Breiman's "Two Cultures" Revisited and Reconciled

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

03-09-2021 / 09-09-2021

CITATION

Mukhopadhyay, Deep; Wang, Kaijun (2021): Breiman's "Two Cultures" Revisited and Reconciled. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.16569201.v1>

DOI

[10.36227/techrxiv.16569201.v1](https://doi.org/10.36227/techrxiv.16569201.v1)

Breiman's "Two Cultures" Revisited and Reconciled

Subhadeep Mkhopadhyay*, Kaijun Wang

deep@unitedstatalgo.com, kwang2@fredhutch.org

Abstract—In a landmark paper published in 2001, Leo Breiman described the tense standoff between two cultures of data modeling: parametric statistical and algorithmic machine learning. The cultural division between the two dominant data modeling frameworks has been growing at an alarming rate in recent years. To make real progress, we are forced to consider the question: how do these two different cultures fit together? This paper introduces a new data analysis framework, called Integrated Statistical Learning (ISL) theory, which offers solutions to blend the 'two cultures' into a coherent whole by establishing a link between them. Through examples, we describe the challenges and potential gains of this new *integrated* statistical thinking.

Index Terms—Integrated statistical learning theory, Exploratory machine learning, Uncertainty prediction machine, ML-powered modern applied statistics, Robustness.

1 INTRODUCTION

'Some time around the year 2000 a split opened up in the world of statistics.'

— Brad Efron (2020)

BY the dawn of the 21st century, the statistics community was clearly divided into two distinct camps—namely, the (parametric) statistics and the (algorithmic) machine learning. A vivid description of the cultural polarization was described by Leo Breiman (2001) in his iconic essay on "Statistical Modeling: The Two Cultures." But what has triggered this split?

ML>STAT: Breiman (2001) argued that algorithmic models are far more flexible, scalable, and accurate for complex big data problems. The traditional statistical methods, by contrast, are based on *a priori assumed* parametric models that are mainly suitable for small datasets.

STAT>ML: This algorithmic-supremacy standpoint had been furiously debated by several eminent statisticians in the discussion of the original paper. They found Breiman's claim problematic on several counts. First, despite convincing prediction results, the absence of an explicit data-generating model (a probabilistic base) can make algorithmic methods less useful for scientific investigation; see the commentaries by Cox (2001) and Parzen (2001). Efron (2020) articulated this admirably: "Abandoning mathematical models comes close to abandoning the historic scientific goal of understanding nature." Second, the black-box nature of algorithmic models makes them inscrutable, opaque, and not easily interpretable. In fact, regarding machine learning models, Breiman himself agreed that "their mechanism for producing a prediction is difficult to understand...So on interpretability, they rate an F."

Where do we stand now? To a large extent, the cultural division between these two statistical learning frameworks has been growing over the past decades, especially in the

post-deep-learning era. What options are we left with? 1) Let's get rid of black-boxes and use parametric models, at the cost of accuracy and flexibility; 2) Let's stick to model-free black-boxes that prioritizes predictive accuracy at the cost of interpretability and robustness.

What is a realistic way forward? Perhaps unsurprisingly, neither of these two extreme positions is tenable. To make real progress in developing next-generation ML technology, we are forced to consider the question: how do these 'two cultures' fit together into a coherent whole? Can we build a framework of data analysis that combines the best ideas from both sides and offer some way to make two into one? In this work, we take steps towards addressing these fundamental questions by introducing a new *Integrated Statistical Learning* (ISL) framework. Figure 1 shows a schematic description of blueprint of our research plan, which has five major components:

(1) **Uncertainty Prediction Machine:** The black-box machine learning (ML) algorithms need ways to assess their own confidence in their predictions. Without attaching some degree of uncertainty, we can't ensure fail-safe decision making with predictive models. Unfortunately, conventional ML, which is a sophisticated curve-fitting exercise, only returns point-predictions. Naturally, the question arises: Given $(x_1, y_1), \dots, (x_N, y_N)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, how can we design a simple and universal 'interface' that can convert *any* 'first generation' ML procedure into an uncertainty distribution prediction machine—one that is capable of predicting the underlying data generating mechanism in the form of conditional distribution $f_{Y|X=x}(y)$. Section 2 presents the theoretical underpinnings, principles, and methods that enable such a design.

(2) **Interpretable Learning:** We like to distill an explicit and simple statistical model out of a black-box technique for easy interpretation. Section 2 lays out such a mechanism to allow algorithmic synthesis of probabilistic models with interpretable statistical parameters.

* Corresponding author. This research was motivated by a conversation with Jerry Friedman, to whom I'm very grateful. We have also benefited from discussions with Brad Efron.

Fig. 1: Integrated statistical learning framework. ‘ML’ stands for (an arbitrary) machine learning algorithm. The power of ISL lies in producing robust and interpretable statistical models that are as expressive and scalable as ML models. For interested readers and ML-practitioners, a step-by-step guide on the mechanics of ISL is presented in the Supplementary A.4 using a stylized example.

(3) Robust Machine Learning: Current ML systems are extremely vulnerable to small perturbations of the training data, which makes them ineffective for critical applications. Having noted that Alvarez-Melis and Jaakkola (2018) raised a very important question: How to enforce robustness into existing ML methods, or how to (re)design new ones that are robust by construction? Section 2.3 develops a new *robust* learning theory to address this question.

(4) Exploratory Machine Learning: The scope of exploratory learning machine goes far beyond just prediction. It provides tools to: judge the efficacy of the starting ML model, allow model criticism to discover the ‘blind spots’ (unexplained patterns) using graphical tools, prescribe mechanism to rectify the starting predictive model, and finally verify the ‘goodness’ of the new model. In contrast to the ‘press the button’ black-box ML culture, our goal is to make it an interactive, exploratory, and testable (falsifiable) learning framework. Sec. 3 introduces several new tools to aid exploratory machine learning.

(5) Fundamentals of Applied Statistics: The new mechanics of data modeling that is presented has deep connections with the basic fundamental statistics. In section 4, we demonstrate how one can systematically derive and generalize many traditional and modern statistical methods as a special case of our framework. This could be especially useful for the practice and pedagogy of ML-powered modern applied statistics. Ultimately, interlinking Statistics and Machine Learning will enrich and revitalize both the communities, by creating excitement to work across the boundary.

Instead of tackling these five major components one at a time using separate tricks—which is the current trend—our ‘Integrated Statistical Learning’ framework seeks to develop a unified strategy.

2 INTEGRATED STATISTICAL LEARNING THEORY

2.1 Entropy, Uncertainty, and Prediction

The marginal distribution $f_Y(y)$ embodies uncertainty of Y . Now, once we observe $X = x$, the uncertainty is updated to $f_{Y|X=x}(y)$. We say that some information is received through the observation $X = x$, if the conditional distribution $f_{Y|X=x}(y)$ differs from the initial $f_Y(y)$. Therefore,

one can measure the *additional* information gain on Y after observing $X = x$ by examining the *change* in the probability distributions $f_Y(y)$ and $f_{Y|X=x}(y)$. Information theory provides an elegant framework for quantifying this “change.”

Definition 1. For a continuous response variable Y , the conditional information density function between F_Y and $F_{Y|X=x}$ is defined as

$$d(u; F_Y, F_{Y|X=x}) := d_x(u) = \frac{f_{Y|X=x}(F_Y^{-1}(u))}{f_Y(F_Y^{-1}(u))}, \quad 0 < u < 1 \quad (1)$$

which satisfies: $\int_0^1 d(u; F_Y, F_{Y|X=x}) du = 1$. To improve readability, $d(u; F_Y, F_{Y|X=x})$ will be abbreviated as $d_x(u)$, with the understanding that this *compares* marginal with conditional.

Definition 2. The Kullback-Leibler (KL) divergence from $f_Y(y)$ to $f_{Y|X=x}(y)$ is defined as

$$\text{KL}(f_{Y|x}; f_Y) = \int f_{Y|x}(y) \log \frac{f_{Y|x}(y)}{f_Y(y)} dy. \quad (2)$$

Theorem 2.1. The KL-divergence between $f_Y(y)$ and $f_{Y|X=x}(y)$ can be rewritten as KL-divergence between $d_x(u)$ (1) and uniform density over $[0, 1]$:

$$\text{KL}(f_{Y|x}; f_Y) = \text{KL}(d_x; \mathbb{1}_{[0,1]}). \quad (3)$$

Proof: First substitute $F_Y(y) = u$ in (2), and then apply the definition (1) to deduce:

$$\text{KL}(d_x; \mathbb{1}_{[0,1]}) = \int_0^1 d_x(u) \log d_x(u) du. \quad (4)$$

Remark 1. Theorem 2.1 formally implies that the deviation or proximity of $d_x(u)$ from uniformity (see Fig. 12(c) and visually compare the blue curve with the black dotted uniform line) captures the amount of (predictive) information gained on Y by virtue of a priori knowing $X = x$. For that reason, we call $d_x(u)$ the “conditional information density.” Consequently, it is not surprising that $d_x(u)$ will play a fundamental role in developing our statistical theory of prediction.

For each x_i in our dataset, we get a different d_{x_i} and a different KL-divergence value (4). Can we measure the overall predictive power of X by averaging all these KL-divergence values? To satisfactorily answer that, we have to introduce the concept of entropy.

Definition 3. Entropy is the most natural and fundamental measure of uncertainty. For a continuous random variable with distribution g , it is defined as $H(g) = -\int g \log g$.

Predictability-index can be defined as the *difference* between the entropy of Y and $Y|X$, which measures how much of our uncertainty about Y *decreased* after observing X . The following theorem establishes a beautiful connection between predictability-index and our conditional information density function.

Theorem 2.2. The average predictability of Y with respect to X (i.e., reduction in entropy) can be interpreted as how different $d_x(u)$ is from uniform distribution on average:

$$H(Y) - H(Y|X) = \mathbb{E}_X[\text{KL}(d_X; \mathbb{1}_{[0,1]})]. \quad (5)$$

The proof is given in Supplementary A.3. The important point here is that predictive information can be expressed

solely in terms of d_x , which distills the discrepancy (the “gap”) between the marginal f_Y and the conditional $f_{Y|x}$. Accordingly, conditional information density (CID) $d_x(u)$ acts as a *natural starting point* for developing a statistical theory of predictive modeling.

2.2 d -Modulated Conditional Density Representation

Definition 4 (d -modulation). Any arbitrary conditional density can be expressed as

$$f_{Y|X=x}(y) = f_Y(y) \cdot d_x(F_Y(y)), \quad (6)$$

where $f_Y(y)$ is the marginal distribution of Y and d_x is defined in Eq. (1).

A few key points about this conditional density formula:

- This new representation of conditional density is universally valid and does not impose *any* distributional assumptions on the outcome (e.g., dispersion, symmetry, skewness, heavy-tailedness, etc.). Most importantly, it decouples the conditional density into two components: the marginal f_Y (the ‘pivot’ density) and the conditional information density d_x (which encapsulates all the essential knowledge for predicting Y from $X = x$).

- We call the density perturbation formula of Eq. (6) the d -modulation of marginal Y . The function d_x *filters* out the directions by which the initial $f_Y(y)$ departs most significantly from the true $f_{Y|X=x}(y)$. For that reason, we call it, alternatively, the *contrast* density function.

We now provide another interpretation of $d_x(u)$ by introducing the notion of conditional probability integral transform (shortly, cPIT).

Definition 5. For continuous $Y \sim F_Y$ and $Y_{|x} \sim F_{Y|X=x}$, the following random variable

$$U_{|x} = F_Y(Y_{|x}) \quad (7)$$

is defined as conditional probability integral transform. Note that when $F_Y = F_{Y|X=x}$, i.e., under homogeneity, $U_{|x}$ follows uniform distribution.

Theorem 2.3. $d_x(u)$ is the distribution of $F_Y(Y_{|x})$, where, recall, F_Y is the marginal cdf of Y and $Y_{|x}$ denotes the conditional random variable $Y|X = x$.

Proof: $D_x(u) = \Pr(F_Y(Y_{|x}) \leq u) = F_{Y_{|x}}(F_Y^{-1}(u))$. Taking derivative we get the density:

$$D'_x(u) \equiv d_x(u) = \frac{f_{Y_{|x}}(F_Y^{-1}(u))}{f_Y(F_Y^{-1}(u))}, \quad 0 < u < 1.$$

Remark 2. Definition 4 uses marginal distribution as a natural pivot for representing conditional density of Y given $X = x$. However, one can choose *any* reasonable density function (parametric, nonparametric, etc.) as a pivot, as long its support¹ contains the support of $f_{Y|X=x}(y)$. The generalization of the formula (6) for arbitrary pivot $f_0(y)$ is straightforward:

$$f_{Y|X=x}(y) = f_0(y) \cdot d_x^0(F_0(y)), \quad (8)$$

where $d_x^0(u)$ is the contrast density between F_0 and $F_{Y|X=x}$; also see Supplementary A. 5.

