random draw many times and rank our resulting covariate sets by I, removing any sets that have values less than 1.

4. With the resulting covariate list, ranked by predictivity, we can proceed to Step 5 where we apply the preferred prediction model to the reduced list of predictive covariates.

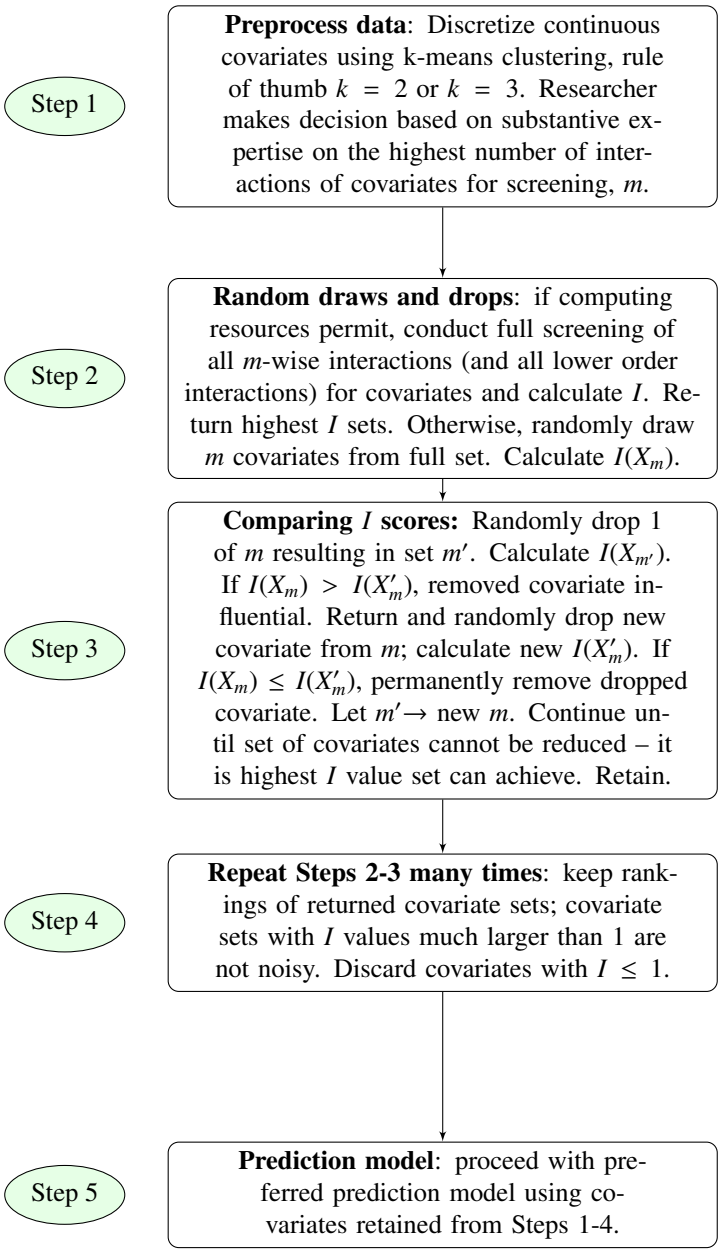

Figure 1: **Workflow for applied researcher**.

3.4 Asymptotics

In this section, we provide details on some of the asymptotic properties of the I statistic. Theorem 1 demonstrates that the I statistic approaches the L_2 norm form of predictivity up to a constant, in probability. Theorem 2 provides the asymptotic lower bound for predictivity using the I statistic, as well as the estimated lower bound. The I statistic

can be shown to converge asymptotically to a constant multiple of:

$$\theta_I(\mathbf{X}) = \sum_{j \in \Pi_{\mathbf{X}}} [\Pr(j|Y=1) - \Pr(j|Y=0)]^2 \tag{4}$$

Equation (4) is a lower bound of the main term in Equation (2) (see Appendix Section # for related proof):

$$\sqrt{2 \sum_{j \in \Pi_{\mathbf{X}}} [\Pr(j|Y=1) - \Pr(j|Y=0)]^2} \leq \sum_{j \in \Pi_{\mathbf{X}}} |\Pr(j|Y=1) - \Pr(j|Y=0)| \tag{5}$$

Thus while estimating the sample analog of the right hand side of Equation (2) leads to the naïve training rate estimator that indiscriminately increases with the addition of more covariates, a phenomenon akin to the behavior of the R^2 of a model with the addition of new covariates, searching for covariate sets with large measured influence values can simultaneously lead to covariate sets with higher predictivity. We proceed with several assumptions, and then present the asymptotic lower bound for the correct prediction rate:

Assumption 1: Sparsity We make the usual sparsity assumption whereby we expect that the true function f of covariates that influence the outcome only depends on a small number of total covariates. Let $Y_i = f(X_i) + \varepsilon_i$, $i = 1, ..., n$ where $X_i = (X_i(1), ..., X_i(p) \in \mathbb{R}^p$ is a p-dimensional covariate, $f : \mathbb{R}^p \to \mathbb{R}$ is an unknown function and $\varepsilon_i \sim N(0, \sigma^2)$. Suppose f satisfies such a sparseness condition, so that

$$f(x) = f(x_R)$$

where $x_R = (x_l : l \in R), R \subset 1, ..., p$ is a subset of the total p covariates of size $r = |R| \ll p$.

Assumption 2: Stable proportion We assume a stable proportion of each class outcome. That is, $\frac{n_{Y=1}}{n} \to \lambda$, $\lambda \in (0, 1)$.

Assumption 3: Equal priors We assume the prior for $Y = 1$ is equal to the prior for $Y = 0$, $\pi(Y = 0) = \pi(Y = 1) = \frac{1}{2}$. We present results where we relax this assumption in the Appendix.

Assumption 4: 0/1 Loss We assign a cost (loss) of one to the failure of predicting the correct class. In some situations, it may be appropriate to have differing prediction errors associated with each class outcome (for instance, a false negative may merit a different loss to a false positive in a medical diagnosis for a patient). We present results with different loss functions in the Appendix.

Theorem 1 Under Assumptions 1-4,

$$\lim_{n \to \infty} \frac{s_n^2 I(\mathbf{X})}{n} \xrightarrow{p} \lambda^2(1-\lambda)^2 \sum_{j \in \Pi_{\mathbf{X}}} (\Pr(j|Y=1) - \Pr(j|Y=0))^2 \tag{6}$$

where $s_n^2 = \frac{n_{Y=1} \cdot n_{Y=0}}{n}$.

Theorem 2 Under Assumptions 1-4, an asymptotic lower bound for the predictivity of $\mathbf{X}$ covariates is:

$$\theta_c(\mathbf{X}) \geq \frac{1}{2} + \frac{1}{4}\sqrt{2 \lim_{n\to\infty} \frac{I(\mathbf{X})}{n\lambda(1-\lambda)}} \tag{7}$$

Then the estimated lower bound for $\theta_c(\mathbf{X})$ is:

$$\frac{1}{2} + \frac{1}{4}\sqrt{\frac{2I(\mathbf{X})}{n\lambda(1-\lambda)}} \tag{8}$$

We refer the interested reader to Appendix 7 for proofs of Theorems 1 and 2 and derived additional properties of the influence statistic, such as the first and second moments.

We note that the true predictivity θ_c of a covariate set $\mathbf{X}$ is an absolute difference, an L_1 norm, of the form:

$$\text{constant} \cdot \sum_{j \in \Pi_{\mathbf{X}}} |\Pr(j|Y=1) - \Pr(j|Y=0)|$$

And that the reason the influence statistic serves as a lower bound lies in how it asymptotically approaches the L_2 norm version of θ_c:

$$\text{constant} \cdot \sum_{j \in \Pi_{\mathbf{X}}} (\Pr(j|Y=1) - \Pr(j|Y=0))^2$$

Intuitively, we might notice that the closer we approach that L_1 norm form the better we can approximate it. As noted earlier however, the sample analog form for the L_1 norm itself suffers from an inability to differentiate between influential and noisy covariates. Should we attempt to design a statistic meant to estimate something of the form $L_1 - p$, $0 < p < 1$ we would lack in convexity.

Suppose we were to design a statistic for L_p, $p = 2 + \alpha$: Consider the case where $p = 2 + \alpha$, $\alpha \in \mathbb{R}_{>0}$. The L_p norm in this case is convex. However, we know the following is true for any function $f(x)$:

$$|f(x)|^1 \geq |f(x)|^2 \geq |f(x)|^{2+\alpha}$$

where $0 < \alpha < \infty$. That is, the more the p increases, the more convex (and further away from L_1) the function becomes. For this reason, we can eliminate $p = 2 + \alpha$ as an alternative that is better than using an L_2 norm. This leaves us with $p = 1 + \varepsilon$, $\varepsilon \in (0, 1)$, which is a difficult space of norm for which to design statistics. As such, we opt for the closest convex norm to Equation 1 for which we can design a statistic.

4 Simulations

4.1 Distinguishing noise and influential covariates, finding interactions

To address some of the benefits and drawbacks of our proposed approach, we provide a variety of simulations that lay out several important characteristics of the influence statistic. First, in order to demonstrate that the statistic can differentiate between noise and influential covariates when assessing the joint predictivity of a group of covariates, we show that after ranking all covariates and covariate sets by their I values, the 'true' sets of influential covariates are still retrieved when the number of noise covariates increase, even when sample size remains modest in comparison. Second, to demonstrate that the statistic can find jointly influential covariates even in the absence of marginal relationships with the outcome variable, we present simulations with data generating processes with a) marginal relationships between a few covariates and the outcome, and no joint effects, b) marginal and joint effects, and c) only joint effects. A limitation which we discuss in Section 4 in the approach is the need to discretize continuous covariates in order to calculate the influence statistic (though original continuous covariates are utilized during any modeling stages). In simulations, we illustrate how discrete and continuous covariates are still able to be retrieved when both are in discrete form.

We first demonstrate how the I statistic can distinguish noise covariates from influential ones, and thereby retain a smaller subset of truly important covariates, with a series of simulations. As noted in Section 2, sure independence screening (SIS) is a comparable alternative approach in the screening literature as it is also a well-known filtering approach designed for high dimensional data.

The first set of simulations are conducted with data that have observations $n \in \{200, 400, 1600\}$. The total number of covariates, both noise and influential, are $k \in \{10, 100, 200\}$, with either 3 or 4 "influential" covariates, or covariates that contribute to the outcome in the data generating process. We consider two types of models, linear and nonlinear, where nonlinear models include sin and cos functions. We conduct covariate screening with the I statistic and report the proportion of retained influential covariates out of the actual influential covariates for each scenario. As a baseline comparison, we also present covariate capture results for SIS. Full details of the simulations are available in the Appendix.

Figure (2) illustrates our first set of simulation findings. Overall, I performs better than SIS for finding influential covariates. This holds when varying the number of observations, the number of total covariates to screen, and whether the model is linear or nonlinear. Both approaches show marked improvements in capture rates of influential covariates as the number of observations n increases, while holding the total number of covariates k constant. When the sample size is small, $n = 200$, compared to the number of total covariates to screen, screening with the I statistic demonstrates the greatest advantage against SIS.

