

# Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation

ALEX GOLDSTEIN\*, ADAM KAPELNER<sup>†</sup>, JUSTIN BLEICH<sup>‡</sup>, AND EMIL PITKIN<sup>§</sup>

*The Wharton School of the University of Pennsylvania*

March 21, 2014

## Abstract

This article presents Individual Conditional Expectation (ICE) plots, a tool for visualizing the model estimated by any supervised learning algorithm. Classical partial dependence plots (PDPs) help visualize the average partial relationship between the predicted response and one or more features. In the presence of substantial interaction effects, the partial response relationship can be heterogeneous. Thus, an average curve, such as the PDP, can obfuscate the complexity of the modeled relationship. Accordingly, ICE plots refine the partial dependence plot by graphing the functional relationship between the predicted response and the feature for *individual* observations. Specifically, ICE plots highlight the variation in the fitted values across the range of a covariate, suggesting where and to what extent heterogeneities might exist. In addition to providing a plotting suite for exploratory analysis, we include a visual test for additive structure in the data generating model. Through simulated examples and real data sets, we demonstrate how ICE plots can shed light on estimated models in ways PDPs cannot. Procedures outlined are available in the R package `ICEbox`.

## 1 Introduction

The goal of this article is to present Individual Conditional Expectation (ICE) plots, a toolbox for visualizing models produced by “black box” algorithms. These algorithms use training data  $\{\mathbf{x}_i, y_i\}_{i=1}^N$  (where  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})$  is a vector of predictors and  $y_i$  is the

---

\*Electronic address: alexg@wharton.upenn.edu; Principal Corresponding author

<sup>†</sup>Electronic address: kapelner@wharton.upenn.edu; Corresponding author

<sup>‡</sup>Electronic address: jbleich@wharton.upenn.edu; Corresponding author

<sup>§</sup>Electronic address: pitkin@wharton.upenn.edu; Corresponding author

response) to construct a model  $\hat{f}$  that maps the features  $\mathbf{x}$  to fitted values  $\hat{f}(\mathbf{x})$ . Though these algorithms can produce fitted values that enjoy low generalization error, it is often difficult to understand how the resultant  $\hat{f}$  uses  $\mathbf{x}$  to generate predictions. The ICE toolbox helps visualize this mapping.

ICE plots extend Friedman (2001)’s Partial Dependence Plot (PDP), which highlights the average partial relationship between a set of predictors and the predicted response. ICE plots disaggregate this average by displaying the estimated functional relationship for each observation. Plotting a curve for each observation helps identify interactions in  $\hat{f}$  as well as extrapolations in predictor space.

The paper proceeds as follows. Section 2 gives background on visualization in machine learning and introduces PDPs more formally. Section 3 describes the procedure for generating ICE plots and its associated plots. In Section 4 simulated data examples illustrate that ICE plots can be used to identify features of  $\hat{f}$  that are not visible in PDPs, or where the PDPs may even be misleading. Each example is chosen to illustrate a particular principle. Section 5 provides examples of ICE plots on real data. In Section 6 we shift the focus from the fitted  $\hat{f}$  to a data generating process  $f$  and use ICE plots as part of a visual test for additivity in  $f$ . Section 7 concludes.

## 2 Background

### 2.1 Survey of Black Box Visualization

There is an extensive literature that attests to the superiority of black box machine learning algorithms in minimizing predictive error, both from a theoretical and an applied perspective. Breiman (2001b), summarizing, states “accuracy generally requires more complex prediction methods ...[and] simple and interpretable functions do not make the most accurate predictors.” Problematically, black box models offer little in the way of interpretability, unless the data is of very low dimension. When we are willing to compromise interpretability for improved predictive accuracy, any window into black box’s internals can be beneficial.

Authors have devised a variety of algorithm-specific techniques targeted at improving the interpretability of a particular statistical learning procedure’s output. Rao and Potts (1997) offers a technique for visualizing the decision boundary produced by bagging decision trees. Although applicable to high dimensional settings, their work primarily focuses on the low dimensional case of two covariates. Tzeng (2005) develops visualization of the layers of neural networks to understand dependencies between the inputs and model outputs and yields insight into classification uncertainty. Jakulin et al. (2005) improves the interpretability of support vector machines by using a device called “nomograms” which provide graphical representation of the contribution of variables to the model fit. Pre-specified interaction effects of interest can be displayed in the nomograms as well. Breiman (2001a) uses randomization of out-of-bag observations to compute a variable importance metric for Random Forests (RF). Those variables for which predictive performance degrades the most vis-a-vis the original model are considered the strongest contributors to forecasting accuracy. This method is also applicable to stochastic gradient boosting (Friedman, 2002). Plate et al. (2000) plots neural network predictions in a scatterplot for each variable by sampling points from covari-

ate space. Amongst the existing literature, this work is the most similar to ICE, but was only applied to neural networks and does not have a readily available implementation.

Other visualization proposals are model agnostic and can be applied to a host of supervised learning procedures. For instance, Strumbelj and Kononenko (2011) consider a game-theoretic approach to assess the contributions of different features to predictions that relies on an efficient approximation of the Shapley value. Jiang and Owen (2002) use quasi-regression estimation of black box functions. Here, the function is expanded into an orthonormal basis of coefficients which are approximated via Monte Carlo simulation. These estimated coefficients can then be used to determine which covariates influence the function and whether any interactions exist.

