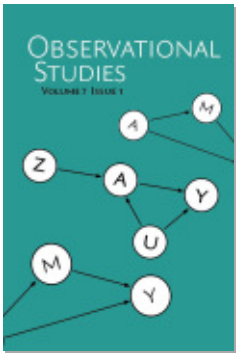
Comment on 'Statistical Modelling: the Two Cultures' by Leo Breiman



Efrén Cruz-Cortés, Fan Yang, Elizabeth Juaréz-Colunga, Theodore Warsavage, Debashis Ghosh

Observational Studies, Volume 7, Issue 1, 2021, pp. 41-57 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0021*

Observational Studies Vol. 7.1 (2021): Published July 2021

Comment on 'Statistical Modelling: the Two Cultures' by Leo Breiman

Efrén Cruz-Cortés **eqc5419@psu.edu**
Department of Statistics
Penn State University
University Park, PA 16802

Fan Yang **fan.3.yang@cuanschutz.edu**
Department of Biostatistics and Informatics
Colorado School of Public Health
Aurora, CO 80045, USA

Elizabeth Juaréz-Colunga **elizabeth.juarez-colunga@cuanschutz.edu**
Department of Biostatistics and Informatics
Colorado School of Public Health
Aurora, CO 80045, USA

Theodore Warsavage **theodore.warsavage@cuanschutz.edu**
Department of Biostatistics and Informatics
Colorado School of Public Health
Aurora, CO 80045, USA

Debashis Ghosh **debashis.ghosh@cuanschutz.edu**
Department of Biostatistics and Informatics
Colorado School of Public Health
Aurora, CO 80045, USA

Abstract

The discussion paper "Statistical Modeling: the Two Cultures" (Statistical Science, Vol 16, 2001) by the late Leo Breiman sent shockwaves throughout the statistical community and subsequently redirected the efforts of much of the field towards machine learning, high-dimensional analysis and data mining approaches. In this discussion, we discuss some of the implications of this work in the sphere of causal inference. In particular, we define the concept of comparability, which is fundamental to the ability to draw causal inferences and reinterpret some concepts in high-dimensional data analysis from this viewpoint. One of the points we highlight in this discussion is the need to consider data-adaptive estimands for causal effects with high-dimensional confounders. We also revisit matching and develop some mathematical formalism for matching algorithms.

Keywords: Causal effects; High-dimensional data; Margin; Overlap condition; Treatment Positivity.

1. Introduction

We thank the previous and current editors of *Observational Studies*, Dr. Dylan Small and Dr. Nandita Mitra, for the opportunity to comment on the article 'Statistical modelling; the two cultures' by Leo Breiman (Breiman, 2001b); we will refer to it as the 'Two Cultures

paper' here and in the sequel. There has been a great deal of attention of the Two Cultures paper, which has impacted the research and practice of statistics in the two decades since its publication in 2001. First, Breiman challenges the tradition of statistical modeling and inference by provocatively suggesting in his abstract that "this commitment has led to irrelevant theory, questionable conclusions and has kept statisticians from working on a large range of interesting current problems." Breiman suggests algorithmic culture, as embodied by data mining and machine learning, as a powerful alternative that can handle large and complex datasets. In a way, some of the discussion in the Two Cultures paper foreshadows the general discussion in the statistical profession about the role of data science (Donoho, 2017). While the Two Cultures paper had discussants defending the statistical modeling and inference tradition in Professors D. R. Cox and Bradley Efron, in a later paper by Efron (2020), there is a concession made to Breiman: "Breiman turned out to be more prescient than me (Efron): pure prediction algorithms have seized the statistical limelight in the twenty-first century, developing much along the lines Leo suggested."

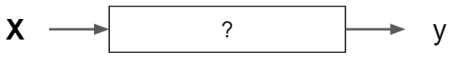
Practically, the Two Cultures paper has had a major impact in the field in terms of bringing machine learning and related procedures to the forefront of theoretical study in the statistical literature. This has also been aided by the emergence of datasets of increasing scientific and data complexity from fields such as genomics, climate sciences, mobile health and imaging. There has been tremendous interest in the statistical literature on techniques and generalization of methods such as the LASSO (Tibshirani, 1996), random forests (Breiman, 2001a), and support vector machines (Cristianini et al., 2000). This in turn has necessitated the development and enhancement of techniques such as empirical process theory (van der Vaart and Wellner, 1996; Kosorok, 2007), concentration inequalities (Boucheron et al., 2013) and random matrix theory (Mehta, 2004; Bai and Silverstein, 2010; Tao, 2012; Erdos and Yau, 2017) in order to understand the theoretical basis behind machine learning methods.