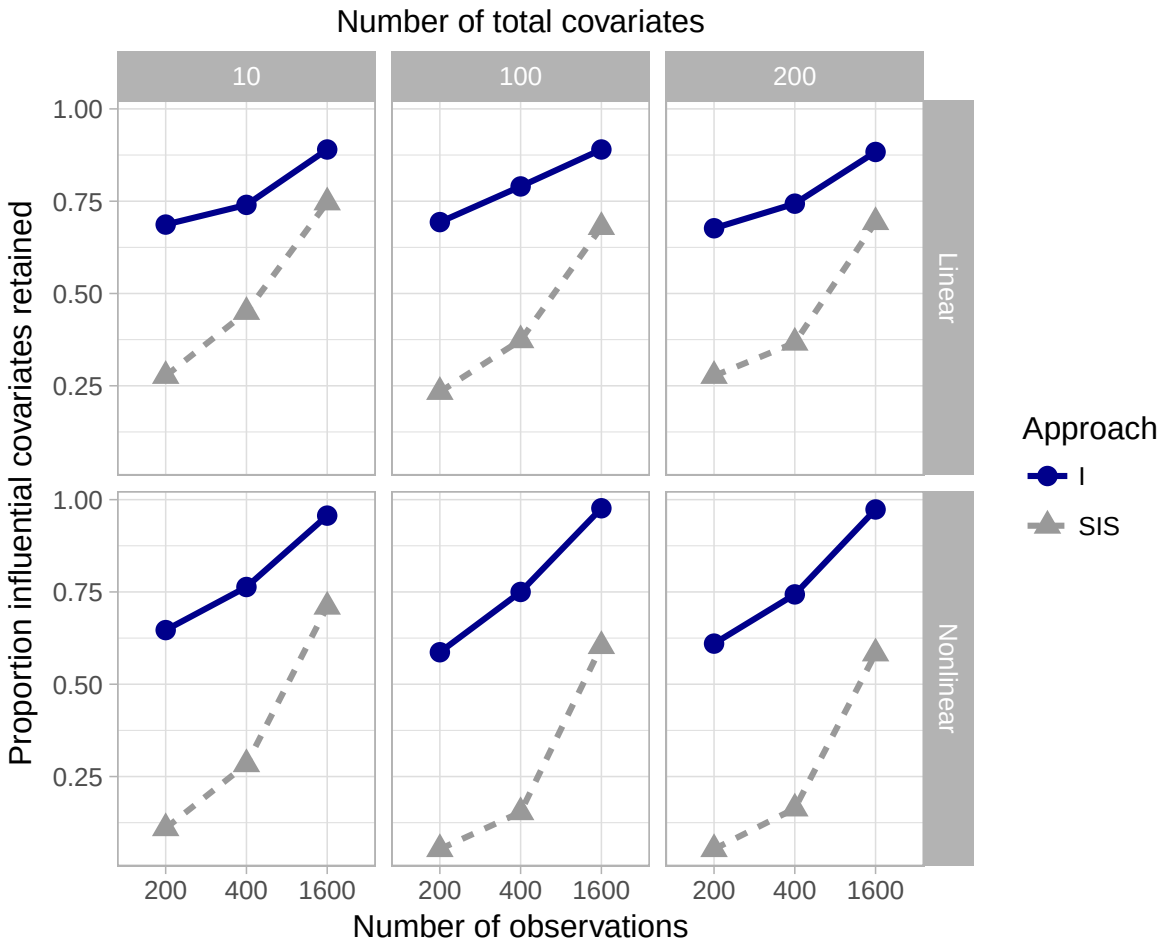

Figure 2: **Proportion of retained influential variables**. Comparison of covariate screening using the I statistic and using SIS. The y-axis reflects the proportion of influential covariates successfully discovered by the screening method from the total true influential covariates. Number of observations n in the simulated data are noted by the x-axis. Columns from left to right give the total number of covariates (influential and noisy) in the data, where $k = 10$ on the left and $k = 200$ on the right most column. The first row of graphs are for data generated with a linear model, while the second row of graphs are for data generated with a nonlinear model. Blue solid trend lines represent screening via the I statistic; gray dotted trend lines representing screening via SIS. Each data point represents results from 1000 data simulations.

The second set of simulations demonstrate the I statistic's screening capabilities for covariate interactions, even in the hard case when covariates that compose the interactions do not have marginal relationships with the outcome but do have a joint relationship. We thus cannot rely on "piggybacking" off of marginal signals to discover joint ones. Again, we vary $n \in \{200, 400, 1600\}$, $k \in \{10, 100, 200\}$ and use a linear and a nonlinear model. We compare *covariate set* capture for SIS with our preferred approach.

Results from our second set of simulations are presented in Figure (3). Again, screening with the I statistic leads to better capture of influential covariate sets when compared with sure independent screening, across all scenarios. Both approaches improve in proportion of covariate sets found as the sample size grows, though SIS fares more poorly when the n/p ratio grows smaller. While the I statistic capture rates remain consistent across

linear and nonlinear models, SIS performance under the linear model outstrips that under nonlinear model. Finally we note that in screening for all marginal and two-way interactions between covariates (20,100 total covariate sets), the right column of simulations represents examples of working with high dimensional data; even when the sample size is small (n = 200) or modest (n = 400), the I statistic is still able to distinguish influential covariate sets.

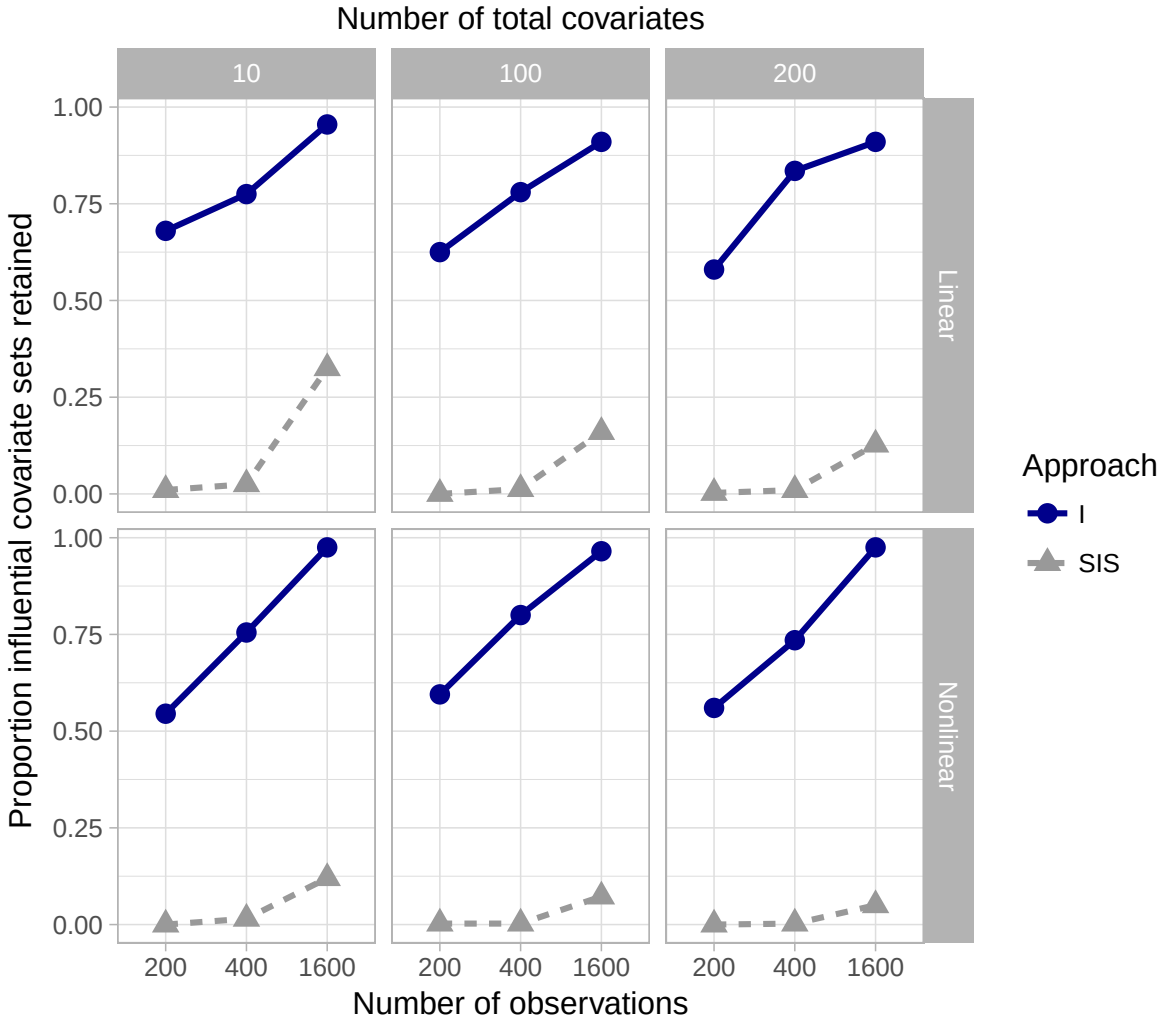

Figure 3: **Retaining influential covariates with joint but no marginal effects on outcome: proportion of retained influential variables**. Comparison of covariate screening using the I statistic and using SIS. Y-axis is the proportion of influential covariate sets successfully discovered by the screening method from the total true influential covariate sets. Covariate sets are jointly but not marginally related to outcome. Number of observations n are denoted by the x-axis. Columns from left to right are total number of covariates (influential and noisy) in the data, where k = 10 on the left and k = 200 on the right most column. The first row of graphs are for data generated with a linear model, while the second row of graphs are for data generated with a nonlinear model. Blue solid trend lines represent screening via the I statistic; gray dotted trend lines representing screening via SIS. Each data point represents results from 1000 data simulations.

4.2 Limitations with discretization

A clear limitation to the proposed approach is that the approach of analyzing partitions created by discrete covariates requires either covariates to be discrete to begin with, or for the researcher to discretize the covariates prior to evaluating covariate predictivity. In practice, we find that while discretization of continuous covariates via data-driven approaches such as k-means clustering inherently results in some loss of fine grained information, obtaining information on covariate set predictivity is still possible. To illustrate, we provide simulations using set of covariates drawn from 1) normal and Bernoulli distributions, and 2) only normal distributions. We retain similar simulation parameters with simulations 1 and 2 otherwise.

Results from this last set of simulations are presented in Figure (4). Again, applying the I statistic as a criterion to select covariates tends to retain more of the truly important covariates than SIS. Both approaches perform demonstrably better under a true linear model than a nonlinear one; however I-led covariate discovery can still recover most of the influential covariates even in nonlinear settings, with improved performance as the sample size-covariate size proportion grows. Most important to note however, is that even when continuous covariates (drawn from a normal distribution) are discretized, our approach can still recover important covariates, faring especially well when the underlying model is linear. Further work along the lines of creating I estimators adapted to continuous covariates are still of pressing importance, however. We discuss possible extensions of such work in the Discussion Section.

We next turn to two high dimensional data applications, each of which has either explicit or implicit goals of prediction.

5 U.S. Congressional text application

5.1 Data

Congressional speeches are a good example of big text data, full of important information that may be difficult to extricate given the sheer number of phrases per speaker, and likely overabundance of procedural and otherwise 'noise' phrases. We consider an application by Gentzkow, Shapiro and Taddy (2016) that proposes measuring partisanship in the U.S. Congress over the last century using Congressional speeches to demonstrate the possible advantages of screening out the noise phrases and reducing the number of phrases to consider when measuring partisanship. We compare incorporating a phrase screening step into the preprocessing steps Gentzkow, Shapiro and Taddy (2016) (henceforth 'GST') take prior to estimating partisanship using with GST's model. The full corpus of data covers the speeches delivered by representatives on the floor of the 43rd to 114th United States Congress (1873-2016), which includes both the House of Senate and House of Representatives, as recorded in the text of the *United States Congressional Record*.