## 2.2 Friedman’s PDP

Another particularly useful model agnostic tool is Friedman (2001)’s PDP, which this paper extends. The PDP plots the change in the average predicted value as specified feature(s) vary over their marginal distribution. Many supervised learning models applied across a number of disciplines have been better understood thanks to PDPs. Green and Kern (2010) use PDPs to understand the relationship between predictors and the conditional average treatment effect for a voter mobilization experiment, with the predictions being made by Bayesian Additive Regression Trees (BART, Chipman et al., 2010). Berk and Bleich (2013) demonstrate the advantage of using RF and the associated PDPs to accurately model predictor-response relationships under asymmetric classification costs that often arise in criminal justice settings. In the ecological literature, Elith et al. (2008), who rely on stochastic gradient boosting, use PDPs to understand how different environmental factors influence the distribution of a particular freshwater eel.

To formally define the PDP, let  $S \subset \{1, \dots, p\}$  and let  $C$  be the complement set of  $S$ . Here  $S$  and  $C$  index subsets of predictors; for example, if  $S = \{1, 2, 3\}$ , then  $\mathbf{x}_S$  refers to a  $3 \times 1$  vector containing the values of the first three coordinates of  $\mathbf{x}$ . Then the partial dependence function of  $f$  on  $\mathbf{x}_S$  is given by

$$f_S = \mathbb{E}_{\mathbf{x}_C} [f(\mathbf{x}_S, \mathbf{x}_C)] = \int f(\mathbf{x}_S, \mathbf{x}_C) dP(\mathbf{x}_C). \quad (1)$$

Each subset of predictors  $S$  has its own partial dependence function  $f_S$ , which gives the average value of  $f$  when  $\mathbf{x}_S$  is fixed and  $\mathbf{x}_C$  varies over its marginal distribution  $dP(\mathbf{x}_C)$ . As neither the true  $f$  nor  $dP(\mathbf{x}_C)$  are known, we estimate Equation 1 by computing

$$\hat{f}_S = \frac{1}{N} \sum_{i=1}^N \hat{f}(\mathbf{x}_S, \mathbf{x}_{C_i}) \quad (2)$$

where  $\{\mathbf{x}_{C_1}, \dots, \mathbf{x}_{C_N}\}$  represent the different values of  $\mathbf{x}_C$  that are observed in the training data. Note that the approximation here is twofold: we estimate the true model with  $\hat{f}$ , the output of a statistical learning algorithm, and we estimate the integral over  $\mathbf{x}_C$  by averaging over the  $N$   $\mathbf{x}_C$  values observed in the training set.

This is a visualization tool in the following sense: if  $\hat{f}_S$  is evaluated at the  $\mathbf{x}_S$  observed in the data, a set of  $N$  ordered pairs will result:  $\{(\mathbf{x}_{S\ell}, \hat{f}_{S\ell})\}_{\ell=1}^N$ , where  $\hat{f}_{S\ell}$  refers to the estimated partial dependence function evaluated at the  $\ell$ th coordinate of  $\mathbf{x}_S$ , denoted  $\mathbf{x}_{S\ell}$ . Then for one or two dimensional  $\mathbf{x}_S$ , Friedman (2001) proposes plotting the  $N$   $\mathbf{x}_{S\ell}$ 's versus their associated  $\hat{f}_{S\ell}$ 's, conventionally joined by lines. The resulting graphic, which is called a partial dependence plot, displays the average value of  $\hat{f}$  as a function of  $\mathbf{x}_S$ . For the remainder of the paper we consider a single predictor of interest at a time ( $|S| = 1$ ) and write  $x_S$  without boldface accordingly.

As an extended example, consider the following data generating process with a simple interaction:

$$Y = 0.2X_1 - 5X_2 + 10X_2\mathbb{1}_{X_3 \geq 0} + \mathcal{E}, \quad (3)$$

$$\mathcal{E} \stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad X_1, X_2, X_3 \stackrel{iid}{\sim} U(-1, 1).$$

We generate 1,000 observations from this model and fit a stochastic gradient boosting model (SGB) via the R package `gbm` (Ridgeway, 2013) where the number of trees is chosen via cross-validation and the interaction depth is set to 3. We now consider the association between predicted  $Y$  values and  $X_2$  ( $S = X_2$ ). In Figure 1a we plot  $X_2$  versus  $Y$  in our sample. Figure 1b displays the fitted model's partial dependence plot for predictor  $X_2$ . The PDP suggests that on average,  $X_2$  is not meaningfully associated with the predicted  $Y$ . In light of Figure 1a, this conclusion is plainly wrong. Clearly  $X_2$  is associated with  $Y$ ; it is simply that the averaging inherent in the PDP shields this discovery from view.

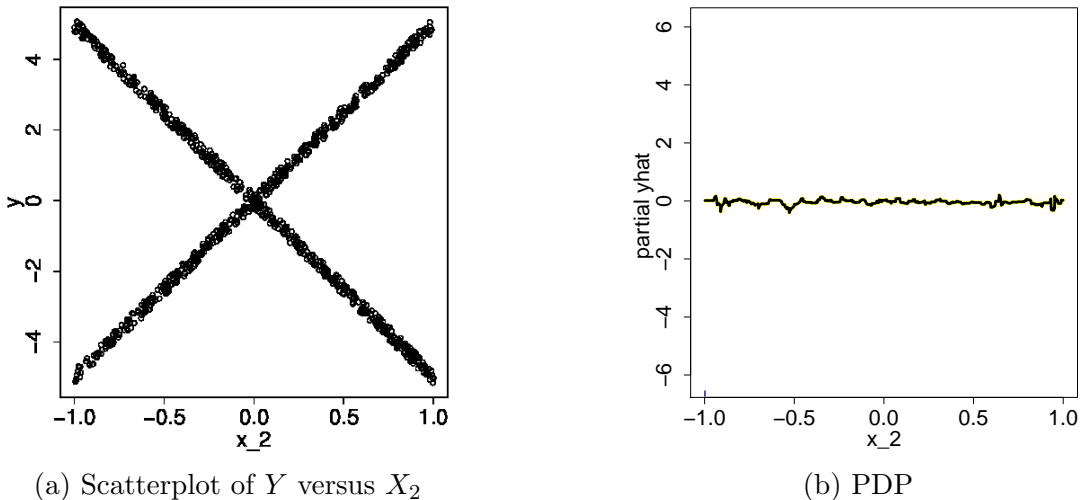


Figure 1: Scatterplot and PDP of  $X_2$  versus  $Y$  for a sample of size 1000 from the process described in Equation 3. In this example  $\hat{f}$  is fit using SGB. The PDP incorrectly suggests that there is no meaningful relationship between  $X_2$  and the predicted  $Y$ .

