Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results

Beau Coker, Cynthia Rudin, Gary King

To cite this article:
Beau Coker, Cynthia Rudin, Gary King (2021) A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results. Management Science

Published online in Articles in Advance 25 Mar 2021

. https://doi.org/10.1287/mnsc.2020.3818 http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
*Articles in Advance*, pp. 1–24
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

# A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results

**Beau Coker,[a] Cynthia Rudin,[b] Gary King[c]**

[a] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115; [b] Department of Computer Science and Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina 27708; [c] Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138

**Contact:** beaucoker@g.harvard.edu, https://orcid.org/0000-0003-3811-5674 (BC); cynthia@cs.duke.edu, https://orcid.org/0000-0003-4283-2780 (CR); king@harvard.edu, https://orcid.org/0000-0002-5327-7631 (GK)

**Received:** April 22, 2018
**Revised:** September 9, 2019; April 23, 2020
**Accepted:** July 6, 2020
**Published Online in Articles in Advance:** March 25, 2021

**https://doi.org/10.1287/mnsc.2020.3818**

**Abstract.** Inference is the process of using facts we know to learn about facts we do not know. A theory of inference gives assumptions necessary to get from the former to the latter, along with a definition for and summary of the resulting uncertainty. Any one theory of inference is neither right nor wrong but merely an axiom that may or may not be useful. Each of the many diverse theories of inference can be valuable for certain applications. However, no existing theory of inference addresses the tendency to choose, from the range of plausible data analysis specifications consistent with prior evidence, those that inadvertently favor one's own hypotheses. Because the biases from these choices are a growing concern across scientific fields, and in a sense the reason the scientific community was invented in the first place, we introduce a new theory of inference designed to address this critical problem. We introduce *hacking intervals*, which are the range of a summary statistic one may obtain given a class of possible endogenous manipulations of the data. Hacking intervals require no appeal to hypothetical data sets drawn from imaginary superpopulations. A scientific result with a small hacking interval is more robust to researcher manipulation than one with a larger interval and is often easier to interpret than a classical confidence interval. Some versions of hacking intervals turn out to be equivalent to classical confidence intervals, which means they may also provide a more intuitive and potentially more useful interpretation of classical confidence intervals.

**History:** Accepted by J. George Shanthikumar, big data analytics.
**Supplemental Material:** The data files and online appendix are available at https://doi.org/10.1287/mnsc.2020.3818.

**Keywords:** robustness • replicability • observational data • model dependence • causal inference • matching

## 1. Introduction

The numerous choices in even "best practice" data analysis procedures lead to high levels of unmeasured and unreported uncertainty in research publications. These choices include, among others, variable selection and transformations, data subsetting, identification and elimination of outliers, functional forms, distributional assumptions, priors, estimators, nonparametric preprocessing (such as matching), and procedures to control for unmeasured confounders (such as difference in differences or instrumental variables). See Wicherts et al. (2016) for an attempt to enumerate a complete list. Classical statistical inference conditions on whichever choices the analyst makes and focuses on uncertainty induced by observing only one possible sample of data. This is uncertainty *across hypothetical data sets*, where one is observed and the rest might have come from an imagined superpopulation. However, within the single observed data set, the often considerable variation *across potential "plausible" analysis choices* can lead to a wide range of empirical estimates, a range that is often considerably larger than the uncertainty induced by hypothetical sampling.

We thus propose that researchers (and readers) ask a simple question that gets to the heart of whether a quantitative conclusion can be trusted: "Would another honest researcher, choosing different but still reasonable analysis techniques, come to a different conclusion?" The best way to answer this question is the very process of science, where numerous researchers work in cooperation and competition in pursuit of a common goal. If one researcher publishes a result that can be questioned by another, a healthy scientific community will ensure that will happen, and together with others, they will be more likely to find the right answer. But what happens in the interim when we write or read a paper today? How do we increase the likelihood that the conclusions in this paper could not be overturned by minor changes in the analysis methods that another reasonable researcher might choose? We offer a quantitative framework for answering these questions.

We use the term *hacking* to describe an earnest researcher working hard to choose appropriately among many data analysis choices. Although this term is sometimes used to describe dishonest manipulation of results, we use it solely (in the positive sense of a "hackathon") to refer to honest scientists genuinely trying to get the right answer by making analysis choices among many reasonable alternatives. For a given model class and loss function, a *hacking interval* is the smallest and largest value of a summary statistic (e.g., a coefficient in a regression, first difference, risk ratio, or other quantity of interest) that can be achieved over a set of constraints for which the researcher, readers, and the scientific community would like robustness. It quantifies the extent to which a different, also reasonable, analyst could come to different conclusions. Researchers who report hacking intervals are being more transparent about the evidence available to support their hypotheses. Hacking intervals are designed to reveal information that any research publication should provide to make it less likely to mislead researchers and readers of their work. A major benefit of hacking intervals is that they are easy to understand, interpret, and teach: we think much easier than introducing hypothetical draws from imaginary superpopulations. They can be taught alongside, before, or even without reference to classical confidence intervals or any other theory of inference. They do not require knowledge of probability.

The hacking intervals we propose come in two varieties. *Prescriptively constrained* hacking intervals allow for an explicit definition of the analysis choices reasonable researchers make, and they identify the range of a summary statistic over these choices. They are useful when one can limit which analysis choices are valid. The second type, *tethered hacking intervals*, avoids the explicit enumeration of analysis choices and requires only that the predictive model chosen by the researcher has a small-enough loss on the observed data. Each type of hacking interval is a consequence of the defined set of researcher constraints. In a maximum likelihood scenario, tethered hacking intervals are mathematically equivalent to profile likelihood confidence intervals (as shown in Online Appendix C.1). Our work therefore provides a new interpretation of profile likelihood confidence intervals that requires no understanding of probability.

Quantifying the potential impact of hacking is especially—but not only—important if researchers are (inadvertently) biased toward a favored hypothesis. This is crucial because standard data analysis procedures leave researchers in a situation that meets all the conditions social psychologists have identified that lead to biased choices. In the presence of high levels of discretion, many analysis choices, little objective way to know which is best, and access to the estimates that each choice results in, even honest, hardworking, earnest researchers are likely to inadvertently bias results toward their favorite hypotheses (Gilbert 1998, Kahneman 2011, Banaji and Greenwald 2013). If a researcher (or reader) is concerned that analysis choices were only chosen because they yielded results consistent with the bias of the researcher, a hacking interval informs them of the degree to which this can matter. A small hacking interval says that *any* researcher making choices within our defined constraints, whether biased toward a conclusion or not, could only have a small impact on the result. Hacking intervals, defined via specific norms such as the ones we suggest here, are a natural solution for conveying the impact analysis choices can have for any one publication, without the costly, time-consuming, and sometimes dubious or tendentious process of ad hoc sensitivity testing designed anew for each article. Hacking intervals characterize the space of analysis choices systematically with precise computational and mathematical tools. This process can also provide insight into the state of researcher bias in an entire literature: if the hacking interval is large and the range of conclusions from many published studies is small, then this suggests researchers may be collectively biased toward a specific conclusion.

There exist some formalized procedures that aim to mitigate the impact of bias: for example, preregistration, lists of "best practices," enforced ignorance (e.g., double-blinding experiments and journal reviews), or requiring replication data sets (King 1995); however, the problem of reasonable researchers being able to reach a different conclusions would still exist even if researchers were each unbiased. The sheer number of possible analysis choices leaves unchecked uncertainty in scientific results unless the space of choices is rigorously defined and explored.

The prescriptive constraints and amount of loss tolerated for tethered hacking are up to the user to choose, so one could argue that these choices are themselves subject to hacking. Although we argue that *any* choice of hacking interval constraints is better than none at all, a set of best practices cannot only remove the burden of making this choice but facilitate comparison of hacking intervals across studies. Accompanying this paper, we provide the R package hackint, which computes constraint-based and tethered hacking intervals for linear models. Like R's built-in function for standard confidence intervals confint, hackint requires as input only a model fitted with lm. This linear model represents a "base" model in the sense that hacking is defined relative to this model. That is, the threshold for tethered hacking is a percentage of the base model's loss, and the prescriptive constraints are specified as modifications of the base model (e.g., removing a feature from the base model). The package itself is available on GitHub,[1] and a quick demonstration is available in Online Appendix A. The code used to produce results in the paper is also available on GitHub.[2]

Throughout this work, we offer examples and illustrations of hacking intervals, in the context of $k$ nearest neighbors ($k$-NNs), matching, variable selection, support vector machines (SVMs), and in more detail, linear regression. In Section 6, we present an analysis of recidivism prediction, where we find evidence that the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) score, which is a commonly used risk-scoring system used in bail and parole decisions, is often miscalculated. This can lead in practice to high-risk criminals being released, as well as low-risk individuals being unfairly sentenced or denied bail or parole. Our evidence for this conclusion is a set of individuals for which *all* reasonable models (by our definition of reasonable and according to our data set) from a particular model class disagree with their COMPAS score. This is followed by related work and further discussion in Section 7. All proofs are available in Online Appendix D.

## 2. Theories of Inference

Each of the diverse theories of inference is united by a common goal—to understand if an observed effect is robust over counterfactual worlds imagined to have occurred. These theories can be distinguished by which set of counterfactual worlds is assumed to be of interest. For example, $p$-values consider if an effect is robust to counterfactual *data* from a superpopulation. Fisher's exact $p$-values fix the data and measure if an effect is robust to counterfactual *treatment assignments* from every possible randomization. Causal sensitivity analysis considers if an effect is robust to counterfactual *unmeasured confounders* from a defined set (Liu et al. 2013, Ding and VanderWeele 2016). Bayesian credible intervals define results as robust to counterfactual *worlds*, generated by redrawing the data from the same data-generating process, given the observed data and assumed prior and likelihood model.

In part because the sum of uncertainties from different forms of inference is usually too large to be able to conclude almost anything at all, current practice is to present, in every applied publication, intervals or another summary from *only one* chosen form of uncertainty, stemming from a single theory of inference, and to temporarily assume away other forms of uncertainty. Another reason for temporarily ignoring all but one form of uncertainty is that one theory of inference may seem to be of more use than another depending on context. For example, despite studies showing a strong correlation between smoking and lung cancer, the question of whether smoking caused lung cancer was unsettled in the 1950s because of the possibility of an unmeasured confounding genetic variable. The Cornfield Conditions assumed that the causal effect was zero and deduced properties of the unmeasured confounding genetic variable, properties that were deemed biologically infeasible (Cornfield et al. 1959). This approach to inference was vital to taking the scientific community from facts that were known (smoking correlates with lung cancer, and there is an approximate biological limit on how much a genetic variable and smoking could be related) to a fact that was unknown (smoking causes lung cancer). Many other sources of uncertainty also afflicted this inference, but confounding bias was the largest perceived threat to validity; therefore, it was well worth it for researchers to at least temporarily set aside other sources of uncertainty.

We introduce our hacking theory of inference to address the growing crisis in science across fields based on the mistrust of published scientific results because of high degrees of researcher discretion. As such, our theory of inference considers if a substantive result is robust to counterfactual *researchers* making counterfactual *analysis choices* from a defined set larger than any one researcher would normally consider. We try to define this set of analytical decisions based on what all reasonable researchers from the entire scientific community might choose. Results from our theory of inference, like all others, are based on a set of counterfactual worlds, but it is designed precisely to respond to the current concern in the community.

We hypothesize which analysis choices reasonable researchers might make, either by explicitly constraining their choices or by allowing a tolerance in the loss function. From this, we then deduce the range of effects—the hacking interval—of results that would have been found within these constraints. A hacking interval can therefore be used to judge whether the observed effect is robust to researcher choices. Although a hacking interval is designed to estimate the range of conclusions that *reasonable* researchers could report, *any* researcher acting within the constraints will report results within the hacking interval. Because hacking intervals are designed to characterize conservatively all reasonable researcher choices, any researcher should report almost the same hacking interval.

An alternative to our approach is a greatly expanded Bayesian model (perhaps via robust Bayes combined with Bayesian model averaging) that formally specifies all possible modeling decisions, enables a choice of priors or classes of priors and the many associated hyperprior values over this large set, and computes classes of posteriors as a result. We do not recommend this approach because it adds numerous researcher choices for which prior information is rarely available and thus, may exacerbate the very problem of hacking we seek to address. Our preferred theory of inference explicitly gives up the goal of full posterior distributions or classes of posterior distributions. In their place, it seeks the more limited goal of an interval as a summary of uncertainty. What we get in return for limiting our goal to intervals is clearer ways of specifying assumptions, more effective ways of limiting researcher discretion, and easy to interpret results.

Hacking intervals, classical frequentist confidence intervals, Bayesian credible intervals, and others each convey important but different components of the strength of evidence in the observed data. However, hacking intervals may offer an especially natural starting point in analysis and in teaching. When researchers calculate numerical results of scientific interest, they need to quantify how strongly the observed data support their result. As with *p*-values, classical confidence intervals quantify the robustness of the result to sampling variability. If the result could be reversed under different data sets that are likely to have occurred under a specific sampling scheme, the result is not robust. Similarly, if a result could be reversed under different but also reasonable analysis choices, then the result is not robust. A large interval of either type should be regarded as lack of robustness of a type. However, hacking intervals may be a more natural starting point. Compared with classic confidence intervals, hacking intervals

1. represent uncertainty that always exists,
2. are easier to understand and explain,
3. are natural even when the superpopulation imagined in classical inference is not, and
4. are often wider than classical confidence intervals.

On the second point, hacking intervals are the solutions to an optimization problem that requires no understanding of probability. In contrast, despite repeated clarifications of their interpretation (Wasserstein and Lazar 2016), frequentist confidence intervals are routinely misinterpreted and misexplained, to the point where they have even been banned in some circles (Trafimow and Marks 2015) (see Section 7).

In regard to the third point in the list, consider problems from the political science fields of comparative politics and international relations, where country-level or time series cross-sectional data are available. The cause of (for instance) civil wars is deeply important for understanding the past, and we may like to determine patterns that characterize political situations that have led to civil wars. One might hypothesize that countries with many people in poverty, having many young men, with neighboring countries in civil war, and with no strong government could be prone to have civil wars. The data are observational, randomization is impossible for events that happened in the past, and no more relevant data may ever be collected (at least until more civil wars of the same type occur). In situations like these, researchers often use some type of regression to estimate causal relationships. If the researcher learns that a variable has a large coefficient in the regression for predicting aspects relating to a civil war, then she may use confidence intervals to determine whether this result is robust—robust across possible model specifications. She may use traditional inference notation (confidence intervals, null hypotheses), but because the idea of a superpopulation may not even make sense, the null hypothesis does not exist, and she may find it more natural to compute a hacking interval. Researchers in this field are not interested in constructing an imaginary superpopulation of world systems with different countries; we really only care about the actual countries and their real civil wars. The question of interest, which hacking intervals address, is whether the researcher can claim a robust empirical relationship or whether she demonstrated only that it was merely possible to find one of a million model specifications that was consistent with her causal hypothesis. In this case, the researcher may wish to focus on the uncertainty in a hacking interval, rather than a classic confidence interval. However, to do this requires a specific mathematical framework for this interval, a subject to which we now turn.

Given these four relative advantages of hacking intervals and that the analyst simply wants to find patterns in the data that are robust, we recommend that researchers calculate a hacking interval first and then decide if calculating a classical interval adds value.

## 3. Prescriptively Constrained Hacking Intervals

Denote $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{n\times p}$ as covariates, $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^{n}$ as outcomes, $\mathbf{Z} \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as data sets, and $f : \mathcal{X} \to \mathcal{Y}$ as prediction functions from a class $\mathcal{F}_{\psi}$, where $\psi \in \Psi$ denotes a vector of hyperparameters. For example, $\mathcal{F}_{\psi}$ could be the space of all binary decision trees of maximum depth $\psi$. Let $L : \mathcal{Z} \times \mathcal{F}_{\psi} \times \Psi \to \mathbb{R}$ be a loss or regularized loss function and $t : \mathcal{Z} \times \mathcal{X} \times \mathcal{F}_{\psi} \to \mathbb{R}$ be a summary statistic of interest. The loss function may or may not depend on the hyperparameters $\psi$, so if not, we omit writing $\psi$. Similarly, although the summary statistic must depend on $f$, it may or may not depend on $\mathbf{Z}$, which is the observed training data, or $\mathbf{X}^{(\text{new})}$, which are covariates for observations the model is not trained on. Depending on the context, we may omit writing $\mathbf{Z}$ and/or $\mathbf{X}^{(\text{new})}$ in the definition of $t$.