The main data $\mathbf{C}_t$ is structured in the following way; each row corresponds to a speaker, while columns repre-

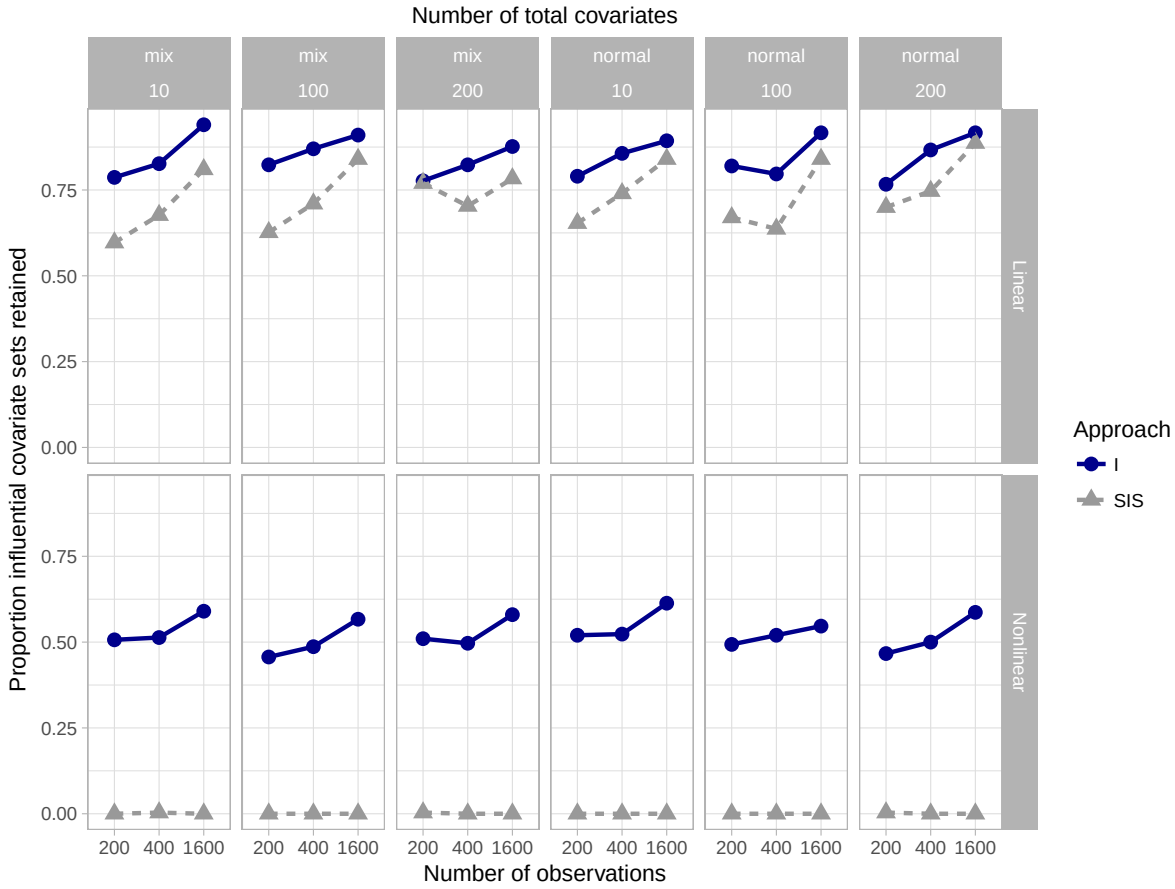

Figure 4: **Proportion of retained influential variables with covariates drawn from normal and mix of normal and Bernoulli distributions**. Comparison of covariate screening using the I statistic and using SIS. Y-axis is the proportion of influential covariate sets successfully discovered by the screening method from the total true influential covariate sets. Number of observations n are denoted by the bottom x-axis. First three columns are for simulated data with covariates drawn from a mix of normal and Bernoulli distributions. Last three columns are for simulated data with covariates drawn from normal distributions. Total number of covariates (influential and noisy) in the data are the values below distribution type. The first row of graphs are for data generated with a linear model, while the second row of graphs are for data generated with a nonlinear model. Blue solid trend lines represent screening via the I statistic; gray dotted trend lines representing screening via SIS. Each data point represents results from 1000 data simulations.

sent distinct two-word phrases or bigrams so that each element c_{ijt} corresponds to the number of times speaker i has spoken the phrase j in session t. To attain the data from the full scripts of speeches we follow GST's preprocessing steps up to the point of the authors dropping phrases that "are likely to be procedural or have low semantic meaning". From there we conduct removal of phrases in two ways. See Figure (5) for steps taken to replicate GST's approach as well as introducing covariate screening in lieu of GST's phrase-removal rules. In pink are the steps taken to replicate GST's approach of removal via the series of rules and cut offs in their Online Appendix. In blue is our data-driven approach to screen phrases (only Step 2 differ between the two approaches). After phrases are reduced to a manageable size, we apply a multinomial inverse regression with a penalization as proposed by GST to estimate partisanship in each session.

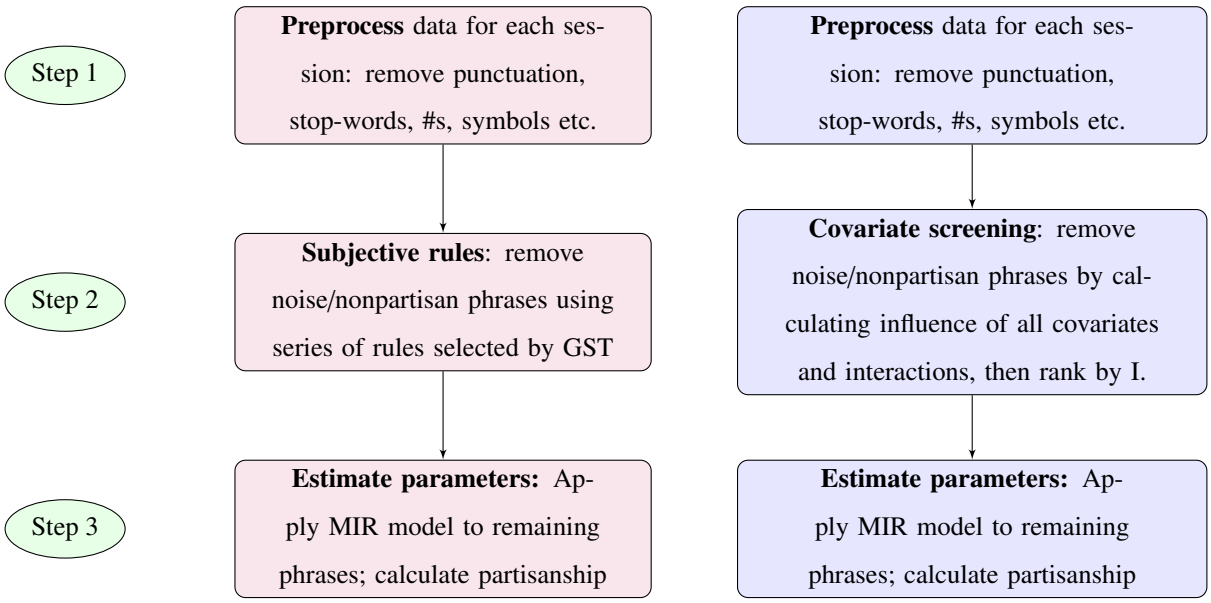

Figure 5: **Workflow for processing and conducting covariate screening of U.S. Congressional Records data**. In pink are the steps taken to follow GST's approach as closely as possible. In blue are steps taken to incorporate covariate screening in order to remove noise phrases and non partisan phrases.

5.2 Discussion of selected phrases

After Step 1, we use our covariate screening approach to remove noise phrases and rank phrases for influence in Step 2.

Before moving to Step 3 in Figure (5), however, we review some of the phrases that survive the covariate screening and conduct a qualitative sanity check of sorts – does covariate screening return phrases that are reflective of the types of debated topics throughout the different Congressional sessions? To do so, we review some of the top phrases returned in a few sessions throughout our dataset (see Table (1)).

We inspect the phrases returned from screening from the 43rd session (1873-1875), a session which convened during the postwar era of Reconstruction. The divide between Republicans and Democrats had been widening over the topic of federal authority; Republicans supported an expansion of federal power ('feder author', 'power state', 'state feder', 'right power') to provide civil rights for African Americans in the Civil Rights Act of 1866 (see phrases such as 'oppress peopl'), the Fourteenth Amendment and the formation of the Department of Justice, while Democrats opposed these proposals. With the passage of the Civil Rights Act and the 14th Amendment, Republican lawmakers claimed the sovereignty of the national government ('sovereign state'), particularly with respect to determining the status of and defending the rights of Americans regardless of race (Kaczorowski, 1987).

By the 80th Congressional session (1947-1949), the U.S. was emerging from the aftermath of World War II. During this time, what became known as the Truman doctrine began to take form, a series of political measures taken to counter Soviet geopolitical expansion in the Cold War. These measures included ending racial segregation

	Sessions			
	43	**50**	**60**	**70**
	1873-1875	**1887-1889**	**1907-1909**	**1927-1929**
1	power state	system taxat	reserv balanc	repres great
2	oppress peopl	great question	repres people	presid veto
3	feder author	free trade	campaign fund	everi dollar
4	state feder	import industri	sovereign state	cotton crop
5	sovereign state	countri time	common sens	class peopl
6	wish state	industri unit	cotton crop	respect state
7	entitl receiv	tax collect	privat citizen	interest farmer
8	expenditur money	duti sugar	equal right	fair just
9	govern pay	remov duti	white peopl	right govern
10	right power	mill bill	state legislatur	state power
	Sessions			
	80	**90**	**100**	**114**
	1947-1949	**1967-1969**	**1987-1989**	**2015-2017**
1	back hous	present administr	good intent	violenc never
2	great man	commend presid	support substitut	commonsens gun
3	forc unit	fiscal crisi	tax increas	invest infrastructur
4	live standard	administr fail	dollar go	lgbt communiti
5	blank check	public life	fiscal respons	gun violenc
6	govern right	public servic	tax incent	expand background
7	foreign aid	civil liberti	job creat	trillion debt
8	militari naval	financi assist	freedom nicaragua	access health
9	cut fund	social justic	moral respons	victim gun
10	system free	fix incom	invest children	immigr reform

Table 1: **Top influential phrases from every 10th session (114th session as last)**

in the armed forces and providing foreign aid for Greece and Turkey to prevent the two countries from falling to the hands of communism during the Greek Civil War (captured by phrases such as "foreign aid", "militari naval").

At the convening of the 114th session (2015-2017), Congress was debating gun control with renewed interest – after the armed attack on a historic black church in Charleston, South Carolina on June 17, 2015, followed by shootings in Tennessee, Oregon, Colorado, California, and Florida among others. Sharp partisan differences exist still over several issues: waiting periods for those who wish to purchase guns legally (Democrats largely opposed proposals intending to shorten the periods while Republicans are divided), the degree to which legally obtained guns contribute to gun violence (Democrats are much more likely than their Republican counterparts to support the view that legal guns contribute a great deal to gun violence), and whether imposing more difficult for people to legally acquire guns would ameliorate the likelihood of mass shootings in the future (a majority of Democrats believe this such hurdles would diminish the likelihood, while Republicans are more skeptical) (Oliphant, 2017). The preoccupation of debate with gun violence and gun control can be evidenced by the selection of phrases "violenc never", "commensens gun", "gun violenc", "expand background", and "victim gun".