In fact, the original work introducing PDPs argues that the PDP can be a useful summary for the chosen subset of variables if their dependence on the remaining features is not too strong. When the dependence is strong, however – that is, when interactions are present – the PDP can be misleading. Nor is the PDP particularly effective at revealing extrapolations in  $\mathcal{X}$ -space. ICE plots are intended to address these issues.

### 3 The ICE Toolbox

#### 3.1 The ICE Procedure

Visually, ICE plots disaggregate the output of classical PDPs. Rather than plot the target covariates’ *average* partial effect on the predicted response, we instead plot the  $N$  estimated conditional expectation curves: each reflects the predicted response as a function of covariate  $x_S$ , conditional on an observed  $\mathbf{x}_C$ .

Consider the observations  $\{(x_{Si}, \mathbf{x}_{Ci})\}_{i=1}^N$ , and the estimated response function  $\hat{f}$ . For each of the  $N$  observed and fixed values of  $\mathbf{x}_C$ , a curve  $\hat{f}_S^{(i)}$  is plotted against the observed values of  $x_S$ . Therefore, at each x-coordinate,  $x_S$  is fixed and the  $\mathbf{x}_C$  varies across  $N$  observations. Each curve defines the conditional relationship between  $x_S$  and  $\hat{f}$  at fixed values of  $\mathbf{x}_C$ . Thus, the ICE algorithm gives the user insight into the several variants of conditional relationships estimated by the black box.

The ICE algorithm is given in Algorithm 1 in Appendix A. Note that the PDP curve is the average of the  $N$  ICE curves and can thus be viewed as a form of post-processing. Although in this paper we focus on the case where  $|S| = 1$ , the pseudocode is general. All plots in this paper are produced using the R package `ICEbox`, available on CRAN.

Returning to the simulated data described by Equation 3, Figure 2 shows the ICE plot for the `SGB` when  $S = X_2$ . In contrast to the PDP in Figure 1b, the ICE plot makes it clear that the fitted values *are* related to  $X_2$ . Specifically, the `SGB`’s predicted values are approximately linearly increasing or decreasing in  $X_2$  depending upon which region of  $\mathcal{X}$  an observation is in.

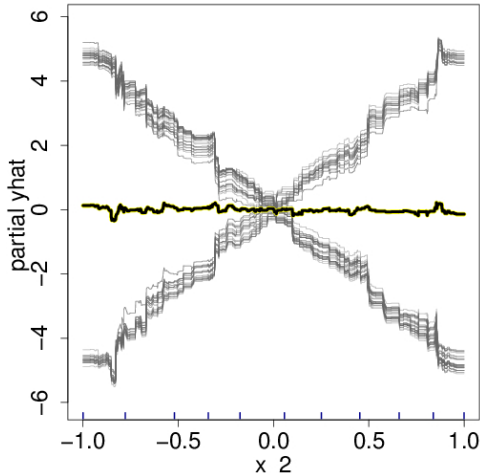


Figure 2: `SGB` ICE plot for  $X_2$  from 1000 realizations of the data generating process described by Equation 3. We see that the `SGB`’s fitted values are either approximately linearly increasing or decreasing in  $X_2$ .

Now consider the well known Boston Housing Data (BHD). The goal in this dataset is to predict a census tract’s median home price using features of the census tract itself. It is important to note that the median home prices for the tracts are truncated at 50, and hence one may observe potential ceiling effects when analyzing the data. We use Random Forests

(RF) implemented in R (Liaw and Wiener, 2002) to fit  $\hat{f}$ . The ICE plot in Figure 3 examines the association between the average age of homes in a census tract and the corresponding median home value for that tract ( $S = \text{age}$ ). The PDP is largely flat, perhaps displaying a slight decrease in predicted median home price as **age** increases. The ICE plot shows those observations for which increasing **age** is actually associated with higher predicted values, thereby describing how individual behavior departs from the average behavior.

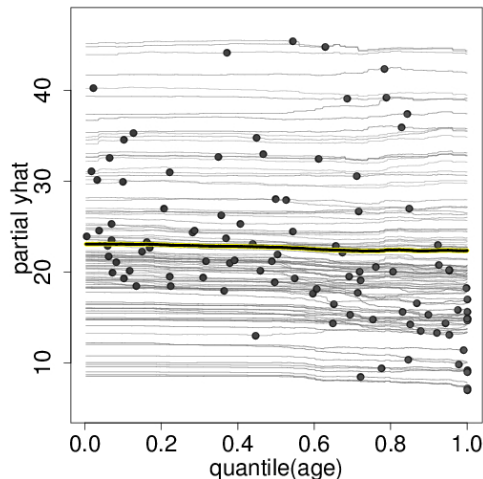


Figure 3: RF ICE plot for BHD for predictor **age**. The highlighted thick line is the PDP. For each curve, the location of its observed **age** is marked by a point. For some observations, higher **age** is associated with a higher predicted values. The upper set of tick marks on the horizontal axis indicate the observed deciles of **age**.