2. Modeling paradigms and causal inference

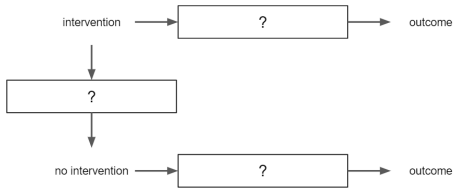
In Pearl and Mackenzie (2018), the learning that a cognitive agent can do is taxonomized into three levels with fundamentally different qualities: seeing, doing, and imagining. A linear hierarchy exists, going from simple (seeing) to complex (imagining), hence it is described as a "ladder". The authors argue that mere association cannot yield the same deep understanding of a given phenomenon that causal thought can. The methods described in the Two Cultures paper fall in the first rung of the ladder, however, integrating causal inference with these methods leads analysis further up.

Indeed, most analysis contains at least a drop of causal assumption. Consider Breiman's data model, as seen in Figure 1a. Breiman focused on the modeling assumptions pertaining to what is inside or replaces the black box, however, the relationship $\mathbf{x} \to \square \to y$ is a model itself of the true mechanism through which Nature generates $(\mathbf{x}, y)$. Certainly Breiman is cautious and avoids the word "causation", stating instead that, inside the box, Nature creates an "association". However, the directionality of the arrows and the structure of association itself already require a level of modeling assumptions.

In particular, we can observe the following set of non-exhaustive assumptions made in the Black-Box-and-Arrow model:



(a) Breiman's original diagram. There is a mechanism connecting X to y. We can choose a generative or an algorithmic approach to model this mechanism.



(b) Simple schematic for a causal inference problem. In the top section a mechanism connects an intervention to an outcome, and in the bottom section a mechanism connects a different or no intervention to another outcome. Causal modeling, in a sense, corresponds to these two boxes being indeed the same. The matching mechanism, connecting top to bottom, can be taken to be generative or algorithmic.

Figure 1: Comparison of modeling paradigms

1. $\mathbf{x}$ and y are ontologically different.
2. $\mathbf{x}$ pre-exists y. In a sense, $\mathbf{x}$ is already available for natural processes to manipulate in such a way as to produce y.
3. The mathematical objects $(\mathbf{x}, y)$ are faithful representations of their real counterpart, and are different to the process that associates them.
4. The mechanism through which y is obtained from $\mathbf{x}$ is mathematizable. Even if arrows were to switch directions, or have no direction, we would assume the process of association is describable by our mathematical tools.

The key idea is that the model described in the Two Cultures paper is an anthropocentric model of reality. It reflects how we, humans, understand and frame natural phenomena; it does not necessarily model nature itself. These pseudo-causal modeling assumptions do not fall comfortably in either the "data" modeling or "algorithmic" modeling paradigms. Instead, it serves as an infrastructural support in which these finer scale models can be developed.

Let us consider the discourse surrounding these two cultures, starting with algorithmic modeling, often criticized for being a black box. As Breiman states in the paper, algorithmic modeling is concerned mostly with prediction, or more specifically, generalization error. Often, prediction and explanation are framed as contesting nemeses, one excluding the other. In reality this is not the case. Designing a highly predictive algorithm that does not readily exhibit explanatory power does not mean it does not have the potential to do so, just that we do not know yet how to extract that information. The more exposition and effort we place on understanding such a successful association mechanism, the more it becomes explainable and interpretable.

Explanation and interpretation are, indeed, cultural processes. Critics of algorithmic modeling often cite regression and decision trees as ideals of transparency. However, their

interpretation and "meaning" are also constructed through specialized training and linguistic consensus. A concept only seems intuitive after we are familiar with it, and at that point it may be the only way through which we interpret reality.

To further this analogy, consider the case of data tables. By a table we just mean the object that organizes information in rows and columns. Tables can be viewed as a high-dimensional version of a list. Their second dimension allows us to delinearize language and information flow. When humans started using tables, only trained professionals were able to interpret and manipulate them. With time, as we became more familiar with their use, and as they became more ubiquitous, mastery of tables came as a result of cultural upbringing, without the need to learn it explicitly (see Anderson et al., 2019). These new algorithmic methods contain high-dimensional operations that may seem obscure or complex for the uninitiated. However, they are highly successful in conveying useful information for us. Their constant use and refinement, and their eventual implementation to everyday life, will render them trivial for future generations. We do note, however, that there is a tension between complexity and simplicity, as humans naturally tend to prefer the simple over the complex.