5.3 Multinomial inverse regression Model

Gentzkow et al. (2018) build on methods developed by Taddy (2013, 2015) and propose a structural choice model and impose an $L1$ penalty via a lasso in a multinomial inverse regression (approximated with the likelihood of a Poisson model) to the data and find that partisanship has grown slightly over the years.

We briefly review the model proposed by the authors (for further details, see Section 3, Gentzkow et al. (2018)) as well as their approach to estimating the parameters of interest and conducting inference.

Let $R_t = i : P(i) = R, m_{it} > 0$, $D_t = i : P(i) = D, m_{it} > 0$ denote the sets of Republicans and Democrats, respectively, for each session t. $\mathbf{x}_{it}$ is $\mathbf{K}$-vector of speaker characteristics: state, chamber, gender census region, and whether the political party is in majority in session t.

Assume:

$$\mathbf{c}_{it} \sim \text{MN}\Big(m_{it}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{it})\Big)$$

where $\mathbf{q}_t^P(\mathbf{x}_{it}) \in (0,1)^J$ for all P, i, and t. m_{it} is the verbosity of individual i in time t and the probability of speaking each phrase is $\mathbf{q}_t^P(\cdot)$, the combination of which characterize the speech-generatign process. Each speaker speaks to maximize the utility payoff associated with speaking their phrases in each session, u_{it}, which in turn enters into the probabilities of speaking each phrase.

$$q_{jt}^{P(i)}(\mathbf{x}_{it}) = e^{u_{ijt}} / \sum_l e^{u_{ilt}}$$

$$u_{ijt} = \alpha_{jt} + \mathbf{x}_{it}'\gamma_{jt} + \varphi_{jt}\mathbf{1}_{i \in R_t}$$

where $\varphi_{\mathbf{t}}$ is J-vector mapping speech to public opinion, α_t J-vector for baseline popularity of phrase at time t, γ_j $K \times J$ matrix mapping speaker characteristics into utility of using each phrase.

Finally, define partisanship of speech at $\mathbf{x}$:

$$\pi_t(\mathbf{x}) = \frac{1}{2}\mathbf{q}_t^R(\mathbf{x}) \cdot \rho_t(\mathbf{x}) + \frac{1}{2}\mathbf{q}_t^D(\mathbf{x}) \cdot (1 - \rho_t(\mathbf{x}))$$

$$\rho_{jt}(\mathbf{x}) = \frac{\mathbf{q}_t^R(\mathbf{x})}{\mathbf{q}_t^R(\mathbf{x}) + \mathbf{q}_t^D(\mathbf{x})}$$

Then, the average partisanship in session t (and our main statistic of interest) can be denoted as:

$$\bar{\pi}_t = \frac{1}{|R_t \cup D_t|} \sum_{i \in R_t \cup D_t} \pi_t(\mathbf{x}_{it})$$

where $\rho_{jt}(\mathbf{x})$ is posterior belief that an observer with a neutral prior assigns to a speaker being Republican if speaker chooses phrase j in session t and has characteristics $\mathbf{x}$. Partisanship ($\pi_t(\mathbf{x})$) averages the posterior over the possible phrases and parties. Average partisanship ($\bar{\pi}$) averages partisanship over characteristics of speakers active in same session. Estimate parameters of interest $\{\alpha_t, \gamma_t, \varphi_t\}_{t=1}^T$ by minimizing the following penalized objective function:

$$\sum_j \left\{ \sum_t \sum_i \left[m_{it} \exp(\alpha_{jt} + \mathbf{x}_{it}'\gamma_{jt} + \varphi_{jt}\mathbf{1}_{i \in R_t}) - c_{ijt}(\alpha_{jt} + \mathbf{x}_{it}'\gamma_{jt} + \varphi_{jt}\mathbf{1}_{i \in R_t}) + \psi(|\alpha_{jt} + \|\gamma)jt\|_1 + \lambda_j|\varphi_{jt}|) \right] \right\} \quad (9)$$

The parameters are estimated via a multinomial inverse regression (MIR) with the appropriate penalization incorporated for the party-phrase coefficients. Inference is done via subsampling (Politis et al. 1999).

5.4 Results

The estimated average partisanship of speech is presented in Figure (6). The x-axis depicts the time period from the 43rd session to the latest 114th session. The y-axis reflects the average estimated partisanship. A baseline random series is plotted in grey, whereby party labels are randomized for the data from each session before estimating partisanship. The magenta line reflects the estimation process presented by Gentzkow et al., which does not utilize our proposed screening step (and is thus referred to as 'w/o screening' in the Figure). The blue trend line represents the same estimation process as the magenta line, with our screening approach incoroprated. Overall, there is an increase in partisanship over time. Estimating partisanship with screening incorporated to remove noisy phrases however, leads us to find a much higher level of partisanship throughout history, though with a similar upwards surge of partisanship in recent sessions. Taking polarization as a proxy to partisanship, research has shown that rises in polarization in more recent years is not an exceptional phenomenon. Rather, as Han and Brady (2007) puts it, it is "a delayed return to historical norms". Should we approach estimating partisanship through the interpretation

of the degree of accuracy with which we can predict a speaker's party label, then a model that predicts well is of paramount importance. As such, our proposed screening step can be considered highly necessary as illustrated in the patterns of higher partisanship detected.

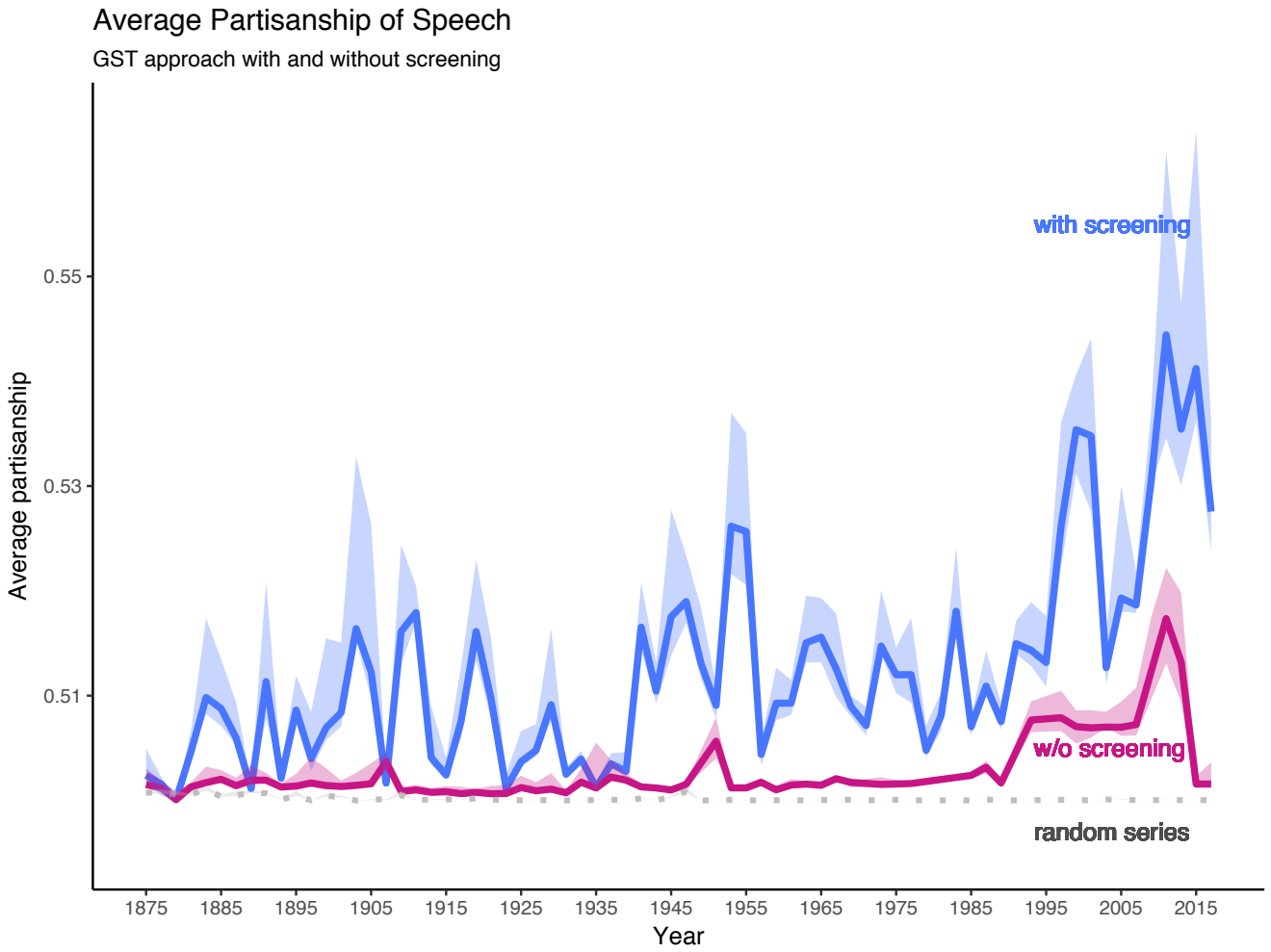

Figure 6: **Average partisanship of speech ($\bar{\pi}_t$) for each session** . The magenta trend line depicts the approach taken by Gentzkow et al. using subjective rules to remove nonpartisan phrases before estimation and inference using MIR with penalization for average partisanship. The blue trend line depicts the estimated average partisanship for each session using the exact same approach as the magenta line, except the removal of nonpartisan phrases is conducted by data-driven phrase-screening. The grey dotted line represents estimation of the model in the same fashion as the magenta line, except for each session, all party labels are reassigned at random. Thus, the grey dotted trend serves as a baseline comparison after accounting for similar $\mathbf{x}$ characteristics aside from party label, as well as overal amounts of speech produced in each session.

6 Forecasting state failure

Being able to forecast peace and conflict is a major motivation for peace research: early warning systems can make it possible to prepare for and intervene against conflicts. We focus our next application on forecasting state failure using data from King and Zeng (2001) drawn from the Political Instability Task Force (PITF). PITF is a

U.S. government financed research group tasked with the goal of predicting political instability for countries across the world in advance and the data it collects has been widely studied in the peace and conflict literature (Goldstone et al., 2010, 2000). There are a total of 8580 country-years (1955-1998) with a total of 1231 covariates. We split the data into training and testing – training data span from 1955-1995 (n = 7995), while testing data span 1996-1998 (n = 585); thus we are attempting to forecast the state failure outcomes of countries three years into the future.

We compare against three theory-guided approaches to selecting covariates – King and Zeng (2001), Goldstone et al. (2010) and Bates (2008). King & Zeng use the same dataset used here to forecast political instability with neural networks. While neural networks are a popular tool in the machine learning tool kit, in King & Zeng's usage they are not used for the purpose of covariate selection; rather, the authors select their covariates guided by the literature on state failure, and rely on neural networks to model possible nonlinearities in the selected covariates flexibly. Goldstone et al. (2010) and Bates (2008)rely on logistic regressions for their forecasting efforts, and differ in their approaches to selecting covariates.