1. Support of a distribution is the region where probability density (or mass function) is strictly positive.

2.3 A Robust Learning Theory

It is evident from the decomposition (6) that we need to estimate $d_x(F_Y(y))$ to obtain the conditional density of Y given $X = (X_1, \dots, X_p)$. In this section, we describe a special technique for estimating d_x that enjoys the following desirable properties:

(a1.) Robust as well as efficient: One of the unsolved dilemmas is to construct learning algorithms that are simultaneously robust against noise and leaks very little, if any, efficiency under the standard (unperturbed) situation². Here we will present a new *class* of statistical learning methods to balance robustness and efficiency.

(a2.) Scalable to massive high-dimensional data: An important practical goal is to have a method that works for high dimensional problems with large- (n, p) . We achieve this goal by developing a new *class* of machine learning powered nonparametric estimation method.

(a3.) Interpretable: It would be nice to have a tractable (and preferably smooth) model for the contrast density $d_x(u)$ with a few interpretable parameters that can capture increasingly complex shapes of the conditional distributions.

It's a challenging task to reliably learn the “true” relationship between the input features X and output Y from big noisy datasets. To ensure that we succeed at it, our method has to be ‘doubly-robust,’ and thus tackle possible contamination in *both* response and covariates. We will approach the estimation of d_x in two stages: first, we will make it robust with respect to the outcome variable Y , and then we will discuss defence strategies to guard against corrupted input X .

Y-Robustness. Substituting $u = F_Y(y)$ in the definition (1), allows us to represent the ratio $f_{Y|X=x}(y)/f_Y(y)$ as a function of the rank-transform (i.e., probability integral transform) $F_Y(Y)$ through $d_x(F_Y(y))$. This is the first *major* step towards ensuring the Y -robustness of our method. As a result, one can expand $d_x(F_Y(y))$ in the orthonormal basis of $F_Y(y)$. One such orthonormal system is the LP-family of rank-polynomials (Mukhopadhyay and Parzen, 2020, Mukhopadhyay and Wang, 2020, Mukhopadhyay, 2017), which we denote $\{T_j(y; F_Y)\}$ to emphasize that they are polynomials of $F_Y(Y)$, hence extremely robust. As the true F_Y is unknown, we instead use the empirical LP-bases (eLP) $\{T_j(y; \hat{F}_Y)\}$ for our data analysis. Supplementary A.2 describes its construction and properties.

Summing up the above discussion, we have the following LP-series representation:

$$d_x(F_Y(y)) = 1 + \sum_j \text{LP}_{j|x} T_j(y; F_Y). \quad (9)$$

The problem of estimating the function d_x now boils down to estimating the orthogonal coefficients $\text{LP}_{j|x}$. To get an expression for the conditional LP-coefficients, first recall that

2. This ‘robustness-efficiency’ dilemma (also known as ‘rank-and-size’ duality) is not new. It is at least 60 years old unsettled issue of Statistics; see Box (1953) and Tukey (1960).

T_j 's are mean zero random variables and are orthonormal with respect to measure F_Y , i.e.,

$$\int T_j(y; F_Y) T_k(y; F_Y) f_Y(y) dy = \delta_{jk}, \quad (10)$$

where δ is the Kronecker delta function. Consequently, we multiply both sides of the equation (9) with $T_j(y; F_Y) \cdot f_Y(y)$ and then integrate to produce the following identity:

$$LP_{j|x} = \int_{-\infty}^{\infty} d_x(F_Y(y)) T_j(y; F_Y) f_Y(y) dy, \quad \text{for } j = 1, 2, \dots \quad (11)$$

This immediately yields the following important theorem.

Theorem 2.4. The j th conditional LP-Fourier coefficient admits the following representation

$$LP_{j|x} = \mathbb{E}[T_j(Y; F_Y) | X = x]. \quad (12)$$

Theorem 2.4 allows us to interpret the LP-coefficients $LP_{j|x}$ (which are surely function of x) as a specialized regression function, where we regress $T_j(y; F_Y)$ on X . This is extremely handy, because one can now use *any general* machine learning method to estimate these coefficients.

X-Robustness. Nevertheless, the computational theory (12) is still not fully satisfactory, as corrupted (or adversarially perturbed) X can substantially influence the coefficients. To provide further defense, we need one last piece of idea, which is a little-known but fundamental fact about ‘regression *via* rank-transform.’ Also see Supplementary A5.2.

Theorem 2.5. For any general random variable X (whether discrete or continuous) and a measurable function $\Psi(\cdot)$, the following result holds

$$\mathbb{E}[\Psi(Y) | X] = \mathbb{E}[\Psi(Y) | F_X(X)], \quad \text{with probability 1} \quad (13)$$

Proof: We start with a basic quantile mechanics fact: for any random variable X we have

$$Q_X(F_X(X)) = X, \quad \text{with probability 1.}$$

where $Q_X(u) = \inf_x \{F_X(x) \geq u\}$, $0 < u < 1$ denotes the quantile function of X . Accordingly, any function of X , say $h(X)$, can be equivalently rewritten as a function of the rank-transform $F_X(X)$, since $h(X)$ is equal to $h \circ Q_X(F_X(X))$ with probability 1.

In plain language, this result proves the ‘sufficiency’ of rank-transform (retaining *all* the information) for estimating conditional mean function. Therefore, we can estimate $LP_{j|x}$ (12) by regressing $T_j(Y; \tilde{F}_Y)$ on the LP-bases of X . This will allow doubly-robust estimation of d_x *without sacrificing* an iota of efficiency in the process. Next, we will go over some examples, to demonstrate how this theory actually translates into practice (algorithm).

Example 1. *The Butterfly data.* The butterfly data superposition of $N = 350$ samples, generated from $y = x + \epsilon$ and $y = -x + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$; see Supp. Fig. 11. The scatter plots of $T_j(y; \tilde{F}_Y)$ against $u = \tilde{F}_X(x)$ for $j = 1, \dots, 4$ are displayed in Fig. 2. We refer these graphs as LP-scatter plots. The red curves denote knn-estimated regression curves. In the R computing language, this is simply `knn(T_j(y; \tilde{F}_Y) ~ T_X)`, ($j = 1, \dots, 4$) where $T_k(x; \tilde{F}_X)$ is the k th col of

Fig. 2: First 4 LP-scatter plots for the butterfly data. It illustrates the steps for obtaining $\hat{LP}_{j|x}$ using generic machine learning-program, by operating on the LP-transformed domain.

Fig. 3: BUPA liver data: (a) The ‘wiggly’ blue line denotes the fitted gbm regression curve; (b) the estimated contrast density function $d_{x_0} \equiv d(u; F_0, F_{Y|X=x_0})$. It is asymmetric and inverted ‘U’-shaped; (c) the estimated $f_{Y|x_0}(y)$ along with the starting pivot $f_0 = \mathcal{N}(3.35, .5^2)$.

LP-transformed feature matrix T_X . Of course, instead of knn method, one can choose any machine learning routine; the next example uses gradient boosting algorithm (gbm). To get the estimated conditional LP-coefficients at $x_0 = 2$ (indicated by the blue dots) we record the predicted values at $u_0 = \hat{F}_X(x_0)$: $\hat{LP}_{2|x_0} = -.020$, $\hat{LP}_{4|x_0} = 0.47$, and the rest are practically zero. Substituting them into (9) yields the estimated contrast function $\hat{d}_{x_0}$, as was shown in Fig. 12(c).

Example 2. *BUPA liver disorders data.* This is a popular dataset (available in the UCI ML repository) used by many researchers, including Breiman (2001). It consists of measurements of gamma-glutamyl transpeptidase (GGT) and alanine-aminotransferase (ALT) extracted from $N = 345$ male individuals’ blood samples. High levels of ALT and GGT indicate a damaged liver. Fig. 3 (a) shows the scatter plot of $y = \log(\text{ALT})$ versus $x = \log(\text{GGT})$. We seek to predict the distribution of log-ALT levels given $x = 3.5$; see the red triangle in panel (a) of Fig. 3.

Our data analysis scheme can be simply summarized as follows:

1) We start by modeling the conditional mean function using gradient-boosting algorithm. The rugged blue line denotes the estimated regression curve $\hat{\mu}(x)$.

2) To estimate the conditional density at $x_0 = 3.5$, we choose $\mathcal{N}(\hat{\mu}(x_0), \sigma_y^2)$ as our pivot model, where $\hat{\mu}(x_0) = 3.35$ and $\sigma_y = 0.5$.

3) Following theorems 4 and 5, we estimate the conditional LP-Fourier coefficients $LP_{j|x_0}$ and the associated contrast density function: $\hat{d}_{x_0}(F_0(y)) \approx 1 - 0.72 T_2(y; F_0) + 0.10 T_3(y; F_0)$. This is shown in the middle panel of Fig. 3.

4) Interestingly, we can interpret the coefficients of $\hat{d}_{x_0}$: the large value of $\hat{LP}_{2|x_0}$ indicates that the selected pivot needs scale-correction ("second-order" correction) to get close to the true conditional density at x_0 . In addition, the negative sign implies the variability has to be reduced, which is vividly apparent from Fig. 3(c).

5) Similarly, the presence of third-order (positive) LP-coefficient indicates the symmetric normal pivot needs to be tilted to the right. The positive skewness of $f_{Y|x_0}(y)$ is probably justified due to the presence of few large y -values³ around $x_0 = 3.5$, as indicated by squares.

6) Finally, the shape of $\hat{d}_x$ provides quick exploratory guidance for choosing appropriate families of parametric conditional distributions. For the BUPA liver-disorders data, one could select skew-normal or asymmetric logistic distribution (Friedman, 2020) as a model for $Y_{|x}$.

2.4 The Basic Algorithm: ISL

We are now in a position to provide the blueprint of the algorithmic interface in Fig. 1 that converts a user-selected machine learning method into an integrated statistical learning machine (ISL) in a robust and interpretable manner. The theory in the previous section, makes it absolutely straightforward and extremely easy to implement, which essentially involves three steps. The following algorithm presents a flowchart of the core computational steps in pseudo-code, so that one can implement it using *any* programming language or ML-platform. We have found that the H2O open-source (see <https://github.com/h2oai>) is extremely efficient. But users can choose any ML environment (e.g., TensorFlow, Azure, AutoGluon, etc.) to operationalize the following workflow.

ISL: A flowchart of the core computational steps in pseudo-code

Input Data (X_i, y_i) for $i = 1, \dots, N$; user-selected machine learning method (ML); values of m_y and m_x . And the target test-point $x_0 \in \mathbb{R}^p$.

```

Step 1.  $T_Y \leftarrow LP.basis(y, m_y); T_X \leftarrow LP.basis(X, m_x)$ 
for( $j = 1, \dots, m_y$ )
{
Step 2.  $Fitted.ML_j \leftarrow ML(T_Y[, j] \sim T_X)$ 

```

3. Classical robust methods treat them as "outliers" and devise mechanisms for removing them from analysis. This is incorrect. A careful examination will reveal that these large- y values (indicated by squares) have some pattern; they are not randomly scattered (like outliers). The real question of what is an outlier and what is a "real" pattern, finally boils down to artfully balancing robustness with efficiency.

```

Step 3.  $\hat{LP}_{j|x_0} \leftarrow predict(Fitted.ML_j, x_0)$ 
}
Return  $\hat{d}_{x_0} \leftarrow 1 + \sum_j \hat{LP}_{j|x_0} T_Y[, j]$ 

```

The LP-transformation is implemented by the function `LP.basis`; T_X is the stacked matrix $[T_{X_1} | T_{X_2} | \dots | T_{X_p}]$, where the j th column of the matrix T_{X_ℓ} is simply $T_j(x; \tilde{F}_\ell)$. Step 1 makes the whole procedure robust, non-linear, and *invariant* under monotone transformations. Step 2 performs the training of the ML method and step 3 predicts the fitted value at the desired $X = x_0$. Based on our experience, $m_x = 4$ and $m_y = 6$ work reasonably well for most problems, including all the datasets analyzed in this paper. Finally, one can generate the uncertainty distribution $f_{Y|x_0}$ by applying the d -modulation formula of equation (6).

Remark 3. In this section we have showed: how to actualize the vision sketched in Fig. 1; how to compress-down large complicated algorithmic methods into concise parametric forms with few interpretable coefficients⁴; how to derive a smoothness-promoting model for the response Y from black-box prediction methods; and how to robustify algorithmic techniques⁵. We have achieved each of these ends by devising a generic mechanism that holds for *any* machine learning method. Furthermore, Section 4 shows how these same modeling principles can unify a wide range of traditional and novel statistical methods.