We also note that too much faith in models we are familiar with might lead us astray. Consider planetary orbits; classical Greeks struggled for generations to explain the motion of the sky in terms of circles. When a single circle did not suffice, they devised circles upon circles, arriving at models that, besides being erroneous, had unnecessary complexity. The reason this happened is because they deemed the circle the perfect geometric shape, so they projected their own conceptions into the heaven. Now we know they could have easily find a better explanation in the ellipse, if they had just taken a different perspective. While we have 20/20 hindsight in this case, in practice it is difficult to detach ourselves from our scientific assumptions and find a new perspective that possibly contradicts principles that we grasp.

These arguments also do not fully support or reject one of the two cultures of modeling in favor of the other. We will instead complexify the paradigm. It is also the case that some scientific endeavours are performed as an interplay between these two frameworks. One notable example is causal inference. To perform causal inference, one must make assumptions about the cause-effect direction, then, depending on the task, we can either take an algorithmic or statistical approach, or a mix of both. We can even use different paradigms, for example agent-based modeling, as our basis assumption. In this note we discuss a particular principle of causal inference, comparability, that does not comfortably fit in either paradigm, but is a mix of both, as shown in Figure 1b. We wish to make several points about comparability:

1. The principle of greatest importance in causal modeling is that of ensuring comparability of treated and untreated observations, which is a different metric than those optimized by standard machine learning algorithms;

2. Performing causal inference in high-dimensional data brings challenges and requires reinterpreting results in a new way;

3. One approach we advocate for is using data-adaptive estimands in the high-dimensional data case, and in particular, we study matching algorithms and provide some new theoretical arguments for their use.

Before getting into this discussion, we briefly review the potential outcomes framework we will use.

3. Background: Potential outcomes and assumptions

We will adopt the potential outcomes framework of Rubin (1974) and Holland (1986). Let Y denote the response of interest and $\mathbf{X}$ be a p-dimensional vector of covariates. Let Z be a binary indicator of treatment exposure. We assume that Z takes the values $\{0, 1\}$: $Z = 1$ if treated, $Z = 0$ if control. Let the observed data be represented as $(Y_i, \mathbf{X}_i, Z_i)$, $i = 1, \ldots, n$, a random sample from the triple $(Y, \mathbf{X}, Z)$.

When Y refers to the observed outcome under the receipt of a certain level of the treatment, we further define $\{Y_i(0), Y_i(1)\}$, $i = 1, \ldots, n$ to be the potential outcomes for subject i if control or treated. What we observe is $Y_i = Y_i(Z_i)$ $(i = 1, \ldots, n)$, which implies that $Y(0)$ and $Y(1)$ can not be observed simultaneously, i.e. one of them is missing. The relationship between Y and $\{Y(0), Y(1)\}$ can be summarized as

$$Y = Y(1)Z + Y(0)(1 - Z). \tag{1}$$

The typical parameter of interest is the average causal effect:

$$ACE = E[Y(1) - Y(0)], \tag{2}$$

although some authors have considered as an alternative estimand the average causal effect among the treated:

$$ACET = E[Y(1) - Y(0)|Z = 1]. \tag{3}$$

There are several assumptions necessary to perform causal inference and identify causal effects such as (2) and (3). The first is the strongly ignorable treatment assignment (SITA) assumption:

$$Z \perp\!\!\!\perp \{Y(0), Y(1)\} \mid \mathbf{X}.$$

This assumption says that treatment assignment is conditionally independent of the set of potential outcomes given the covariates. In other words, conditioning on the same value of $\mathbf{X}$, we can pretend that the observed outcomes are from a randomized trial. However, conditioning on a p-dimensional vector suffers from the "curse of dimensionality", especially when the dimension is high.

Another assumption is the Stable Unit Treatment Value Assumption (SUTVA): There is only a single version of each treatment level and there is no interference between subjects so that a subject's potential outcomes won't be affected by other individuals' treatment assignments. The notion of SUTVA is violated in infectious disease studies, where transmission of disease can lead to dependence of one individual's potential outcomes on the treatment received by another individual.

Finally, we also assume the Treatment Positivity Assumption (TP):

$$1 > P(Z = 1 \mid \mathbf{X}) > 0$$

for all values of $\mathbf{X}$. The TP assumption means that for any individual in a study, he/she has a positive probability of receiving either level of the treatment. This assumption can be violated in settings in which there are contraindications against providing treatments to subjects. In Khan and Tamer (2010), the authors showed that violations of TP assumption lead to irregularities regarding identification and inference about causal effects. This was also seen in Luo et al. (2017), who demonstrated super-efficiency for causal effect estimators if one made a weaker covariate overlap assumption. This phenomenon also occurs in the collaborative targeted maximum likelihood estimator of van der Laan and Gruber (2010).