### 3.2 The Centered ICE Plot

When the curves have a wide range of intercepts and are consequently “stacked” on each other, heterogeneity in the model can be difficult to discern. In Figure 3, for example, the variation in effects between curves and cumulative effects are veiled. In such cases the “centered ICE” plot (the “c-ICE”), which removes level effects, is useful.

c-ICE works as follows. Choose a location  $x^*$  in the range of  $x_S$  and join or “pinch” all prediction lines at that point. We have found that choosing  $x^*$  as the minimum or the maximum observed value results in the most interpretable plots. For each curve  $\hat{f}^{(i)}$  in the ICE plot, the corresponding c-ICE curve is given by

$$\hat{f}_{\text{cent}}^{(i)} = \hat{f}^{(i)} - \mathbf{1}\hat{f}(x^*, \mathbf{x}_{C_i}), \quad (4)$$

where the unadorned  $\hat{f}$  denotes the fitted model and  $\mathbf{1}$  is a vector of 1’s of the appropriate dimension. Hence the point  $(x^*, \hat{f}(x^*, \mathbf{x}_{C_i}))$  acts as a “base case” for each curve. If  $x^*$  is the minimum value of  $x_S$ , for example, this ensures that all curves originate at 0, thus removing the differences in level due to the different  $\mathbf{x}_{C_i}$ ’s. At the maximum  $x_S$  value, each centered curve’s level reflects the cumulative effect of  $x_S$  on  $\hat{f}$  relative to the base case. The result is a plot that better isolates the combined effect of  $x_S$  on  $\hat{f}$ , holding  $\mathbf{x}_C$  fixed.

Figure 4 shows a c-ICE plot for the predictor `age` of the BHD for the same RF model as examined previously. From the c-ICE plot we can now see clearly that the cumulative effect of `age` on predicted median value increases for some cases, and decreases for others. Such divergences of the centered curves suggest the existence of interactions between  $x_S$  and  $\mathbf{x}_C$  in the model. Also, the magnitude of the effect, as a fraction of the range of  $y$ , can be seen in the vertical axis displayed on the right of the graph.

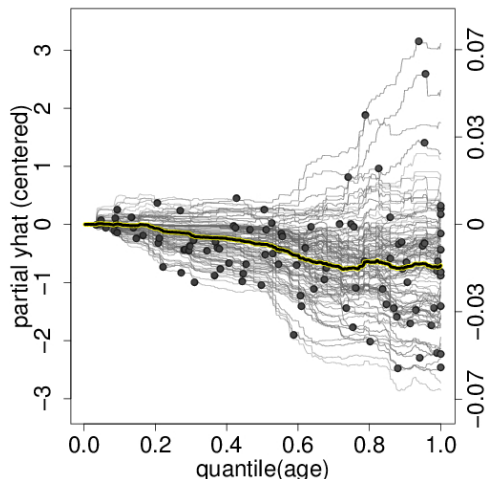


Figure 4: c-ICE plot for `age` with  $x^*$  set to the minimum value of `age`. The right vertical axis displays changes in  $\hat{f}$  over the baseline as a fraction of  $y$ 's observed range. In this example, interactions between `age` and other predictors create cumulative differences in fitted values of up to about 14% of the range of  $y$ .

### 3.3 The Derivative ICE Plot

To further explore the presence of interaction effects, we develop plots of the partial derivative of  $\hat{f}$  with respect to  $x_S$ . To illustrate, consider the scenario in which  $x_S$  does not interact with the other predictors in the fitted model. This implies  $\hat{f}$  can be written as

$$\hat{f}(\mathbf{x}) = \hat{f}(x_S, \mathbf{x}_C) = g(x_S) + h(\mathbf{x}_C), \quad \text{so that} \quad \frac{\partial \hat{f}(\mathbf{x})}{\partial x_S} = g'(x_S), \quad (5)$$

meaning the relationship between  $x_S$  and  $\hat{f}$  does not depend on  $\mathbf{x}_C$ . Thus the ICE plot for  $x_S$  would display a set of  $N$  curves that share a single common shape but differ by level shifts according to the values of  $\mathbf{x}_C$ .

As it can be difficult to visually assess derivatives from ICE plots, it is useful to plot an estimate of the partial derivative directly. The details of this procedure are given in Algorithm 2 in Appendix A. We call this a “derivative ICE” plot, or “d-ICE.” When no interactions are present in the fitted model, all curves in the d-ICE plot are equivalent, and the plot shows a single line. When interactions do exist, the derivative lines will be heterogeneous.

As an example, consider the d-ICE plot for the RF model in Figure 5. The plot suggests that when `age` is below approximately 60,  $g' \approx 0$  for all observed values of  $\mathbf{x}_C$ . In contrast, when `age` is above 60 there are observations for which  $g' > 0$  and others for which  $g' < 0$ , suggesting an interaction between `age` and the other predictors. Also, the standard deviation of the partial derivatives at each point, plotted in the lower panel, serves as a useful summary to highlight regions of heterogeneity in the estimated derivatives (i.e., potential evidence of interactions in the fitted model).

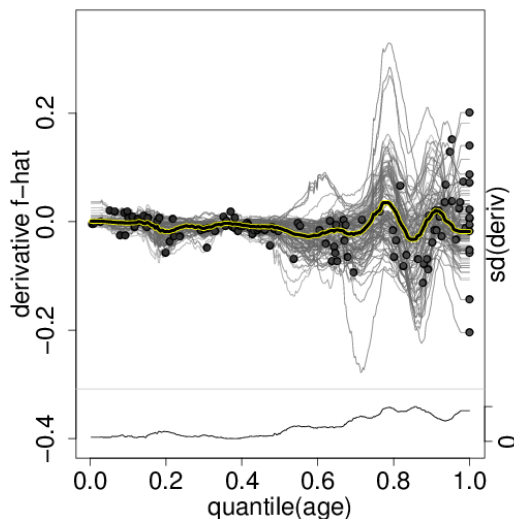


Figure 5: d-ICE plot for `age` in the BHD. The left vertical axis' scale gives the partial derivative of the fitted model. Below the d-ICE plot we plot the standard deviation of the derivative estimates at each value of `age`. The scale for this standard deviation plot is on the bottom of the right vertical axis.