For hyperparameters $\psi$, training data $\mathbf{Z} \in \mathcal{Z}$, and optionally, test data $\mathbf{X}^{(\text{new})}$, we assume the user finds $f^*$ that minimizes the loss $L(\mathbf{Z}, f^*, \psi)$ and then, computes the summary statistic $t^* := t(\mathbf{Z}, \mathbf{X}^{(\text{new})}, f^*)$ based on this result. For instance, in linear regression, the user finds the linear function $f^*(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}^*$ that minimizes the quadratic loss on the data set $\mathbf{Z} = [\mathbf{X}, \mathrm{y}]$. Possible summary statistics include an estimate of a single regression coefficient $t(f^*) = \beta_j^*$, a goodness-of-fit measurement of $f^*$ on $\mathbf{Z}$, or a prediction $t(\mathbf{x}^{(\text{new})}, f^*) = f^*(\mathbf{x}^{(\text{new})}) = \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^*$ on a single test observation $\mathbf{x}^{(\text{new})} \in \mathcal{X}$. Our interest is in the range of summary statistics $t^*$ that could be achieved if the researcher was permitted to adjust the data set $\mathbf{Z}$ and hyperparameters $\psi$.

The approach to this problem is to explicitly constrain data adjustments $\phi : \mathcal{Z} \to \mathcal{Z}$ to a set $\Phi$ and hyperparameters to a set $\Psi$. We assume that $\phi$ can be separated into two functions $\phi_{\mathbf{X}}$ and $\phi_{\mathbf{Y}}$ such that for any $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ we have $\phi(\mathbf{Z}) = [\phi_{\mathbf{X}}(\mathbf{X}), \phi_{\mathbf{Y}}(\mathbf{Y})]$. We then wish to calculate the minimum and maximum summary statistics over these two sets, $\Psi$ and $\Phi$, which constrain researcher choices:

$$a_{\min} := \min_{\psi\in\Psi,\phi\in\Phi} t\left(\phi(\mathbf{Z}), \phi_{\mathbf{X}}\left(\mathbf{X}^{(\text{new})}\right), \operatorname*{argmin}_{f\in\mathcal{F}_\psi} L\big(\phi(\mathbf{Z}), f, \psi\big)\right), \tag{1}$$

$$a_{\max} := \max_{\psi\in\Psi,\phi\in\Phi} t\left(\phi(\mathbf{Z}), \phi_{\mathbf{X}}\left(\mathbf{X}^{(\text{new})}\right), \operatorname*{argmin}_{f\in\mathcal{F}_\psi} L\big(\phi(\mathbf{Z}), f, \psi\big)\right). \tag{2}$$

Notice that hyperparameters $\psi$ impact $a_{\min}$ and $a_{\max}$ through $\mathcal{F}_\psi$ (e.g., by controlling the maximum depth of decision trees) as well as through the loss directly (e.g., by controlling the regularization). In other words, $\psi$ is assumed to contain all relevant hyperparameters to determine hard constraints on the function class as well as soft constraints through regularization. We define the interval $[a_{\min}, a_{\max}]$ as the *prescriptively constrained hacking interval*. For example, if the summary statistic $t$ is a prediction of $f$ on a new point $\mathbf{x}^{(\text{new})}$, then Equations (1) and (2) can be written as

$$\begin{aligned}
a_{\min} = \min_{\psi\in\Psi,\phi\in\Phi} \quad & f\left(\phi_{\mathbf{X}}\left(\mathbf{x}^{(\text{new})}\right)\right)\\
\text{s.t.} \quad & f \in \operatorname*{argmin}_{\mathcal{F}_\psi} L\big(\phi(Z), f, \psi\big)\\
a_{\max} = \max_{\psi\in\Psi,\phi\in\Phi} \quad & f\left(\phi_{\mathbf{X}}\left(\mathbf{x}^{(\text{new})}\right)\right)\\
\text{s.t.} \quad & f \in \operatorname*{argmin}_{\mathcal{F}_\psi} L\big(\phi(Z), f, \psi\big),
\end{aligned}$$

where "s.t." stands for "subject to."

Although a prescriptively constrained hacking interval is designed for a single loss function, one could include in $\psi$ a hyperparameter that switches between more than one loss function, allowing for specification of the loss function to be among the researcher choices.

Instead of using her own discretion, a researcher may pick all or some of the hyperparameters based on cross validation. To compute the prescriptively constrained hacking interval in this case, the objective function $t$ in Equations (1) and (2) is evaluated by first computing the optimal hyperparameters based on cross validation (which will depend on the data adjustment function $\phi$ and the remaining non-cross-validated hyperparameters, if any) and then plugging them into $t$.

### 3.1. Examples

We present examples of prescriptively constrained hacking intervals for, first, the simple example of $k$ nearest neighbors ($k$-NN) (where the researcher chooses $k$ within a reasonable range) and then, the more complex example of adding a new feature (where the researcher adds a new feature constrained by its relationship to existing data). Using results from Morucci et al. (2018), we also show the example of matching for causal inference (where the researcher chooses a matching algorithm) in Online Appendix B.

**3.1.1. *k*-NN.** This is a simple example. Suppose we have observed data $\mathbf{Z} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, and we wish to predict on a new point $\mathbf{x}^{(\text{new})}$ by averaging nearby observations. In this example, we will keep the data $\mathbf{Z}$ fixed but allow the researcher to choose the hyperparameter $k$, the number of nearest neighbors over which to average. To construct a simple prescriptively constrained hacking interval, we define a subset of reasonable hyperparameter choices $\Psi$, which in this case, we can write as a range $[k_{\min}, k_{\max}]$, and find the range of predictions on a new point $\mathbf{x}^{(\text{new})}$ subject to the constraint that $k \in [k_{\min}, k_{\max}]$:

$$\max/\min_{k\in[k_{\min},k_{\max}]} \frac{1}{k}\sum_j \eta_{i^{(\text{new})}j}^{(k)} y_j,$$

where $\eta_{ij}^{(k)}$ is an indicator that is one if $\mathbf{x}_j$ is within the $k$ nearest neighbors of $\mathbf{x}_i$ and zero otherwise. This range of predictions is the prescriptively constrained hacking interval. Notice that there is no loss function. The hyperparameter $k$ allows for only one function in the function space $\mathcal{F}_k$ (namely, the one that averages over the $k$ nearest neighbors). To solve this problem, we evaluate the nearest neighbor average for each $k$ within the range $\Psi = [k_{\min}, k_{\max}]$.

Prescriptively constrained hacking intervals require that the researchers justify to readers their choice of $\Psi$, and we recommend that this discussion be briefly included in every paper. This approach therefore does not remove all research discretion and arguably, not all hacking, but it changes the nature of scholarly papers from a justification of a single specification to one where they justify a definition for the range of reasonable specifications.

One possibility for this choice is to center $\Psi = [k_{\min}, k_{\max}]$ around a fixed value and calculate the hacking interval over $[k_{\min}, k_{\max}]$ constraints of increasing width. For example, find $k^* \in [1, n-1]$ that minimizes the training error and then find the hacking interval over $\Psi(m) := [k^* - m, k^* + m] \cap [1, n-1]$ for each $m = 1, 2, 3, \ldots, m_{\max}$. Figure 1 shows the results of such a procedure for a data set in two dimensions and $\mathbf{x}^{(\text{new})} = (\mathbf{0.5}, \mathbf{0.5})$. We find that $k^* = 5$ minimizes the training error, and the resulting prediction on $\mathbf{x}^{(\text{new})}$ is 0.6. However, if the researcher is allowed to pick any $k$ in $[k^* - 2, k^* + 2] = [3, 7]$, for example, then the prediction ranges from 0.57 to 0.70. This is the hacking interval for $m = 2$. Displaying the hacking interval as a function of $m$ illustrates the sensitivity of the hacking interval to the freedom given to the researcher.

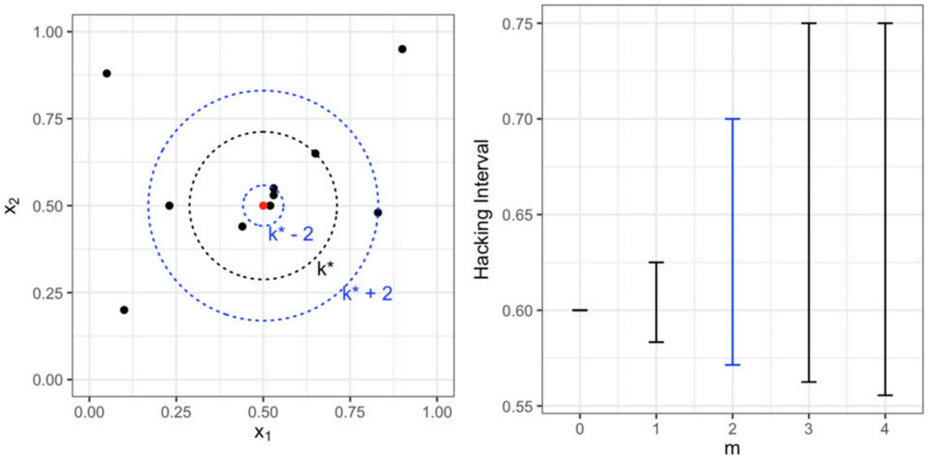
Another choice for the range of $k$ could be to use prior information of acceptable past researcher choices. We might choose the range of $k$ large enough to include the smallest and largest $k$ values used in $k$-NN in any article in the last five years in that field. In practice, that interval may actually be the smallest and largest values that would not be objected to by reviewers.

Other researcher choices for $k$-NN that we did not consider in this example could include the distance function or the weighting of the $k$ nearest neighbors. The use of $k$-NN as the predictive function class could also be considered a researcher degree of freedom. We could use a binary hyperparameter to switch between $k$-NN and any other regression algorithm.

**3.1.2. Adding a New Feature.** The addition of a new feature to a given collection of features, possibly from existing features (such as an interaction term) or from new data, is a data adjustment that can impact the conclusions about prediction models. A prescriptively constrained hacking interval in this context is the range of a summary statistic that can be achieved over all of the possible choices made by the researcher about new features, subject to explicit constraints on those choices. If the researcher is given the freedom to choose each of $n$ feature values (one for each observation), then solving this problem requires optimizing over a potentially large space because there are $n$ choices made by the researcher. Fortunately, it may only be necessary to specify a small number of attributes about the new feature to calculate its impact on the summary statistic. The prescriptively constrained hacking interval would then be an optimization problem over a smaller space of attributes, subject to explicit constraints on those attributes.

In a causal inference setting, where the researcher observes a treatment feature among other possibly confounding features, sensitivity analysis deals with this exact problem. The goal is to find the impact on a causal effect (the summary statistic) of an unmeasured confounder $u$ (the new feature).[3] To do this, one needs to choose a value for several attributes about the unmeasured confounder. There are a number of approaches to this problem that require different attributes of $u$ to be chosen (see Liu et al. 2013 for a review), but generally, only a few attributes are required: its distribution, its relationship to the outcome, and its relationship to the treatment. In applications of causal sensitivity analysis, a researcher

**Figure 1.** (Color online) (Left panel) Observed Data with Distance to $k = 3, 5,$ and 7 Nearest Neighbors Highlighted, Where $k^* = 5$



*Notes.* (Right panel) Hacking intervals as a function of the hyperparameters space width $m$. $m = 2$ corresponds to a hacking interval over researcher choice $[k^* - 2, k^* + 2] = [3, 7]$.

will often display the adjusted causal effect for each of a few choices of these attributes. If we explicitly define a range of choices for each attribute, then the maximum and minimum causal effects over these ranges are a prescriptively constrained hacking interval.

The motivations of causal sensitivity analysis and prescriptively constrained hacking are different. In causal sensitivity analysis, $u$ exists but is unmeasured by the researcher. Constraints on the values of the attributes of $u$ are based on what we believe is scientifically reasonable. In prescriptively constrained hacking, $u$ is created by the researcher. Constraints on the values of the attributes of $u$ are based on what we believe is a reasonable amount of researcher freedom.

We now define our approach in more detail. Let $\mathbf{Y} = (y_1, \ldots, y_n)^T \in \{-1, 1\}^n$ be a $n \times 1$ matrix of observed binary outcomes, $\mathbf{X} = (\mathbf{x}_i^T, \ldots, \mathbf{x}_n^T)^T$ be an $n \times p$ matrix of observed covariates, and $\mathbf{W} = (w_1, \ldots, w_n)^T \in \{0, 1\}^n$ be an $n \times 1$ matrix of observed binary covariates. In a causal inference setting, $\mathbf{W}$ is the treatment. The researcher degrees of freedom constitute the choice of an additional binary covariate $\mathbf{U} = (u_1, \ldots, u_n)^T \in \{0, 1\}^n$. This is equivalent to the choice of a data adjustment function $\phi : [\mathbf{Y}, \mathbf{X}, \mathbf{W}] \mapsto [\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{U}]$. After $\phi$ has been chosen, we assume the researcher finds a model $f$ from a set of linear functions $\mathcal{F}$ of the form $f([\mathbf{x}, w, u]) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}_{\mathrm{x}} + \beta_w w + \beta_u u$ by minimizing the logistic loss function:

$$L([\mathbf{X}, \mathbf{W}, \mathbf{U}], f) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i f([\mathbf{x}_i, w_i, u_i])}\right).$$

Notice that this is equivalent to maximizing the likelihood under the model: logit $\Pr(Y = 1 \mid \mathbf{x}, w, u) = f([\mathbf{x}, w, u])$, where $Y$ is the random variable corresponding to the observed $y$. In other words, the researcher performs logistic regression. (For simplicity, the objective is fixed, and there are no user choices except to add the extra feature.) We further assume the researcher is interested in the odds ratio of $y$ and $w$ controlling for covariates $\mathbf{x}$ and $u$:

$$OR_{yw|\mathbf{x},u} := \frac{\Pr(Y = 1 \mid \mathbf{x}, w = 1, u)}{\Pr(Y = 1 \mid \mathbf{x}, w = 0, u)},$$

so we set the test statistic to be $t(f) := e^{\beta_w}$. The steps followed by the researcher can be summarized as follows:

- Step 1a. Choose a data adjustment $\phi \in \Phi$ (we discuss $\Phi$ later).
- Step 1b. Find $\hat{f}([\mathbf{x}, w, u]) = \hat{\beta}_0 + \mathbf{x}^T\hat{\boldsymbol{\beta}}_{\mathbf{x}} + \hat{\beta}_w w + \hat{\beta}_u u$ that minimizes the logistic loss on the adjusted data, $(\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{U}) = \phi(\mathbf{Y}, \mathbf{X}, \mathbf{W})$.
- Step 1c. Calculate the summary statistic $\widehat{OR}_{yw|\mathbf{x},u} = t(\hat{f}) = e^{\hat{\beta}_w}$.

The prescriptively constrained hacking interval is the maximum and minimum values of $t(\hat{f})$ that can be achieved over all of the possible researcher choices of $\phi \in \Phi$. There are no hyperparameters in this example.

Interestingly, we can calculate $\widehat{OR}_{yw|\mathbf{x},u}$ without knowing the researcher-created covariate $u$ exactly. We need only know the relationship of $u$ to the binary covariate $w$, specified by $p_0 := p(U \mid w = 0)$ and $p_1 := \Pr(U \mid w = 1)$ (where $U$ is the random variable corresponding to $u$), and the relationship of $u$ to the binary outcome $y$, specified by $OR_{yu} := \Pr(Y = 1 \mid u = 1)/\Pr(Y = 1 \mid u = 0)$. When $p_0, p_1$, and $OR_{yu}$ are known, Lin et al. (1998) show[4] that we can derive the odds ratio adjusting for $\mathbf{x}$ and $u$, $\widehat{OR}_{yw|\mathbf{x},u} = t(\hat{f})$, from the odds ratio that only adjusts for $\mathbf{x}$, $\widehat{OR}_{yw|\mathbf{x}}$, by the following formula:

$$\widehat{OR}_{yw|\mathbf{x},u} = \frac{1}{AF}\widehat{OR}_{yw|\mathbf{x}}, \text{ where}$$
$$AF = \frac{(OR_{yu} - 1)p_1 + 1}{(OR_{yu} - 1)p_0 + 1}. \tag{3}$$

We write $OR_{yu}$ rather than $\widehat{OR}_{yu}$ because the former quantity is the true odds ratio, not one estimated from the data.