4. Comparability and high-dimensional data

4.1 Comparability

At its essence, the goal of causal inference is to take a population that self-selects into treatment groups and attempt to derive an inference that might be more in line with an experimental design in which the treatment was randomized. We refer to the latter scenario as one in which observations are comparable. A major assumption we need to attempt to make noncomparable observations comparable is (4). This relates to the discussion of Figure 1 in Ghosh et al. (2015), in which it is described that the major region in which we can reliably perform causal inference is in the region where there is overlap of the propensity scores. This will also roughly correspond to regions of covariate space where we cannot reliably predict treatment. In Ghosh (2018), an attempt to characterize this region using margin theory from support vector machines was used, which leads to a relaxation of the overlap and related TP condition defined in §2.

We now describe several standard approaches in causal inference and observe how they enforce comparability:

1. Inverse weighting: Comparability is maintained by reweighting the treatment and control populations to create a pseudo-population that mimics the population on which we can perform causal inference;

2. Propensity score regression adjustment: if we include the propensity score as a covariate in a regression model of outcome on treatment, we enforce comparability by estimating a propensity score-adjusted causal effect;

3. Combining inverse weighting with propensity score regression adjustment: this approach enforces comparability by the arguments in 1. and 2.

4. Matching: we will discuss this further in §4. It enforces comparability by only including observations from both groups that have similar distribution of confounders.

Other more recent approaches include enforcing covariate balance as part of the causal effect estimation process (Imai and Ratkovic, 2014; Chan et al., 2016; Josey et al., 2020).

4.2 High-dimensional findings and implications

We next want to consider what role high-dimensional data concepts have with respect to comparability and the potential outcomes framework. We begin with a canonical example. If we take $\mathbf{X}_1, \ldots, \mathbf{X}_n$ to be multivariate normal $d-$dimensional vectors, then it is known that for n fixed and d approaching infinity, the vectors will lie on the surface of a $d-$dimensional hypersphere with high probability. While this is often taken as limitation of high-dimensional data analysis regarding the inability to find nearest neighbors with high-dimensional data, one can reinterpret this to mean that as the dimensionality increases, there is an innate tendency towards greater comparability between observations $\mathbf{X}_i$ $(i = 1, \ldots, n)$.

The example in the previous paragraph dealt with a one-sample setting. The optimism about comparability there is counterbalanced by the fact that as the dimension of confounders increases, the chances that a random classifier (i.e., a classifier that is non-informative for prediction) perfectly predicts treatment approaches one. What this leads to is a violation of the treatment positivity assumption from §2. This has been explored in detail in D'Amour et al. (2020) and Ghosh and Cruz Cortés (2019). As noted by D'Amour et al. (2020), a stronger assumption than TP that typically is made is termed 'strict overlap': there exists $\eta \in (0, 1/2)$ such that

$$\eta < P(Z = 1|\mathbf{X}) < 1 - \eta$$

for all $\mathbf{X}$. With the strict overlap assumption, one can guarantee, for example, that inverse probability weighted estimators of the ACE will exhibit regular asymptotics. It is this strict overlap assumption that becomes problematic in higher dimensions (Ghosh and Cruz Cortés, 2019; D'Amour et al., 2020). If this assumption is violated, then estimators of causal effects will exhibit nonregular behavior asymptotically.

Our last example has to do with another condition commonly assumed in the literature on high-dimensional theory and analysis, termed the 'low-noise condition' (Tsybakov et al., 2004). Defining $e(\mathbf{X}) = P(Z = 1|\mathbf{X})$, the low-noise condition is satisfied if there exists $\gamma \in (0, 1)$, $C > 0$ and $t_0 \in (0, 0.5]$ such that

$$P(|e(\mathbf{X}) - 0.5| \leq t) \leq Ct^{\frac{\gamma}{1-\gamma}}$$

for all $t \in [0, t_0]$. Under this assumption, classifiers will have fast rates of convergence of their associated empirical errors to population-level quantities (e.g., Audibert et al. 2007; Srebro et al. 2010). However, the low-noise condition implies that propensity scores will tend to be far away from 0.5. This property will in turn correlate with a greater tendency for propensity scores to be closer to zero or one, which will mean some likelihood of violation of the overlap condition (Ghosh and Cruz Cortés, 2019; D'Amour et al., 2020).

5. Data-driven estimands and Matching

5.1 Matching: definition

One approach to handling the problems of potential outcomes assumptions is to focus on data-driven estimands. One such approach is matching; a comprehensive review describing

their use being found in Stuart (2010). We will use the term matching to allude to the use of an algorithm to balance the distribution of confounders between treatment groups.