2.5 d -Kernel Smoothing and Importance Weighting

In many applications, we are often interested in estimating quantities like $\mathbb{E}[\Psi(Y)|X = x]$ for general $\Psi(\cdot)$ function. This can be rewritten as

$$\begin{aligned} \mathbb{E}[\Psi(Y)|X = x] &= \int_y \Psi(y) f_{Y|X=x}(y) dy \\ &= \int_y \Psi(y) f_Y(y) d_x(F_Y(y)) dy, \end{aligned}$$

by applying the formula (6). This leads to the following theorem, which expresses the (local) conditional mean as a weighted average of the global samples $\{y_1, \dots, y_N\}$.

Theorem 2.6 (d -kernel smoothing). Conditional expectation of $\Psi(Y)$ given $X = x$ can alternatively be expressed as:

$$\mathbb{E}[\Psi(Y)|X = x] = \mathbb{E}[\Psi(Y) \cdot d_x(F_Y(Y))], \quad (14)$$

the rhs expectation is taken over the marginal distribution of Y .

Remark 4. The above representation implies the following d -weighted estimator:

$$\hat{\mathbb{E}}[\Psi(Y)|X = x] = \frac{1}{N} \sum_{i=1}^N w_i(x) \Psi(y_i), \quad (15)$$

where the empirical weighting function $w_i(x) = \hat{d}_x(\tilde{F}_Y(y_i))$. The formula (15) can be interpreted as a

4. This materialize the objective of "Trend 1" as proposed by Efron (2020): "Trend 1 aims to make the output of a prediction algorithm more interpretable, that is, more like the output of traditional statistical methods."

5. Friedman (2020) advocates strongly for this.

specially-designed kernel smoothing, which could be extremely handy in some cases. Three particular examples are given below:

- 1) $\Psi(Y) = \mathbb{I}(Y > k)$ gives the conditional probability $\Pr(Y > k|X = x)$; This can be used to construct different risk measures based on tail events;
- 2) $\Psi(Y) = Y$ recovers the conditional mean $\mathbb{E}[Y|X = x]$;
- 3) $\Psi(Y) = (Y - \mathbb{E}(Y))^2$ yields conditional variance $\text{Var}(Y|X = x)$, whose square-root, i.e., the standard error of $Y_{|x}$, is often used as an uncertainty quantification measure for a prediction. However, we should point out that standard error is not always a prudent choice for quantifying reliability of a prediction, specially when the conditional distribution is heavy-tailed, or skewed, or multi-modal; see Section 4.6.

3 EXPLORATORY LEARNING MACHINE

Our goal here is to develop some graphical diagnostic methods to make a transition from a *black-box* answer machine to an *exploratory* learning machine that can drill deeper and generate previously unanticipated, interesting questions to facilitate data-driven scientific discovery and understanding. This can be viewed as a pre-deployment “health check-up” for statistical learning methods that comes with a set of nonparametric exploratory tests.

3.1 Heterogeneity Component Analysis

Heterogeneity component analysis (HCA) is a technique for identifying *dynamic* components of $Y_{|x}$ —the aspects of the conditional distribution $f_{Y|X=x}(y)$ (such as location, scale, skewness, tail, etc.) that vary with covariates. We illustrate the main idea using the *butterfly* example. The procedure starts with LP-scatter plots, as was shown in Fig. 2. The first scatter plot captures how the location of $f_{Y|X=x}(y)$ (conditional mean) changes with x ; the second one captures the change of scale (the nature of heteroscedasticity); the third one provides information on the changing skewness and so on. Thus, an obvious approach for detecting the heterogeneity components would be to:

Step 1. Compute a goodness-of-fit statistic for the j th LP-scatter plot by its coefficient of multiple correlation: $R_j^2 = \text{Proportion of variance in } T_j(y; F_Y) \text{ that is explained by } \{T_1(x; F_X), \dots, T_m(x; F_X)\}$.

Step 2. Assess the significance of the R_j^2 by constructing F-statistic for the j th LP-scatter

$$F_j = \frac{R_j^2}{1 - R_j^2} \times \frac{N - m - 1}{m}$$

whose null distribution is $F_{m, N-m-1}$. Fig. 12(a) displays the significant (whose pvalue < 0.05) heterogeneity components for the *butterfly* data.

HCA Diagnostic Plot. For real data analysis, we apply the above method on the *residuals* to determine what structure of the data (relationship between y and x) remains to be described after the ML-regression. The magnitude of the heterogeneity components reveals which shape parameters of the residuals are changing with x . If all the components of HCA are small, then it confirms to the practitioner that

Fig. 4: Heterogeneity component analysis (HCA): Top row shows the bone mineral density data. Beyond conditional mean (regression), the scale and skewness (symmetry) of the error distribution are also changing with x . Bottom row is the online news popularity data. Left plot shows the distribution of the standardized residuals and the right one is the associated HCA plot. The variability needs to be modelled as a function of the covariates for better fitting.

errors are already homogeneous (we call this scenario: residual “flatland”) with respect to the covariates and nothing left to be modeled.

Example 3. Bone mineral density data (Bachrach et al., 1999). It contains measurements on the relative change in spinal bone density (over one year period) of $N = 485$ North American adolescents, as a function of age. This dataset was previously analyzed by Hastie et al. (2009, p. 152). Fig. 4(a) shows the data along with the estimated regression smoother. We like to understand whether there is any “excess” heterogeneity beyond the 1st-order regression modeling. To check the homogeneity of the residuals with respect to the covariate x , we apply our HCA procedure on the residuals $y_i - \hat{y}_i$. The result is shown in the top row of Fig. 4, which says that the regression model has not taken into account the heteroscedasticity of the data (furthermore, there is some evidence of varying skewness in the conditional distribution). These unmodelled (dynamic) heterogeneity components should be modeled to improve the prediction. The heterogeneity is partly due to the presence of both male and female youngsters’ in the dataset.

Example 4. Online news popularity data (Fernandes et al., 2015). The objective of this study is to predict the popularity of online news. The data⁶ consist of $N = 39,644$ articles,

6. The article in row # 31,038 seems to be an outlier (in fact it is an erroneous data: check its 4th and 5th feature values—rate of unique and non-stop words—which can’t be more than 1). Due to the *doubly-robust* property, our method can handle this easily. But standard machine learning methods might get unduly influenced by this one data point (article); see Fig. 17 in the Supplementary.

and for each article we have y ($\log_{10}$ of number of shares in social networks, a measure of popularity) and $p = 59$ extracted features (such as number of words in the title, number of links, number of images, etc.) x . This dataset was recently analyzed by Friedman (2020).

We start by fitting gradient boosting regression to the data. At this point, it is a natural question to ask whether the gbm captured most of the important information in the data. If not, what is missing? To spot the unmodelled dynamic components, we start from the residuals, as shown in Fig. 4(c). The HCA diagnostic plot of panel (d) is constructed based on the Lasso-selected features, which clearly indicates the presence of a strong heteroscedasticity. However, it is also true that the conditional distribution will be asymmetric and long-tailed, but our analysis shows that they are not swiftly changing with covariates; they are more or less static in nature, not dynamically evolving shape-parameters. This further corroborates the findings of Friedman (2020, Sec. 6.2).

3.2 Pivot Uncertainty Modeling

The parameters $LP_{j|x}$ of the contrast density d_x can be viewed as “generalized coordinates” that describe the shape of $f_{Y|X=x}(y)$ relative to some reference distribution (pivot) $f_0(y)$.

In fact, these shape-coefficients $LP_{j|x}$ capture the whole dynamics of how the conditional distribution $f_{Y|X=x}(y)$ evolves with x .⁷ Naturally, a large value of

$$\text{qPivot}(x) = \sum_j |LP_{j|x}|^2 \quad (16)$$

indicates higher uncertainty of the pivot f_0 for modeling the conditional $f_{Y|X=x}$. The ‘q’ in qPivot stands for quantification of lack-of-fit. The above formula can also be interpreted as measuring the deviation of $d_x(u)$ from uniformity (follows from Parseval’s identity):

$$\int_0^1 (d_x(u) - 1)^2 du = \sum_j |LP_{j|x}|^2 \quad (17)$$

Example 5. *Movie box-office revenue data* (Voudouris et al., 2012). The goal of this study is to build a forecasting model for film revenues. The data contain $N = 4031$ pairs of observations (x_i, y_i) on the 1990s film data, where x_i is the log of opening box-office revenues and y_i is the log of box office revenues after the first week. We are interested in predicting the distribution of the outcome y for $x = 12$ and 14—marked with the blue and green dotted line in Supp. Fig. 18 (a). Our starting pivot $f_0 = \mathcal{N}(14, \sigma_y = 3.25)$ is shown in the panel (b). To check whether the conditional density at x follows this presumed parametric law, we estimate the respective contrast densities; see the middle panel of Supp. Fig. 18. The bimodality of $\hat{d}_{x=12}$ indicates that the Gaussian pivot failed to capture the presence of films with two kinds of box-office earnings. This is also reflected in the pattern of the scatter plot, which is divided into *two* branches. On the other hand, the inverted ‘U’ shape (quadratic) of $\hat{d}_{x=14}$ implies that the

variability of the Gaussian pivot needs to be adjusted (2nd-order correction) to go from $f_0(y) \mapsto f_{Y|X=14}(y)$. The final refurbished conditional density models at $x = 12$ and 14 are shown in the bottom row of Supp. Fig. 18. They are estimated using the d -modulation technique, described in Sec. 2.2.

3.3 Goodness-of-fit Diagnostics

The flexibility of our approach (as depicted in Fig. 1) enables us to choose *any* machine learning method to design the integrated learning machine. Nonetheless, a user may want to know which ML-powered distribution prediction algorithm(s) better explains the pattern in the data. To answer that here we introduce a graphical exploratory procedure to evaluate the overall goodness-of-fit of the method. Deploying statistical models without knowing how well it actually fits the data, is a statistical sin. We illustrate our idea using the *butterfly* data.

ML-driven ISL system. To construct an ISL for the butterfly data, we start with three base learners: k-nearest neighbors (knn), random forest (RF), and gradient-boosting algorithm (GBM). We train each model using the generic ISL-procedure described in section 2.4. To check whether these models are congruent with the observed data, we proceed as follows:

Step 1. Compute the generalized quantile-residuals for all three models on a hold-out validation dataset (separate from the training samples) of size n_v

$$U_{\ell i} = \hat{F}_{Y|X=x_i}^\ell(y_i), \text{ for } i = 1, \dots, n_v; \text{ and } \ell = 1, 2, 3. \quad (18)$$

Note that if the model accurately captures the pattern in the validation data, then we expect the distribution of the U_ℓ ’s to be close to uniform.

Step 2. Display the histograms and QQ-plot of $\{U_{\ell 1}, \dots, U_{\ell n_v}\}$. This is done in Fig. 19 for the *butterfly* data, which shows the knn model satisfactorily captures the observed reality.

Step 3. To quantify the model-fit we propose the following test statistic for the ℓ -th model:

$$\text{qDIV}_\ell = \sum_{j=1}^m \left| n_v^{-1} \sum_{i=1}^{n_v} \text{Leg}_j(U_{\ell i}) \right|^2, \quad (19)$$

where Leg_j denotes the j th orthonormalized shifted Legendre polynomial on unit interval. The qDIV statistic measures the distance between $\text{Uniform}[0, 1]$ and the sample distribution of U_ℓ . We have used $m = 6$ in all our examples. One can also compute pvalues using the χ_m^2 null distribution of qDIV. For more details see Mukhopadhyay and Wang (2020).

Step 4. It is also interesting to note that one can *directly* compute these generalized quantile-residuals from the contrast distribution function, since (by substituting $F_0(t) = u$):

$$\begin{aligned} F_{Y|X=x}(y) &= \int_{-\infty}^y f_0(t) d_x(F_0(t)) dt \\ &= \int_0^{F_0(y)} d_x(u) du = D_x(F_0(y)). \end{aligned}$$

7. Indeed, one can even construct an m -dimensional phase-space (analogous to statistical mechanics), consisting of N points $\{LP_{1|x_i}, \dots, LP_{m|x_i}\}_{i=1}^N$ to graphically represent the evolving shape of $f_{Y|X=x_i}(y)$. Each point in LP-phase-space corresponds to a particular shape for a particular value of $x = (x_1, \dots, x_p)$.

For the butterfly data qDIV values (with p-values in the parentheses) for knn, random forest, and gbm respectively are 0.0466 (0.82), 1.12 (1.4×10^{-10}), and 0.386 (0.0026). These statistic values can be used as a general measure of “lack-of-fit” for comparing (ranking or even tuning hyperparameters) the performances of different ML-engine based ISL models.

4 SOME CONNECTIONS AND UNIFICATION

The theoretical constructs of our integrated learning framework are not based on an isolated trick, but are fundamental to statistics. Our formalism and principles of data modeling unify many traditional and modern statistical methods. Its significance on the practice and pedagogy of modern applied statistics is described in Supplementary A.6. In this section, we will report some surprising findings to highlight this fact.