Because $\widehat{OR}_{yw|\mathbf{x}}$ can be estimated from the observed data, Equation (3) implies the impact of the researcher choice of $u$ is completely summarized by $p_1$, $p_0$, and $OR_{yu}$ because they determine $AF$. Conversely, if we knew the data adjustment $\phi$, we could estimate $p_1, p_0$, and $OR_{yu}$, calling the estimates $\hat{p}_1$, $\hat{p}_0$, and $\widehat{OR}_{yu}$, respectively, from the adjusted data. Steps *1a-1c* are therefore equivalent to Steps 2a-2d defined by

- Step 2a. Calculate $\widehat{OR}_{yw|\mathbf{x}}$.
- Step 2b. Choose a data adjustment $\phi \in \Phi$.
- Step 2c. Calculate $\widehat{OR}_{yu}$, $\hat{p}_1$, and $\hat{p}_0$ using the adjusted data $[\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{U}] = \phi([\mathbf{Y}, \mathbf{X}, \mathbf{W}])$.
- Step 2d. Calculate the summary statistic $\widehat{OR}_{yw|\mathbf{x},u} = \frac{1}{\widehat{AF}}\widehat{OR}_{yw|\mathbf{x}}$, where $\widehat{AF}$ is analogous to Equation (3) but depends on the estimated quantities $\widehat{OR}_{yu}$, $\hat{p}_1$, and $\hat{p}_0$:

$$\widehat{AF} = \frac{\left(\widehat{OR}_{yu} - 1\right)\hat{p}_1 + 1}{\left(\widehat{OR}_{yu} - 1\right)\hat{p}_0 + 1}. \tag{4}$$

Notice that the researcher's choice of a data adjustment $\phi$ implies a value for $u$ and the three attributes about $u$—$\widehat{OR}_{yu}$, $\hat{p}_1$, and $\hat{p}_0$—but it is through these three attributes only that $\phi$ impacts the summary statistic. If we instead allow the researcher to choose only the three attributes, we can find the impact on the summary statistic without ever knowing $u$. We just need to define the space of allowable data adjustments $\Phi$ in terms of its impact on these three attributes:

$$\Phi := \Big\{\phi : (\mathbf{Y},\mathbf{X},\mathbf{W}) \mapsto (\mathbf{Y},\mathbf{X},\mathbf{W},\mathbf{U}) \mid \widehat{OR}_{yu} \in [a,b],$$
$$|\hat{p}_1 - \hat{p}_0| \le c, \hat{p}_0 > d\Big\},$$

for constants $a$, $b$, and $c < d$ (the reason for these exact constraints will become clear later). Then, Steps 2a-2d can be replaced with Steps 3a-3c defined by

- Step 3a. Calculate $\widehat{OR}_{yw|\mathbf{x}}$.
- Step 3b. Choose $OR_{yu}$, $p_0$, and $p_1$ such that $OR_{yu} \in [a,b]$, $|p_1 - p_0| \le c$, and $p_0 \ge d$.
- Step 3c. Calculate the summary statistic $\widehat{OR}_{yw|\mathbf{x},u} = \frac{1}{AF}\widehat{OR}_{yw|\mathbf{x}}$, where $AF$ depends on $OR_{yu}$, $p_0$, and $p_1$.

For any equivalent choice of constraints, the maximum and minimum values of $\widehat{OR}_{yw|\mathbf{x},u}$ that could be achieved by any of the three sequences of steps (Steps 1a–1c, 2a–2d, and 3a–3c) are all equal. We can think of finding the maximum and minimum values of $\widehat{OR}_{yw|\mathbf{x},u}$ for each of the three sequences as the three following optimization problems (each solved for the maximum and minimum):

$$\text{Steps 1a–1c}: \quad \max_{\phi\in\Phi}/\min \left\{OR_{yw|\mathbf{x},u}\right\} \qquad (5)$$

$$\text{Steps 2a–2d}: \quad \max/\min_{\phi \text{ s.t.} \begin{cases}\widehat{OR}_{yu}\in[a,b]\\ |\hat{p}_1-\hat{p}_0|\le c\\ \hat{p}_0>d\end{cases}} \left\{\frac{1}{\widehat{AF}}\widehat{OR}_{yw|\mathbf{x}}\right\} \qquad (6)$$

$$\text{Steps 3a–3c}: \quad \max/\min_{\substack{OR_{yu}\in[a,b]\\ |p_1-p_0|\le c\\ p_0\ge d}} \left\{\frac{1}{AF}\widehat{OR}_{yw|\mathbf{x}}\right\}. \qquad (7)$$

Optimization problem (7) will prove the most useful as it does not require knowledge of $u$. Because $\widehat{OR}_{yw|\mathbf{x}}$ is estimated from the observed data, solving optimization problem (7) is equivalent to solving for the maximum and minimum values of $AF$ subject to the same constraints and dividing $\widehat{OR}_{yw|\mathbf{x}}$ by each value. Using Equation (3) for $AF$, we find the maximum and minimum values of $AF$ by solving the following optimization problem:

$$\max_{OR_{yw},p_1,p_0}/\min \frac{(OR_{yu}-1)p_1+1}{(OR_{yu}-1)p_0+1} \quad \text{s.t.} \begin{cases} OR_{yu}\in[a,b]\\ |p_1-p_0|\le c\\ p_0\ge d\end{cases}. \qquad (8)$$

Dividing $\widehat{OR}_{yw|\mathbf{x}}$ by the maximum and minimum values given by optimization problem (8) gives the minimum and maximum values, respectively, of $OR_{yw|\mathbf{x},u}$, which define the hacking interval in this case.

We can solve Equation (8) for the case where $OR_{yu}$ is fixed greater than 1 (implying $\Pr(Y = 1 \mid u = 1) > \Pr(Y = 1 \mid u = 0)$). In this case, the maximization problem in Equation (8) (i.e., the hacking interval upper bound) becomes

$$\max_{\substack{|p_1-p_0|\le c\\ p_0\ge d}} \frac{(OR_{yu}-1)p_1+1}{(OR_{yu}-1)p_0+1} = \max_{p_0\ge d} \frac{(OR_{yu}-1)(p_0+c)+1}{(OR_{yu}-1)p_0+1}$$
$$= \max_{p_0\ge d} 1 + \frac{(OR_{yu}-1)c}{(OR_{yu}-1)p_0+1},$$

whereas the minimization problem (i.e., the hacking interval lower bound) becomes

$$\min_{\substack{|p_1-p_0|\le c\\ p_0\ge d}} \frac{(OR_{yu}-1)p_1+1}{(OR_{yu}-1)p_0+1} = \min_{p_0\ge d} \frac{(OR_{yu}-1)(p_0-c)+1}{(OR_{yu}-1)p_0+1}$$
$$= \min_{p_0\ge d} 1 - \frac{(OR_{yu}-1)c}{(OR_{yu}-1)p_0+1}.$$

In each case, the optimum occurs at $p_0 = d$. Therefore, Equation (8) can be solved when $OR_{yu}$ is fixed greater than one. We apply this result in Section 6.1.

This section shows how results from causal sensitivity analysis can be leveraged to solve problems where the researcher is permitted to hack a new feature. Here, we have been in a noncausal inference setting of logistic regression modeling. In Section 6.1, we apply these results to a recidivism data set.

## 4. Tethered Hacking Intervals

In prescriptively constrained hacking intervals, discussed in Section 3, we optimize over a data adjustment function $\phi$ and hyperparameters $\psi$ constrained to be in sets $\Phi$ and $\Psi$, respectively. An advantage of this approach is that we can clearly define acceptable researcher adjustments. A disadvantage is that the possible adjustments may be difficult to enumerate or optimize over efficiently. One way to circumvent this requirement is to allow *any* choice of $\psi$ and $\phi$ so long as the loss using the unadjusted data $\mathbf{Z}$ and a set of default hyperparameters $\psi_d$ is not too large. The *tethered hacking interval* is the minimum and maximum summary statistics under this constraint. In other words, it is given by the interval $[b_{\min}, b_{\max}]$,

$$b_{\min} := \min_{f\in\mathcal{F}_{\psi_d}} t\left(\mathbf{Z},\mathbf{X}^{(\text{new})},f\right) \quad \text{s.t.} \quad L\left(\mathbf{Z},f,\psi_d\right) \le \theta, \quad (9)$$

$$b_{\max} := \max_{f\in\mathcal{F}_{\psi_d}} t\left(\mathbf{Z},\mathbf{X}^{(\text{new})},f\right) \quad \text{s.t.} \quad L\left(\mathbf{Z},f,\psi_d\right) \le \theta, \quad (10)$$

given a fixed, chosen value of $\theta$. The default hyperparameters $\psi_d$ could be specified based solely on problem-specific standards, cross validation, or a combination of both. To do the combination in the case that there are multiple viable values of the problem-specific hyperparameters, we would first choose values for the problem-specific hyperparameters and perform cross validation on the rest, repeating this procedure for every viable value of the problem-specific hyperparameters. After a single set of hyperparameters $\psi_d$ is selected, we can proceed with computing the tethered hacking interval by Equations (9) and (10). In contrast to the computation of prescriptively constrained hacking intervals in the case of cross validation, as described in Section 3, any cross validation of the hyperparameters is done prior to solving the optimization problems.

For example, suppose $\mathcal{F}$ is the set of constant functions $f(x) = \lambda$, $t(\mathbf{Z}, \mathbf{X}^{(\text{new})}, f) = \lambda$ is the parameter $\lambda$ that defines $f$, and $L$ is the quadratic loss for each of $n$ observations in data set $\mathbf{Z}$. There are no hyperparameters $\psi_d$, so we suppress their notation. Then, Equations (9) and (10) become

$$b_{\min} = \min_{\lambda} \quad \lambda \quad \text{s.t.} \quad \sum_{i=1}^{n} (\lambda - y_i)^2 \le \theta$$

$$b_{\max} = \max_{\lambda} \quad \lambda \quad \text{s.t.} \quad \sum_{i=1}^{n} (\lambda - y_i)^2 \le \theta.$$

For another example, if $\mathcal{F}$ is the set of linear functions $f(x) = \lambda_0 + \lambda_1 x$, $t(\mathbf{Z}, \mathbf{x}^{(\text{new})}, f) = \lambda_0 + \lambda_1 \mathrm{x}^{(\text{new})}$ is a prediction of $f$ on a new point $\mathrm{x}^{(\text{new})}$, and $L$ is the same quadratic loss, then Equations (9) and (10) become

$$\begin{aligned} b_{\min} = \min_{\lambda_0, \lambda_1} &\quad \lambda_0 + \lambda_1 \mathrm{x}^{(\text{new})} \\ \text{s.t.} &\quad \sum_{i=1}^{n} (\lambda_0 + \lambda_1 x_i - y_i)^2 \le \theta \\ b_{\max} = \max_{\lambda_0, \lambda_1} &\quad \lambda_0 + \lambda_1 \mathrm{x}^{(\text{new})} \\ \text{s.t.} &\quad \sum_{i=1}^{n} (\lambda_0 + \lambda_1 x_i - y_i)^2 \le \theta. \end{aligned}$$

In general, when the summary statistic is a prediction on a new point $\mathbf{x}^{(\text{new})}$, Equations (9) and (10) become

$$b_{\min} = \min_{f \in \mathcal{F}_{\psi_d}} \quad f\left(\mathbf{x}^{(\text{new})}\right) \quad \text{s.t.} \quad L(\mathbf{Z}, f, \psi_d) \le \theta$$

$$b_{\max} = \max_{f \in \mathcal{F}_{\psi_d}} \quad f\left(\mathbf{x}^{(\text{new})}\right) \quad \text{s.t.} \quad L(\mathbf{Z}, f, \psi_d) \le \theta.$$

The interpretation of a tethered hacking interval is that a researcher could have hacked the data or adjusted the hyperparameters to obtain values of the test statistic in the interval. In other words, for each point $b' \in [b_{\min}, b_{\max}]$ there could exist a data adjustment function $\phi'$ and a set of hyperparameters $\psi'$ such that $b'$ is the output of the summary statistic when applied to the minimum loss predictive model $f$ using $\phi'$ and $\psi'$. That is,

$$b' = t\left(\phi'(\mathbf{Z}), \phi'_{\mathbf{X}}\left(\mathbf{X}^{(\text{new})}\right), \operatorname*{argmin}_{f \in \mathcal{F}_{\psi'}} L(\phi'(\mathbf{Z}), f, \psi')\right).$$

This interpretation describes how results are hacked in practice. A researcher first chooses how to adjust a data set and which hyperparameters are appropriate and then, summarizes the resulting best function in a class. The purpose of a tethered hacking interval is to bound the results of this procedure by specifying a single constraint on the loss function.

The set of models achieving small loss is also called the *Rashomon set* (Fisher et al. 2019), based on terminology originally from Leo Breiman's analogy to the 1950 Akira Kurosawa film *Rashomon* (Breiman 2001). Fisher et al. (2019) introduce a measure of variable importance for a class of prediction functions based on the Rashomon set. Although the computation and interpretation of their "empirical model class reliance" measure of variable importance could be viewed as similar to those of hacking intervals, their ultimate goal is to study the population version of this quantity in order to study the Rashomon set for the population.

We note two things about tethered hacking intervals. First, when the loss function corresponds to a likelihood function, tethered hacking intervals are equivalent to profile likelihood confidence intervals for an appropriate choice of the loss threshold $\theta$. See Online Appendix C.1 for details. Second, as with prescriptively constrained hacking intervals, a tethered hacking interval is a statement about the degree to which summaries of a single observed data set could be hacked by a researcher. It does not require an assumption about a true data-generating procedure. If we make such an assumption about the true data-generating procedure, we can derive an appropriate generalization bound in order to unite traditional inference with our new inference paradigm. See Online Appendix C.2 for details.

Next, we discuss tethered hacking intervals for predictions made by SVM. The examples of predictions made by kernel regression and features selected using Principal Component Analysis (PCA) can be found in Online Appendices C.3 and C.4, respectively.

### 4.1. Example: SVM

In this section, we demonstrate how hacking intervals can be calculated in the context of SVMs with a linear kernel. Recall that SVM is trained by minimizing the following loss function:

$$L(\mathbf{Z}, f, \psi_d) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2 + \psi_d \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+,$$

where $f(\mathbf{x}) := \boldsymbol{\lambda}^T \mathbf{x} + \lambda_0$ is the scaled distance of $\mathbf{x}$ to the separating hyperplane and $\psi_d \in \mathbb{R}^+$ is a hyperparameter that controls the degree of regularization. Here, we define the summary statistic as the distance of a new point $\mathbf{x}^{(\text{new})}$ to the separating hyperplane. The hacking interval is then given by

$$\begin{aligned}&\max_{\boldsymbol{\lambda},\lambda_0}/\min\ \boldsymbol{\lambda}^T\mathbf{x}^{(\text{new})}+\lambda_0\\ &\text{s.t.}\quad \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2+\psi_d\sum_{i=1}^{n}\left(1-y_i\left(\boldsymbol{\lambda}^T\mathbf{x}_i+\lambda_0\right)\right)_+\le\theta,\end{aligned}\tag{11}$$

where $\theta$ controls the loss tolerance. Figure 2 illustrates this problem.

For simplicity, we can write both the minimum and maximum problems from Equation (11) as a single minimization problem that depends on the choice of a binary variable $s\in\{-1,+1\}$ ($s=1$ for minimum, $s=-1$ for maximum). If we also write the loss constraint in terms of slack variables $\xi$, then Equation (11) becomes

$$\min_{\boldsymbol{\lambda},\lambda_0,\xi} s\boldsymbol{\lambda}^T\mathbf{x}^{(\text{new})}+s\lambda_0\quad \text{s.t.}\quad \begin{cases} y_i\left(\boldsymbol{\lambda}^T\mathbf{x}_i+\lambda_0\right)\ge 1-\xi_i, & \forall i\\ \xi_i\ge 0,\ \forall i\\ \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2+\psi_d\sum_{i=1}^n\xi_i\le\theta.\end{cases}\tag{12}$$

This is a convex optimization problem. The objective is linear. The first two constraints are the same as in nonseparable SVM and are linear. The last constraint is the sum of a norm (always convex) and a linear function in $\xi$, so it is convex; also, it is the objective function for nonseparable SVM. Therefore, we can apply the Karush-Kuhn-Tucker (KKT) conditions to obtain the dual problem.