We also compare against common data-guided, machine learning approaches that have gained popularity in the conflict forecasting literature: neural networks, random forests, and lasso. For each model, we add a covariate screening step. Figure (7) presents the workflow associated with this application. First, we split the data into training and testing sets. This step is designed safeguard our screening and fitted models from overfitting; all of our reported prediction outcomes (here area under the receiver operating curve (AUC) values) are from held-out testing data from future years. Given the preponderance of missing values across observations and covariates and the low-likelihood that the missing values are missing at random, we proceed with treating missingness as information itself. To do this, we create binary indicators for covariates that indicate whether there is missingness in the covariate. In Step 3, we discretize continuous covariates. Given the highly interactive nature of conflict and instability data, we search up to 5-way interactions of covariates.[7]

In Steps 4 and 5, we rank covariate sets by highest I values and discard any sets with $I < 1$ as noise covariates. Using the (non-discretized) original forms of each covariate, we input high I scoring sets into the models of choice and fit our model on the training data. In Step 6 we evaluate our model's forecasts with testing data.

The next section presents results from our state failure forecasting exercise.

[7] Selection of covariates with up to 4-way and up to 3-way covariate interactions as the search space return similar highest ranked variable sets.

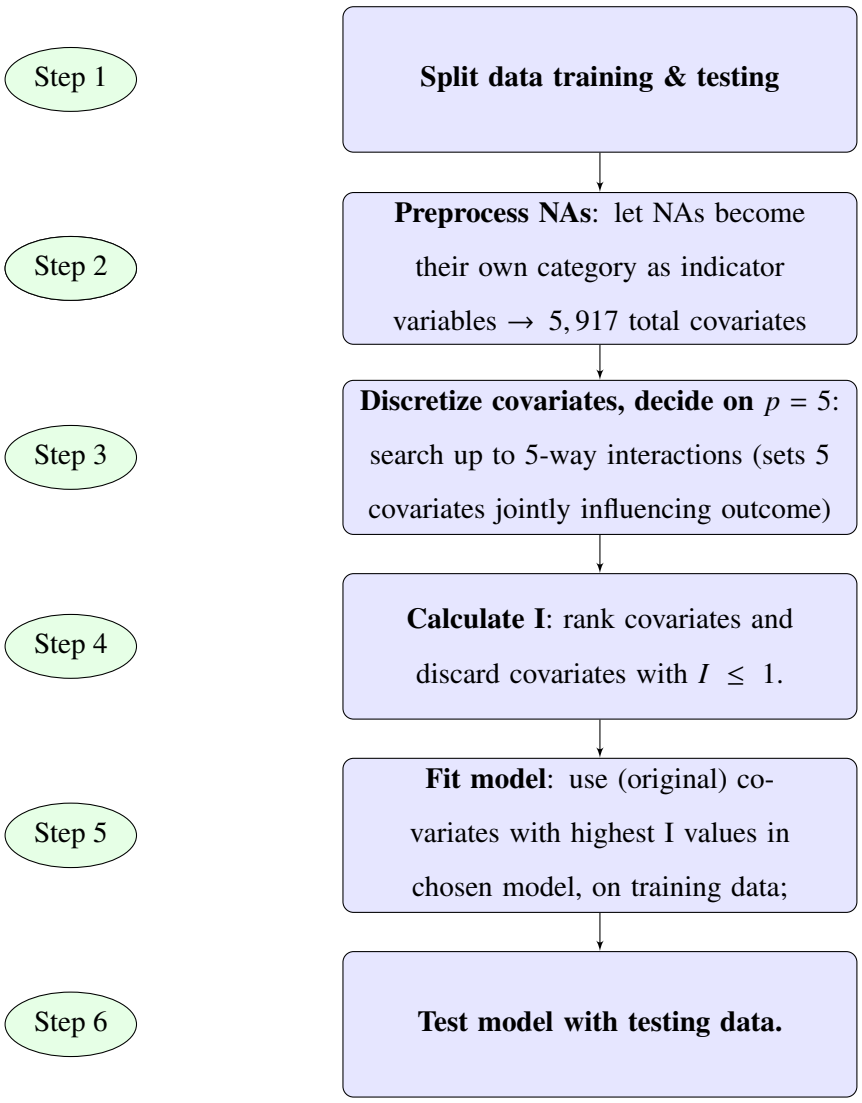

Figure 7: **Workflow for processing and conducting covariate screening of PITF data**

6.1 Results

We present first the covariates chosen for each theory-guided model and the results of covariates passing the screening stage using the I statistic in Table (2), represented as Steps 3 and 4 in Figure (7). Covariates are ranked by magnitude from highest to lowest. For approaches that use a prediction model based on the logistic regression (Bates, Goldstone et al.) these are absolute magnitudes of logistic coefficients. For neural nets (King & Zeng), this is a measure of variable relative importance based on the model weights as proposed by Garson (1991) and Goh (1995). Covariate sets selected by our screening approach are ranked by influence score.

From Table (2) it is evident that there is some overlap in types of covariates found to be important using screening and guided by state failure theory. For instance, discrimination is identified by Goldstone et al. (2010) and also identified as a highly important, predictive covariate by I statistic value. Likewise, whether a country neighbor

is experiencing conflict – thereby making a state more vulnerable to spillover effects among other destabilizing forces – is identified as being of key importance by Goldstone et al. (2010) as well as by *I* score. However, it is interesting to note that not only is a neighbor's experience important, so must that experience interact with a state's own history of instability. Given that this is a forecasting exercise, we must remain vigilant in our interpretation of *how* these covariates interact – we do not know, for instance, whether the causal arrow points from the one's own historical state instability to a neighbor's and then rebounding back, or whether perhaps a third force has affected both states at different time periods and then resulted in regional contagion.

We also find an interaction between problem country (as defined by members of the State Failure Task Force[8]) and whether there exists a large, politically significant group within the borders of the state. Indeed, many of the selected covariates relate to the existence of groups with cleavages along political, ethnic or linguistic dimensions. This shouldn't be overly surprising to scholars of political instability; following the work of Horowitz (1985) we expect that increased fractionalization of ethnic groups (large numbers of ethnic groups that each makes up a relatively small proportion of the overall population), especially more politically salient ones, can lead to a more politically contentious and unstable state. Noteable mechanisms through which ethnic diversity can lead to greater instability are diverse in the literature; these range from impediments to economic development (Easterly and Levine, 1997) to retaliation from exclusion from state power (Cederman and Weidmann, 2017).

Unsurprisingly, the predictive ability of the historical trajectory of instability for a country is substantial. What is perhaps surprising and an important discovery from the covariate screening stage is that this trajectory interacts with other influential factors, such as the history of conflict in a neighboring state.

With these covariates in hand, we turn to the prediction models, trained on our training data, and then tested for prediction error on our testing data (Step 5 in Figure (7)).

We compare the area under the receiver operating curve (AUC) for each of the models with covariates selected with guidance from theories of state failure, as well as our approach of data-driven screening, combined with a simple logistic regression (see Table (3)). The AUC is a commonly applied performance measure to prediction exercises (see Bradley (1997) for an explanation of some of its desirable properties), and is widely applied in forecasting in peace and conflict research (Hegre et al., 2017).

We note that of what we term the 'theoretically-driven' approaches, Bates' logistic regression has the largest AUC value (0.7493), when compared to Goldstone et al.'s logistic regression and King & Zeng's neural network. However this is still a smaller AUC value than that achieved by covariate-screening with the *I* statistic, followed by a logistic regression (0.929). Our approach is particularly comparable to Bates' and Goldstone et al.'s as the same model is applied to all three approaches. We explore a variety of machine learning models next.

[8]See Sources section of variable definition files associated with PITF data: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/RPQIODIANR`.

Covariate (set)	Bates	Goldstone et al.
1	Competitiveness of political participation	Economic or political discrimination
2	Legislative coalitions	Regime
3	Energy exports	Conflict neighbor
4	GDP	Infant mortality rate
Covariate (set)	**King & Zeng**	**I**
1	Legislative effectiveness	Ethnic group
2	Democracy	Linguistic group
3	Infant mortality rate	Politically significant group
4	Trade openness	Economic discrimination
5	Population density	Problem country *politically significant group
6	Autocracy	Religious group
7	Military population	Lag 1 yr state failure
8		Lag 2 yr state failure
9		Lag 2 yr state failure*conflict neighbor
10		Change in political discrimination
11		Conflict neighbor

Table 2: **Selected covariates across theory-driven approaches and covariate-screening with *I* statistic**: importance of covariates are ranked in order of magnitude.

Theoretically-driven approaches	AUC
Bates (Logistic)	0.7493
Goldstone et al. (Logistic)	0.5909
King & Zeng (Neural nets)	0.6596
I + Logistic	**0.929**

Table 3: **AUC comparisons, theory**

To determine whether screening can help the applied researcher interested in forecasting state instability, we compare direct estimation of three different machine learning approaches, neural networks, random forests and lasso, with and without an *I*-guided screening step immediately prior to estimating the model.AUC results are presented in Table (4).

Since we are exploring the space of interactions up to 5-way interactions (see Figure (7), Step 3), this results in over 2.3×10^{13} covariate sets to screen. While the *I* statistic can comfortably screen these[9] down to 11 covariate sets, this is computationally prohibitive for direct application to the ML techniques at hand. As such, we only explore up to 2-way interactions for the non-screening neural network, random forests and lasso; the last approach still suffers

[9]We conduct screening of all marginal and 2-way interactions. Above 2-way interactions we utilize the random draw and drop approach presented in Figure (1).

from over-parameterization and cannot produce a tuned model without even further dimension reduction. As such for "lasso (alone)" we do not have AUC results to provide. We find that covariate screening with the I statistic can range from marginally improving (in the case of random forests) to drastically increasing (in the case of neural networks and lasso) AUC values compared to direct application of the ML approach.

ML-driven approaches		AUC
Neural networks	(alone)	0.5996
	+ I	**0.928**
Random forests	(alone)	0.9293
	+ I	**0.9294**
Lasso	(alone)	N/A
	+ I	**0.9513**

Table 4: **AUC comparisons, machine learning approaches**

7 Discussion

In this article we demonstrate the importance of covariate selection for dimension reduction in high dimensional data and introduce a screening approach that takes into account the goal of prediction directly through a screening statistic that has a direct relationship with covariate predictivity. This is especially important now that high dimensional data is more prevalent in the social sciences; nearly all of text data is inherently high dimensional and availability of online and technological resources make data more likely to grow than shrink in the years to come. A key question for applied researchers interested in using this data to predict or forecast some political phenomenon of interest is how to select the important covariates, and remove noisy ones, to focus attention on the factors that are likely to do the most work in our prediction models and to make our data more manageable for off-the-shelf approaches, machine-learning or otherwise.

We discuss how theory-guided approaches to selecting covariates risk missing important covariates that are (as yet) untheorized, and can select covariates based on a different criterion (statistical significance) and arrive at different covariates than we would have under the criterion of predictivity. We further elaborate on how off-the-shelf machine learning approaches can still face computational challenges as the data grows in covariates faster than in observations. Dimension reduction or regularization are common approaches adopted for this precise problem. Our proposed approach can be classified as a filter-based one, and differs from other filter techniques in that the statistic with which we evaluate covariates and sets of covariates is shown to have a direct relationship to predictivity itself. In asymptotics, it is a lower bound to the correct prediction rate. We argue that this makes our approach more likely

to capture covariates important for prediction and demonstrate in a series of simulations how the influence statistic performs in various sample-constrained settings. We find that the influence statistic captures important covariates quite well, and that this performance is reasonably stable throughout different types of data distributions, sample-to-variable ratios, and models. When compared against sure independence screening, a filtering approach that also screens covariates, albeit through correlation-based magnitudes, the I screening approach always performs better.