4.1 K-sample Comparison Problems

Given random samples from k distributions $F_i(y)$, $i = 1, \dots, k$ we seek to test the following null hypothesis:

$$H_0 : F_1(y) = \dots = F_k(y) \text{ for all } y, \quad (20)$$

with the alternative hypothesis that at least two distributions are different. Before attacking this general problem of equality of distributions, we start with a simpler problem of testing the equality of means. The Kruskal–Wallis statistic provides a nonparametric test for it.

Definition and notation. We have k mutually independent random samples with sizes $n_1, \dots, n_k$ with the combined sample size N ; R_i denotes the rank of Y_i in the pooled samples; G_ℓ is the set of indices for the ℓ -th group. The Kruskal–Wallis test statistic is defined as:

$$KW = \frac{12}{N(N+1)} \sum_{\ell=1}^k n_\ell \left\{ \frac{1}{n_\ell} \sum_{i \in G_\ell} R_i - \frac{N+1}{2} \right\}^2. \quad (21)$$

One may go one step further and ask how to test the equality of scale parameters (variability over different groups), which can be tested using Mood statistic, defined as

$$\text{Mood} = \frac{180}{N(N+1)(N^2-4)} \sum_{\ell=1}^k n_\ell \left\{ \frac{1}{n_\ell} \sum_{i \in G_\ell} \left(R_i - \frac{N+1}{2} \right)^2 - \frac{N^2-1}{12} \right\}^2 \quad (22)$$

For testing high-order differences like skewness (not to mention kurtosis, which will take almost two full lines to write), the formula of the test statistic gets even more complicated.

$$\text{SKEW} = \frac{7}{N(N+1)(N^2-4)(N^2-9)} \sum_{\ell=1}^k n_\ell \left\{ \frac{20}{n_\ell} \sum_{i \in G_\ell} \left(R_i - \frac{N+1}{2} \right)^3 - \frac{3N^2-7}{n_\ell} \sum_{i \in G_\ell} \left(R_i - \frac{N+1}{2} \right) \right\}^2 \quad (23)$$

A practitioner at this point might wonder how to make sense of these monstrous formulae? Is any intuitive or systematic derivation possible? The answer is yes. We just have to look at the problem from a different perspective. A modern introduction to the k -sample comparison problem is given below.

Step 1. View it as a (X, Y) problem: Define X to be the group-index variable taking values $\{1, \dots, k\}$. We now have $\{(x_i, y_i)\}_{i=1}^N$ data, which can be displayed as a scatter plot.

Step 2. Reformulate H_0 : The k -sample null-hypothesis of equality of distributions can now be viewed as a test of the homogeneity problem (see Section 3.1) $f_{Y|X=\ell}(y) = f_Y(y)$ for $\ell = 1, \dots, k$. Now since

$$f_{Y|X=\ell}(y) = f_Y(y) \left\{ 1 + \sum_j \text{LP}_{j|\ell} T_j(y; F_Y) \right\}, \quad (24)$$

the original k -sample hypothesis (20) can alternatively be rewritten as

$$H'_0 : \text{LP}_{j|\ell} = 0, \text{ for all } j.$$

Hence, intuitively, it makes sense to apply the HCA diagnostic test of section 3.1 on the (x, y) data. Doing so we get almost all of the known results of k -sample modeling in one-shot. To justify the statement, we start with the following theorem.

Theorem 4.1. The Kruskal–Wallis statistics can be expressed as

$$KW = \sum_{\ell=1}^k n_\ell |\widetilde{\text{LP}}_{1|\ell}|^2 = NR_1^2. \quad (\text{change in location}) \quad (25)$$

where the first-order LP-multiple correlation R_1^2 is defined in the step 1 of Section 3.1.

The proof is in Supplementary A.3. The beauty of this result is that it seamlessly generalizes to high-order cases (proofs are tedious but not difficult and are left to the reader):

$$\begin{aligned} \text{MOOD} &= \sum_{\ell=1}^k n_\ell |\widetilde{\text{LP}}_{2|\ell}|^2 = NR_2^2; \quad (\text{change in scale}) \\ \text{SKEW} &= \sum_{\ell=1}^k n_\ell |\widetilde{\text{LP}}_{3|\ell}|^2 = NR_3^2; \quad (\text{change in skewness}) \\ \text{KURT} &= \sum_{\ell=1}^k n_\ell |\widetilde{\text{LP}}_{4|\ell}|^2 = NR_4^2, \quad (\text{change in tail}) \end{aligned}$$

and so on. The practical benefit of these results is that they provide a systematic and comprehensive k -sample analysis program that is radically simple to implement, since LP-multiple correlations R_j^2 can be computed in a single line R-code: `summary(lm( $T_j(y) \sim T_X$ ))$r.squared`.

Example 6. LDL cholesterol of Quail. This is a randomized experiment investigating LDL (low-density lipoprotein) cholesterol in quails; $N = 39$ quails were randomly assigned to $k = 4$ diets for a specific period of time, where each diet mixed with a different drug compound. The boxplot is shown in Fig. 5. The scientific question is: whether or not different drug compounds change the LDL

Fig. 5: LDL cholesterol data. (a) shows the boxplots of cholesterol values over four different drug components; (b) this is essentially the stretched HCA-plot. R_j^2 were multiplied by the sample size. The asterisk-sign $*$ indicates the associated test statistic has p-value < 0.10 . In R, `kruskal.test(y ~ x)` yields the value 7.18, which is practically same as ours using (25).

cholesterol levels. The right panel of Fig. 5 shows the rescaled HCA-plot, where we multiply the j th LP-multiple-correlations R_j^2 by the sample size N . It shows that the distribution of LDL cholesterol levels is changing in scale and location.

Remark 5. For this data, the classical anova F-test yields p-value 0.345, thus fails to detect any significant location differences—clearly contradicting the boxplots. Why has the F-test failed? There are two primary reasons for that: (i) the highly non-Gaussian nature of the distribution of Y along with a few outlying values were enough to kill the parametric F-test; and (ii) the assumption of equal variance is also seems inaccurate. The takeaway: we need *both* robustness and efficiency to reliably detect high-order distributional differences.

4.2 Probabilistic Index Model

Do baseball players gain weight as they get older? Can a larger dose of an antipsychotic drug reduce the severity of depression (after adjusting for the confounders)? All of these problems are concerned with modeling the *comparison probability regression*: $\text{CPR}(x) = \Pr(Y \leq Y_{|x})$, as a function of x , instead of as a usual conditional mean regression. $\text{CPR}(x)$ encodes the probability that a randomly selected subject from the population with $X = x$ has a higher response value than a randomly selected subject from the full data. This concept was pioneered by Olivier Thas under the name of probabilistic index model; see Thas et al. (2012). To understand how $\text{CPR}(x)$ is related to our framework, we start with the following integral representation:

$$\Pr(Y \leq Y_{|x}) = \int F_Y(y) dF_{Y|X=x}(y). \quad (26)$$

This leads to the following important result.

Theorem 4.2. $\Pr(Y \leq Y_{|x})$ can be expressed as a regression of probability integral transform $F_Y(Y)$ on X :

$$\Pr(Y \leq Y_{|x}) = \mathbb{E}[F_Y(Y)|X = x] = \frac{1}{\sqrt{12}} \text{LP}_{1|x} + 1/2. \quad (27)$$

Fig. 6: (color) Baseball data. (a) Scatter plot of age versus weight of $N = 1015$ major-league baseball players; (b) the density of weights for four different age groups.

The first equality follows from (26), and the second equality follows from supplementary equation (E5.4). The representation (27) justifies the name comparison probability regression, which is simply a standardized 1st-order LP-regression function. As a consequence of Theorem 4.2, the whole procedure can be implemented in one line of R-code: `lm( $T_1(y) \sim T_X$ )`.

Example 7. Baseball data (Matloff, 2017). Fig. 6 shows the age versus weight scatter plot of $N = 1015$ major-league baseball players. We would like to investigate whether baseball players gain weight as they age. In other words, our task is to check whether $\Pr(Y \leq Y_{|x})$ increases as a function of age or not. We proceed as follows:

Step 1. Compute the LP-basis functions: $T_1(y; \tilde{F}_Y)$ for the response and $T_X = [T_1(x; \tilde{F}_X), \dots, T_m(x; \tilde{F}_X)]$ for the predictor variable. Here we have picked $m = 4$.

Step 2. (penalized) LP-regression: Perform the multiple linear regression $T_1(y; \tilde{F}_Y)$ on T_X . Return the Bayesian information criterion (BIC) selected model. If no variable is selected then it supports the null hypothesis $H_0 : \Pr(Y \leq Y_{|x}) = 0.5$. For the baseball data we get:

$$\tilde{\text{LP}}_{1|x} = \mathbb{E}[T_1(y; \tilde{F}_Y)|X = x] = 0.17 T_1(x; \tilde{F}_X). \quad (28)$$

There is no intercept as $\mathbb{E}[T_1(y; \tilde{F}_Y)] = 0$, by construction. Although athletes strive to keep physically fit, eq. (28) seems to suggest that the baseball players do gain some weight (in a slow linear rate) over time. Fig. 6(b) makes this somewhat clear to understand how.

4.3 Shape Predictors: Generalized Feature Selection

When dealing with a large number of covariates, it is often important to understand which are the most *predictive* for the response variable Y . Currently, the most popular feature selection method is L_1 -penalized lasso regression (Hastie et al., 2009, Ch. 3.4), which minimizes the residual sum of squares $\|y - X\beta\|_2^2$ subject to $\|\beta\|_1 \leq \lambda$. This will be denoted algorithmically as `lasso( $y \sim X$ )`. It is important to note that, when we perform lasso-regression of y on X , we only get ‘first-order’ informative features—the variables that influence the conditional mean $\mathbb{E}[Y|X = x]$. Naturally, the question is: can we go beyond mean, and find the most relevant attributes to describe the *shape* of the conditional distribution $f_{Y|X=x}(y)$? Can we better utilize the lasso-technique to find those “shape predictors”? The following

Fig. 7: Online news data: Top ten generalized shape predictors are shown. They are classified based on how they impact the response distribution $f_{Y|X=x}(y)$.

is an attempt to construct such a generalized feature selection program, which is extremely easy to implement. The algorithmic logic proceeds as follows:

Step 1. Let's start by recalling our basic model: $f_{Y|X=x}(y) = f_Y(y) \cdot d_x(F_Y(y))$. The covariates $x = (x_1, \dots, x_p)$ can influence the distribution of the response Y only through $d_x \circ F_Y(y) = 1 + \sum_j LP_{j|x} T_j(y; F_Y)$, which is governed by the following system of equations:

$$\begin{aligned} LP_{1|x} &= \mathbb{E}[T_1(Y; F_Y)|X = x] \\ LP_{2|x} &= \mathbb{E}[T_2(Y; F_Y)|X = x] \\ &\vdots \end{aligned}$$

The first equation captures the dynamics of conditional mean as a function of x , while the second one describes how the variability (scale) changes with x , and so on.

Step 2. Thus, it is quite obvious that one can extract the j th order features, denoted by $\mathcal{O}_j = \{\ell : \hat{\beta}_\ell^{(j)} \neq 0\}$, where $\hat{\beta}^{(j)}$ is the output of $\text{lasso}(T_j(y) \sim X)$, which efficiently computes the solution of

$$\hat{\beta}^{(j)} = \arg \min_{\beta^{(j)} \in \mathbb{R}^p} \left\{ \|T_j(y) - X\beta^{(j)}\|_2^2 + \lambda \|\beta^{(j)}\|_1 \right\}. \quad (29)$$

Accordingly, $\mathcal{O}_1$ is the collection of location-informative variables, $\mathcal{O}_2$ contains all the scale-informative ones, $\mathcal{O}_3$ variables are responsible for predicting change in skewness of Y , etc.

Step 3. To make the analysis doubly-robust and to allow for non-linearity of the features, simply substitute X by the LP-transformed matrix T_X , as we did in Sec. 2.4.

Definition 6. Let $\hat{\beta}_{\ell k}^{(j)}$ be the estimated lasso regression-coefficient for $T_k(x; \tilde{F}_\ell)$. Then define the Omnibus Variable Importance Score (Ovis) for X_ℓ as

$$\text{Ovis}_\ell = \sum_j \sum_k |\hat{\beta}_{\ell k}^{(j)}|^2 \quad \text{for } \ell = 1, \dots, p. \quad (30)$$

Remark 6 (Portability). Our procedure for finding generalized shape predictors is extremely flexible: one can easily integrate it with *any* machine learning algorithm. Train the model $\text{ML}(T_j(y) \sim X)$, and plug that into a feature-importance wrapper e.g., `h2o.varimp()` from `h2o` R-package, `varImp()` from `caret` R-package, or `FeatureImp()` from `iml` R-package, etc.

$\mathcal{O}_1$ (location)	$\mathcal{O}_2$ (scale)	$\mathcal{O}_3$ (skewness)
kw_avg_avg is_weekend kw_min_max self_reference_avg_shares data_channel_is_tech	is_weekend LDA_04 timedelta kw_min_min LDA_03	kw_max_max LDA_03 data_channel_is_socmed data_channel_is_tech weekday_is_saturday

TABLE 1: Online news data: Top five variables for each of the three categories are shown. Investigators can use this tool to identify different sources of heterogeneity.