**Proposition 1** (Hacking Intervals for SVM). *The solution to optimization problem* (12) *is given by*

$$\boldsymbol{\lambda}^*=\frac{1}{\beta^*}\left(-s\mathbf{x}^{(\text{new})}+\sum_{i=1}^n\alpha_i^*y_i\mathbf{x}_i\right)\quad \textit{and}$$
$$\lambda_0^*=y_{i_{sv}}-\boldsymbol{\lambda}^{*T}\mathbf{x}_{i_{sv}},$$

*where* $i_{sv}$ *is such that* $0<\alpha_{i_{sv}}^*<\beta^*\psi_d$ *and the optimal dual variables* $(\boldsymbol{\alpha}^*,\beta^*)$ *are the solutions to the following dual problem:*

$$\begin{aligned}\max_{\boldsymbol{\alpha},\beta}-\frac{1}{2\beta}\Bigg[&\mathbf{x}^{(\text{new})T}\mathbf{x}^{(\text{new})}-2s\sum\alpha_iy_i\mathrm{x}_{\mathrm{i}}^{\mathrm{T}}\mathbf{x}^{(\text{new})}\\ &+\sum_i\sum_k\alpha_i\alpha_ky_iy_k\mathbf{x}_i^T\mathbf{x}_k\Bigg]+\sum\alpha_i-\beta\theta\\ &\text{s.t.}\ \begin{cases}0\le\alpha_i\le\beta\psi_d,\ \forall i\\ \sum_{i=1}^n\alpha_iy_i=s\\ \beta\ge 0\end{cases}.\end{aligned}\tag{13}$$

In Section 6.2, we apply SVM hacking intervals to a recidivism data set.

## 5. Tethered Hacking Intervals for Linear Regression

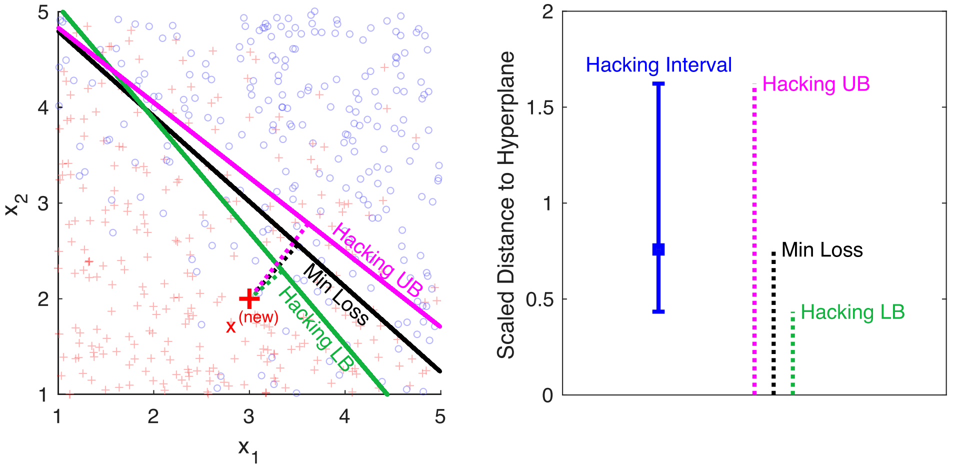
We develop hacking intervals in detail for two linear regression scenarios.

- Scenario 1: average treatment effect (ATE). We assume a class of linear functions $\mathcal{F}$ with $p$ confounders and an indicator covariate for the treatment (one if treatment, zero if control). We write $f\in\mathcal{F}$ as

$$\begin{aligned}&f(\mathbf{x},\text{treated or control})\\ &\quad=\beta_1x_{.1}+\beta_2x_{.2}+\ldots\beta_px_{.p}+\beta_0 1_{\text{treated}}.\end{aligned}$$

- The goal is to construct a tethered hacking interval for $\beta_0$, the coefficient of the treatment indicator. In other words, the test statistic is $t(\mathbf{Z},f)=\beta_0$. The coefficient $\beta_0$ represents the average treatment effect. Section 5.1 develops this in detail.
- Scenario 2: individual treatment effect (TE). We assume a class of linear functions $\mathcal{F}$ with $p$ confounders

**Figure 2.** (Color online) Lower Bound (LB) and Upper Bound (UB) of the Hacking Interval for an SVM Prediction



*Notes.* The summary statistic being hacked is the distance from the separating hyperplane to a new observation, $\mathbf{x}^{(\text{new})}$. For a default regularization trade-off of $\psi_d=1$ and a 5% tolerance on the loss relative to the minimum loss solution, SVM will always predict a +1 label, but the scaled distance to the hyperplane, $\boldsymbol{\lambda}^T\mathbf{x}^{(\text{new})}+\lambda_0$, can range from about 0.4 to about 1.6.

for both the treatment and control groups. We write $f \in \mathcal{F}$ as

$$\begin{aligned} &f(\mathbf{x}, \text{treated or control}) \\ &\quad = 1_{\text{control}}\left[\beta_1^c x_{\cdot 1} + \beta_2^c x_{\cdot 2} + \ldots \beta_p^c x_{\cdot p}\right] \\ &\quad + 1_{\text{treated}}\left[\beta_1^t x_{\cdot 1} + \beta_2^t x_{\cdot 2} + \ldots \beta_p^t x_{\cdot p}\right], \end{aligned}$$

- where $1_{\text{control}}$ is 1 only for the control group and $1_{\text{treated}}$ is 1 only for the treatment group. The goal is to construct a tethered hacking interval for a prediction of $f$ on a new point $[\mathbf{x}^{(\text{new})}$, treated or control. In other words, the test statistic is $t(\mathbf{Z}, [\mathbf{x}^{(\text{new})}, \text{treated or control}], f) = f(\mathbf{x}^{(\text{new})}, \text{treated or control})$. The value $f(\mathbf{x}^{(\text{new})}, \text{treated or control})$ represents the prediction for a person with covariates $\mathbf{x}^{(\text{new})}$. Section 5.2 develops this in detail.

In both scenarios, we maintain the canonical assumptions of overlap, Stable Unit Treatment Value Assumption (SUTVA), and conditional ignorability, and we use a quadratic loss function $L(\mathbf{Z}, f) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i, 1_{[i\,\text{treated}]}))^2$, where $\mathbf{Z} = \{[\mathbf{x}_i, 1_{[i\,\text{treated}]}, y_i]\}_{i=1}^n$ is the observed data. There are no hyperparameters, so we suppress their notation in the loss function.

### 5.1. Scenario 1: Average Treatment Effect

The goal is to find the range of treatment effects, $\beta_0$, corresponding to all possible ways the analyst can hack the observed data subject to a constraint on the loss. Thus, our goal is to solve

$$\max_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \beta_0 \quad \text{s.t.} \quad \sum_{i=1}^n \left(y_i - \mathbf{x}_i^T \beta - \beta_0 1_{[i\,\text{treated}]}\right)^2 \le \theta \quad \text{and} \tag{14}$$

$$\min_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \beta_0 \quad \text{s.t.} \quad \sum_{i=1}^n \left(y_i - \mathbf{x}_i^T \beta - \beta_0 1_{[i\,\text{treated}]}\right)^2 \le \theta. \tag{15}$$

This is a convex quadratically constrained linear program. Because there are inequality constraints, we require the full KKT conditions (the method of Lagrange multipliers does not handle inequality constraints). As it turns out, answers to these problems can be found analytically. This is one of the rare problems for which a subset of the KKT conditions can be used to find a closed form solution. The proof is available in Online Appendix D.

**Theorem 1** (Hacking Interval for Least Squares ATE)**.** *Define the following:*

- $\boldsymbol{\beta}_{LS}^* := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, *the optimal least square solution from regressing* $\mathbf{Y}$ *on* $\mathbf{X}$.
- $\tilde{\boldsymbol{\beta}}_{LS}^* := (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}$, *the optimal least square solution from regressing* $\mathbf{Y}$ *on* $\tilde{\mathbf{X}} := [\mathbf{X}, \mathbf{1}_{[treated]}]$. *The coefficient within this vector for the treatment variable is denoted* $\tilde{\beta}_{0,LS}^*$.
- $\gamma_{LS}^* := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{1}_{[treated]}$, *the optimal least square solution from regressing* $\mathbf{1}_{[treated]}$ *on* $\mathbf{X}$.
- $V_{tt} := ([\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}]^{-1})_{tt}$, *the diagonal entry corresponding to the treatment variable of* $[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}]^{-1}$.
- $SSE := (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{LS}^*)^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{LS}^*)$, *the sum of squared errors of the optimal least square solution.*

*Then, the solutions of the optimization problem* (14) *are*

$$\begin{aligned} \beta_{0,\max}^* &= \tilde{\beta}_{0,LS}^* + \sqrt{V_{tt}}\sqrt{\theta - SSE}, \\ \boldsymbol{\beta}_{\max}^* &= \boldsymbol{\beta}_{LS}^* - \beta_{0,\max}^*\gamma_{LS}^*, \end{aligned} \tag{16}$$

*and the solutions of the optimization problem* (15) *are*

$$\begin{aligned} \beta_{0,\min}^* &= \tilde{\beta}_{0,LS}^* - \sqrt{V_{tt}}\sqrt{\theta - SSE}, \\ \boldsymbol{\beta}_{\min}^* &= \boldsymbol{\beta}_{LS}^* - \beta_{0,\min}^*\gamma_{LS}^*. \end{aligned} \tag{17}$$

From this theorem, one can see that the range $\beta_{0,\max}^* - \beta_{0,\min}^*$ scales as the square root of the permitted level of optimality $\theta$. The solution is not difficult to find if the relevant KKT conditions are substituted into each other in a particular order.

Next, we relate the new confidence intervals to the standard ones and then produce new interpretations for confidence intervals, based on in-sample error increases. In the process, we will discuss possible meanings for the user-defined parameter $\theta$.

#### 5.1.1. Relationship to Classical Confidence Intervals.

We have just produced a confidence interval for $\beta_0$. How does that compare with a typical confidence interval produced using the standard approach where we assume a null distribution? The confidence interval is symmetric in both cases around the least squares solution, so we must be able to equate them. We next equate traditional confidence intervals with our confidence intervals, which relates $\alpha$ for a significance test with $\theta$ for our robust confidence interval.

**Theorem 2** (ATE Hacking Intervals and Standard Confidence Intervals)**.** *Start with a standard confidence interval for* $\beta_0$ *under usual assumptions (normality of errors given a linear model), which is given by*

$$\begin{aligned} &\left[\tilde{\beta}_{0,LS}^* - t_{(1-\alpha/2),(n-p-1)}\sqrt{\frac{SSE}{n-p-1}}\sqrt{V_{tt}},\right. \\ &\left.\tilde{\beta}_{0,LS}^* + t_{(1-\alpha/2),(n-p-1)}\sqrt{\frac{SSE}{n-p-1}}\sqrt{V_{tt}}\right], \end{aligned} \tag{18}$$

*where* $t_{(1-\alpha/2),(n-p-1)}$ *is the* $1 - \alpha/2$ *quantile of a t distribution with* $n - p - 1$ *degrees of freedom (we estimate p coefficients plus the treatment variable). Then, in order to keep the new confidence interval from Theorem* 1 *the same* *as the standard one, we would take the following value for* $\theta$:

$$\theta = SSE\left(1+\frac{t_{(1-\alpha/2),(n-p-1)}}{n-p-1}\right).$$

Thus, for teaching purposes, rather than explaining the $t$ distribution or the meaning of $\alpha$ to a student unfamiliar with these topics, we can explain $\theta$ first and later convert to $\alpha$ for those who want this interpretation.

### 5.1.2. Nonclassical Choices for $\theta$.

In classical hypothesis testing, one would choose the significance level $\alpha$ and say that if the data were drawn repeatedly from the true model, the probability that an estimated value of $\beta_0$ would be within the confidence interval with probability at least $1-\alpha$. We propose *in-sample* alternatives based on the meaning of $\theta$. Here are some natural choices.

- Choose $\theta$ as a percentage of the $SSE$. Assume the user would not allow a model that would achieve more than 10% higher error than the $SSE$. Then, we set $\theta = 1.1 \cdot SSE$. Generally, if we do not tolerate more than $r$% error higher than the $SSE$, we would choose $\theta = (1+r)SSE$.

To use this, we would ask questions like: "If we were to tolerate any type of change to the data or model that would incur an additional error of 10%, what are the largest and smallest treatment effects one could estimate?" If the answer is that the treatment effect estimate is robust to 10% error because of user hacking, then the estimate is reliable inside the hacking interval.

- Choose $\theta$ as the minimum loss suffered to allow the treatment effect coefficient to be zero. Let us say without loss of generality that the estimated treatment effect coefficient is negative. Then, the upper confidence interval is (using Theorem 1)

$$\beta^*_{0,\max} = \beta^*_{0,LS} + \sqrt{V_{tt}}\sqrt{\theta - SSE}.$$

- We can set this value to zero, which would provide the minimum sacrifice in least square error necessary for that coefficient to become zero. We thus need to solve for $\theta_0$ in the following:

$$0 = \beta^*_{0,LS} + \sqrt{V_{tt}}\sqrt{\theta_0 - SSE} \Longleftrightarrow \theta_0 = \frac{\left(\beta^*_{0,LS}\right)^2}{V_{tt}} + SSE.$$

- In other words, we would need to sacrifice a least squared error of at least $(\beta^*_{0,LS})^2/V_{tt}$ beyond that of the optimal solution in order that the regression coefficient could be zero.

To use this, we would ask questions like: "How much loss would need to be sacrificed in order for the treatment to have the opposite estimated effect?"

### 5.1.3. Combining with Data Variance.

The bounds of the hacking interval, $\beta^*_{0,\max}$ and $\beta^*_{0,\min}$, are deterministic functions of a fixed data set $[\tilde{\mathbf{X}}, \mathbf{Y}]$. If we assume the outcomes $\mathbf{Y}$ are one possible realization of a ground truth linear process given by

$$\mathbf{Y} \sim N(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 I), \tag{19}$$

then the bounds of the hacking interval are random variables. The following theorem gives their variance.

**Theorem 3.** (Variance of Least Squares ATE Hacking Interval Bounds). *If outcomes* $\mathbf{Y}$ *are generated by* Equation (19) *and the threshold* $\theta$ *is set to* $(1+r)SSE$ *for any* $r>0$, *then the variance of both hacking interval bounds* $\beta^*_{0,\min}$ *and* $\beta^*_{0,\max}$ *given by Equations* (17) *and* (16), *respectively, is:*

$$\mathbb{V}\left[\beta^*_{0,\min} \mid \tilde{\mathbf{X}}\right] = \mathbb{V}\left[\beta^*_{0,\max} \mid \tilde{\mathbf{X}}\right] = \sigma^2 V_{tt}\left(1 + r\left(n-p-1-\mu^2\right)\right), \tag{20}$$

*where*

$$\mu = \left(\frac{\sqrt{2}\Gamma\left((n-p)/2\right)}{\Gamma\left((n-p-1)/2\right)}\right). \tag{21}$$

### 5.1.4. Illustration.

Let us consider an illustrative example. We suppose a ground truth with two covariates called $v_{\cdot 1}$ and $v_{\cdot 2}$ chosen uniformly and independently over the interval [1,5], a 1/2 probability of treatment assignment for each observation, and outcomes generated by the following process:

$$y_i = 2 \times 1_{[\text{treated}]} + v_{i1} + v_{i2} + \epsilon_i, \tag{22}$$

where $\epsilon_i \sim N(0,1)$. In this illustration, the researcher observes more than $v_{i1}$, $v_{i2}$, and the treatment indicator, $1_{[i \text{ treated}]}$. We assume they observe monomials $\mathbf{x}_i = (v_{i1}, v_{i2}, v_{i1}^2, v_{i2}^2, v_{i1}v_{i2}, v_{i1}v_{i2}^2, v_{i1}^2 v_{i2}, v_{i1}^2 v_{i2}^2)$ and $1_{[i \text{ treated}]}$. In the language of Online Appendix C.2, $\{[v_{i1}, v_{i2}, 1_{[i \text{ treated}]}, y_i]\}_{i=1}^n$ is the pristine data, and $\{[\mathbf{x}_i, 1_{[i \text{ treated}]}, y_i]\}_{i=1}^n$ is the observed data. This puts the researcher at risk for overfitting the observed covariates in $\mathbf{x}_i$ that are not part of the ground truth.