The idea of matching is to find for each subject in the treatment group the subject(s) from the control group whose confounders are the most similar in nature. In many practical situations, the confounder gets summarized via the propensity score (Rosenbaum and Rubin, 1983), and matching proceeds with the caliper metric approach described in Rosenbaum and Rubin (1985). An appeal to Theorem 1 from Rosenbaum and Rubin (1983) can be used to justify the procedure of matching individuals from the treatment and control groups leading to covariate balance on $\mathbf{X}$.

An alternative to matching would involve fitting a regression model of the outcome variable on Z incorporating the estimated propensity score either as a covariate and/or a weight. Regression estimators typically involve some degree of extrapolation. By contrast, matching estimators do not involve any such extrapolation and thus can be thought to exhibit a certain type of robustness. As described in Stuart (2010), matching comes with the following benefits:

1. Matching methods allow for the analyst to create effectively randomized block designs (Box et al., 1978), which can potentially lead to efficiency gains in the estimation of treatment effects;

2. Matching methods are quite flexible and can be used in combination with regression, weighting and subclassification approaches (Imbens and Rubin, 2015);

3. Matching methods target the region of the covariate space where there is sufficient overlap between treatment groups. We formalize this notion mathematically in §5.4.

4. There exist diagnostics that are available to check for covariate balance in the distribution of confounders after the matching is performed.

There have been many approaches developed for matching. These include nearest neighbor matching (Rubin, 1973), optimal matching (Rosenbaum, 1989), and more recently, graph-based matching approaches using the cross-match statistic as well as the minimal spanning tree approach (Rosenbaum, 2005).

5.2 Overlap and Bayes Error

We show analytically how certain matching algorithms directly target the overlap of confounders analytically. Recall that $e(\mathbf{X}) \equiv P(Z = 1|\mathbf{X})$ is the propensity score (Rosenbaum and Rubin, 1983). Define the Bayes decision function (Devroye et al. (2013), p. 10) by $\tilde{e}(\mathbf{X}) = 1$ if $e(\mathbf{X}) > 0.5$ and zero otherwise. We begin by proving a simple lemma, which is Proposition 4 from D'Amour et al. (2020).

Proposition 1 *For any fixed $\eta \in (0, 0.5)$,*

$$P(\eta < e(\mathbf{X}) < 1 - \eta) \leq \eta^{-1} P(\tilde{e}(\mathbf{X}) \neq Z). \tag{4}$$

Proof Define the event $B = \{\mathbf{X} : \eta < e(\mathbf{X}) < 1 - \eta\}$. Then

$$P(\tilde{e}(\mathbf{X}) \neq Z) \geq P(B) P(\tilde{e}(\mathbf{X}) \neq Z|B).$$

It is straightforward to see that $P(\tilde{e}(\mathbf{X}) \neq Z|B) \geq \eta$. Plugging this in the previous equality concludes the proof. ■

Remark 1. In D'Amour et al. (2020), the authors use Proposition 1 to show that it is impossible to satisfy (4) for all $\eta \in (0, 0.5)$ so as to emphasize the restrictiveness of the strict overlap assumption. In what we pursue here, we will treat η as fixed.

An interpretation of Proposition 1 is as follows. The left-hand side of (4) is the probability of the strict overlap assumption being satisfied. The right hand-side of (4) is nothing more than the Bayes error, scaled by η.

Before describing how we can utilize results from the information theory literature on divergence functions (Amari, 2016) to link strict overlap to matching procedures, we describe some graph-based matching procedures.

5.3 Matching, algorithms and graph-based asymptotics

There are many issues in how to apply matching, which are described further in the monographs by Rosenbaum (2002, 2010). Here, we formulate the observations from the treated and control populations as a bipartite graph and view matching as assigning edges between the nodes. We assume that there exists a distance metric $d : R^p \times R^p \to (0, \infty]$ that computes the distance between $\mathbf{X}_i$ and $\mathbf{X}_j$. We assume that $d(\mathbf{X}_i, \mathbf{X}_j) = \infty$ if $Z_i = Z_j$, which guarantees that two observations from the same treatment group will not be matched to each other. There are several ways in which we can envision performing matching.

1. $k-$nearest neighbor matching: In this approach, the k observations from the control group whose distances are the smallest relative to the ith treated observation are matched to that observation. In the case of $k = 1$, this is referred to as nearest neighbor matching.

2. Minimum spanning tree matching: Friedman and Rafsky (1979) proposed the use of minimal spanning trees as a means of constructing multivariate tests of comparing two-sample distributions. Their algorithm consists of constructing minimum spanning tree for the pooled set of observations between the treatment and control groups. We can then retain the edges connecting observations from different treatment groups.