Example 8. *Online news popularity data, continued.* Which attributes are most important for predicting the popularity of a given article? More specifically, which variables affect the changing shape of the conditional distribution? Our approach consists of four main steps:

- First, determine which shape parameters of $f_{Y|X=x}(y)$ are changing with x ? In Fig. 4, we have already seen that scale and skewness are the two principal dynamic components, in addition to location.

- Second, to determine the variables that impact the changing location, scale, and skewness we perform the following LP-lasso regression⁸

$$\text{lasso}(T_j(y) \sim T_1(x_1) + \dots + T_1(x_p)), \quad \text{for } j = 1, 2, 3$$

where $T_k(x_\ell) := T_k(x; \tilde{F}_\ell)$. Store the selected features in the set $\mathcal{O}_j$.

- Third, we summarize the findings by ranking the variables using the `Ovis` index; here we use $k = 1$ and $j = 3$ in the formula (30). The top 10 features are displayed in Fig. 7, which shows the *nature* of contributions of different variables. For example, the variable 'is_weekend' (whether the news article was published on the weekend or not) plays a dual role of being both a location and scale informant. Consequently, our multi-layered robust method provides a refined understanding of the "which and how" aspects of feature selection.

- Finally, one can perform *targeted* feature screening. An investigator can specifically query which variables are mostly responsible for heteroscedasticity, or change in symmetry, etc. Table 1 shows the top five variables in each category: location, scale, and skewness. These insights will ultimately help the applied researchers to better understand the *nature* of the association between response Y and $X = (X_1, \dots, X_p)$.

4.4 The XYZ Problem: Distributional Impact Analysis

In application fields such as healthcare, economics, and social sciences, data often arise in XYZ format, where Z is the binary treatment variable, Y is the response variable, and X is the (possibly large) collection of covariates. One such problem is discussed below.

Example 9. *Rosner's FEV data* (Rosner, 1995). Table 2 (in Supplementary) describes the data. The main interest lies in understanding the impact of smoking on the *distribution* of FEV as a function of age. This is important for individualized custom-tailored decision-making, where a treatment might be *locally*-effective for a sub-population (characterized by certain values of x) without being globally effective for the whole heterogeneous population.

8. see Supplementary A.5.4 for a related discussion on 'robust lasso.'

Fig. 8: Rosner’s FEV data. Our analysis reveals the heterogeneous nature of the impact of smoking on the distribution of FEV, across different age groups. The findings of DIF barplot excellently corroborated by the kernel density estimates of the right panel.

Our model for conditional distribution of Y given $X = x$ and $Z = z$ is

$$f_{Y|x,z}(y) = f_Y(y) \left\{ 1 + \sum_j \text{LP}_{j|x,z} T_j(y; F_Y) \right\}. \quad (31)$$

Distributional impact analysis aims to quantify (how much) and characterize (in which ways) the differences between $f_{Y|x,1}(y)$ and $f_{Y|x,0}(y)$. To develop a measure of distributional impact, first note that from equation (31), both $f_{Y|x,1}$ and $f_{Y|x,0}$ have the same pivot $f_Y(y)$. Thus the deviation between them directly depends on the distance between the contrast densities $d_{x,0}$ and $d_{x,1}$.

Theorem 4.3. The L^2 distance between the $d_{x,1}(u)$ and $d_{x,0}(u)$ can be expressed as ℓ^2 distance between the corresponding LP-Fourier coefficients.

$$\int_0^1 \left(d_{x,0}(u) - d_{x,1}(u) \right)^2 du = \sum_j \left| \text{LP}_{j|x,0} - \text{LP}_{j|x,1} \right|^2. \quad (32)$$

The proof is in Supplementary A.3. This result motivated us to define the following measure.

Definition 7. Define the distributional impact function for a given $X = x$:

$$\text{DIF}(Y, Z | X = x) = \sum_j \left| \text{LP}_{j|x,0} - \text{LP}_{j|x,1} \right|^2. \quad (33)$$

The components of the DIF capture the amount of change in the location, scale, etc. For example, the location difference can be estimated from the first-order LP-difference statistic: $\text{LP}_{1|x,0} - \text{LP}_{1|x,1}$.

We are now ready to apply this procedure on the FEV data. To estimate the LP-coefficients in (31) we apply our ISL framework (algorithm of Section 2.4) with gradient-boosting machine as the base learner. We chose gbm, as it efficiently computes the interactions⁹ between the treatment and covariates. The left panel of Fig. 8 shows the DIF statistic for ages 12 to 15. The stunning fact about this graph is that it simultaneously answers *which* x -groups are

most impacted by the treatment and *how* they are impacted. For example, at age 14 variability is the dominant mode of difference, whereas at age 15, location is the main source of difference. Overall, the DIF barplot does a modest job of capturing the heterogeneous impact of teenage smoking on lung function across different age groups; see Remark A.5.5 in the Supplementary.

4.5 Quantile Regression

Quantile regression has become a pervasive tool in a host of real-world application areas. Despite the great progress made in the last 40 years (Koenker, 2017), some practical challenges still remain: how to develop a *nonparametric* quantile regression method that can scale to high-dimensional problems? How to efficiently incorporate non-linearity? How to ensure that the estimated conditional quantile curves will not cross each other? Our statistical learning theory can offer some realistic solutions to these frontier problems of quantile regression. But before getting into that, let us start with a slightly broader context. Our integrated statistical learning architecture (as depicted in Fig. 1) has two notable consequences for the practice of nonparametric data modeling:

- Our technology opens up brand-new ways of building “ML-powered” statistical models that are simultaneously flexible and scalable for large(n, p) problems, where classical nonparametric methods are not a viable option.
- Our ISL learning architecture is designed using a high-level universal language that is agnostic to the specific ML method, to ensure the portability and durability of the technology.

To perform ISL-based quantile regression, simply extract the appropriate percentiles from the conditional density estimate $f_{Y|X=x}(y)$. The beauty of our method is that we don’t have to develop theories specialized to a certain kind of machine learning method like random forest (Meinshausen, 2006) or neural network (Cannon, 2011)—it’s all done using a single generic procedure; see Supplementary A.5.3 and Fig. 16 for more details.

Example 10. *Dutch Boys data* (Fredriks et al., 2000). This dataset is a part of the Fourth Dutch Growth Study, which comprised of observations on age and BMI of $N = 7294$ Dutch boys. The goal is to estimate age-specific reference growth curves based on conditional quantile function $\hat{Q}_{Y|X=x}(u)$. Supp. Fig 20 displays our result where we have used `lasso` (as implemented in the `glmnet` R-package) as the base learner inside ISL engine. Following the current World Health Organisation (WHO) recommendation, we have used $u = (.03, .15, .50, .85, .97)$ for constructing the cross-sectional reference growth charts; see Remark A.5.6 in the Supplementary.

4.6 Prediction Interval Estimator

One way to summarize the conditional density is through prediction interval (PI), which provides a concise view of the most likely values of the response variable¹⁰. For a specified

9. Treatment $\times$ covariates interactions allow the treatment effect to vary among individuals with covariates.

10. Knowing the range of values of Y as opposed to a single-point estimate, helps decision-makers choose proper action by carefully evaluating the degree of uncertainty (length of the PIs).

level α , the goal is construct an interval that covers no less than $(1 - \alpha)$ of the probability mass of $f_{Y|X=x}(y)$. How to construct such PIs? We discuss three different procedures. The real challenge is to make it as narrow as possible, while maintaining the desired coverage. • *Quantile-based PI*. As noted by Meinshausen (2006), one can use quantile regression to build PIs. A $100(1 - \alpha)\%$ prediction interval for Y given $X = x$ can be estimated as

$$\text{qPI}_{1-\alpha}(x) = [Q_{Y|X=x}(\alpha/2), Q_{Y|X=x}(1 - \alpha/2)]. \quad (34)$$

Naturally, the width (precision) of the quantile-PIs vary over the covariate space.

- *Standard-error-based PI*. If we are brave enough to assume some parametric form of the conditional density, then we can greatly simplify the expression for the PIs. For example, under Gaussianity assumption, one can express (34) in the following compact form, just involving conditional mean $\mu(x)$ and standard-deviation:

$$\text{gPI}_{1-\alpha}(x) = [\mu(x) \pm z^{(\alpha/2)} \sigma_{|x}] \quad (35)$$

where $\Phi^{-1}(\alpha) = z^{(\alpha)}$, $\text{Var}(Y|X = x) = \sigma_{|x}^2$, and ‘g’ denotes the Gaussian-PI.

- *Highest-density PI*. This is the shortest possible interval with a given coverage probability (Box and Tiao, 1973), defined as

$$\text{hdPI}_{1-\alpha}(x) = \{y : f_{Y|x=x}(y) > \tau_{1-\alpha}\} \quad (36)$$

where $\tau_{1-\alpha}$ is the largest number satisfying $\int_{y \in \text{hdPI}_{1-\alpha}(x)} f_{Y|X=x}(y) dy = 1 - \alpha$. Another nice property of hdPIs is that any point within the interval has a higher density than any other point outside it, which justifies their name.

Example 11 *Butterfly data, continued*. All three prediction intervals are shown in Fig. 9(a). Three main conclusions: (i) The shortest among all is the highest-density PI, which take the form of disjoint subintervals—one for each local mode. (ii) The Gaussian assumption of gPI seems too restrictive for real-world data analysis. (iii) Both gPI and qPI include a low-density valley area around $x = 0$ at the cost of other, more-likely values around $x = \pm 2$.

Example 12. *The Auto-MPG data*. The task is to predict the miles-per-gallon (MPG) gas consumption of an automobile based on $p = 7$ features including horsepower, weight, and acceleration. The dataset was used in the 1983 American Statistical Association Exposition and is available in the UCI Machine Learning Repository. Here, we are mainly interested in predicting the MPG of a specific instance—row number 390, which corresponds to the 1982 Dodge Rampage car. The ISL-predicted uncertainty distribution is displayed in Fig. 9(b). The bimodality of the estimated conditional density reflects the fact that 1982 Dodge Rampage is a *hybrid* two-door car—a mix between a passenger car and a truck. The next notable thing is the length of the PIs: The standard error-based gPI is almost three times wider than the hdPI, and thus provides an overly pessimistic assessment of an accurate prediction.

Fig. 9: (color) Left: Butterfly data. Three different prediction intervals for $Y|X = 2$ along with the conditional mode and mean estimates. All are 68% PIs, equivalent to $\pm 1\sigma$ uncertainty. Right: the curious case of 1982 Dodge Rampage. Two notable things: the shape of the conditional density estimate and the contrasting lengths of different PIs. hdPIs tend to produce narrow, more precise prediction intervals while ensuring the desired coverage. qPI seems to be little misaligned, due to the long right-tail.

5 70 YEARS OF STATISTICAL MACHINE LEARNING

Breiman’s 2001 paper was highly influential and hailed as a heroic effort to challenge the status quo in a “David and Goliath” style, where two diametrically opposite modeling cultures—machine learning and traditional parametric statistics—battled for the throne. However, the origin of machine learning has strong historical roots in (nonparametric) Statistics. In an unpublished US Air Force School of Aviation Medicine report in 1951, two UC Berkeley Statisticians (students of Jerzy Neyman) Evelyn Fix and J.L. Hodges, Jr., introduced a non-parametric method for pattern classification, which is now called the k-nearest neighbor (knn) method. This was a groundbreaking paper that marked the beginning of statistical machine learning. Fix and Hodges’s work was a precursor to the celebrated kernel-based methods and decision tree-based algorithms, which revolutionized the world of machine learning. In this paper, we have initiated the “re-unification” process with the hope to bring machine learning back to its statistical roots. Two broad messages of this research:

1. *United statistical learning culture*: The algorithmic models should not be feared because of their complexity, parametric models should not be shunned because of their simplicity, we can *interlink* them to create a more powerful and coherent learning technology. This research shows how one can build such an *integrated learning system* where different cultures of data modeling (embracing: algorithmic, parametric, nonparametric, information-theoretic, robust, and exploratory) can be “glued” together by some underlying deeper principles.

2. *There’s plenty of room in the middle*: Tremendous opportunities lie at the boundaries between the ‘two cultures.’ This paper describes a new theory for constructing a *universal interface* between statistics and machine learning, as a means to close the gap. What we have proposed is by no means the only possible or the “best” theory. As we go forward, it will be very important to come up with new innovative ways to develop “integrated statistical learning” theory that can reconcile and build new links between various cultures of data analysis in order to maintain a

diverse, yet united front of statistical science. All of this is still in its nascent stages, but definitely poised for rapid progress in the coming decade.

ACKNOWLEDGMENTS

This paper celebrates the 200th anniversary of Gauss-Markov least-square regression and 70th anniversary of statistical machine learning¹¹, by putting forward a modern unified perspective.