### 3.4 Visualizing a Second Feature

Color allows overloading of ICE, c-ICE and d-ICE plots with information regarding a second predictor of interest  $x_k$ . Specifically, one can assess how the second predictor influences the relationship between  $x_S$  and  $\hat{f}$ . If  $x_k$  is categorical, we assign colors to its levels and plot each prediction line  $\hat{f}^{(i)}$  in the color of  $x_{ik}$ 's level. If  $x_k$  is continuous, we vary the color shade from light (low  $x_k$ ) to dark (high  $x_k$ ).

We replot the c-ICE from Figure 4 with lines colored by a newly constructed predictor,  $x = 1(\mathbf{rm} > \text{median}(\mathbf{rm}))$ . Lines are colored red if the average number of rooms in a census tract is greater than the median number of rooms across all all census tracts and are colored blue otherwise. Figure 6 suggests that for census tracts with a larger number of average rooms, predicted median home price value is positively associated with `age` and for census tracts with a lesser number of average rooms, the association is negative.



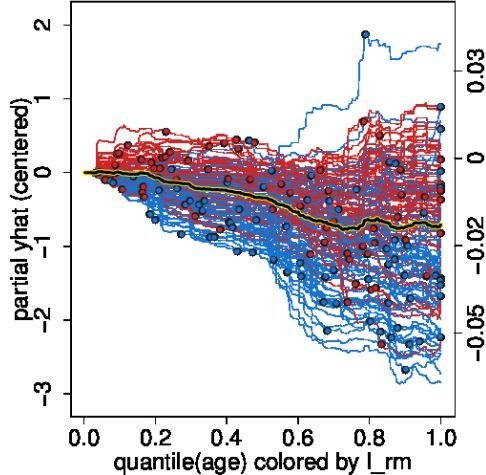


Figure 6: The c-ICE plot for `age` of Figure 4 in the BHD. Red lines correspond to observations with `rm` greater than the median `rm` and blue lines correspond to those with fewer.

## 4 Simulations

Each of the following examples is designed to emphasize a particular model characteristic that the ICE toolbox can detect. To more clearly demonstrate given scenarios with minimal interference from issues that one typically encounters in actual data, such as noise and model misspecification, the examples are purposely stylized.

### 4.1 Additivity Assessment

We begin by showing that ICE plots can be used as a diagnostic in evaluating the extent to which a fitted model  $\hat{f}$  fits an additive model.

Consider again the prediction task in which  $\hat{f}(\mathbf{x}) = g(x_S) + h(\mathbf{x}_C)$ . For arbitrary vectors  $\mathbf{x}_{C_i}$  and  $\mathbf{x}_{C_j}$ ,  $\hat{f}(x_S, \mathbf{x}_{C_i}) - \hat{f}(x_S, \mathbf{x}_{C_j}) = h(\mathbf{x}_{C_i}) - h(\mathbf{x}_{C_j})$  for all values of  $x_S$ . The term  $h(\mathbf{x}_{C_i}) - h(\mathbf{x}_{C_j})$  represents the shift in level due to the difference between  $\mathbf{x}_{C_i}$  and  $\mathbf{x}_{C_j}$  and is independent of the value of  $x_S$ . Thus the ICE plot for  $x_S$  will display a set of  $N$  curves that share a common shape but differ by level shifts according to the unique values of  $\mathbf{x}_C$ .

As an illustration, consider the following additive data generating model

$$Y = X_1^2 + X_2 + \mathcal{E}, \quad X_1, X_2 \stackrel{iid}{\sim} U(-1, 1), \quad \mathcal{E} \stackrel{iid}{\sim} \mathcal{N}(0, 1).$$

We simulate 1000 independent  $(\mathbf{X}_i, Y_i)$  pairs according to the above and fit a generalized additive model (GAM, Hastie and Tibshirani, 1986) via the R package `gam` (Hastie, 2013). As we have specified it, the GAM assumes

$$f(\mathbf{X}) = f_1(X_1) + f_2(X_2) + f_3(X_1 X_2)$$

where  $f_1$ ,  $f_2$  and  $f_3$  are unknown functions estimated internally by the procedure using smoothing splines. Because  $f_3$  appears in the model specification but not in the data gener-

ating process, any interaction effects that **GAM** fits are spurious.<sup>1</sup> Here, ICE plots inform us of the degree to which interactions were fit. Were there no interaction in  $\hat{f}$  between  $X_1$  and  $X_2$ , the ICE plots for  $X_1$  would display a set of curves equivalent in shape but differing in level.

Figure 7a displays the ICE plots for  $X_1$  and indicates that this is indeed the case: all curves display a similar parabolic relationship between  $\hat{f}$  and  $X_1$ , shifted by a constant, and independent of the value of  $X_2$ . Accordingly, the associated d-ICE plot in Figure 7b displays little variation between curves. The ICE suite makes it apparent that  $f_3$  (correctly) contributes relatively little to the **GAM** model fit. Note that additive structure cannot be observed from the PDP alone in this example (or any other).