3. Cross-match statistic: Rosenbaum (2005) recently proposed a graph-based statistic for performing multivariate two-sample tests that he termed the cross-match statistic. The graph induced by the cross-match statistic comes from solving a certain relaxation of a combinatorial optimization problem. One of the advantages of the cross-match statistic, in contrast to the multivariate statistics of Friedman and Rafsky (1979), is that the exact null distribution is distribution-free and does not require any knowledge of the topology of the underlying graph.

In Arias-Castro and Pelletier (2016), the authors note that these graph-based statistics can be represented using a very general framework. In particular, we can consider $\mathcal{G}$ to be a directed graph with the node set being $\mathbf{V} = (\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*)$, a permutation of the original confounders in which the first n_0 entries are from the $Z = 0$ population and the remaining $n_1 = n - n_0$ entries are from the $Z = 1$ population. Following Arias-Castro and Pelletier

(2016), define $\mathbf{X}_i \to \mathbf{X}_j$ when node i points to node j in a graph. Then the statistics corresponding to the matching scheme in 1) – 3) above can all be expressed as

$$W_{\mathcal{G}}(\mathbf{V}) \equiv \#\{i \leq n_0, j > n_0 : \mathbf{X}_i^* \to \mathbf{X}_j^*\} + \#\{i \leq n_0, j > n_0 : \mathbf{X}_j^* \to \mathbf{X}_i^*\},$$

which in words represents the number of neighbors in the graph that come from different treatment groups. An asymptotic consistency result is presented in Arias-Castro and Pelletier (2016) that requires the following assumptions:

A1. $n_0/(n_0 + n_1) \to \pi \in (0, 1)$ as both n_0 and n_1 approach infinity;

A2. All the vertices in $\mathcal{G}$ have constant out-degree;

A3. The graph $\mathcal{G}$ has bounded degree averaged over all the vertices;

A4. The outdegree, suitably normalized, converges to a constant c, which we will treat as known.

A5. Edges of length greater than $O(t^{-1/d})$ are unlikely.

We refer the reader to Arias-Castro and Pelletier (2016) for a more mathematical description of these assumptions. Based on these assumptions, Arias-Castro and Pelletier (2016) prove the following result:

Theorem: (Theorem 1 from Arias-Castro and Pelletier (2016)) Let $\mathbf{X}|T = 0$ have density f and $\mathbf{X}|T = 1$ have density g. Then assuming A1. – A5. above, as $n_0, n_1 \to \infty$,

$$\frac{W_{\mathcal{G}}(\mathbf{V})}{n_0 + n_1} \to 2c \int \frac{\pi(1-\pi)f(\mathbf{x})g(\mathbf{x})}{\pi f(\mathbf{x}) + (1-\pi)g(\mathbf{x})} d\mathbf{x} \tag{5}$$

almost surely.

While the proof of the theorem can be found in Arias-Castro and Pelletier (2016), we will make a few remarks about it. As mentioned earlier, all three of the graph-based matching schemes satisfy assumptions A1.) - A5.), so the theorem applies to the nearest neighbor, minimal spanning and cross-match statistics. It is effectively a consistency result about the convergence of the empirical distribution of the graph-based statistic to a limit that depends on the following factors: c (from assumption A4.), π, f and g. In the situation where $f \neq g$, the theorem guarantees that tests based on $W_{\mathcal{G}}(\mathbf{V})$ will be consistent (i.e., the tests will have power converging to one as n_0 and n_1 get large). We also note that the theorem is a synthesis of the work of several authors who have focused on limits for specific graph-based statistics. In particular, it generalizes the results of Schilling (1986) for nearest-neighbor graphs and Henze and Penrose (1999) for minimal spanning tree-based graphs. The result for the cross-match statistic was new in Arias-Castro and Pelletier (2016). As Arias-Castro and Pelletier (2016) write on page 186, "the proof of Theorem 1 is exactly the same as that of Theorem 2 in Henze and Penrose (1999), treating out-edges and in-edges separately." As we will see in the next section, our interest is in the right-hand side of (5) and how it relates to the strict overlap criterion.

5.4 Divergence functions and synthesis

We now introduce the concept of divergence functions, which is a foundational concept in information geometry (Amari, 2016). The area treats distributions as points in a manifold with an associated coordinate stucture as defined by a chart system. A divergence function M comparing two points in the manifold P and Q with associated coordinates $\mathbf{h}_P$ and $\mathbf{h}_Q$ satisfies the following properties:

1. $M(\mathbf{h}_P, \mathbf{h}_Q) \geq 0$;
2. $M(\mathbf{h}_P, \mathbf{h}_Q) = 0$ iff $P = Q$;
3. If $\mathbf{h}_Q = \mathbf{h}_P + d\psi$, then M can be expanded using a Taylor series as

$$M(\mathbf{h}_P, \mathbf{h}_Q) = \frac{1}{2} \sum_{i,j} g_{ij}(\mathbf{h}_P) d\psi_i d\psi_j + O(\|d\psi\|^3),$$

   and $\mathbf{G} = [g_{ij}]$ is positive definite.