REFERENCES

- Alvarez-Melis, D. and T. S. Jaakkola (2018). On the robustness of interpretability methods. In *ICML Workshop in Human Interpretability in Machine Learning; On the robustness*.
- Bachrach, L. K., T. Hastie, M.-C. Wang, B. Narasimhan, and R. Marcus (1999). Bone mineral acquisition in healthy asian, hispanic, black, and caucasian youth: a longitudinal study. *The journal of clinical endocrinology & metabolism* 84(12), 4702–4712.
- Box, G. E. (1953). Non-normality and tests on variances. *Biometrika* 40(3/4), 318–335.
- Box, G. E. and G. C. Tiao (1973). *Bayesian inference in statistical analysis*. John Wiley & Sons.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* 16, 199–231.
- Cannon, A. J. (2011). Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & geosciences* 37(9), 1277–1284.
- Cox, D. R. (2001). Comment on Leo Breiman’s “Statistical Modeling: The Two Cultures”. *Statistical Science*. 16(3), 216–218.
- Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association* 115(530), 636–655.
- Fernandes, K., P. Vinagre, and P. Cortez (2015). A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*, pp. 535–546. Springer.
- Fix, E. and J. Hodges (1951). Discriminating analysis, non-parametric discrimination, consistency problems. *Technical Report, US Air Force, School of Aviation Medicine*.
- Fredriks, A. M., S. van Buuren, J. M. Wit, and S. Verloove-Vanhorick (2000). Body index measurements in 1996–7 compared with 1980. *Archives of disease in childhood* 82(2), 107–112.
- Friedman, J. (2020). Predicting regression probability distributions with imperfect data through optimal transformations. *Stanford University Technical Report., arXiv:2001.10102*.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Prediction, Inference and Data Mining* (Second ed.). Springer Verlag, New York.
- Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics* 9, 155–176.
- Matloff, N. (2017). *Statistical regression and classification: From linear models to machine learning*. Chapman and Hall/CRC.
- Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research* 7, 983–999.
- Mukhopadhyay, S. (2017). Large-scale mode identification and data-driven sciences. *Electronic Journal of Statistics* 11(1), 215–240.
- Mukhopadhyay, S. and E. Parzen (2020). Nonparametric universal copula modeling. *Applied Stochastic Models in Business and Industry, special issue on “Data Science”* 36(1), 77–94.
- Mukhopadhyay, S. and K. Wang (2020). A nonparametric approach to high-dimensional k-sample comparison problem. *Biometrika* 107(3), 555–572.
- Parzen, E. (2001). Comment on Leo Breiman’s “Statistical Modeling: The Two Cultures”. *Statistical Science*. 16(3), 224–226.
- Rosner, B. (1995). *Fundamentals of biostatistics*. Duxbury Press: New York.
- Thas, O., J. D. Neve, L. Clement, and J.-P. Ottoy (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (with discussion)* 74(4), 623–671.
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin et al., eds.) 2, 448–485.
- Voudouris, V., R. Gilchrist, R. Rigby, J. Sedgwick, and D. Stasinopoulos (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics* 39(6), 1279–1293.

Subhadeep Mukhopadhyay Dr. Subhadeep Mukhopadhyay received his Ph.D. degree in Statistics from Texas A&M University, USA in 2013. He previously held positions at Temple University, as an Assistant Professor, and Stanford University, as a visiting Assistant Professor. He is currently an AI advisor and Research Consultant at H2O.ai. His current research interests include Nonparametric data science, Admissible machine learning, Empirical Bayes, and Network science.

Kaijun Wang Dr. Kaijun Wang received his M.S. in Mathematical Finance at University of Southern California and Ph.D. in Statistics at Fox School of Business, Temple University. He is currently a post-doctoral fellow at Fred Hutchinson Cancer Research center in Vaccine and Infectious Disease Division. His research interests includes applied nonparametric statistics, biomedical data science, and large-scale inference.

11. These two historical moments symbolize the birth of ‘two cultures’ of data modeling—linear parametric and flexible algorithmic modeling cultures.

Supplementary Material

A.1. SOFTWARE

We have developed an R Software `LPMachineLearning`¹² to perform all the tasks outlined in the paper. We hope this software will encourage applied data scientists to apply our method for their real prediction problems.

A.2. LP-ORTHONORMAL SYSTEM: CONSTRUCTION AND PROPERTIES

LP-basis functions are specially-designed orthonormal system of rank-polynomials. They are inherently ‘nonparametric,’ constructed in a fully data-driven way, not pre-defined like classical polynomials. For a mixed (either discrete or continuous) random variable Z with distribution function F_Z , we describe the construction of LP-basis functions.

Construction Algorithm. A fully-automated construction of LP-orthonormal rank-polynomials of Z consists of the following steps:

Step 1. Mid-Distribution Transform: The mid-distribution function of Z is defined as

$$F_Z^{\text{mid}}(z) = F_Z(Z) - \frac{1}{2}p_Z(z), \quad (\text{E5.1})$$

where $p_Z(z)$ is the probability mass function.

Step 2. Standardizing $F_Z^{\text{mid}}(Z)$: The random variable $F_Z^{\text{mid}}(Z)$ has mean $\mathbb{E}[F_Z^{\text{mid}}(Z)] = .5$ and variance $\text{Var}[F_Z^{\text{mid}}(Z)] = \frac{1}{12}(1 - \sum_z p_Z^3(z))$. Define the 1st-order LP-basis function as

$$T_1(z; F_Z) = \frac{\sqrt{12}\{F_Z^{\text{mid}}(z) - 1/2\}}{\sqrt{1 - \sum_z p_Z^3(z)}}, \quad (\text{E5.2})$$

Step 3. Gram-Schmidt Orthonormalization: Construct high-order polynomials by applying Gram-Schmidt orthonormalization (see Appendix A of Mukhopadhyay and Parzen (2020) for a quick refresher on Gram-Schmidt process) on the set of functions of the power of $T_1(Z; F_Z)$.

Some properties. Here we list some important properties of the They satisfy the following properties of the LP-orthonormal systems:

1) LP-bases are orthonormal with respect to the measure (weighting function) F_Z :

$$\int T_j(z; F_Z) dF_Z = 0,$$

and

$$\int T_j(z; F_Z) T_k(z; F_Z) dF_Z = \delta_{jk}.$$

For related discussions, see Remark 4 of Mukhopadhyay and Parzen (2020, p. 82).

2) For empirical data analysis, construct empirical LP basis (in short eLP basis) $\{T_j(z; \tilde{F}_Z)\}_{j=1,2,\dots,m}$, where m is strictly less than the number of unique values in the sample $\{z_1, \dots, z_n\}$. eLP bases are orthonormal system of functions with respect to the discrete empirical measure $\tilde{F}_Z$.

12. A preliminary version is available from the authors. A more upgraded one is currently under development.

Fig. 10: First four LP-basis functions for the covariates (a) weight and (b) acceleration of Auto-MPG data. They are plotted with respect to $\tilde{F}_X(x)$ over unit interval $[0, 1]$, following the equation (E5.11).

3) For Z continuous, it is not difficult to show that the LP-bases reduces to the following universal shape:

$$T_j(z; F_Z) := \text{Leg}_j \circ F_Z(z), \quad (\text{E5.3})$$

where $\text{Leg}_j(u)$ denotes j -th shifted orthonormal Legendre polynomials on $[0, 1]$ and ‘ $\circ$ ’ denotes composition of functions.¹³ Few top LP-polynomials for Z continuous are given below:

$$T_1(z; F_Z) = \text{Leg}_1(F_Z(Z)) = \sqrt{12}(F_Z(z) - 1/2) \quad (\text{E5.4})$$

$$\begin{aligned} T_2(z; F_Z) &= \text{Leg}_2(F_Z(Z)) \\ &= \sqrt{5}(6F_Z^2(Z) - 6F_Z(Z) + 1) \end{aligned} \quad (\text{E5.5})$$

$$\begin{aligned} T_3(z; F_Z) &= \text{Leg}_3(F_Z(Z)) \\ &= \sqrt{7}(20F_Z^3(Z) - 30F_Z^2(Z) + 12F_Z(Z) - 1). \end{aligned} \quad (\text{E5.6})$$

and so on. For more information, see Mukhopadhyay and Fletcher (2018), Mukhopadhyay (2017). This will be useful for constructing LP-polynomials for any general pivot density f_0 of the response variable.

4) Derivation of empirical $T_1(z; \tilde{F}_Z)$ from a sample of all unique $z_1, \dots, z_N$ (i.e, realization from a continuous Z). In this case, the empirical probability mass function (pmf) is $\tilde{p}_Z(z_i) = 1/N$, and the empirical cdf is $\tilde{F}_Z(z_i) = R_i/N$, where $R_i = \text{rank}(z_i)$. This immediately implies (following eq. E5.2)

$$T_1(z; \tilde{F}_Z) = \sqrt{\frac{12}{N^2 - 1}} \left(R_i - \frac{N+1}{2} \right). \quad (\text{E5.7})$$

Since,

$$\tilde{F}_Z^{\text{mid}}(z_i) = \tilde{F}_Z(z_i) - \frac{1}{2}\tilde{p}_Z(z_i) = \frac{1}{N}(R_i - 1/2).$$

and the correction factor

$$1 - \sum_i \tilde{p}_Z^3(z_i) = 1 - \frac{1}{N^2}.$$

Eq. (E5.7) makes it clear why we call LP-bases are polynomials of ranks.

13. In our regression context the response variable Y is continuous, thus admits this analytic LP-basis form. However, this one-to-one correspondence between LP-polynomials and Legendre polynomials of rank-transform (also know as probability integral transform) $F_Z(Z)$ is not true for discrete Z , i.e., for the covariates X .

5) The shapes of LP-polynomials are data-adaptive. The left panel of Fig. 10 displays the top four ϵ LP-basis functions for the covariate *weight*, taken from the Auto-MPG data (Example 12 in the main paper). The *weight* variable has very little ties, which is the reason why the shapes of LP-bases completely match with eq. (E5.4-E5.6). Contrast this with the right panel of Fig. 10, which shows top four ϵ LP bases for the variable *acceleration*. Due to the presence of a large number of ties, it takes a unique piecewise-constant shape.

A.3. REMAINING PROOFS

Theorem 2.2: Since the conditional entropy $H(Y|X) = \int H(Y|X = x) dF_X(x)$, we have

$$\begin{aligned} H(Y) - H(Y|X) &= - \int f_Y(y) \log f_Y(y) dy \\ &\quad + \iint f_{Y|x}(y) \log f_{Y|x}(y) dy dF_X(x) \\ &= \int \left\{ \int f_{Y|x}(y) \log \frac{f_{Y|x}(y)}{f_Y(y)} dy \right\} dF_X(x) \\ &= \mathbb{E}_X \text{KL}(d_X; \mathbb{I}_{[0,1]}). \end{aligned}$$

The last equality follows from Theorem 2.1.

Theorem 4.1: The Kruskal–Wallis statistic (21) is given by (up to a negligible factor $(n-1)/n$)

$$\text{KW} = \sum_{\ell=1}^k n_{\ell} \left\{ \sqrt{\frac{12}{N^2-1}} \left(n_{\ell}^{-1} \sum_{i \in G_{\ell}} R_i - \frac{N+1}{2} \right) \right\}^2 \quad (\text{E5.8})$$

Now note that

$$\begin{aligned} \widetilde{\text{LP}}_{1|\ell} &= \mathbb{E}[T_1(Y; \widetilde{F}_Y) | X = \ell] \\ &= \frac{1}{n_{\ell}} \sum_{i \in G_{\ell}} \sqrt{\frac{12}{N^2-1}} \left(R_i - \frac{N+1}{2} \right). \end{aligned} \quad (\text{E5.9})$$

The first equality follows from (11) and the second one from (E5.7). Combining (E5.8) and (E5.9) we get the first desired result

$$\text{KW} = \sum_{\ell=1}^k n_{\ell} |\widetilde{\text{LP}}_{1|\ell}|^2. \quad (\text{E5.10})$$

Next, to show that (E5.10) is actually equals to NR_1^2 , note that

$$R_1^2 = \frac{\text{Var}(E[T_1(Y; F_Y) | X])}{\text{Var}(T_1(Y; F_Y))},$$

where the denominator is 1 by the construction of LP-basis functions, and the numerator is $\sum_{\ell=1}^k \pi_{\ell} |\widetilde{\text{LP}}_{1|\ell}|^2$, π_{ℓ} is $\Pr(X = \ell) = n_{\ell}/N$. This is because

$$\mathbb{E}[\text{LP}_{1|X}] = \mathbb{E}[\mathbb{E}[T_1(Y; F_Y) | X]] = \mathbb{E}[T_1(Y; F_Y)] = 0.$$

Theorem 4.3: Define the LP unit-basis functions as

$$S_j(u; F_Y) = T_j(Q_Y(u); F_Y) \text{ for } 0 < u < 1, \quad (\text{E5.11})$$

which satisfy the following orthonormality relations (immediate from eq. 10)

$$\int_0^1 S_j(u; F_Y) S_k(u; F_Y) du = \delta_{jk}. \quad (\text{E5.12})$$

Fig. 11: The butterfly prediction problem: A superposition of $N = 350$ samples, generated from $y = x + \epsilon$ and $y = -x + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$. The red line is the estimated regression function, whose 90% bootstrap confidence band is shown by the blue dotted line.