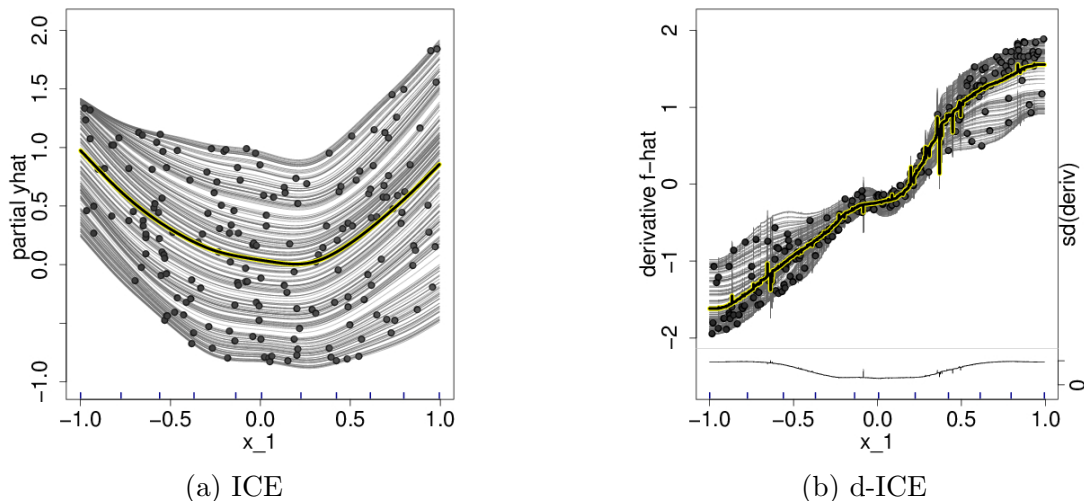


Figure 7: ICE and d-ICE plots for  $S = X_1$  when  $\hat{f}$  is a **GAM** with possible interaction effects between  $X_1$  and  $X_2$ . So as to keep the plot uncluttered we plot only a fraction of all 1000 curves. In the ICE plots the dots indicate the actual location of  $X_1$  for each curve.

## 4.2 Finding interactions and regions of interactions

As noted in Friedman (2001), the PDP is most instructive when there are no interactions between  $x_S$  and the other features. In the presence of interaction effects, the averaging procedure in the PDP can obscure any heterogeneity in  $\hat{f}$ . Let us return to the simple interaction model

$$\begin{aligned}
 Y &= 0.2X_1 - 5X_2 + 10X_2\mathbb{1}_{X_3 \geq 0} + \mathcal{E}, \\
 \mathcal{E} &\stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad X_1, X_2, X_3 \stackrel{iid}{\sim} U(-1, 1)
 \end{aligned}
 \tag{6}$$

to examine the relationship between **SGB**'s  $\hat{f}$  and  $X_3$ . Figure 8a displays an ICE plot for  $X_3$ . Similar to the PDP we saw in Section 1, the plot suggests that averaged over  $X_1$  and  $X_2$ ,

<sup>1</sup>If we were to eliminate  $f_3$  from the **GAM** then we would know a priori that  $\hat{f}$  would not display interaction effects.

$\hat{f}$  is not associated with  $X_3$ . By following the non-parallel ICE curves, however, it is clear that  $X_3$  modulates the fitted value through interactions with  $X_1$  and  $X_2$ .

Where in the range of  $X_3$  do these interactions occur? The d-ICE plot of Figure 8b shows that interactions are in a neighborhood around  $X_3 \approx 0$ . This is expected; in the model given by Equation 6, being above or below  $X_3 = 0$  changes the response level. The plot suggests that the fitted model’s interactions are concentrated in  $X_3 \in [-0.025, 0.025]$  which we call the “region of interaction” (ROI).

Generally, ROIs are identified by noting where the derivative lines are variable. In our example, the lines have highly variable derivatives (both positive and negative) in  $[-0.025, 0.025]$ . The more heterogeneity in these derivative lines, the larger the effect of the interaction between  $x_S$  and  $\mathbf{x}_C$  on the model fit. ROIs can be seen most easily by plotting the standard deviation of the derivative lines at each  $x_S$  value. In this example, the standard deviation function is plotted in the bottom pane of Figure 8b and demonstrates that fitted interactions peak at  $X_3 \approx 0$ .

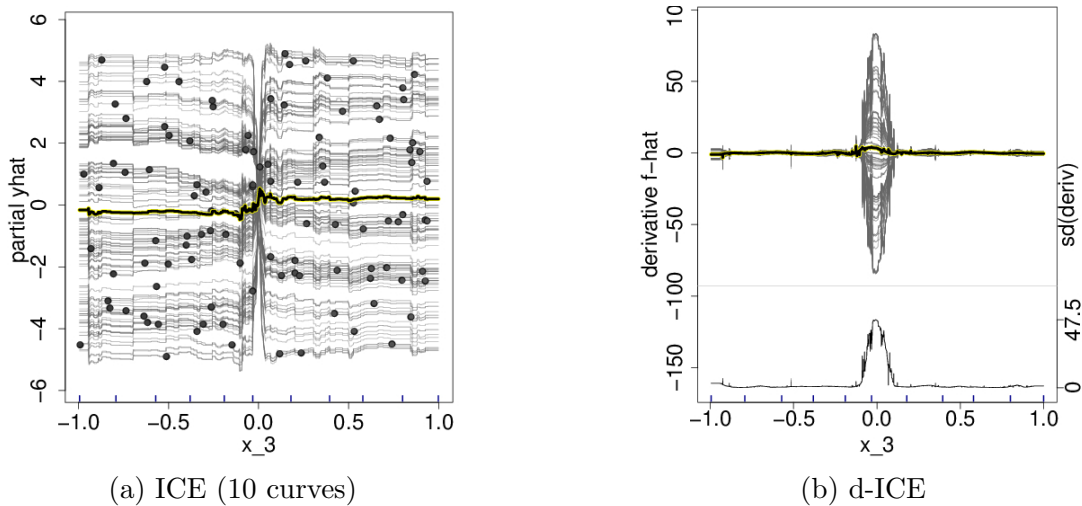


Figure 8: ICE plots for an SGB fit to the simple interaction model of Equation 6.