One of the drawbacks to the proposed approach is that it requires discretization of covariates that are continuous, leading to some loss of information. Further research on extending the influence statistic to continuous settings is required and outside the scope of this current article. A promising direction would be to approach redefining the statistic through k-nearest neighbors constructions. To explore how "roughening" continuous covariates into discrete ones may affect capture rates of the correct predictive variables, we conduct a series of simulation experiments whereby covariates are drawn from continuous distributions as well as both continuous and discrete distributions and then discretized appropriately, prior to screening. We find that the I statistic is still able to achieve high rates of correct capture, especially when the underlying model is linear. It is also still able to outperform SIS, despite the latter approach not requiring discretization.

We explore whether and how screening can improve prediction in two applications; the first delves into the measurement of partisanship in the U.S. Congress from 1873-2017. We find that our phrase-screening results in qualitatively reasonable returned phrases, indicative of some of the most inter party dividing debates throughout history. We also find that partisanship has grown throughout the years, though not as dramatically as Gentzkow, Shapiro and Taddy (2016) find. Part of the reason is because we estimate higher measured values of partisanship in earlier sessions of Congress, thus making the rise in partisanship levels less steep. We argue that this is due to the more data-driven removal of noise phrases resulting in a smaller subset of relevant phrases to input into model estimation stage.

Our second exploration looks into a classic case of forecasting political instability – can screening for predictive covariates (and interactions of covariates) improve our out-of-sample correct prediction rates? Can we discover new and interesting (possibly nonlinear) interactions of covariates? Our screening returns covariate sets that contribute to better predictions. The types of covariate sets returned also suggest that the conflict histories of neighboring states may be important, as well within-country cleavages that result in large groups of people aligned on ethnic, religious or political dimensions. Furthermore, these factors may interact with one another and need to be accounted for in forecasting models. We find that screening can help but not harm prediction rates as measured by AUC.

Taken together, we suggest there are gains to be made in the area of high dimensional prediction with the simple incorporation of a covariate screening step. It is important that such a screening process utilize criteria that are adapted specifically to the goal of prediction, however. As such, we propose the influence statistic as a fitting candidate, based on its theoretical relationship to predictivity itself, and its ability to screen between noise and

influential covariates in sample-constrained settings. We imagine that other data contexts with naturally interactive covariates and possibly limited sample sizes may find similar applicability of such a screening approach. For instance, a popular experimental design, the conjoint experiment, inherently considers the effects of profiles of varying attributes, often randomized by the researcher, on some respondent outcome. While the experiment may be aptly powered to test whether specific attributes are likely to affect respondent outcomes, we easily envision difficulties of powering tests for combinations of attributes; a full exploration, without some prior knowledge of likely candidate combinations, would be prohibitively costly. As such, a screening of such combinations during a pilot would provide an extremely straight-forward capture of such a prior.

8 References

Bates, Robert H. 2008. "The logic of state failure: Learning from late-century Africa." *Conflict Management and Peace Science* 25(4):297–314.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *The American Political Science Review* 94(1):21–35.

Bernard, Henri and Stefan Gerlach. 1998. "Does the Term Structure Predict Recessions? The International Evidence." *International Journal of Finance and Economics* 3(3):195–215.

Bien, Jacob, Jonathan Taylor and Robert Tibshirani. 2013. "A lasso for hierarchical interactions." *Annals of Statistics* 41(3):1111–1141.

Bradley, Andrew P. 1997. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognition* 30(7):1145–1159.

Brandt, Patrick T., John R. Freeman and Philip A. Schrodt. 2014. "Evaluating forecasts of political conflict dynamics." *International Journal of Forecasting* 30(4):944–962.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45(1):5–32.

Cederman, Lars Erik and Nils B. Weidmann. 2017. "Predicting armed conflict: Time to adjust our expectations?" *Science* 355(6324):474–476.

Chernoff, Herman, Shaw-Hwa Lo and Tian Zheng. 2009. "Discovering influential variables: A method of partitions." *The Annals of Applied Statistics* 3(4):1335–1369.

Colaresi, Michael and Zuhaib Mahmood. 2017. "Do the robot: Lessons from machine learning to improve conflict forecasting." *Journal of Peace Research* 54(2):193–214.

Dominguez, Kathryn M E and Matthew D Shapiro. 2016. "Forecasting the Recovery from the Great Recession : Is This Time Different?" *American Economic Review* 103(3):147–152.

Easterly, William and Ross Levine. 1997. "Africa's Growth Tragedy : Policies and Ethnic Divisions." *The Quarterly Journal of Economics* 112(4):1203–1250.

Estrella, Arturo and Frederic S Mishkin. 2016. "Predicting U . S . Recessions : Financial Variables as Leading Indicators Author (s): Arturo Estrella and Frederic S . Mishkin Source : The Review of Economics and Statistics , Vol . 80 , No . 1 (Feb ., 1998), pp . 45-61 Published by : The MIT Press S." 80(1):45–61.

Esty, Daniel C., Jack A. Goldstone, Ted Robert Gurr, Barbara Harff, Marc Levy, Geoffrey D. Dabelko, Pamela T. Surko and Alan N. Unger. 1999. "State failure task force report: Phase II findings." (January).

Fan, Jianqing and Jinchi Lv. 2008. "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 70(5):849–911.

Fearon, JD and DD Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American political science review* 97(1):75–90.
URL: *http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=142718*

Flom, Peter L and David L Cassell. 2009. "NESUG 2009 Statistics & Analysis Stopping stepwise : Why stepwise and similar selection methods are bad , and what you should use." *Nesug* pp. 1–7.

Flynn, Cheryl J, Clifford M Hurvich and Jeffrey S Simonoff. 2017. "On the Sensitivity of the Lasso to the Number of Predictor Variables." *Statistical Science* 32(1):88–105.
URL: *https://projecteuclid.org/download/pdfview_1/euclid.ss/1491465629*

Garson, G.D. 1991. "Interpreting neural network connection weights." *Artificial Intelligence Expert* 6(4):46–51.

Gentzkow, Matthew, Jesse Shapiro and Matt Taddy. 2016. "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." pp. 1–46.
URL: *http://www.nber.org/papers/w22423.pdf*

Goh, A.T.C. 1995. "Back-propagation neural networks for modeling complex systems." *Artificial Intelligence in Engineering* 9(3):143–151.

Goldsmith, Benjamin E, Charles R Butcher, Dimitri Semenovich and Arcot Sowmya. 2013. "Forecasting the onset of genocide and politicide: Annual out-of-sample foreasts on a global dataset, 1988-2003." *Journal of Peace Research* 50(4):437–452.

Goldstone, Jack A., Robert Bates, David Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder and Mark Woodward. 2010. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* 54(1):190–208.
URL: *http://papers.ssrn.com/abstract=1531942*

Goldstone, Jack a, Ted Robert Gurr, Barbara Harff, Marc a Levy, Monty G Marshall, Robert H Bates, David L Epstein, Colin H Kahl, Pamela T Surko, John C Ulfelder and Alan N Unger. 2000. "State Failure Task Force Report : Phase III Findings." *Science* .

Gransbo, K, P Almgren, M Sjogren, J G Smith, G Engstrom, B Hedblad and O Melander. 2013. "Chromosome 9p21 genetic variation explains 13% of cardiovascular disease incidence but does not improve risk prediction." *Journal of internal medicine* 274(3):233–40.

Han, Hahrie and David W. Brady. 2007. "A delayed return to historical norms: Congressional party polarization after the second World War." *British Journal of Political Science* 37(3):505–531.

Hegre, Havard, Nils W. Metternich, Havard Mokleiv Nygard and Julian Wucherpfennig. 2017. "Introduction: Forecasting in peace research." *Journal of Peace Research1* 54(2):113–124.

Horowitz, Donald L C N B Canaday GN496 .H67 1985 D U E 10-27-09 S McCabeHnrs 4wk Pols 111 Swarthmore c.2 AVAILABLE H Magill GN496 .H67 1985 D U E 02-15-10. 1985. *Ethnic groups in conflict*. Berkeley and Los Angeles: University of California Press.

Jakobsdottir, Johanna, Michael B Gorin, Yvette P Conley, Robert E Ferrell and Daniel E Weeks. 2009. "Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers." *PLoS genetics* 5(2):e1000337.

Kaczorowski, Robert J. 1987. "To Begin the Nation Anew: Congress, Citizenship, and Civil Rights after the Civil War." *The American Historical Review* 92(1):45–68.

King, Gary and Langche Zeng. 2001. "Improving Forecasts of State Failure." *World Politics* 53(04):623–658.
URL: *http://www.journals.cambridge.org/abstract_S0043887100019171*

Lewis-Beck, Michael S; Tien, Charles. 2012. "Election Forecasting for Turbulent Times." *Political Science & Politics* 45(October):625 – 629.

Lo, Adeline, Herman Chernoff, Tian Zheng and Shaw-Hwa Lo. 2015. "Why significant variables aren't automatically good predictors." *Proceedings of the National Academy of Sciences* 112(45):13892–13897.

Lo, Adeline, Herman Chernoff, Tian Zheng and Shaw-Hwa Lo. 2016. "Framework for making better predictions by directly estimating variables' predictivity." *Proceedings of the National Academy of Sciences* 113(50):14277–14282.
URL: *http://www.pnas.org/lookup/doi/10.1073/pnas.1616647113*

Lockerbie, Brad. 2008. "Election forecasting: The future of the presidency and the house." *PS - Political Science and Politics* 41(4):713–716.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-assisted text analysis for comparative politics." *Political Analysis* 23(2):254–277.

Muchlinski, David, David Siroky and Matthew Kocher. 2015. "Comparing Random Forest with Logistic Regression for Predicting Class-imbalanced Civil War Onset Data." *Political Analysis* 24:87–103.

Nyberg, Henri. 2010. "Dynamic probit models and financial variables in recession forecasting." *Journal of Forecasting* 29(1-2):215–230.

Olden, Julian D. and Donald A. Jackson. 2002. "Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks." *Ecological Modelling* 154(1-2):135–150.

Oliphant, Baxter. 2017. Bipartisan support for some gun proposals, stark partisan divisions on many others. Technical report Pew Research Center.
URL: *http://www.pewresearch.org/fact-tank/2017/06/23/bipartisan-support-for-some-gun-proposals-stark-partisan-divisions-on-many-others/*

Rost, Nicolas, Gerald Schneider and Johannes Kleibl. 2009. "A global risk assessment model for civil wars." *Social Science Research* 38(4):921–933.
URL: *http://dx.doi.org/10.1016/j.ssresearch.2009.06.007*

Saldana, Diego Franco and Yang Feng. 2018. "SIS : An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models." *Journal of Statistical Software* 83(2).
URL: *http://www.jstatsoft.org/v83/i02/*

Stock, James H. and Mark W Watson. 2002. "Macroeconomic forecasting using diffusion indexes." *Journal of Business & Economic Statistics* 20(2):147–162.

Taylor, Jonathan and Robert J. Tibshirani. 2015. "Statistical learning and selective inference." *Proceedings of the National Academy of Sciences* 112(25):201507583.
URL: *http://www.pnas.org/content/112/25/7629.abstract?ijkey=a2fba6551c0765bd9cbb2465e1687e238807bea8&keytype2=tf_ip*

Tomandl, Dirk and Andreas Schober. 2001. "A Modified General Regression Neural Network (MGRNN) with new, efficient training algorithms as a robust 'black box'-tool for data analysis." *Neural Networks* 14(8):1023–1034.