We note that a divergence function is not necessarily a distance metric, as M is not required to be symmetric or satisfy the triangle inequality.

While there are many choices of divergence function to use, we will select as our divergence measure the Henze-Penrose divergence (Sekeh et al., 2019), defined as

$$M_{HP}(f_0, f_1) = \frac{1}{4\pi(1-\pi)} \left[\int \frac{[\pi f_0(\mathbf{x} - (1-\pi) f_1(\mathbf{x})]^2}{\pi f_0(\mathbf{x}) + (1-\pi) f_1(\mathbf{x})} d\mathbf{x} - (2\pi - 1)^2 \right].$$

It can be shown that M_{HP} is bounded and symmetric in addition to satisfying the three properties listed above. Sekeh et al. (2019) also reexpress M_{HP} as

$$M_{HP}(f_0, f_1) = 1 - A_{HP}(f_0, f_1),$$

where

$$A_{HP}(f_0, f_1) = \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x})}{\pi f_0(\mathbf{x}) + (1-\pi) f_1(\mathbf{x})} d\mathbf{x}.$$

Utilizing the theorem from Section 5.3 in conjunction with the arguments in the proof of Theorem 1 of Sekeh et al. (2019), we can show that

$$\frac{W_{\mathcal{G}}(\mathbf{V})}{n_0 + n_1} \to 2c\pi(1-\pi) A_{HP}(f, g)$$

almost surely.

In addition, we have from chapter 3.4 of Fukunaga (2013) that the Bayes error can be upper bounded by divergences. Using the notation of (4), we have that

$$\begin{aligned} P(\tilde{e}(\mathbf{X}) \neq Z) &\leq M_{HP}(f, g) \\ &= 1 - A_{HP}(f, g) \end{aligned} \tag{6}$$

Synthesizing all the results across the paper, we can make the following findings:

1. The probability of strict overlap is majorized (upper bounded) by the divergence between the densities of $\mathbf{X}|T=1$ and $\mathbf{X}|T=0$. Mathematically, this is represented as

$$\begin{aligned} P(\eta < e(\mathbf{X}) < 1-\eta) &\leq P(\tilde{e}(\mathbf{X}) \neq Z)/\eta \\ &\leq M_{HP}(f,g)/\eta \\ &= (1 - A_{HP}(f,g))/\eta \end{aligned} \tag{7}$$

2. An empirical estimator of (7) is

$$\left(1 - \frac{nW_{\mathcal{G}}(\mathbf{V})}{2cn_0n_1}\right)\eta^{-1}$$

These findings provide new analytical insights on the role of matching algorithms. First, for any associated matching algorithm, there is an induced graph structure connecting treated subjects to their matched control subjects. Given the matching algorithms discussed in Section 5.3, we see that the graphs are attempting to minimize the divergence between f, the density of $\mathbf{X}|Z=0$ and g, that of $\mathbf{X}|Z=1$. Second, this divergence serves as a surrogate criterion for the strict overlap criteria, and through this derivation, it becomes explicit how matching targets the overlap criteria. In much of the presentation of matching procedures, the focus is more on algorithms without defining what are potential estimands or probabilistic targets/criteria that are being optimized. The results presented here are a quantification of qualitative observations about matching made in Rosenbaum (1989) and (Stuart, 2010).

6. Conclusion

We would like to thank Drs. Small and Mitra again for the opportunity to revisit Breiman (2001b) and consider it through the lens of what has transpired in the two decades since its publication. Many of the predictions Breiman put forward in his article have come to fruition, such as the role of machine learning and algorithmic modelling now occupies a central role in statistical research.

It is curious to ponder a bit what Breiman might think of the whole endeavor of causal modeling. Here are elements of the Two Cultures paper that we feel align with causal inference:

1. The practice of statistics in the wild: causal inference methods are used through much of scientific and medical research, such as sociology, psychology, economics and health services research and applied to "real-life" data sets used to inform policy making. The mainstreaming of these methods into leading statistical journals speaks to Breiman's desire to see papers with an emphasis on substantive applications spotlighted there.

2. In his rejoinder to Efron and Cox, Breiman does say (p. 229) that many statistical investigations have 'the identification of causal factors as their ultimate role'.

3. Ideas surrounding causality have also been seeping into machine learning research, at least informally. Much of machine learning takes place in an industrial setting,

where the intention is to influence the behavior of users, making it an interventionist practice.