We simulated a data set of $n = 500$ observations and used Theorem 1 to find the values of $\boldsymbol{\beta}^*_{\max}$, $\beta^*_{0,\max}$, $\boldsymbol{\beta}^*_{\min}$, and $\beta^*_{0,\min}$, where $\theta$ was set to 10% higher than the least squares loss of $\boldsymbol{\beta}^*_{LS}$. Table 1 gives the results for the coefficient on treatment indicator, $\beta_0$. To illustrate these results, on a grid of $v_{\cdot 1}^{\text{new}}$ and $v_{\cdot 2}^{\text{new}}$, we found the vector of monomials that would be observed by the researcher, $\mathbf{x}^{(\text{new})}$, and evaluated the four possible

**Table 1.** Minimum, Least Squares, and Maximum Coefficients on the Treatment Indicator

| $\beta^*_{0,\min}$ | $\beta^*_{0,LS}$ | $\beta^*_{0,\max}$ |
|---|---|---|
| 1.52 | 2.16 | 2.80 |

*Notes.* $[\beta^*_{0,\min}, \beta^*_{0,\max}]$ is the tethered hacking interval. The ground truth is $\beta_0 = 2$.

outcome predictions (maximum and minimum, treatment and control):

$$\hat{y}_{\max,\text{treated}} = \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^*_{\max} + \mathbf{1} \times \beta^*_{\mathbf{0},\max} \quad (23)$$

$$\hat{y}_{\min,\text{treated}} = \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^*_{\min} + \mathbf{1} \times \beta^*_{\mathbf{0},\min} \quad (24)$$

$$\hat{y}_{\min,\text{untreated}} = \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^*_{\min} + \mathbf{0} \times \beta^*_{\mathbf{0},\min} \quad (25)$$

$$\hat{y}_{\max,\text{untreated}} = \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^*_{\max} + \mathbf{0} \times \beta^*_{\mathbf{0},\max}. \quad (26)$$

Equations (23)–(26) are ordered by value, from largest to smallest. This gives four surface plots, shown in Figure 3 from different rotations. Asymptotically, or if we had a larger number of points, the curves would be hyperplanes because the ground truth in Equation (22) depends linearly on $v_{\cdot 1}$ and $v_{\cdot 2}$. As it stands, the curves are very close to the optimal hyperplanes, overfitting only slightly.

We would like to consider *individual* treatment effects, where the treatment effects can differ between units. The simple regression setting will predict a constant treatment effect for all units, so we need to have a more flexible modeling approach.

## 5.2. Scenario 2: Individual Treatment Effect

For the second regression scenario we consider, the regression model is more flexible, including separate terms for treatment and control. Our goal is to find the range of treatment effects for a particular point $\mathbf{x}^{(\text{new})}$.

To explain the motivation for this problem, let us consider a new patient receiving a prediction of the expected treatment effect for a drug. Before taking the drug, the patient might want to know whether there are other reasonable models that give different predictions. That is, the patient might want to know the answer to the following: *considering all reasonable models for predicting treatment effects, what are the largest and smallest possible predicted treatment effects for this drug on me?*

To determine the range, we solve

$$\max_{\beta,\beta_0} f\left(\mathbf{x}^{(\text{new})}\right) \text{ s.t. } \sum_{i=1}^{n}\left(f\left(\mathbf{x}_i^{(\text{new})}\right) - y_i^{(\text{new})}\right)^2 \le \theta.$$

$$\min_{\beta,\beta_0} f\left(\mathbf{x}^{(\text{new})}\right) \text{ s.t. } \sum_{i=1}^{n}\left(f\left(\mathbf{x}_i^{(\text{new})}\right) - y_i^{(\text{new})}\right)^2 \le \theta.$$

The model is

$$\begin{aligned} f(\mathbf{x}, \text{treated or control}) &= 1_{\text{control}}\left[\beta_1^c x_{\cdot 1} + \beta_2^c x_{\cdot 2} + \ldots \beta_p^c x_{\cdot p}\right] \\ &\quad + 1_{\text{treated}}\left[\beta_1^t x_{\cdot 1} + \beta_2^t x_{\cdot 2} + \ldots \beta_p^t x_{\cdot p}\right]. \end{aligned}$$

Using notation $w_i = 1$ for treatment points and $w_i = 0$ for control points, the least squares loss thus decouples, leading to separate regression problems for the treatment and control points:

$$\begin{aligned} &\sum_{i=1}^{n}\left(f(\mathbf{x}_i, w_i) - y_i\right)^2 \\ &\quad = \sum_{i:w_i=1}\left(f(\mathbf{x}_i, 1) - y_i\right)^2 + \sum_{i:w_i=0}\left(f(\mathbf{x}_i, 0) - y_i\right)^2 \\ &\quad = \sum_{i:w_i=1}\left(\left[\beta_1^c x_{i1} + \beta_2^c x_{i2} + \ldots \beta_p^c x_{ip}\right] - y_i\right)^2 \\ &\qquad + \sum_{i:w_i=0}\left(\left[\beta_1^t x_{i1} + \beta_2^t x_{i2} + \ldots \beta_p^t x_{ip}\right] - y_i\right)^2. \end{aligned}$$

Because the first sum involves only the control observations and control coefficients and the second sum involves only treatment observations and treatment coefficients, this decouples as two separate regressions, one for the control group and one for the treatment group. We will assume that the user wants neither of the regressions to be too suboptimal, so we will have separate constraints $\theta$ on the quality of each regression. We will find the maximum and minimum values for the control regression and the treatment regressions (four values). All of these optimization problems are very similar, so for simplicity, we solve the optimization problem on a generic regression problem, for point $\mathbf{x}^{(\text{new})}$. Here, $\mathbf{x}^{(\text{new})}$ does not need to be one of the training observations.

**Theorem 4** (Hacking Intervals for Least Squares Individual TE)**.** *Consider the hacking interval optimization problems:*

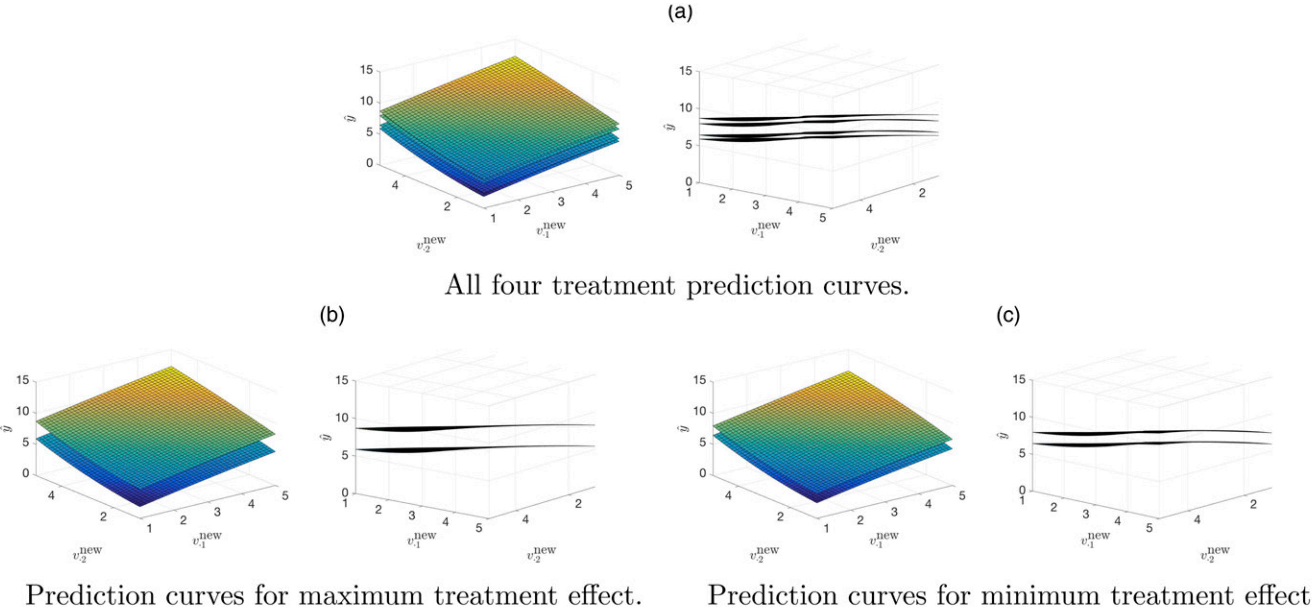
$$\max_{\beta}\left(\mathbf{x}^{(\text{new})T}\boldsymbol{\beta}\right) \text{ s.t. } \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 \le \theta,$$

$$\min_{\beta}\left(\mathbf{x}^{(\text{new})T}\boldsymbol{\beta}\right) \text{ s.t. } \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 \le \theta.$$

*Define* $\boldsymbol{\beta}^*_{LS} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$*; define* $\Upsilon = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{(\text{new})}$*, which is a vector of size p, SSE* $= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*_{LS}\|^2$*, and*

$$\tilde{\mu} = \frac{\sqrt{\theta - SSE}}{\|\mathbf{X}\Upsilon\|}.$$

**Figure 3.** (Color online) For Each of the Three Panels, the Left and Right Panels Are Two Different Vantage Points of the Same Figure



All four treatment prediction curves.

Prediction curves for maximum treatment effect.

Prediction curves for minimum treatment effect

*Notes.* Because the true data generation process in Equation (22) depends linearly on only $v_{\cdot 1}$ and $v_{\cdot 1}$, the optimal prediction curve as a function of $v_{\cdot 1}$ and $v_{\cdot 1}$ is a hyperplane. The addition of monomials to the observed $\mathbf{x}$ causes some overfitting. (a) All four prediction curves (max/min, treatment/control). (b) Prediction curves that yield the maximum treatment effect. The upper curve shows $\hat{y}_{\text{max,treated}}$, and the lower curve shows $\hat{y}_{\text{max,untreated}}$. The difference between the curves is the maximum treatment effect, $\beta^*_{0,\text{max}}$. These curves correspond to the top and bottom curves in panel (a). (c) Prediction curves that yield the minimum treatment effect. The upper curve shows $\hat{y}_{\text{min,treated}}$, and the lower curve shows $\hat{y}_{\text{min,untreated}}$. The difference between the curves is the minimum treatment effect, $\beta^*_{0,\text{min}}$. These curves correspond to the middle two curves in panel (a).

*The solutions to the optimization problems are*

$$\boldsymbol{\beta}^*_- = \boldsymbol{\beta}^*_{LS} - \tilde{\mu}\Upsilon, \qquad \boldsymbol{\beta}^*_+ = \boldsymbol{\beta}^*_{LS} + \tilde{\mu}\Upsilon.$$

**Theorem 5** (Individual TE Hacking Intervals and Standard Confidence Intervals). *Start with a standard confidence interval for $\mathbf{x}^{(new)T}\boldsymbol{\beta}$ under usual assumptions (normality of errors given a linear model), which is given by the boundary points:*

$$\boldsymbol{\beta}^*_{LS} \pm t_{(1-\alpha/2),(n-p-1)}\sqrt{\frac{SSE}{n-p-1}}\sqrt{\mathbf{x}^{(\text{new})T}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}^{(\text{new})}},$$

*where $t_{(1-\alpha/2),(n-p-1)}$ is the $1-\alpha/2$ quantile of a t distribution with $n-p-1$ degrees of freedom. Then, in order to keep the hacking interval from Theorem 4 the same as the standard one, we would take the following value for $\theta$:*

$$\theta = SSE\left(1 + \frac{t^2_{(1-\alpha/2),(n-p-1)}}{n-p-1}\right).$$

We can use the result of Theorem 4 to determine the hacking interval, which in this case, is the range of causal effect estimates for $\mathbf{x}^{(\text{new})}$. Let us apply Theorem 4 to the treatment regression and the control regression separately. We thus obtain $\beta_+^{t*}$, $\beta_-^{t*}$, $\beta_+^{c*}$, and $\beta_-^{c*}$. To find the maximum of the causal effect estimate, use

$$\max\left(\mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{t}*}_+, \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{t}*}_-\right) - \min\left(\mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{c}*}_+, \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{c}*}_-\right).$$

To find the minimum of the causal effect estimate, use

$$\min\left(\mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{t}*}_+, \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{t}*}_-\right) - \max\left(\mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{c}*}_+, \mathbf{x}^{(\text{new})T}\boldsymbol{\beta}^{\mathbf{c}*}_-\right).$$

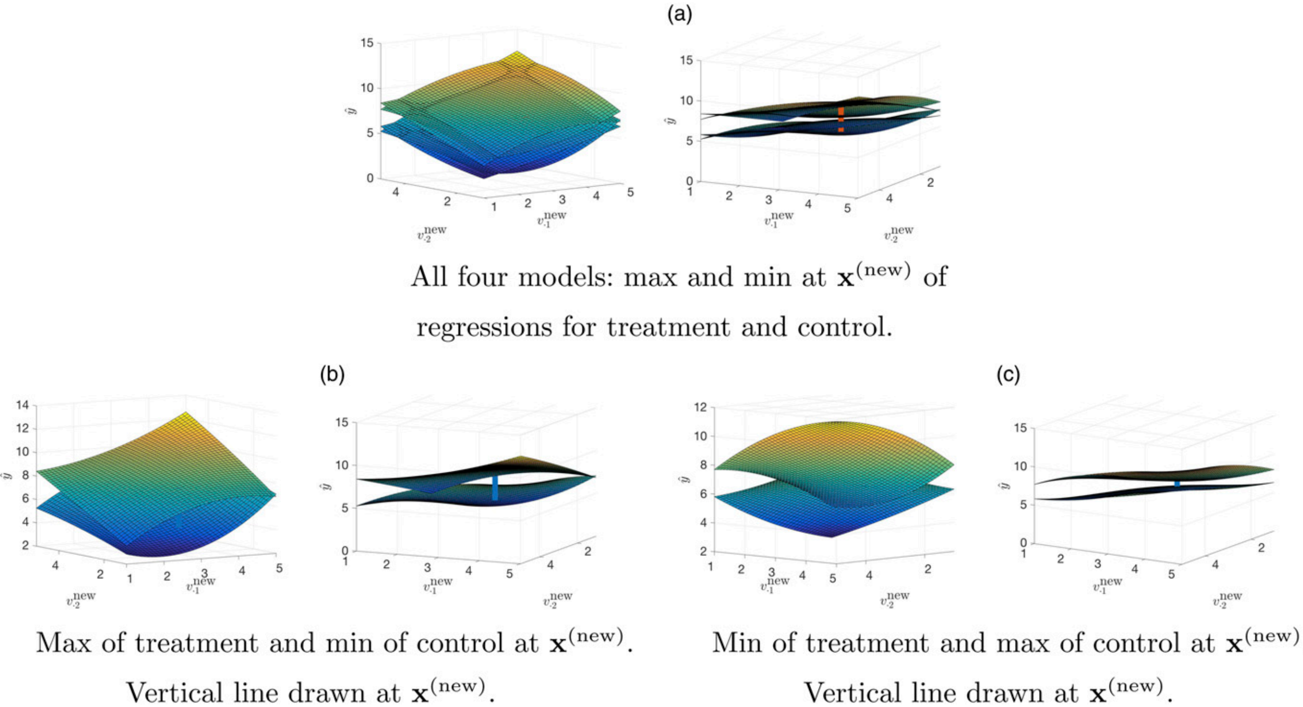
#### 5.2.1. Illustration.