Wang, Hansheng. 2009. "Forward regression for Ultra-High dimensional variable screening." *Journal of the American Statistical Association* 104(488):1512–1524.

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

Welch, Ivo and Amit Goyal. 2008. "A comprehensive look at the empirical performance of equity premium prediction." *Review of Financial Studies* 21(4):1455–1508.

Wilson, Robert E, Samuel D Gosling and Lindsay T Graham. 2012. "A Review of Facebook Research in the Social Sciences." *Perspectives on Psychological Science* 7(3):203–220.

Appendix

Naïve estimator for $\theta_c(\mathbf{X})$

We illustrate the sample analog's (naïve estimator) tendency to prefer adding any and all covariates to a set of covariates by always increasing with the addition of covariates.

Suppose $\mathbf{X}_m = \{X_1, ..., X_m\}$ and $\mathbf{X}_{m+1} = \{X_1, ..., X_m, X_{m+1}\}$. The partition formed by $\mathbf{X}_m$ is

$$\Pi_{\mathbf{X}_m} = \{A_1, ..., A_{m_1}\},$$

while the partition formed by $\mathbf{X}_{m+1}$ is

$$\begin{aligned} \Pi_{\mathbf{X}_{m+1}} &= \{A_1 \cap B,\ ...,\ A_{m_1} \cap B,\ A_1 \cap B^c,\ ...,\ A_{m_1} \cap B^c\} \\ &= \{\Pi_{\mathbf{X}_m} \cap B, \Pi_{\mathbf{X}_m} \cap B^c\} \end{aligned}$$

where $B = \{\mathbf{X}_{m+1} = 1\}$. Let

$$\begin{aligned} \Pi^1_{\mathbf{X}_m} &= \Pi_{\mathbf{X}_m} \cap \{X_{m+1} = 1\} \\ \text{and } \Pi^0_{\mathbf{X}_m} &= \Pi_{\mathbf{X}_m} \cap \{\mathbf{X}_{m+1} = 0\}, \end{aligned}$$

where $\Pi^1_{\mathbf{X}_m}$ and $\Pi^0_{\mathbf{X}_m}$ form two subpartitions of $\Pi_{\mathbf{X}_{m+1}}$, i.e. , $\Pi_{\mathbf{X}_{m+1}} = \Pi^0_{\mathbf{X}_m} \cup \Pi^1_{\mathbf{X}_m}$. Then

$$\begin{aligned} |\hat{p}_{\Pi_{\mathbf{X}_m}}(d) - \hat{p}_{\Pi_{\mathbf{X}_m}}(u)| &\leq |\hat{p}_{\Pi^0_{\mathbf{X}_m}}(d) - \hat{p}_{\Pi^0_{\mathbf{X}_m}}(u)| \\ &\quad + |\hat{p}_{\Pi^1_{\mathbf{X}_m}}(d) - \hat{p}_{\Pi^1_{\mathbf{X}_m}}(u)|, \end{aligned}$$

where $\hat{p}(\cdot)$ is the sample estimator. We see that the sample analog, much like the R^2 statistic, inherently favors an increase in number of partition cells, i.e. adding more variables.

Properties of I

Proof Theorem 1

Theorem 1 follows directly from Lo et al. (2016). The proof requires showing that the I statistic asymptotically approaches a constant multiple of θ_I.

$I_{\Pi_{\mathbf{X}}}$ can be expressed as a weighted average of chi-squares under the null hypothesis of no association between

$\mathbf{X} = X_k, k = 1, ..., m$ and Y. That is, $I_{\Pi_\mathbf{X}}$ can be expressed as $\sum_{j=1}^{J} \lambda_j \chi_j^2$, where $\lambda_j \in (0, 1)$ and $\sum_{j=1}^{J} \lambda_j = 1 - \sum_{j=1}^{J} p_j^2$ where p_j is the probability of cell j.

For notational simplicity, let $n_{Y=1}$ be n_1 and $n_{Y=0}$ be n_0. Thus $n = n_1 + n_0$. We can then express $ns_n^2 I_{\Pi_\mathbf{X}}$ as:

$$\begin{aligned} ns_n^2 I_{\Pi_\mathbf{X}} &= \sum_{j \in \Pi_\mathbf{X}} n_j^2 (\bar{Y}_j - \bar{Y})^2 \\ &= \sum_{j \in \Pi_\mathbf{X}} \Big(n_{0,j} + n_{1,j}\Big)^2 \cdot \Big(\frac{n_{1,j}}{n_{0,j} + n_{1,j}} - \frac{n_1}{n_0 + n_1}\Big)^2 \\ &= \Big(\frac{n_0 n_1}{n_0 + n_1}\Big)^2 \cdot \sum_{j \in \Pi_\mathbf{X}} \Big(\frac{n_{1,j}}{n_1} - \frac{n_{1,j}}{n_1}\Big)^2 \end{aligned}$$

where $n_{1,j}$ and $n_{0,j}$ denote the number of observations where $Y = 1$ and number of observations where $Y = 0$ in cell j. We can decompose $ns_n^2 I_{\Pi_\mathbf{X}}$ into:

$$ns_n^2 I_{\Pi_\mathbf{X}} = \sum_{j \in \Pi_\mathbf{X}} n_j^2 (\bar{Y})_j - \bar{Y})^2 = A_n + B_n + C_n$$

where:

$$\begin{aligned} A_n &= \sum_{j \in \Pi_\mathbf{X}} n_j^2 (\bar{Y}_j - \mu_j)^2 \\ B_n &= \sum_{j \in \Pi_\mathbf{X}} n_j^2 (\bar{Y} - \mu_j)^2 \\ C_n &= \sum_{j \in \Pi_\mathbf{X}} -2n_j^2 (\bar{Y}_j - \mu_j)(\bar{Y} - \mu_j) \end{aligned}$$

$E(\bar{Y}_j) = \mu_j$ and $E(\bar{Y}) = \mu = \frac{n_1}{n_0 + n_1}$ for fixed n. A_n and C_n converge in probability to 0 as $n \to \infty$ when divided by the term n^2. What remains is simply B_n. Note:

$$\lim_n \frac{B_n}{n^2} =_{\text{prob}} \lim_n \sum_{j \in \Pi_\mathbf{X}} \Big(\frac{n_j^2}{n^2}\Big)(\mu_j - \mu)^2$$

Since our outcome is binary, we have:

$$\begin{aligned} \mu_j &= \frac{n_1 \mathrm{P}(j|Y = 1)}{n_0 \mathrm{P}(j|Y = 0) + n_1 \mathrm{P}(j|Y = 1)} \\ \mu &= \frac{n_1}{n_0 + n_1} \end{aligned}$$

And for every j, $\frac{n_j}{n}$ converges in probability to $p_j = \lambda \mathrm{P}(j|Y=1) + (1-\lambda)\mathrm{P}(j|Y=0)$ as $n \to \infty$, if $\lim_n = \frac{n_1}{n} = \lambda$, a fixed constant between 0 and 1, it follows that

$$\begin{aligned}
\frac{B_n}{n^2} &= \sum_{j \in \Pi_{\mathbf{X}}} \Big(\frac{n_j^2}{n^2}\Big)(\mu_j - \mu)^2 \\
&\to_{\text{prob}} \sum_{j \in \Pi_{\mathbf{X}}} p_j^2 \Big(\frac{\lambda \mathrm{P}(j|Y=1)}{\lambda \mathrm{P}(j|Y=1) + (1-\lambda)\mathrm{P}(j|Y=0)} - \lambda \Big)^2 \text{ as } n \to \infty \\
&= \sum_{j \in \Pi_{\mathbf{X}}} \Big[\lambda \mathrm{P}(j|Y=1) - \lambda [\mathrm{P}(j|Y=1) + (1-\lambda)\mathrm{P}(j|Y=0)] \Big]^2 \\
&= \sum_{j \in \Pi_{\mathbf{X}}} \Big[\lambda(1-\lambda) \mathrm{P}(j|Y=1) - [\lambda(1-\lambda)\mathrm{P}(j|Y=0)] \Big]^2 \\
&= \lambda^2 (1-\lambda)^2 \sum_{j \in \Pi_{\mathbf{X}}} [\mathrm{P}(j|Y=1) - \mathrm{P}(j|Y=0)]^2
\end{aligned}$$

where the last equation is a constant multiple of $\sum_{j \in \Pi_{\mathbf{X}}} [\mathrm{P}(j|Y=1) - \mathrm{P}(j|Y=0)]^2$.

Proof Theorem 2

Theorem 2 follows directly from Lo et al. (2016). Under Assumptions 1-4, we can produce an asymptotic lower bound for the correct prediction rate of covariate set$\mathbf{X}$:

Starting from Equation (7),

$$\begin{aligned}
\theta_c(\mathbf{X}) &= \frac{1}{2} + \frac{1}{4} \sum_{j \in \Pi_{\mathbf{X}}} |\mathrm{P}(j|Y=1) - \mathrm{P}(j|Y=0)| \\
&\leq \frac{1}{2} + \frac{1}{4} \sqrt{2 \sum_{j \in \Pi_{\mathbf{X}}} (\mathrm{P}(j|Y=1) - \mathrm{P}(j|Y=0))^2} \\
&= \frac{1}{2} + \frac{1}{4} \sqrt{2\theta_I(\Pi_{\mathbf{X}})} \\
&=_{\text{prob}} \frac{1}{2} + \frac{1}{4} \sqrt{2 \lim_{n \to \infty} \frac{s_n^2 I_{\Pi_{\mathbf{X}}}}{n\lambda^2(1-\lambda)^2}} \\
&=_{\text{prob}} \frac{1}{2} + \frac{1}{4} \sqrt{2 \lim_{n \to \infty} \frac{I_{\Pi_{\mathbf{X}}}}{n\lambda(1-\lambda)}}
\end{aligned}$$

To arrive at line 2 from line 1, we use the fact the following (see Lemma 1 from Lo et al. 2016 for proof):

For K real values $\{z_j; 1 \leq i \leq K\}$, $\sum_{j=1}^K z_j = a$ and $\sum_{j=1}^K |z_j| = b$ we have: $\sum_{j=1}^K z_j^2 \leq \frac{a^2+b^2}{2}$.

To arrive at line 4 from line 3, we use Theorem 1.

Generalization to arbitrary priors

In the main text, we proceed with equal priors for $Y = 1$ and $Y = 0$; this may be an inappropriate choice for many social science applications, especially ones that feature rare events, such as state failure. Furthermore, we may not

always be able to acquire prior information or estimate it with empirical data. As such, we present extensions of the correct prediction rate θ_c and the modified I statistic under arbitrary priors (see Lo et al. (2016) for original presentation):

Assuming an arbitrary prior $\pi(Y = 1)$ and $\frac{n_1}{n} \to \lambda$ as $n \to \infty$, the correct prediction rate for covariate set $\mathbf{X}$ $\theta_c(\mathbf{X})$ is:

$$\theta_c(\mathbf{X}) = \frac{1}{2} + \frac{1}{2} \sum_{j \in \Pi_{\mathbf{X}}} |\mathrm{P}(j|Y = 1)\pi(Y = 1) - \mathrm{P}(j|Y = 0)\pi(Y = 0)|$$

Let the modified score $I^*_{\Pi_n}$ be defined as

$$ns_n^2 I^*_{\Pi_n} = \frac{1}{4} \sum_{j \in \Pi_{\mathbf{X}}} n_j^2 \Big[\bar{y}_j (\frac{\pi(Y = 1)}{\lambda}) - (1 - \bar{y}_j)(\frac{\pi(Y = 0)}{1 - \lambda}) \Big]^2.$$

Then we have:

$$\lim_{n \to \infty} \frac{s_n^2 I^*_{\Pi_n}}{n} =_p \frac{1}{4} \sum_{j \in \Pi_{\mathbf{X}}} \big[P(j|Y = 1)\pi(Y = 1) - P(j|Y = 0)\pi(Y = 0) \big]^2. \tag{10}$$

Similar lower bounds can then be derived as:

$$\theta^*_c(\mathbf{X}) = \frac{1}{2} + \frac{1}{2} \sum_{j \in \Pi_{\mathbf{X}}} \Big| P(j|Y = 1)\pi(Y = 1) - P(j|Y = 0)\pi(Y = 0) \Big| \tag{11}$$

$$\geq \frac{1}{2} + \frac{1}{2} \sqrt{\lim_{n \to \infty} \frac{\lambda(1 - \lambda) I^*_{\Pi_n}}{2n} - a^2} \tag{12}$$

where $a = \sum_{j \in \Pi_{\mathbf{X}}} (P(j|Y = 1)\pi(Y = 1) - P(j|Y = 0)\pi(Y = 0)) = \pi(Y = 1) - \pi(Y = 0)$.