Substitute $F_Y(y) = u$ in (9) to express the contrast density function in the quantile domain:

$$d_{x,z}(u) = 1 + \sum_j \text{LP}_{j|x,z} S_j(u; F_Y), \quad 0 < u < 1.$$

It is now easy to see that the L^2 distance between $d_{x,0}$ and $d_{x,1}$ can be written as

$$\|d_{x,0} - d_{x,1}\|_2^2 = \sum_j |\text{LP}_{j|x,0} - \text{LP}_{j|x,1}|^2$$

due to (E5.12). Hence proved.

A.4. THE BUTTERFLY PREDICTION PROBLEM

Here we take a ‘crude look at the whole’ procedure using a simple example, to clearly convey what practical results can be hoped from our statistical learning technology.

We are given (x_i, y_i) , $i = 1, \dots, N$ as shown in the Fig 11, which we refer to the *butterfly data*. The goal is to build a prediction model for the outcome variable Y given $X = x$. In particular, our primary focus is the inference of $Y|x$, where $Y|x$ denotes the conditional random variable $Y|X = x$ with distribution function $F_{Y|X=x}(y)$ and density $f_{Y|X=x}(y)$.

Level 0. ML curve-fitting. It proceeds in two-steps: (i) Estimation: To fit a regression curve, we have a vast number of choices available. On the one hand, we have the simplest parametric model—the linear regression—and on the other, a suit of sophisticated machine learning (ML) methods like k-nearest neighbors (knn), random forest, support-vector machine (svm), gradient boosting machine (gbm), neural net, etc. Irrespective of whether we choose a parametric or algorithmic method, the estimated-regression function takes the form, more or less, of a ‘flat’ zero line, as shown in Fig. 11. (ii) Inference: The obvious next question is, how accurate is the estimate? The blue dotted line in Fig. 11 displays the 90% bootstrap confidence band, which is narrow at the center (around 0) and gets wider as we move

outward—correctly reflecting the uncertainty of the mean function, which indicates that the estimated regression line is impressively accurate!

Even though we have an “accurate” estimate of the underlying regression prediction, it does not tell us anything useful. In fact, it paints a grossly misleading picture of the reality (in the sense that x is not useful for predicting y , which is clearly untrue). To a fuller picture, it necessary to go beyond conventional practice of curve fitting.

Level 1. Heterogeneity diagnostic. A natural question is whether the initial ML_0 captured the essential relationship between y and the predictor variables. This can be checked by searching for (unmodeled) patterns in the residuals. Recall that for the `butterfly` data $\hat{\mu}(x) \approx 0$ for all x ; thus residuals are actually the original response values $y_1, \dots, y_n$.

The critical task boils down to checking whether the residuals are homogeneous, i.e., $f_{Y|X=x}(y) = f_Y(y)$ for all x . This would ensure that ML_0 was able to capture all the important predictive information. However, in most practical real-life problems (including `butterfly` data) the residuals will be heterogeneous. And the main challenge is to identify the components of $f_{Y|X=x}(y)$ that are significantly changing with x . We like to address this by constructing a simple diagnostic plot that can help users to visually examine and detect the prominent *sources* of heterogeneity in the residuals. Fig. 12(a) is one such graphical exploratory plot drawn for the `butterfly` data, which reveals scale (second-order effect) and tail (fourth-order kurtosis effect) of $f_{Y|X=x}(y)$ are changing significantly with x . Accordingly, this tool helps us to *discover* the hidden patterns that, while invisible to ML_0 , still persist and should be modeled for sharpening the prediction quality; also see Appendix A.8.

Level 2. Pivot density. We now turn our attention to the problem of predicting conditional distribution at a particular x , say at $x_0 = 2$ for the `butterfly` example. There are some obvious candidates: $\hat{f}_Y(y)$ the empirical (marginal) distribution of y ; or its nonparametric smooth (say, kernel-smoothed) version $\hat{f}_Y(y)$; or some kind of parametric model motivated from domain-knowledge or statistical convenience, e.g., $\mathcal{N}(\hat{\mu}(x_0), \hat{\sigma}_y^2)$, where $\hat{\sigma}_y$ denotes the standard error of y . We call these initial starting “guesses” the pivot density; they are nothing but a crude first-approximation of the true *unknown* conditional density estimate. For the `butterfly` data, we choose $\mathcal{N}(0, \hat{\sigma}_y^2 = 2.55^2)$ to be the pivot, as shown in Fig. 12(b).

A natural question to ask is whether the presumed pivot model is approximately correct. In most practical cases, the answer will be no, they are not adequate. Thus we have to take some corrective measures to rectify the deficiency of the starting pivot. But before doing so, it is crucial to devise some technique that can reveal the deficiency of the current model. The only problem is that we don’t have enough y -samples at $X = x_0$ to perform traditional goodness-of-fit tests. We have to come up with something *new* that is computable and interpretable.

Level 3. Uncertainty quantification for the pivot. The basic question is whether the starting pivot density $f_0(y)$ is a good model for $Y_{|x_0}$. We answer this question by estimating the

Fig. 12: Analysis of Butterfly data: (a) Heterogeneity component analysis for residuals. It decomposes total deviation from homogeneity into explainable components. Here, the starting ML_0 (which we have selected to be `knn` method with $k = 15$) fails to capture the varying scale and tail. (b) Gaussian pivot $f_0(y)$ shown in blue; black dotted line denotes the true $\frac{1}{2}\mathcal{N}(2, 1) + \frac{1}{2}\mathcal{N}(-2, 1)$. (c) The estimated bimodal $\hat{d}_{x_0}^0(u)$. (d) The shape of d -modulated density function at $X = 2$ eq. (5), which has an inbuilt ‘self-correcting’ mechanism through $\hat{d}_{x_0}^0$. (e) Estimated conditional quantile curves for $u = .05, .30, .50, .70$ and $.95$. (f) QQ-plot for checking the uniformity of generalized quantile-residuals $U_i = \hat{F}_{Y|X=x_i}(y_i)$ on a 15% hold-out set.

following *contrast* density function for $u = F_0(y)$

$$d(u; F_0, F_{Y|X=x}) := d_x^0(u) = \frac{f_{Y|X=x}(F_0^{-1}(u))}{f_0(F_0^{-1}(u))}, \quad 0 < u < 1 \quad (\text{E5.13})$$

where $F_0^{-1}(u)$ is the quantile function for the pivot $f_0(y)$. Note that $d_x^0(u)$ captures the deviation of $f_0(y)$ from the true *unknown* conditional density $f_{Y|X=x_0}(y)$, and by virtue of doing so it helps us to quantify and characterize the uncertainty of the pivot density. By inspecting the shape of $d_x^0(u)$, $0 < u < 1$ (how it deviates from uniformity), practitioners can quickly infer whether $f_0(y)$ needs any repair or not.

Fig. 12(c) displays the estimated contrast function for the `butterfly` data. Looking at the shape, it becomes clear that the initial Gaussian pivot completely missed the bimodality of $f_{Y|X=2}(y)$. The contrast function $d_x^0(u)$ plays a central role in our theory. The fascinating aspect of the algorithm is that it utilizes the power of the initial machine learning

algorithm ML_0 to get the covariate-adaptive $\hat{d}_x^0$.

Level 4. Density modulation. When the contrast function indicates the inadequacy of the starting pivot $f_0(y)$ as a model for the conditional distribution $f_{Y|X=x}(y)$, what to do next? How to modify (or ‘boost’) $f_0(y)$ to make it data-consistent? This can be solved in a remarkably simple way via *density modulation*: $\hat{f}_{Y|X=x} = f_0 \times \hat{d}_x^0$, where $\hat{d}_x^0$ modulates (modifies) the shape of starting $f_0(y)$ to produce the conditional density. The detailed theory of why and how this works are discussed in Section 2 of the main article. Fig. 12(d) displays the estimated predictive density at $X = 2$, which we can write down analytically as

$$\begin{aligned} \hat{f}_{Y|X=2}(y) &= \frac{1}{\hat{\sigma}} \phi\left(\frac{y}{\hat{\sigma}}\right) \times \{1 - 0.20 T_2(y; F_0) - 0.47 T_4(y; F_0)\}, \\ &\text{where } \hat{\sigma} = 2.55 \quad (\text{E5.14}) \end{aligned}$$

which is product of two components: the Gaussian pivot and the contrast density $d_{x_0}^0(F_0(y))$. The basis functions $\{T_j(y; F_0)\}$ in equation (5) are specially-designed rank-based orthonormal polynomials of the pivot. They injects robustness and non-linearity into the data modeling process—but more on this later. We will also show that this new class of d -modulated conditional density models can generate an extraordinarily rich spectrum of shapes.

One notable aspect of (5) is that we have ‘identified’ a parametric model (instead of blindly *assuming*, which Breiman sharply criticized) by algorithmic (fully nonparametric) means. In Section 2.3, we elucidate the process by which we boost and convert the “0th order” machine learning method (ML_0) into a finitely parameterizable uncertainty prediction machine. This allows our learning philosophy to *integrate* the fundamental character of both algorithmic data-driven modeling and parametric statistical modeling. Nevertheless, the predicted distribution (5) compactly represents all possible outcome values along with their respective probabilities, which is essential for decision-making under uncertainty.

Level 5. Predictive inference: What are some of the most likely values of the response variable y , given that we have observed $X = x_0$? For a fixed α , ideally, we would like to find the smallest volume region that covers $100(1 - \alpha)\%$ area of the conditional distribution, aka the highest density prediction region. The orange highlighted part in Fig. 12(d) displays 68% highest-density prediction region for the butterfly data, which consists of two disjoint intervals $[-3.33, -1.40] \cup [1.20, 3.28]$. The volume of this set (the sum of the lengths of the intervals) can be used to quantify the *uncertainty* of the prediction. Additionally, we can extract all the conditional quantiles from the estimated $\hat{F}_{Y|X=x}(y)$; see panel (e). Our quantile regression curves are, by design, non-crossing and nonlinear. Curious readers may contrast this with the traditional quantile regression estimator, shown in Fig. 13 of Appendix A.5.1.

Level 6. Verification: Data modeling typically consists of four stages: estimation (of the initial pivot), exploration (discovering the deficiency of the pivot), rectification (repairing the pivot), and verification (checking whether the corrected pivot emulates the observed reality). The last step remains to be done, which requires *evidence* that the estimated uncer-

Fig. 13: Butterfly data: Performs Roger Koenker’s (1978) quantile regression routine as implemented in `rq()` function of the R-package `quantreg`. Note the difference with Fig. 16.

tainty prediction model provides a satisfactory description of the data. Scientific data analysis, by definition, needs to be testable (falsifiable), and there’s no two ways about it.

For our butterfly example, we compute $U_i = \hat{F}_{Y|X=x_i}(y_i)$ for $i = 1, \dots, N$ using our trained model on a 15% hold-out validation dataset. It is not difficult to see that the distribution of U_i ’s will be ‘close’ to $\text{Uniform}[0,1]$, if the estimated conditional density can explain the essential patterns/variations in the data. The quantiles of the empirical distribution of U_i ’s are plotted against the respective quantiles of the uniform distribution at the bottom right of Fig. 12. The tight cluster of points around the 45-degree reference line strongly indicates that the model is successful in capturing the underlying data generating process, and can be trusted to make predictions. Additionally, the formal goodness-of-fit test (proposed in Section 3.3), when applied for this example, yields a p-value of 0.82.

A.5. ADDITIONAL REMARKS

Remark A.5.1 *Koenker’s Quantile regression for Butterfly data:* Fig. 13 shows the quantile regression curves for the butterfly data, which implements the algorithm proposed in Koenker and d’Orey (1987).

Remark A.5.2 *Smoothness of Regression via Rank-transform.* Fig. 14 shows house prices (median value of in 1000s) in Boston metropolitan area as function of `crime` (per capita crime rate). The data were collected in 1970 and available in the UCI Machine Learning Repository. The left plot shows the data in the xy -domain, and the right plot shows the same data in the quantile domain where we replaced x by its rank-transform $\tilde{F}_X(x)$. Red curves are the spline smoother (estimated using `smooth.spline()` R-function). Few interesting points to note:

- **Smoothness:** Regression *via* rank-transform (see Theorem 2.5) seems to produce much smoother (parsimonious) regression curve compared to the rough zigzag xy -domain estimate. Is it universal that nature reveals herself in a more parsimonious way in the quantile domain? Apart from intuitive understanding, we still don’t know any concrete mathematical explanation for this surprising phenomenon.