We continue with the same data generation process we used in Section 5.1.4, where the ground truth outcomes are created as follows:

$$y_i = 2 \times 1_{[\text{treated}]} + v_{i1} + v_{i2} + \epsilon.$$

We chose $\mathbf{x}^{(\text{new})}$ to be created from the point $v_1^{\text{new}} = 3$, $v_2^{\text{new}} = 2$. Here, we created four separate regressions. One regression maximizes the expected outcome at $\mathbf{x}^{(\text{new})}$ for the treatment observations. Another regression minimizes the expected outcome at $\mathbf{x}^{(\text{new})}$ on the treatment observations. Analogous regressions are created for the control observations. Figure 4 shows these regressions explicitly for $\mathbf{x}^{(\text{new})} = (3,2)^T$. One can see the regressions starting to bend away from each other at $\mathbf{x}^{(\text{new})}$ for the maximization problem and bend toward each other for the minimization

**Figure 4.** (Color online) For All Three Panels, the Left and Right Panels Are Two Different Vantage Points of the Same Figure



All four models: max and min at $\mathbf{x}^{(\text{new})}$ of regressions for treatment and control.

Max of treatment and min of control at $\mathbf{x}^{(\text{new})}$. Vertical line drawn at $\mathbf{x}^{(\text{new})}$.

Min of treatment and max of control at $\mathbf{x}^{(\text{new})}$. Vertical line drawn at $\mathbf{x}^{(\text{new})}$.

*Notes.* (a) All four regressions (max/min, treatment/control). (b) Maximizing the gap between treatment and control at $\mathbf{x}^{(\text{new})}$. The upper curve is the regression for maximizing expected outcomes on the treated at $\mathbf{x}^{(\text{new})}$. The lower curve is the regression for minimizing expected outcomes on the control units at $\mathbf{x}^{(\text{new})}$. One can see how the curves pull away from each other at $\mathbf{x}^{(\text{new})}$ to make the differences between treatment and control as large as possible. (c) Minimizing the gap between treatment and control at $\mathbf{x}^{(\text{new})}$. The upper curve is the regression for minimizing outcomes on the treatment units at $\mathbf{x}^{(\text{new})}$. The lower curve is the regression for maximizing the control outcomes at $\mathbf{x}^{(\text{new})}$. Here, the curves pull toward each other to minimize the estimated treatment effect.

problem. We placed a blue line between the curves at the point $\mathbf{x}^{(\text{new})}$.

## 6. Application: Recidivism Prediction

Understanding the potential impact of researcher choices on machine learning methods becomes especially important when issues of fairness are involved. Although there does not exist a widely accepted mathematical definition of fairness when assessing risk with machine learning (Berk et al. 2018), if a machine learning method could reach opposing conclusions about a person or group of persons were small adjustments to a data set or hyperparameters made, then this could potentially undermine any definition of fairness (one could simply argue a negative decision to be unfair because an equally good model exists that predicts the opposite). A hacking interval quantifies the degree to which this can happen.

In the criminal justice system, algorithms are increasingly being used to make risk assessments about defendants: for example, their risk of failing to appear in court or reoffending. Clearly, issues of fairness are involved. One such algorithm is COMPAS, created by Northpointe, Inc. COMPAS produces three decile scores that indicate the risk that a defendant will fail to reappear in court, reoffend, or violently reoffend. As of October 2017, it was used by 4 of 58 counties in California (Back et al. 2017). It is a proprietary algorithm that bases its assessment on a questionnaire that is either pulled from criminal records or answered by the defendant. The data gathered by the questionnaire are not publicly available. *ProPublica* assembled COMPAS scores and other data—including criminal history and demographic information—on more than 7,000 defendants in Broward County, Florida, from 2013 to 2014 with the help of the Broward County Sheriff's Office (Angwin et al. 2016). Using the same metric used by Northpointe, Inc.—whether a defendant was charged with a crime within two years of the COMPAS score calculation—*ProPublica* concluded that COMPAS was biased against African Americans. For example, they found that of African-American defendants who did not reoffend, 45% were misclassified as higher risk, whereas of Caucasian defendants who did not reoffend, only 23% were misclassified as higher risk. Northpointe, Inc. has issued a rebuttal that argues a definition of fairness based on a false-positive rate is not appropriate in this case (Dieterich et al. 2016). Angelino et al. (2018) argue that, although COMPAS is ostensibly not influenced by race, its dependence on prior record could effectively induce dependence on race because of disproportionate arrest rates that count toward one's prior record. This agrees with the sentiment of other work on interpretable models for recidivism (Zeng et al. 2017). More work on this data set has provided further insight into how COMPAS may depend on prior record as well as age (Rudin et al. 2020).

In our analysis, we use the data collected by *ProPublica*, but our interest is not in comparing a risk assessment score like COMPAS against a given definition of fairness. Rather, we are interested in the impact that researcher choices could have on conclusions made about this data set. In Section 6.1, we use the methods of Section 3.1.2 to assess the impact that a new feature created by the researcher could have on inferences about the population, in this case the odds ratio of reoffending and gender. This is an example of a prescriptively constrained hacking interval because we explicitly constrain researcher choices about the new feature. In Section 6.2, we use the methods of Section 4.1 to assess the impact of researcher choices on the predictions of a support vector machine about individual defendants. This is an example of a tethered hacking interval because we constrain researcher choices only through their impact on the loss function. For both applications, we use the following set of features:

- c_charge_degree_F: binary indicator if the most recent charge prior to the COMPAS score calculation is a felony.
- sex_Male: binary indicator if the defendant is male.
- age_screening: age in years at the time of the COMPAS score calculation.
- age_18_20, age_21_22, age_23_25, age_26_45, and age__45: binary indicators based on age_screening for age groups 18–20, 21–22, 23–25, 26–45, and greater than 45, respectively.
- juvenile_felonies__0, juvenile_misdemeanors__0, and juvenile_crimes__0: binary indicators one or more juvenile felony, misdemeanor, or crime, respectively. We use binary indicators because the counts of each are highly right skewed.
- priors__0, priors__1, priors_2_3, and priors__3: binary indicators of whether the number of priors is zero, one, two to three, or more than three, respectively.

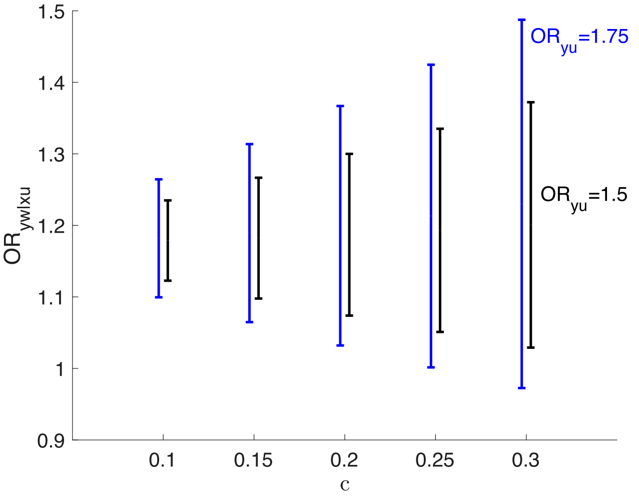
We filtered the data set to include only defendants whose most recent charge prior to the COMPAS score calculation was a felony or misdemeanor and occurred at most 30 days prior to the COMPAS score calculation (otherwise, we assume this charge did not trigger the COMPAS score calculation, so it seems that data about this defendant are missing). The binary indicator variables for age and number of priors were added to the data set because in general, recidivism is highly nonlinear with respect to these features.

### 6.1. Prescriptively Constrained Example: Adding a New Feature

We suppose a researcher is interested in the odds ratio between gender and recidivism but is allowed to create a new binary feature $u$, perhaps as a function of the existing features or by introducing new data. Notice that this is not a valid causal question because gender is not assignable, but we only use the mathematical tools of causal sensitivity analysis. A benefit of this approach is that we do not need to understand exactly what the new feature is, only its relationship to the outcome $y$ (whether a defendant reoffends) and "treatment" $w$ (gender). In the setup described in Section 3.1.2, this means the researcher specifies constraints $OR_{yu} \in [a, b]$, $|p_1 - p_0| \le c$, and $p_0 \ge d$ (by specifying $a$, $b$, $c$, and $d$), where $p_0 := p(U \mid w = 0)$, $p_1 := p(U \mid w = 1)$. We will use a simple version where $OR_{yu}$ is fixed (or equivalently, $a = b = OR_{yu}$). As shown in Section 3.1.2, the hacking interval can be calculated as a function of $c$.

Figure 5 shows hacking intervals for $OR_{yw|\mathbf{x},u}$—the odds ratio between recidivism and gender adjusted for the observed covariates $\mathbf{x}$ and the new feature $u$—for each combination of $c \in (0.1, 0.15, 0.2, 0.25, 0.3)$ and $OR_{yu} \in (1.5, 1.75)$. These constraints are picked arbitrarily for illustration. In practice, the choice of these constraints describes the degree of freedom given to the researcher. For example, if the researcher were permitted to pick any new binary feature $u$ such that the odds ratio between the outcome and the new feature was $OR_{yu} = 1.5$ and the difference between $p_1$ and $p_0$ (the probability of the new feature when the treatment $w$ is present or not present, respectively) was constrained to be less than or equal to $c = 0.3$, then the value of $OR_{yw|\mathbf{x},u}$ they could get would necessarily be in the hacking interval $[1.03, 1.37]$. For the same

**Figure 5.** (Color online) Hacking Intervals for $OR_{yw|\mathbf{x},u}$ for Different Values of Constraints $c$ and $a = b = OR_{yu}$



restriction of $c = 0.3$, if the researcher was permitted to pick $u$ such that $OR_{yu} = 1.75$, indicating a stronger relationship between the new feature and the outcome, then she could obtain a value of $OR_{yw|\mathbf{x},u}$ above or below 1 because Figure 5 shows that the hacking interval in this case overlaps with 1. In other words, with this freedom given to the researchers, they could conclude that the odds ratio between recidivism and gender, after controlling for measured covariates and the new covariate they created, could be above or below one.

## 6.2. Tethered Example: SVM

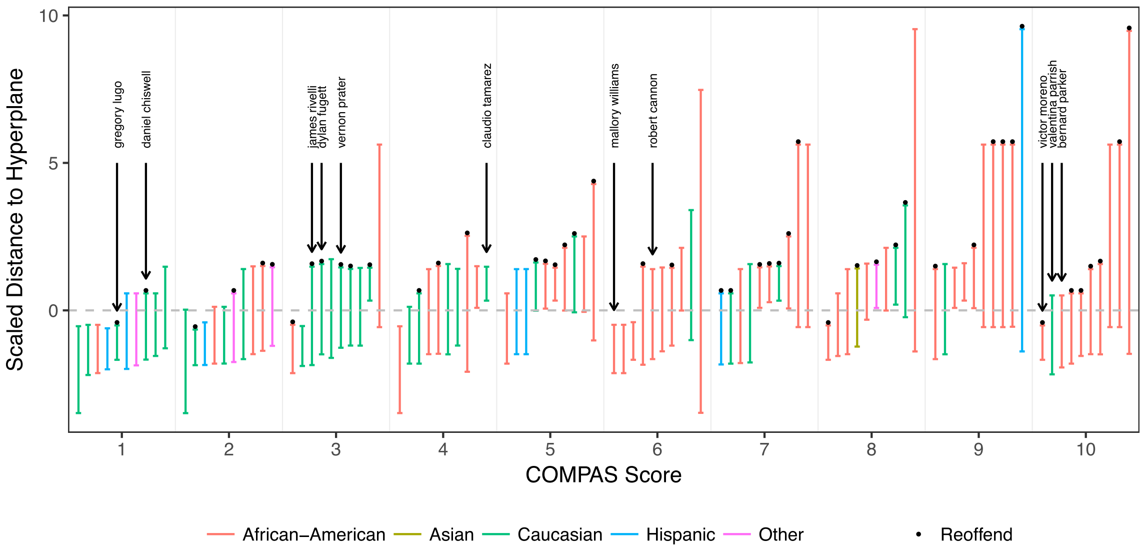
We now consider the impact of researcher hacking on predictions of two-year recidivism for individual defendants. We use an SVM as our predictive model. For prediction on a new defendant represented by $\mathbf{x}^{(\text{new})}$, SVM calculates the distance of $\mathbf{x}^{(\text{new})}$ to the hyperplane that minimizes the hinge loss. If the distance is positive, the model predicts the defendant will reoffend within two years. If the distance is negative, the model predicts the defendant will not reoffend within two years. By adjusting the hyperplane, the tethered hacking interval is the range of distances of $\mathbf{x}^{(\text{new})}$ to the hyperplane that can be achieved within a constraint on the loss. As discussed in Section 4.1, we can find this range of values by solving the dual problem in Equation (13) for $s = -1$ and $s = 1$. We do this using the `fmincon` function in MATLAB. We thus solved two optimization problems for each defendant.

Figure 6 shows the hacking intervals for 10 selected defendants from each group of COMPAS scores. We included a few individuals highlighted in an article by *ProPublica* (Angwin et al. 2016) and randomly selected the rest. The loss is constrained to be within 5% of the minimum loss on a group of 1,000 defendants randomly selected from the remaining defendants (so, each prediction in Figure 6 is out of sample).

Consider three possible cases: (i) the hacking interval is entirely below zero, (ii) the hacking interval is entirely above zero, or (iii) the hacking interval overlaps with zero. In case (i), this means that there does not exist an SVM model such that the loss on the 1,000 training observations is within 5% of the minimum loss, and the model predicts that the defendant will reoffend; all "reasonable" models (i.e., within this loss constraint) predict that the defendant *will not* reoffend. In case (ii), when the hacking interval is entirely above zero, the interpretation is the same except that all reasonable models predict that the defendant *will* reoffend. In case (iii), when the hacking interval overlaps with zero, then reasonable SVM models exist that make either prediction. Although this is only a sample of the data, notice that of the 10 defendants shown here with COMPAS scores of 10—the riskiest possible COMPAS score—9 of them have hacking intervals that overlap with zero. On the other hand, of the 10 people shown here with COMPAS scores of 1— the least risky COMPAS score—5 of them have hacking intervals entirely above zero.

In the *ProPublica* article (Angwin et al. 2016), several pairs of defendants are highlighted. For each pair, one defendant received a low COMPAS score despite a significant criminal history, whereas the other received a high COMPAS score despite a limited

**Figure 6.** (Color online) SVM Hacking Intervals for 10 Defendants for Each COMPAS Score



*Note.* Loss is constrained to be within 5% of the minimum loss on a random sample of 1,000 defendants.