### 4.3 Extrapolation Detection

As the number of predictors  $p$  increases, the sample vectors  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are increasingly sparse in the feature space  $\mathcal{X}$ . A consequence of this curse of dimensionality is that for many  $\mathbf{x} \in \mathcal{X}$ ,  $\hat{f}(\mathbf{x})$  represents an extrapolation rather than an interpolation (see Hastie et al., 2009 for a more complete discussion).

Extrapolation may be of particular concern when using a black-box algorithm to forecast  $\mathbf{x}_{\text{new}}$ . Not only may  $\hat{f}(\mathbf{x}_{\text{new}})$  be an extrapolation of the  $(\mathbf{x}, y)$  relationship observed in the training data, but the black-box nature of  $\hat{f}$  precludes us from gaining any insight into what the extrapolation might look like. Fortunately, ICE plots can cast light into these extrapolations.

Recall that each curve in the ICE plot includes the fitted value  $\hat{f}(x_{Si}, \mathbf{x}_{Ci})$  where  $x_{Si}$  is actually observed in the training data for the  $i$ th observation. The other points on this curve

represent extrapolations in  $\mathcal{X}$ . Marking each curve in the ICE plot at the observed point helps us assess the presence and nature of  $\hat{f}$ 's hypothesized extrapolations in  $\mathcal{X}$ .

Consider the following model:

$$Y = 10X_1^2 + \mathbf{1}_{X_2 \geq 0} + \mathcal{E}, \tag{7}$$

$$\mathcal{E} \stackrel{iid}{\sim} \mathcal{N}(0, .1^2), \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \begin{cases} \text{U}(-1, 0), \text{ U}(-1, 0) & \text{w.p. } \frac{1}{3} \\ \text{U}(0, 1), \text{ U}(-1, 0) & \text{w.p. } \frac{1}{3} \\ \text{U}(-1, 0), \text{ U}(0, 1) & \text{w.p. } \frac{1}{3} \\ \text{U}(0, 1), \text{ U}(0, 1) & \text{w.p. } 0. \end{cases}$$

Notice  $\mathbb{P}(X_1 > 0, X_2 > 0) = 0$ , leaving the quadrant  $[0, 1] \times [0, 1]$  empty. We simulate 1000 observations and fit a RF model to the data. The ICE plot for  $x_1$  is displayed in Figure 9a with the points corresponding to the 1000 observed  $(x_1, x_2)$  values marked by dots. We highlight observations with  $x_2 < 0$  in red and those with  $x_2 \geq 0$  in blue. The two subsets are plotted separately in Figures 9b and 9c.

The absence on the blue curves of points where both  $x_1, x_2 > 0$  confirms that the probability of  $X_1 > 0$  and  $X_2 > 0$  equals zero. From Figure 9c, we see that in this region of  $\mathcal{X}$ ,  $\hat{f}$  increases roughly in proportion with  $x_1^2$  even though no data exists. Ostensibly the RF model has extrapolated the polynomial relationship from the observed  $\mathcal{X}$ -space to where both  $x_1 > 0$  and  $x_2 > 0$ .

Whether it is desirable for  $\hat{f}$  to display such behavior in unknown regions of  $\mathcal{X}$  is dependent on the character of the extrapolations in conjunction with the application at hand. Moreover, different algorithms will likely give different extrapolations. Examining the ICE plots can reveal the nature of these extrapolations and guide the user to a suitable choice.

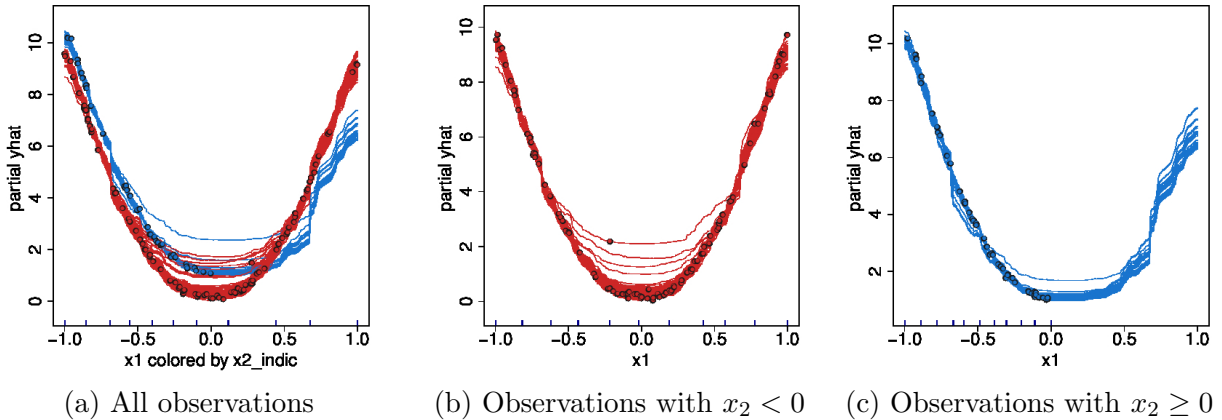


Figure 9: ICE plots for  $S = x_1$  of a RF model fit to Equation 7. The left plot shows the ICE plot for the entire dataset where  $x_2 < 0$  is colored red and  $x_2 \geq 0$  in blue. The middle plot shows only the red curves and the right only the blue. Recall that there is no training data in the quadrant  $[0, 1] \times [0, 1]$ , and so Figure 9c contains no points for observed values when  $x_1 > 0$  (when both  $x_1$  and  $x_2$  are positive). Nevertheless, from Figure 9c's ICE curves it is apparent that the fitted values are increasing in  $x_1$  for values above 0. Here, the ICE plot elucidates the existence and nature of the RF's extrapolation outside the observed  $\mathcal{X}$ -space.