4. There is tremendous emphasis in causal inference placed on elucidating the assumptions needed to (a) perform valid causal inference and (b) identify the causal effect from observed data. Given these assumptions, the discussion of models and estimation then proceeds. Furthermore, if one adopts a propensity score modelling approach, then one separates the model for understanding the science (the causal model or equivalently, the model for the potential outcomes) from the model for treatment assignment (propensity score model). Thus, modelling considerations in causal inference problems are much more nuanced than what is portrayed for statistical modeling in Breiman (2001b).

5. In data science terminology, Breiman is arguing for the use of domain knowledge, and this happens routinely in causal inference analyses through consideration of variables for predicting the potential outcomes.

6. Regarding point # 4, if one can proceed to do estimation, then one can incorporate machine learning approaches in many different ways. This is a subject of massive research interest in statistics (e.g., Chernozhukov et al., 2017; Wager and Athey, 2018; Díaz, 2020).

Here are aspects of causal inference that do not align with the Two Cultures paper:

1. The idea of a generative model for the potential outcomes seems anathema to the viewpoint advocated by Breiman.

2. Breiman praises his colleague, the late David Freedman, for his critiques of path analysis. Freedman was very critical of the causal modelling approach (Freedman, 1999).

Our view is that there is a false dichotomy between the algorithmic and data model cultures put forward by Breiman. Our schematic in Figure 1 and results about matching in §5 align with the algorithmic viewpoint he advocates. However, the complaint by discussants was that the black box approach fails to explain the mechanism, and the potential outcomes outlined here is one means of explanation.

We look forward to developments in machine learning, high-dimensional data analysis, causal inference and observational studies in the next two decades to see how the predictions of Breiman play out.

Acknowledgments

FY was supported in part by the IES grant R305D200031. DG was supported in part by NSF DMS 1914937.

References

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Miranda Anderson, Douglas Cairns, and Mark Sprevak. *Distributed Cognition in Classical Antiquity*. Edinburgh University Press, 2019.

Ery Arias-Castro and Bruno Pelletier. On the consistency of the crossmatch test. *Journal of Statistical Planning and Inference*, 171:184–190, 2016.

Jean-Yves Audibert, Alexandre B Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

George EP Box, William Gordon Hunter, and J Stuart Hunter. *Statistics for experimenters*, volume 664. John Wiley and sons New York, 1978.

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

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001b.

Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2020.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 2020.

Bradley Efron. Prediction, estimation, and attribution. *International Statistical Review*, 88:S28–S59, 2020.

László Erdos and Horng-Tzer Yau. A dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics*, 28, 2017.

David Freedman. From association to causation: some remarks on the history of statistics. *Journal de la société française de statistique*, 140(3):5–32, 1999.

Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

Debashis Ghosh. Relaxed covariate overlap and margin-based causal effect estimation. *Statistics in medicine*, 37(28):4252–4265, 2018.

Debashis Ghosh and Efrén Cruz Cruz Cortés. A gaussian process framework for overlap and causal effect estimation with high-dimensional covariates. *Journal of Causal Inference*, 7 (2), 2019.

Debashis Ghosh, Yeying Zhu, and Donna L Coffman. Penalized regression procedures for variable selection in the potential outcomes framework. *Statistics in medicine*, 34(10): 1645–1658, 2015.

Norbert Henze and Mathew D Penrose. On the multivariate runs test. *Annals of statistics*, pages 290–298, 1999.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

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

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

Kevin P Josey, Elizabeth Juaréz-Colunga, Fan Yang, and Debashis Ghosh. A framework for covariate balance using bregman distances. *Scandinavian Journal of Statistics*, 2020.

Shakeeb Khan and Elie Tamer. Irregular identification, support conditions and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

Wei Luo, Yeying Zhu, and Debashis Ghosh. On estimating regression causal effects using sufficient dimension reduction. *Biometrika*, 104(1):51–65, 2017.

Madan Lal Mehta. *Random matrices*. Elsevier, 2004.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Paul R Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

Paul R Rosenbaum. *Observational studies*. Springer, 2002.

Paul R Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

Paul R Rosenbaum. *Design of observational studies*, volume 10. Springer, 2010.

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

Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Mark F Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

Salimeh Yasaei Sekeh, Morteza Noshad, Kevin R Moon, and Alfred O Hero. Convergence rates for empirical estimation of binary classification bounds. *Entropy*, 21(12):1144, 2019.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Alexander B Tsybakov et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Mark J van der Laan and Susan Gruber. Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, 6(1), 2010.

Aad van der Vaart and Jon A Wellner. Weak convergence and empirical processes. 1996.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.