- **Data-sparsity:** The quantile-domain treatment seems to address the ‘data sparsity’ (that exists for $x > 30$) prob-

Fig. 14: Boston housing data: The house prices drop sharply with high crime rates.

lem quite well. Surprisingly, after quantile-transformation, a data-sparse regression problem turns into a data-dense one.

- **Visualization:** The rank-transform scatter makes it far easy to visualize and understand the relationship between x and y .

- **Robust+Smooth:** This shows an additional benefit of X-robustness (Theorem 2.5) as a tool for promoting smoothness.

Remark A.5.3 *ML-assisted Quantile regression: comparison.* The first row of Fig. 16 shows the generalized random forests-based (Athey et al., 2019) and gradient boosted quantile regression estimates. They are specially-designed ML-algorithms to produce the desired conditional quantile curves. The second row shows our ISL versions, which are derived from a general scheme described in Section 2.3 of the main paper.

Remark A.5.4 *Robust Lasso:* As already noted, the online news data contain an outlier—the news article in row #31,038. How does it impact traditional estimation and feature selection methods? As we will see, a single data point can have a devastating impact on the overall modeling. Fig. 17 (left panel) shows the lasso estimated $\hat{\beta}$ under two scenarios: first, on the outlier-removed data, and second, on the full data (with outlier). Notice the estimated coefficient values (as well as their sparsity-levels) are dramatically affected because of just one data point. One way to address this fragility is to apply lasso on the first-order LP-transformed $T_X = [T_1(x; \tilde{F}_{X_1}), \dots, T_1(x; \tilde{F}_{X_p})]$, instead of raw feature matrix X . The result is displayed on the right panel, which shows extraordinary consistency with (almost) no impact on the estimated coefficients. This mid-rank transform based *robust lasso* method could be a reliable alternative to deal with noisy messy data.

Remark A.5.5 *Connection of XYZ Problem with causal inference:* For randomized experiments and clinical trials our DIF statistic can be interpreted as a ‘distributional’ treatment effect¹⁴, one whose impact vary across sub-populations. For observational studies, one needs to assume

14. The topic of ‘distributional’ treatment effect (as opposed to ‘mean’ effects) carries a major significance in modern econometrics (Bitler et al., 2006, Imbens and Wooldridge, 2009, Banerjee et al., 2015), where it is known that the impact of an intervention on the outcome distribution can be highly heterogeneous—somewhat similar to what we have already seen in Fig. 8. The good news is: $\text{DIF}(Y, Z|X)$ provides a systematic way to characterize the impact of a treatment Z across the entire distribution of Y as a function of X .

the so-called ‘unconfoundedness’ assumption to attach a causal interpretation.

Remark A.5.6 *Personalized reference interval prediction:* The topic of individualized referencing carries special significance in the era of precision medicine. To accurately infer each individual’s reference curves (describing the *normal* range of the outcome y given x) medical researchers are now incorporating more and more variables in the form of genetic and environmental factors. Thus it will be of great practical value to have a method that can tackle a large number of covariates, and do so in a completely nonparametric and nonlinear fashion.

A.6. MODERN APPLIED STATISTICS: THEORY, PRACTICE, AND PEDAGOGY

A theoretical research program that has no relevance to practice and teaching, is not persuasive enough to be taken as fundamental. To us, it is very important to know whether our research was able to link (at least partially) the triad of theory-practice-teaching. In this article, through several examples, we made an attempt to describe ‘the joy of systematic data analysis’ derived from a general theory, which has the following implications for the practice and pedagogy of Statistics:

- Our theory makes it easy to apply (arise from simplification) and see the connections (arise from unification) between different statistical methods.

- Our theory provides a holistic training¹⁵ by introducing a large variety of statistical topics (embracing different cultures: algorithmic, parametric, nonparametric, information-theoretic, robust, exploratory) in an ‘organized’ manner through a common semantics. This is extremely important to produce broad integrator foxes rather than highly specialized hedgehogs¹⁶. Ultimately, what matters is whether our theory provides a faster and easier (lazier?) way to learn the fundamentals of statistics¹⁷. We believe—yes, it does.

A.7. TWO SURPRISING CONNECTIONS: FROM AI AND ECONOMICS

- *Herbert Simon’s theory to simulate human thinking.* In the year of 1956, at the historic conference “Dartmouth Summer Research Project on Artificial Intelligence,” Nobel laureate Herbert Simon, along with Cliff Shaw and Allen Newell, presented a universal problem solver machine called ‘General Problem Solver’ [GPS], which is considered by many to be the first AI program. Their theory is based on the idea of “Difference Machine,” which proceeds as follows:

15. For more discussion on this topic see Mukhopadhyay (2017).

16. Readers may find it worthwhile to read Philip Tetlock’s book on ‘Why Foxes Are Better Forecasters Than Hedgehogs,’ which is based on a 20-year study with 284 experts from diverse fields, including government officials, professors, journalists, and others. As Richard Feynman said “Science is not a specialist business.”

17. As students and applied data science professionals don’t have to go through the struggle of bookkeeping and remembering a long list of disconnected formulae and algorithms. The case-specific methods will flow naturally from the general theory in an automatic way

(i) compare C [initial steady-state] with F [future desired-state] to check whether there exists any difference; (ii) devise some technique (“intelligence”) to rectify the present state to reduce the important differences between C and F —call it the difference-operator; (iii) if this succeeds, then we have created a new object A' , which (one hopes) no longer has the difference when compared with the desired state F . The question remains, how can we transform this hierarchical logic into a computable theory? It’s quite fascinating that the GPS logic is already encoded into our modeling scheme (6), where the contrast function d_x acts according to Herbert Simon’s difference-operator. In our setup, d_x actively works to remove the differences between the present (steady/null) state $f_Y(y)$ and the future (desired) state $f_{Y|X=x}(y)$. The main reference here is Newell et al. (1959). Similar ideas also appeared in the Ph.D. thesis of Patrick Winston (1970), and more recently Marvin Minsky’s (2007).

- *Hansen-Sargent theory of model misspecification.* Lars Hansen and Thomas Sargent, two Nobel laureates, in their book “Robustness,” raised a profound modeling question: Is there any robust way to construct a class of plausible models starting from a misspecified one?

It is an understatement to say that model misspecification is at least as much of a problem in economics as it is in physics and engineering. (Hansen and Sargent, 2008, p. 3)

Our d -modulation based density representation theory (8) creates a class of reasonable models around a possibly misspecified pivot. Here the contrast density d_x acts as a data-driven perturbation function in order to ‘repair’ the starting approximate model f_0 .

A.8. SHARPENING BLACK-BOX GENERATORS

Imagine we are given some black-box generative model that can simulate y -samples for a given $x = (x_1, \dots, x_p)$:

$$\text{MODEL}(x) \mapsto \{\tilde{y}_1^x, \dots, \tilde{y}_s^x\}.$$

Without having any information on the particular MODEL^{18} , our goal is to: (i) check whether the samples are truly generated from the underlying conditional distribution $f_{Y|X=x}(y)$; (ii) if not, explain ‘why’ by displaying the associated $\tilde{d}_x$; and finally, (iii) sharpen the given imperfect samples to make them compatible with the underlying stochastic data generating mechanism.

Butterfly data cont. To illustrate the method, let us consider that we are given all the y_i ’s for which the x_i value is between 0 and 4. The histogram of these samples is shown in the left panel of Fig. 15. The blue curve shows the true $f_{Y|X=2}(y)$. Starting from these *weak* conditional samples, our goal is to refine it so that they comply with the underlying law at $x = 2$.

Step 1. Choose the empirical distribution $\tilde{f}_{y|x}$ of $\{\tilde{y}_1^x, \dots, \tilde{y}_s^x\}$ be the pivot; see Fig. 15(a).

Step 2. Randomly generate one sample $\tilde{y}^*$ from $\tilde{f}_{y|x}$ and U from $\text{Uniform}[0, 1]$.

18. This could be *any* arbitrarily complex computational program, e.g., deep learning method with millions of parameters, an ensemble of 100 machine learning models, or the method proposed in Friedman, J. H. (2020).

Fig. 15: Butterfly data: the histogram at left shows the y_i ’s with $0 \leq x_i \leq 4$. On the right, we display the histogram of d -sharpened samples, which shows remarkable agreement with the underlying true conditional density (the blue line).

Fig. 16: Butterfly data: Quantile Regression comparisons. (a) generalized random forests based quantile regression curves (as implemented in the `grf` R-package), which are too wiggly even after spline smoothing (with default degrees of freedom). (b) Generalized boosted quantile regression curves (as implemented in `gbm` R-package). They suffer from under-smoothing in the central region and over-smoothing at the tails. The bottom row shows our ISL-versions.

Step 3. Let $\tilde{F}_{y|x}$ denotes the cdf of $\tilde{f}_{y|x}$. Then, set $\hat{y}^x = \tilde{y}^*$, if

$$d_x(\tilde{F}_{y|x}(\tilde{y}^*)) > U \max_u d_x(u).$$

Otherwise discard it and go back to the previous step. Here $d_x(u)$ means $d_x(u; \tilde{F}_{y|x}, F_{y|x})$.

Step 4. Repeat steps 2-3 until we have s -sample $\{\hat{y}_1^x, \dots, \hat{y}_s^x\}$, which we call d -sharp samples, since d_x acts as a tool for sharpening the initial weak conditional samples; see Fig. 15(b).

Fig. 17: Online news popularity data. The goal is to assess the impact of a single outlying observation (the news article in row #31,038) on the lasso-estimated coefficients. The left plot shows the result for $\text{lasso}(y \sim X)$ case, and the right one for robustified $\text{lasso}(y \sim T_X)$. For more details see remark A.5.4.

X (Age)	Y (FEV)	Z (Smoking)
9	1.70	0
8	1.72	0
$\vdots$	$\vdots$	$\vdots$
15	3.73	1
18	2.85	0

TABLE 2: We are given $N = 654$ observations on youths aged 3 to 19 from East Boston recorded during 1970's. The outcome variable Y is forced expiratory volume in 1 second—a measure of lung capacity. Z is 1/0 binary treatment (smoker: yes or no) indicator variable.

REFERENCES

- Athey, S., J. Tibshirani, S. Wager, et al. (2019). Generalized random forests. *The Annals of Statistics* 47(2), 1148–1178.
- Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics* 7(1), 22–53.
- Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review* 96(4), 988–1012.
- Friedman, J. H. (2020). Contrast trees and distribution boosting. *Proceedings of the National Academy of Sciences* 117(35), 21175–21184.
- Hansen, L. P. and T. J. Sargent (2008). *Robustness*. Princeton university press.
- Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature* 47(1), 5–86.
- Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*. 96(4), 988–1012.
- Koenker, R. W. and V. d’Orey (1987). Algorithm as 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36(3), 383–393.
- Minsky, M. (2007). *The Emotion Machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster.

Fig. 18: (color) 1990's film data: first row shows a part of the full `film90` data and the starting normal pivot. Notice the “two-branches” of the data cloud. Second row: The estimated contrast densities $\hat{d}_{Y|X=x}(u)$, which describe the conversion factor: how the pivot $f_0(y)$ has to be transformed (perturb) to get to $f_{Y|X=x}(y)$. Third row: The refurbished conditional densities $\hat{f}_{Y|X=x}(y)$ obeying the d -modulation law (8); the red triangles denote the mode of the distributions. Notice the difference in shape between the starting Gaussian pivot (panel b) and the final estimated conditional densities (of the bottom row).

- Mukhopadhyay, S. (2017). Large-scale mode identification and data-driven sciences. *Electronic Journal of Statistics* 11(1), 215–240.
- Mukhopadhyay, S. (2017). Statistics educational challenge in the 21st century. *Biostatistics and Biometrics Journal* 2(2), 1–2. Online version: arXiv:1708.04098.
- Mukhopadhyay, S. and Fletcher, D. (2018). Generalized Empirical Bayes Modeling via Frequentist Goodness-of-Fit *Scientific Reports*. 8, 1–15.
- Mukhopadhyay, S. and E. Parzen (2020). Nonparametric universal copula modeling. *Applied Stochastic Models in Business and Industry, special issue on “Data Science”* 36(1), 77–94.
- Newell, A., J. C. Shaw, and H. A. Simon (1959). Report on a general problem solving program. In *IFIP congress*, Volume 256, pp. 64. Pittsburgh, PA.
- Winston, P. H. (1970). *Learning structural descriptions from examples*. Ph. D. thesis, Massachusetts Institute of Technology.

Fig. 19: Exploratory goodness-of-fit assessment for the butterfly data. The histograms and QQ plots of the generalized quantile-residuals (18) are shown for three ML-models. They are computed based on a 15% hold-out data. Knn (with $k = 15$) seems fits the data well.

Fig. 20: Dutch boys data: The ISL-estimated conditional quantile curves. Our method yields non-crossing quantile curves, as they are extracted from the conditional distributions.