Similar to the situation under equal priors, Equation (11) is a direct consequence of Equation (10).

Generalization to different loss and cost functions

We have utilized a zero-one loss thus far, treating false negatives and false positives equally. In applying to specific applications, the researcher may find other loss functions more fitting; for instance, incorrectly predicting one country facing state failure may be deemed less serious a mistake than mistakenly allaying the fears of another in avoiding political disorder. As such, we generalize the correct prediction rate θ_c and the I statistic to different loss functions.

We define the loss function L as:

$$\mathrm{L}(Y = 1, Y = 0) = l_{Y=1},\ \mathrm{L}(Y = 0, Y = 1) = l_{Y=0} \tag{13}$$

and

$$\mathrm{L}(Y = 1, Y = 1) = L(Y = 0, Y = 0) = 0 \tag{14}$$

where $l_{Y=1}$ and $l_{Y=0}$ are the losses incurred for misclassifying a $Y = 1$ observation as a $Y = 0$ and vice versa.

The optimum Bayes' solution is found by minimizing the expected predicted loss, given its j value. We simply assign a test sample with partition (predictor) j to $Y = 1$ if:

$$P(j|Y = 1)\pi(Y = 1)\mathrm{L}(Y = 1, Y = 0) < P(j|Y = 0)\pi(Y = 0)\mathrm{L}(Y = 0, Y = 1)$$

otherwise, assign to $Y = 0$. Equivalently, choose $Y = 1$ if

$$\mathrm{P}(j|Y = 1)\pi(Y = 1)l_{Y=1} < \mathrm{P}(j|Y = 0)\pi(Y = 0)l_{Y=0}$$

otherwise $Y = 0$. The expected loss of adopting this rule is:

$$e^l = \frac{1}{2}\sum_{j\in\Pi_{\mathbf{X}}} \min\{a_j, b_j\},$$

where $a_j = \mathrm{P}(j|Y = 1)\pi(Y = 1)l_{Y=1}$ and $b_j = \mathrm{P}(j|Y = 0)\pi(Y = 0)l_{Y=0}$. The random rule of classifying an observation to $Y = 0$ or $Y = 1$ has the expected loss:

$$\gamma = \frac{1}{2}\sum(a_j + b_j) = \frac{1}{2}\left(\pi(Y = 1)l_{Y=1} + \pi(Y = 0)l_{Y=0}\right),$$

a constant independent of the partition Π_x. The "gain" in θ_c^l (interpreted as less the expected loss of Bayes' rule) can be defined as:

$$\theta_c^l(\mathbf{X}) = \frac{1}{2}\sum_{j\in\Pi_{\mathbf{X}}} \max\{a_j, b_j\} = \frac{1}{2}\sum_{j\in\Pi_{\mathbf{X}}}(a_j + b_j) - e^l = \gamma - e^l.$$

Since γ is independent of $\mathbf{X}$ and $\Pi_{\mathbf{X}}$, it is desirable to search for $\mathbf{X}$ with larger θ_c^l to achieve better "gains". Again we have

$$\begin{aligned}\theta_c^l(\mathbf{X}) &= \frac{\gamma}{2} + \frac{\theta_c^l - e^l}{2} \\ &= \frac{\gamma}{2} + \frac{1}{4}\sum_{j\in\Pi_{\mathbf{X}}} \left|a_j - b_j\right|\end{aligned}$$

After standardizing by γ, we obtain the improved prediction rate as:

$$\begin{aligned}\theta_c(\mathbf{X}) &= \frac{\theta_c^l}{\gamma} \\ &= \frac{1}{2} + \frac{1}{4\gamma}\sum_{j\in\Pi_{\mathbf{X}}} \left|a_j - b_j\right|\end{aligned}$$

Collecting the above discussion together, let the cost-based I-score $I^c_{\Pi_\mathbf{X}}$ be defined as:

$$
\begin{aligned}
n s_n^2 I^c_{\Pi_\mathbf{X}} &= \frac{1}{4\gamma} \sum_{j \in \Pi_\mathbf{X}} n_j^2 \left[\bar{y}_j \left(\frac{\pi(Y=1)}{\lambda} \right) l_{Y=1} - (1-\bar{y}_j) \left(\frac{\pi(Y=0)}{1-\lambda} \right) l_{Y=0} \right]^2 \\
&\approx \frac{n^2}{4\gamma} \sum_{j \in \Pi_\mathbf{X}} \left[\mathrm{P}(j|Y=1)\pi(Y=1) l_{Y=1} - \mathrm{P}(j|Y=0)\pi(Y=0) l_{Y=0} \right]^2 . \qquad (15)
\end{aligned}
$$

Expectation of influence statistic

$$
I(\mathbf{X}) = 0.5 n^{-1} \sum_{j \in \Pi_\mathbf{X}} (n_{1j} - n_{0j})^2
$$

$$
\begin{aligned}
\mathrm{E}(I(\mathbf{X})/n) &= \frac{1}{2} \sum_{j \in \Pi_\mathbf{X}} \left[(f_1(j) - f_0(j))^2 + \frac{1}{n} f_1(j)(1 - f_1(j)) + \frac{1}{n} f_0(j)(1 - f_0(j)) \right] \\
&\approx 0.5 \sum_{j \in \Pi_\mathbf{X}} (f_1(j) - f_0(j))^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}((I(\mathbf{X})/n)^2) = \frac{1}{4} \Bigg\{ & \frac{(n-1)^2}{n^2} \sum_{j,j'} (f_1(j) - f_0(j))^2 (f_1(j') - f_0(j'))^2 \\
& + \frac{4}{n} \sum_j (f_1(j) - f_0(j))^2 \\
& + \frac{4}{n} \sum_j (f_1(j) + f_0(j))(f_1(j) - f_0(j))^2 \\
& + \frac{4}{n} \left[\sum_x f_1(j)(f_0(j) - f_1(j)) \right] \left[\sum_j f_1^2(j) \right] \\
& + \frac{4}{n} \left[\sum_j f_0(j)(f_1(j) - f_0(j)) \right] \left[\sum_j f_0^2(j) \right] + O(\frac{1}{n^2}) \Bigg\}
\end{aligned}
$$

Variance of influence statistic

$$
\begin{aligned}
\mathrm{var}(I(\mathbf{X})/\mathbf{n}) = \mathbf{\frac{1}{4}} \Bigg\{ & \frac{4}{n} \sum_j (f_1(j) - f_0(j))^2 \\
& - \frac{2}{n} \sum_j (f_1(j)(1 - f_0(j)) \sum_j (f_1(j) - f_0(j))^2
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{2}{n} \sum_j (f_0(j)(1 - f_1(j)) \sum_j (f_1(j) - f_0(j))^2 \\
&+ \frac{4}{n} \sum_j (f_1(j) + f_0(j))(f_1(j) - f_0(j))^2 \\
&+ \frac{4}{n} \left[\sum_j f_1(j)(f_0(j) - f_1(j)) \right] \left[\sum_j f_1^2(j) \right] \\
&+ \frac{4}{n} \left[\sum_j f_0(j)(f_1(j) - f_0(j)) \right] \left[\sum_j f_0^2(j) \right] + O(\frac{1}{n^2}) \Bigg\} \\
= \frac{1}{n} \Bigg\{ &\left[\sum_j f_0(j) f_1(j) \right] \sum_j (f_1(j) - f_0(j))^2 \\
&+ \sum_j (f_1(j) + f_0(j))(f_1(j) - f_0(j))^2 \\
&+ \left[\sum_j f_1(j) f_0(j) \right] \left[\sum_j f_1^2(j) + \sum_j f_0^2(j) \right] - \left[\sum_j f_1^2(j) \right]^2 - \left[\sum_j f_0^2(j) \right]^2 \Bigg\} \\
&+ O(\frac{1}{n^2}) \\
= \frac{1}{n} \Bigg\{ &2 \left[\sum_j f_0(j) f_1(j) \right] \sum_j (f_1(j) - f_0(j))^2 \\
&+ \sum_j (f_1(j) + f_0(j))(f_1(j) - f_0(j))^2 \\
&+ 2 \left[\sum_j f_1(j) f_0(j) \right]^2 - \left[\sum_j f_1^2(j) \right]^2 - \left[\sum_j f_0^2(j) \right]^2 \Bigg\} \\
&+ O(\frac{1}{n^2})
\end{aligned}
$$

Simulation details

We provide further details on the three sets of simulations presented in this article. As noted before, we vary several parameters throughout the simulations. Observations n are drawn from the set $\{200, 400, 1600\}$. Covariates k are drawn from the set $\{10, 100, 200\}$. We generate covariates either from Bernoulli, Gaussian (normal), or mixture of two distributions. Possible outcome models are linear "LN", nonlinear "NL", linear with joint effects "LNJ", and nonlinear with joint effects.

In screening, covariates are only discretized for I screening, not for SIS. Covariates are screened for all marginal and all 2-way interactions for I. To calculate the main statistic of interest, the proportion of true influential covariates captured, we use the following: 1) for marginal effects only cases: all covariates captured by the screening approach, divided by true list of influential covariates in the model. 2) for marginal + joint or joint-only cases: all

covariate *sets*, accurately captured, divided by the true list of influential covariate sets in the model.

The data generating processes for the models are provided below:

1. Linear model (Simulations 1 and 3): proportion of important covariates out of total are (.3, .03, .015)

$$p = \frac{1}{1 + \exp(-\beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3)}$$
$$Y \sim \text{Bern}(p)$$

2. Nonlinear model (Simulations 1 and 3): proportion of important covariates out of total are (.3, .03, .015)

$$p = \frac{1}{1 + \exp(0.3 - sin(X_1 \cdot X_2) + X_3^2}$$
$$Y \sim \text{Bern}(p)$$

3. Linear model with joint only (Simulations 2 and 3): proportion of important covariates out of total are (.4, .04, .02)

$$p = \frac{1}{1 + \exp(-0.2 - \beta_1 X_1 \cdot X_2 + X_3 \cdot X_4}$$
$$Y \sim \text{Bern}(p)$$

4. Nonlinear model with joint only (Simulations 2 and 3): proportion of important covariates out of total are (.4, .04, .02)

$$p = \frac{1}{1 + \exp(0.1 - \sin(X_1 \cdot X_2) + \cos(X_3 \cdot X_4))}$$
$$Y \sim \text{Bern}(p)$$