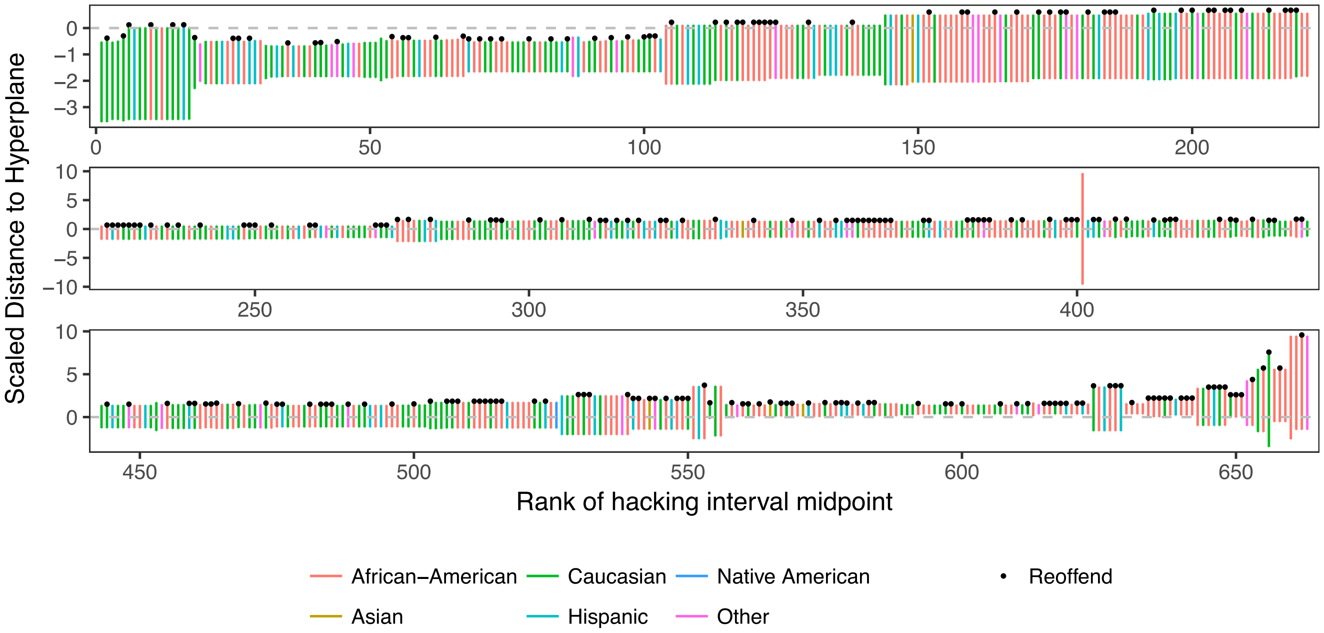
criminal history. For example, James Rivelli and Robert Cannon were both charged with theft, but Rivelli was charged with felony grand theft and possession of heroin, whereas Cannon was charged with misdemeanor petit theft. In addition, Rivelli had three prior arrests, including for felony aggravated assault and felony grand theft, whereas Cannon had none. Despite this, Rivelli—who is white—received a low-risk COMPAS score of 3, whereas Cannon—who is black—received a medium-risk COMPAS score of six. Rivelli later reoffended in Broward County with grand theft again, whereas Cannon did not. Interestingly, the hacking intervals for both defendants overlapped with zero, indicating that justifiable SVM models (on our limited feature set) could have made either prediction. The hacking intervals also overlap with zero for the similarly contrasting pair of Bernard Parker and Dylan Fugett, both arrested on drug charges. For the pair of Vernon Prater and Brisha Borden, both arrested on petty theft charges, the more experienced criminal Prater also has a hacking interval that overlaps with zero, but we do not have data on Borden. The exception is Mallory Williams, who received a medium-risk COMPAS score of six after a Driving Under the Influence (DUI) arrest and only two prior misdemeanors. Her hacking interval is entirely below zero, meaning no justifiable SVM model would predict that she would reoffend in this experiment. She did not reoffend. In general, we see a high degree of uncertainty from SVM models for the individuals discussed in this article. The counterpart to Mallory Williams in the *ProPublica* article, Gregory Lugo, illustrates how offense data can be easily misinterpreted. Gregory Lugo was charged with a DUI but had zero priors according to the data we used in our analysis. Not surprisingly, his COMPAS score was low, and his hacking interval was entirely below zero. However, *ProPublica* claimed he had four priors, including three DUIs, and used this as an example of a poorly calibrated COMPAS score. This seems to be a misinterpretation of the data: all of his supposed prior offenses have the same offense date as the offense related to his COMPAS score calculation, so the supposed prior offenses seem to be rerecordings (perhaps for ordinary bureaucratic reasons) of the same offense.

There are other interesting examples in Figure 6. Claudio Tamarez, a 30-year-old Caucasian male, received a COMPAS score of 4, which means low risk, following a charge for possession of phentermine and despite 9 priors that included battery on an officer. In contrast, his hacking interval was entirely above zero. He did not, however, recidivate within the 2-year follow-up period. Daniel Chiswell, a 41-year-old Caucasian male, was assigned a COMPAS score of only one despite being charged with felony possession of heroin and having previously been charged with felony battery on an officer. His hacking interval overlapped with zero, meaning there exists a reasonable SVM model that would have predicted he would reoffend. He was charged again with felony possession of heroin later that year. Valentina Parrish, a 21-year-old Caucasian female, was charged with driving under the influence and possession of less than 20 grams of cannabis. She was given a COMPAS score of 10. In contrast, her hacking interval, $[-2.16, 0.50]$, was mostly below 0, although not entirely. She did not reoffend. There are also examples that illustrate limitations of our limited feature set. Victor Moreno, a 31-year-old African-American male, received a COMPAS score of 10 despite zero priors. However, the arrest related to his COMPAS score calculation included felony charges of battery, tampering with a victim, tampering with physical evidence, and delivering cocaine. Our SVM model, without access to the content of these charges, not surprisingly gave him a low hacking interval given his lack of prior offenses.

Figures 7 and 8 show the hacking intervals for every defendant in our data set with COMPAS scores of 3 and eight, respectively. The loss constraint is the same (within 5% of the minimum loss on the same 1,000 defendants). Of the 663 people in our data set with COMPAS scores of 3—a "low-risk" score—75 of them had hacking intervals entirely above zero. Again, this means that, had SVM been used for prediction, any reasonable model would have predicted that they would reoffend. These 75 people had an average of about 6.3 priors, and 35 of them reoffended. Conversely, of the 428 people in our data set with COMPAS scores of 8—a "high-risk" score—121 of them had hacking intervals entirely below zero, meaning any reasonable SVM model would predict that they would not reoffend. These 121 people had an average of about 8.75 priors, and 94 of them reoffended. This potentially means we may be missing data on their past criminal history that is not in the data set we use for our analysis. Although it is possible that missing information can explain COMPAS scores that are high, it cannot explain COMPAS scores that are too low.

We also show hacking intervals grouped by race in Figure 9. As before, we allow for a 5% tolerance on the loss on a sample of 1,000 defendants, but for this figure, we use a different sample of defendants. Each hacking interval in Figure 9 is out of sample (i.e., the defendant corresponding to the hacking interval was not included in the 1,000-defendant training sample used for the loss constraint). Some of the COMPAS scores again do not align with the hacking intervals. Consider Edwin Chaj, a 27-year-old Hispanic male with only one prior related to trespassing, who received a COMPAS score of nine following a charge of disorderly intoxication. In contrast to the high-risk

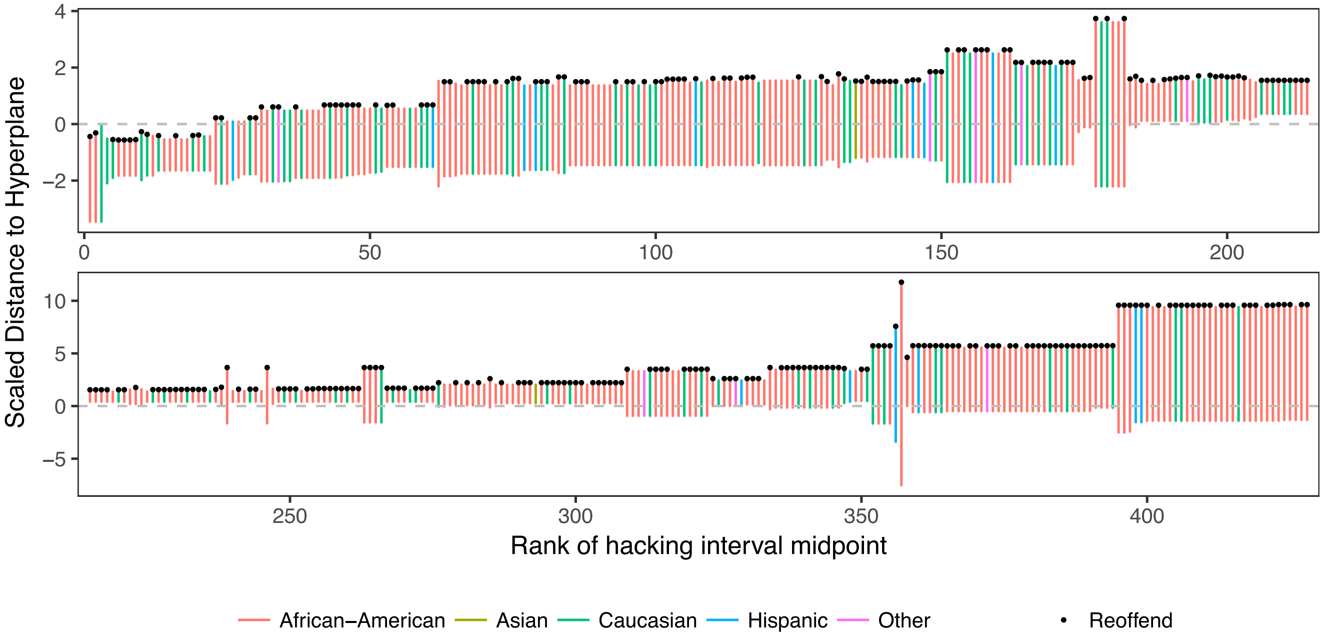
**Figure 7.** (Color online) SVM Hacking Intervals for All Defendants with a COMPAS Score of 3



*Note.* Loss is constrained to be within 5% of the minimum loss on a random sample of 1,000 defendants.

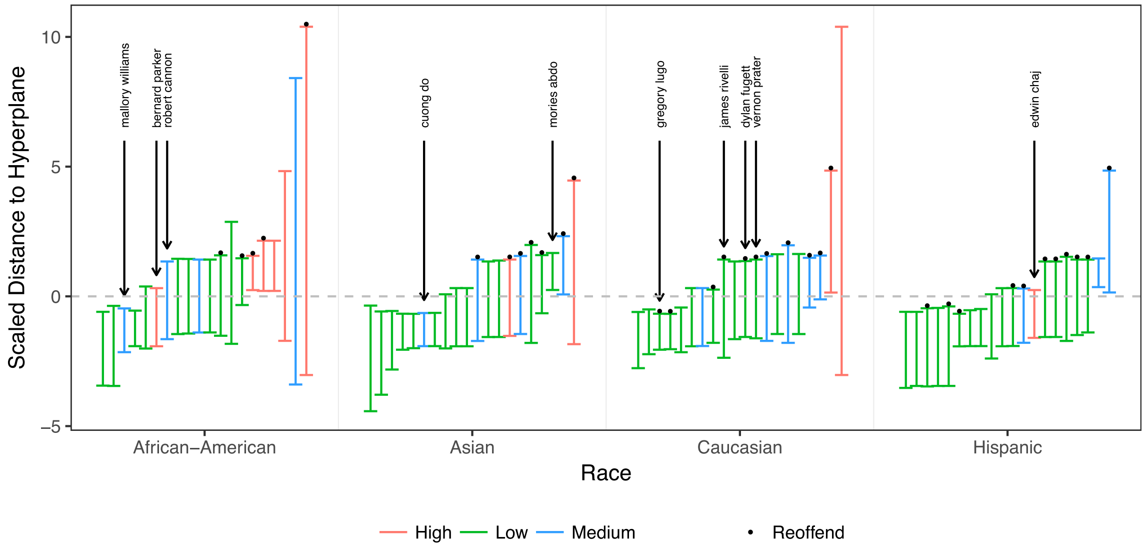
COMPAS score, his hacking interval was low ([−1.59, 0.24]), although not entirely below 0. He did not reoffend. Similarly, Cuong Do, a 32-year-old Asian male with no priors, received a COMPAS score of 8 following charges with felony burglary and misdemeanor petit theft. In contrast to the high-risk COMPAS score, his hacking interval was entirely below zero. He did not reoffend. On the other hand, consider Mories Abdo, a 27-year-old Asian male with six priors, who received a COMPAS score of 3 following a battery charge. In contrast to the low-risk COMPAS score, his hacking interval was entirely above zero. He did not reoffend during the two-year follow-up period but did commit felony aggravated assault with a firearm just after the follow-up period ended according to the Broward County Clerk of

**Figure 8.** (Color online) SVM Hacking Intervals for All Defendants with a COMPAS Score of 8



*Note.* Loss is constrained to be within 5% of the minimum loss on a random sample of 1,000 defendants.

**Figure 9.** (Color online) SVM Hacking Intervals for 10 Randomly Selected Defendants for Each Race in the Data Set (Except Native American as There Are Only 11 in the Data Set)



*Notes.* Loss is constrained to be within 5% of the minimum loss on a random sample of 1,000 defendants. Color indicates COMPAS scores (high/medium/low).

the Courts.[5] Figure 9 also indicates the individuals discussed in the *ProPublica* article. Because the 1,000-defendant training sample is different from Figure 6, the hacking intervals are slightly different, but they are each in the same category (below zero, overlapping with zero, or above zero).

We summarize Figures 6–9 with a couple observations.

- *If we had used SVM on our limited data set rather than the COMPAS score to predict reoffense, then for most people there is enough uncertainty in predictions that we could justifiably predict either reoffend or not reoffend*. This can be seen in Figures 7 and 8, where 75% and 67% of defendants with COMPAS scores of 3 and 8, respectively, have hacking intervals that overlap with zero, meaning justifiable SVM models exist that could make either prediction. Even for the extreme cases discussed in the *ProPublica* article, the hacking intervals often overlapped with zero.
- *There are many individuals for which no justifiable SVM model would agree with the COMPAS score using our feature set*. In the case of an individual with a low COMPAS score, this means the hacking interval is entirely above zero, whereas in the case of an individual with a high COMPAS score, this means the hacking interval is entirely below zero. In either case, this is suggestive of an error in the COMPAS calculation. Figure 7 shows 75 examples of the former case, and Figure 8 shows 121 examples of the latter case.

## 7. Related Work and Discussion

Hacking intervals are designed to quantify a form of uncertainty that is usually ignored in statistical inference. This could have implications for scientific research; let us discuss this first.

### 7.1. Problems with Replication of Scientific Studies and Proposed Solutions

The evidence for $p$-hacking primarily comes from two types of meta-analyses: replication studies and the distribution of $p$-values for a set of independent findings (or "$p$-curve") (Simonsohn et al. 2014). For an example of the former approach, a major 2015 study attempted to replicate 100 studies and found that very few findings could be reproduced (Open Science Collaboration 2015), although a replication of this replication found that the percentage of studies that were replicated was not statistically different from the fraction that would be expected to replicate because of chance alone (Gilbert et al. 2016). Camerer et al. (2016) found a higher initial percentage being replicable in 18 economic studies, but this still reflects a problem in the field. In commercial applications, large corporations are keenly aware of this problem: based on their own comparisons, Bayer HealthCare found that only about 20%–25% of preclinical studies were completely in line with their in-house findings (Prinz et al. 2011). Amgen replicated 11% of 53 scientific findings (Begley and Ellis 2012).

For the $p$-curve approach, a uniform distribution of $p$-values across articles indicates a lack of significant results; a right skew indicates a general existence of significant results; and a left skew, especially near the 0.05 threshold, supposedly indicates $p$-hacking. Head et al. (2015) concluded that the evidence indicates the existence of "widespread" evidence for $p$-hacking after searching all open access papers in the PubMed database (~ 100,000 papers). This type of analysis has also been contested (Bishop et al. 2016), and not all meta-analyses have found evidence of *p*-hacking (Jager and Leek 2014).

Several types of solutions to *p*-hacking have been proposed.

- We could require researchers to "preregister" the details of their study, so that they cannot selectively make choices to achieve significant results, but this rules out learning from the data in any other way.
- Another proposal is to reduce the significance threshold (Simmons et al. 2011, Humphreys et al. 2012, Gelman and Loken 2013, Monogan 2015) because when explicitly considering multiple comparisons, decreasing the threshold for significance is sensible (e.g., the Bonferroni correction). Recently, a group of 72 scientists advocated reducing it to 0.005 (Benjamin et al. 2018), which might lessen false positives but would also invalidate the quantitative meaning of the *p*-value in the first place. This is also a drastic measure, leading to a higher true-negative rate and thus, many important results being dismissed as insignificant.
- We could create Bayesian confidence intervals or Bayesian hierarchical models. In comparison with frequentist hypothesis testing, Bayesian hypothesis testing provides a more comfortable interpretation of the conclusion (the probability that the alternative hypothesis is true), but it is still subject to hacking: the introduction of a prior gives the researcher even more discretion, which may lead to more user choices (see Gelman et al. 2012 for examples of complicated priors leading to bias). If we place a prior on analysts' decisions, it is easy to argue that any given prior is wrong. An example of this, discussed earlier, is the choice of matching algorithm for treatment and control units in a matched pairs experiment. This is a case where uniform priors do not make sense, but any other choice of prior is not defensible either.
- In the case where the researcher does variable selection, post-selection *inference* can be used to adjust classical confidence intervals in order to account for the variables being chosen after examining the data. In the case of linear regression, Tibshirani et al. (2016) present a framework for *specific* variable selection procedures (forward stepwise regression, least angle regression, and the lasso regularization path), and Berk et al. (2013) present a framework that holds for *all* variable selection procedures that is more conservative than Scheffé protection (Scheffé 1959). Hacking intervals differ from post-selection inference in at least two ways. (1) Hacking intervals are more general as they could include uncertainty to many choices made by the analyst for *any* prediction problem (not just regression) and do not necessarily require independent and identically distributed (*i.i.d.*) Gaussian errors. (2) Post-selection is useful when you already have a model selected and you want to do regular inference, whereas hacking intervals consider robustness to other models that *could* have been selected. Post-selection confidence intervals can be combined with hacking intervals to account for other researcher choices.
- The work of Dwork et al. (2015) provides a method to avoid *p*-hacking in a setting where data are provided sequentially, chosen *i.i.d*. from the same distribution. Our setting is very different; in our work, the data could be subject to preprocessing, and the underlying distribution may not exist.

These solutions are obviously sometimes useful but often unfulfilling, highlighting the importance, inherent difficulty, and urgency of the problem.

### 7.2. Problems with Classical Inference That Are Easy to Overlook

Here, we highlight some drawbacks to classical inference, including frequentist, Bayesian, and fiducial inference (see Hannig et al. 2016 for a review of a modern version of fiducial inference), in the way they are used in practice and how hacking intervals can help to fix these issues.

- In cases where a superpopulation exists, the null hypothesis for data analysis is not the correct null hypothesis. The entire confidence interval calculation for an observed data set is conditional on statistical assumptions about measurement, distributions, asymptotics, and modeling, among others. Changes in any of these can greatly impact the resulting substantive conclusions, a problem known as *model dependence* (King and Zeng 2006, Iacus et al. 2011). The null hypothesis used for the analysis depends on the processed data and thus, is subject to model dependence. Let us say we want to know whether a pharmaceutical drug causes a side effect. We might process data by choosing covariates, choosing a match assignment, performing regression with a choice of regularization, and so on. The "true" null hypothesis is that the drug does not have any side effect. Instead, the null hypothesis that is actually tested is that the drug has no effect after the researcher's preprocessing is done to future instances of raw data. It is not clear which preprocessing steps will make the researcher's null hypothesis close to the true null hypothesis on the correct superpopulation. If the researcher's results are robust to a range of possible data-processing options, then this range may include processing that brings the data closer to a sample drawn from the true superpopulation. To analyze the data in this case, we would want a combination of a hacking interval (for the data-processing choices) and a regular confidence interval (for the processed data) to ensure robustness both to user manipulation and to randomness in the sample of data. We discuss such combinations in Section 5.1.3 for regression. To summarize, hacking intervals help to ensure that the conclusions about the true null hypothesis with respect to the true superpopulation are valid.

- It does not make sense to explicitly model analyst choices. In the case of Bayesian model averaging or other decision-making frameworks, one might try to model the way the analyst might treat the data and average over realistic choices an analyst might make. However, this makes little sense. The hypothesis is about the ground truth, not about researcher choices. We would like the result to be robust to *any* choices made by a reasonable researcher.

The example of matching, discussed earlier, is an example where placing a prior on analyst choices of matching method does not make sense.

### 7.3. Enumerative Approaches Similar to Prescriptively Constrained Hacking Intervals

In the social sciences, there are several works that propose enumerating all reasonable model specifications and computing the effect estimate of interest for each specification. The "extreme bound analysis" of Leamer (1983) focuses on covariate combinations; the "specification curve" of Simonsohn et al. (2015) proposes a graphical display of all effect estimates and a method for conducting joint inference across all specifications; and the method of Young and Holsteen (2015) investigates a variety of model specification types, including functional form, and develops a model influence analysis showing how each model component impacts the effect estimate. Each of these approaches proposes brute force calculation of all chosen model specifications, which can be costly. Muñoz and Young (2018) draws on the framework of Young and Holsteen (2015) to a simulated data set, fitting a total of 9 billion linear regression models on a simulated data set, but the computation takes several months. Tethered hacking intervals differ in that they aim only to identify the smallest and largest effect estimates, which permit an optimization-based approach. We also provide a variety of examples for machine learning models in addition to linear models, whereas the approaches mentioned only consider linear or generalized linear models.

### 7.4. Mathematical Equivalence of Hacking Intervals to Other Problems but with Different Meaning

In some contexts, hacking intervals bear mathematical equivalence to other problems, which means we can leverage existing methods in some cases. Prescriptively constrained hacking intervals often fall under a form of sensitivity analysis (Leamer 2010). If we consider uncertainty in the inputs to a mathematical model (usually in an applied math context), they fall under the field of uncertainty quantification. If we consider uncertainty in prior specification, they fall under robust Bayesian analysis. If we consider uncertainty in assumptions for causal inference, they fall under (causal) sensitivity analysis. See Ghanem et al. (2017), Berger et al. (1994), and Liu et al. (2013), for overviews of these fields, respectively. Uncertainty quantification provides useful computational tools, like Monte Carlo simulation and surrogate models (Sudret et al. 2017). In the latter two methods, theoretical bounds on effect estimates have been proven. Berger (1990) determines the range of a posterior quantity for priors contained in a certain class. In linear regression, difference in betas (DFBETAS) and difference in fits (DFFITS) measure the change in a coefficient and a prediction of a linear model, respectively, when a single observation is removed and can be computed without refitting the model (Belsley et al. 1980). These results can be used to compute tethered hacking intervals for linear models when the space of data adjustment functions includes those that remove a single observation. In causal inference, we can find the range of effect estimates subject to an unmeasured confounder being within specified bounds on its relationship to both the treatment and the outcome (Lin et al. 1998, Vanderweele and Arah 2011). If we think of an unmeasured confounder as an additional feature created by a researcher, we can use these results to find the prescriptively constrained hacking interval under this researcher degree of freedom. We applied this idea in Section 3.1.2. Tethered hacking intervals are equivalent to profile likelihood confidence intervals (Bjornstad 1990) when the loss function corresponds to a likelihood. We discuss this in more detail in Online Appendix C.1.

Finding hacking intervals can be viewed as a form of robust optimization. Robust optimization serves as a worst case analysis in decision theory. Uncertainty sets are the primitives for hacking intervals, namely the ranges of user choices we are willing to consider. In prescriptively constrained hacking intervals, the uncertainty set is the range of prescriptive choices the researcher is allowed to make. In tethered hacking intervals, the uncertainty set is determined by the set of functions achieving low loss. If we cannot easily determine the uncertainty set in advance, we may be able to learn the uncertainty sets from related problems if data (from other sources) are available. This is done by Tulabandhula and Rudin (2014b) for machine learning to determine uncertainty sets for decision making.

The "Machine Learning with Operational Costs" framework (Tulabandhula and Rudin 2013, 2014a) computes a tethered hacking interval of the cost that a company might incur to enact an optimal policy in response to any good predictive model. The work of Letham et al. (2016) uses tethered hacking intervals in the setting of uncertainty quantification and optimal experimental design for dynamical systems. They recommend to perform an experiment that would most reduce the hacking interval on the quantity the experimenter wishes to estimate.

### 7.5. Teaching of Hacking Intervals

A major benefit of hacking intervals is that they are easy to explain. Confidence intervals and $p$-values are difficult to teach and interpret, and they are regularly misinterpreted. In response, the American Statistical Association recently issued a document explaining hypothesis testing to users (Wasserstein and Lazar 2016), and the field of basic and applied social psychology banned $p$-values altogether (Trafimow and Marks 2015), but as the authors of these proposals recognize, this does not fully solve the problem.

Hacking intervals are easy to explain, do not require knowledge of probability to understand, and sometimes capture as much, if not more, uncertainty as regular confidence intervals. Teaching hacking intervals first may give a gentle introduction to the effect of uncertainty on conclusions.

## 8. Conclusion

In this work, we presented an alternative theory of inference. It complements existing theories in that it handles a form of uncertainty that arises from analyst choices, rather than from randomness in the data. We presented several examples of hacking intervals stemming from regression and classification, as well as dimension reduction and feature selection. We showed in a real example how hacking intervals can be helpful—in particular, our results indicate that a commonly used model for pretrial risk analysis may sometimes be miscalculated, potentially leading to suboptimal judicial decision making throughout the United States. Our examples indicate that it is possible that these incorrectly computed risk scores could lead (or have led) to high-risk individuals being released or low-risk individuals being detained.

## Acknowledgments

The authors give special thanks to Aaron Fisher for insightful comments and assistance with proofs. B. Coker and C. Rudin contributed equally to the manuscript.

## Endnotes

[1] The repository for the hacking package is at https://github.com/beauCoker/hacking.

[2] The repository for paper results is at https://github.com/beauCoker/hacking_paper_results.

[3] Alternatively, the goal may be to assume that the unmeasured confounder reduced the causal effect to zero and see what this would imply about the unmeasured confounder.

[4] Lin et al. (1998) show this result exactly for log-linear regression, but they argue it should hold approximately for logistic regression.

[5] Mories Abdo also committed a Municipal Ordinance for Possession of a Controlled Substance during the two-year follow-up period, but this charge does not count as a reoffense in our data set (there are many charges, like ordinary traffic violations, that do not count as reoffenses).

## References

Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C (2018) Learning certifiably optimal rule lists for categorical data. *J. Machine Learn. Res.* 18(234):1–78.

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. *ProPublica* (May 23), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Back BJ, Rodriguez LR, Boessenecker M, Calvo A, Castro A, Chittick HA, Eskin GC, et al (2017) Pretrial detention reform—recommendations to the Chief Justice. Technical report, Judicial Branch of California, Sacramento.

Banaji MR, Greenwald AG (2013) *Blindspot: Hidden Biases of Good People* (Random House, New York).

Begley CG, Ellis LM (2012) Raise standards for preclinical cancer research. *Nature* 483:531–533.

Belsley DA, Kuh E, Welsh RE (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley Series in Probability and Mathematical Statistics (John Wiley and Sons, Hoboken, NJ).

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, et al. (2018) Redefine statistical significance. *Nature Human Behav.* 2(1):6–10.

Berger JO (1990) Robust Bayesian analysis: Sensitivity to the prior. *J. Statist. Planning Inference* 25:303–328.

Berger JO, Moreno E, Pericchi L, Bayarri M, Bernardo J, Cano J, Horra J, et al. (1994) An overview of robust Bayesian analysis. *Test* 3:5–124.

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. *Ann. Statist.* 41(2):802–837.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.*, ePub ahead of print July 2, https://doi.org/10.1177/0049124118782533.

Bishop DVM, Chen J, Thompson PA (2016) Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ* 4:e1715.

Bjornstad JF (1990) Predictive likelihood: A review. *Statist. Sci.* 5(2):242–254.

Breiman L (2001) Statistical modeling: The two cultures. *Statist. Sci.* 16(3):199–215.

Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, Kirchler M, et al. (2016) Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–1436.

Cornfield J, Haenszel W, Hammond EC (1959) Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. National Cancer Inst.* 22:173–203.

Dieterich W, Mendoza C, Brennan T (2016) COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe, Traverse City, MI.

Ding P, VanderWeele TJ (2016) Sensitivity analysis without assumptions. *Epidemiology* 27(3):368–377.

Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth AL (2015) Preserving statistical validity in adaptive data analysis. *Proc. Forty-Seventh Annual ACM Sympos. Theory Comput. (STOC), Portland, Oregon*, 117–126.

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Machine Learn. Res.* 20(177):1–81.

Gelman A, Loken E (2013) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Accessed April 23, 2018, http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gelman A, Hill J, Yajima M (2012) Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educational Effectiveness* 5:189–211.

Ghanem R, Higdon D, Owhadi H (2017) *Handbook of Uncertainty Quantification* (Springer International Publishing, Cham, Switzerland).

Gilbert DT (1998) Ordinary psychology. Gilbert DT, Fiske ST, Lindzey G, eds. *The Handbook of Social Psychology*, vol. 2 (McGraw Hill, New York), 89–150.

Gilbert DT, King G, Pettigrew S, Wilson TD (2016) Comment on "estimating the reproducibility of psychological science." *Science* 351(6277):1037.

Hannig J, Iyer H, Lai RC, Lee TCM (2016) Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* 111(515):1346–1361.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. *PLoS Biology* 13(3):e1002106.

Humphreys M, Sanchez De La Sierra R, Van Der Windt P (2012) Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Anal.* 21(1):1–20.

Iacus SM, King G, Porro G (2011) Multivariate matching methods that are monotonic imbalance bounding. *J. Amer. Statist. Assoc.* 106:345–361.

Jager LR, Leek JT (2014) An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15(1):1–12.

Kahneman D (2011) *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).

King G (1995) Replication, replication. *Political Sci. Politics* 28(3):443–499.

King G, Zeng L (2006) The dangers of extreme counterfactuals. *Political Anal.* 14(2):131–159.

Leamer EE (1983) Let's take the con out of econometrics. *Amer. Econom. Rev.* 73(1):31–43.

Leamer EE (2010) Extreme bounds analysis. Durlauf SN, Blume LE, eds. *Microeconometrics*, The New Palgrave Economics Collection (Palgrave Macmillan, London), 49–52.

Letham B, Letham PA, Rudin C, Browne E (2016) Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos* 26(6).

Lin DY, Psaty BM, Kronmal RA (1998) Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* 54(3):948–963.

Liu W, Kuramoto SJ, Stuart EA (2013) An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. *Prevention Sci.* 14(6):570–580.

Monogan JE (2015) Research preregistration in political science: The case, counterarguments, and a response to critiques. *Political Sci. Politics* 48(3):425–429.

Morucci M, Noor-E-Alam M, Rudin C (2018) Hypothesis tests that are robust to choice of matching method. Preprint, submitted December 5, https://arxiv.org/abs/1812.02227.

Muñoz J, Young C (2018) We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociol. Methods Res.* 48(1):1–33.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251).

Prinz F, Schlange T, Asadullah K (2011) Believe it or not: How much can we rely on published data on potential drug targets? *Nature Rev. Drug Discovery* 10:712.

Rudin C, Wang C, Coker B (2020) The age of secrecy and unfairness in recidivism prediction. *Harvard Data Sci. Rev.* 2(1). https://doi.org/10.1162/99608f92.6ed64b30.

Scheffé H (1959) *The Analysis of Variance* (Wiley, New York).

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psych. Methods* 22(11): 1359–1366.

Simonsohn U, Nelson LD, Simmons JP (2014) P-curve: A key to the file-drawer. *J. Experiment. Psych. General* 143(2):534–547.

Simonsohn U, Simmons JP, Nelson LD (2015) Specification curve: Descriptive and inferential statistics on all reasonable specifications. Preprint, submitted November 25, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2694998.

Sudret B, Marelli S, Wiart J (2017) Surrogate models for uncertainty quantification: An overview. Sibille A (chair), *2017 11th Eur. Conf. Antennas Propagation (EUCAP)* (IEEE, Piscataway, NJ), 793–797.

Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R (2016) Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* 111(514):600–620.

Trafimow D, Marks M (2015) Editorial. *Basic Appl. Soc. Psych.* 37:1–2.

Tulabandhula T, Rudin C (2013) Machine learning with operational costs. *J. Machine Learn. Res.* 14:1989–2028.

Tulabandhula T, Rudin C (2014a) On combining machine learning with decision making. *Machine Learn.* 97(1–2):33–64.

Tulabandhula T, Rudin C (2014b) Robust optimization using machine learning for uncertainty sets. Preprint, submitted July 4, https://arxiv.org/abs/1407.1097.

Vanderweele TJ, Arah OA (2011) Unmeasured confounding for general outcomes, treatments, and confounders: Bias formulas for sensitivity analysis. *Epidemiology* 22(1):42–52.

Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: Context, process, and purpose. *Amer. Statist.* 70(2):129–133.

Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, VanAert RCM, vanAssen MALM (2016) Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Front Psychol. Nov 25;7: 1832. doi:10.3389/fpsyg.2016.01832. eCollection 2016.

Young C, Holsteen K (2015) Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, ePub ahead of print October 23, https://doi.org/10.1177/0049124115610347.

Zeng J, Ustun B, Rudin C (2017) Interpretable classification models for recidivism prediction. *J. Royal Statist. Soc.* 180(3): 689–722.