## 5 Real Data

We now demonstrate the ICE toolbox on three real data examples. We emphasize features of  $\hat{f}$  that might otherwise have been overlooked.

### 5.1 Depression Clinical Trial

The first dataset comes from a depression clinical trial (DeRubeis et al., 2014). The response variable is the Hamilton Depression Rating Scale (a common composite score of symptoms of depression where lower scores correspond to being less depressed) after 15 weeks of treatment. The treatments are placebo, cognitive therapy (a type of one-on-one counseling), and paroxetine (an anti-depressant medication). The study also collected 37 covariates which are demographic (e.g. age, gender, income) or related to the medical history of the subject (e.g. prior medications and whether the subject was previously treated). For this illustration, we drop the placebo subjects to focus on the 156 subjects who received either of the two active treatments.

The goal of the analysis in DeRubeis et al. (2014) is to understand how different subjects respond to different treatments, conditional on their *personal* covariates. The difference between the two active treatments, assuming the classic linear (and additive) model for treatment, was found to be statistically insignificant. If the clinician believes that the treatment effect is heterogeneous and the relationship between the covariates and response is complex, then flexible nonparametric models could be an attractive exploratory tool.

Using the ICE toolbox, one can visualize the impact of the treatment variable on an  $\hat{f}$  given by a black box algorithm. Note that extrapolations in the treatment indicator (i.e. predicting at 0 for an observed 1 or vice versa) correspond to counterfactuals in a clinical setting, allowing the researcher to see how the same patient might have responded to a different treatment.

We first modeled the response as a function of the 37 covariates as well as treatment to obtain the best fit of the functional relationship using the black-box algorithm BART (implemented by Kapelner and Bleich, 2014) and obtained an in-sample  $R^2 \approx 0.40$ .

Figure 10a displays an ICE plot of the binary treatment variable, with cognitive therapy coded as “0” and paroxetine coded as “1”, colored by marital status (blue if married and red if unmarried). The plot shows a flat PDP, demonstrating no relationship between the predicted response and treatment when averaging over the effects of other covariates. However, the crossing of ICE curves indicates the presence of interactions in  $\hat{f}$ , which is confirmed by the c-ICE plot in Figure 10b. After centering, it becomes clear that the flat PDP obscures a complex relationship: the model predicts between -3 and +3 points on the Hamilton scale, which is a highly clinically significant range (and almost 20% of the observed response’s range). Further, we can see that BART fits an interaction between treatment and marital status: married subjects are generally predicted to do better on cognitive therapy and unmarried subjects are predicted to do better with paroxetine.

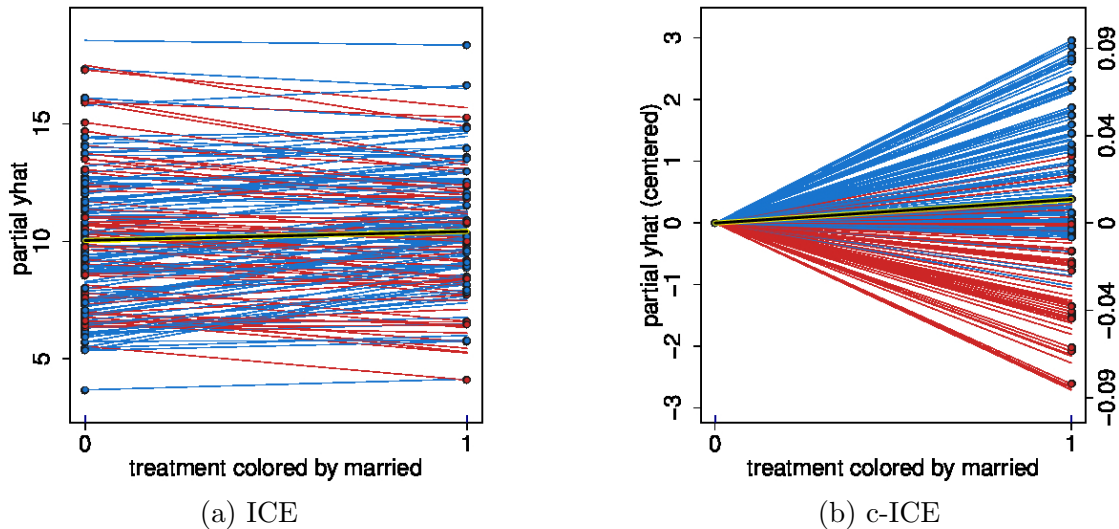


Figure 10: ICE plots of a BART model for the effect of treatment on depression score after 15 weeks. Married subjects are colored in blue and unmarried subjects are colored in red.

## 5.2 White Wine

The second data set concerns 5,000 white wines produced in the *vinto verde* region of Portugal obtained from the UCI repository (Bache and Lichman, 2013). The response variable is a wine quality metric, taken to be the median preference score of three blind tasters on a scale of 1-10, treated as continuous. The 11 covariates are physicochemical metrics that are commonly collected for wine quality control such as citric acid content, sulphates, etc. The model is fit with a neural network (NN) using the R package `nnet` (Venables and Ripley, 2002). We fit a NN with 3 hidden units and a small parameter value for weight decay<sup>2</sup> and achieved an in-sample  $R^2$  of approximately 0.37.

We find the covariate pH to be the most illustrative. The c-ICE plot is displayed in Figure 11a. Wines with high alcohol content are colored blue and wines with low alcohol content are colored red. Note that the PDP shows a linear trend, indicating that on average, higher pH is associated with higher fitted preference scores. While this is the general trend for wines with higher alcohol content, the ICE plots reveal that interaction effects are present in  $\hat{f}$ . For many white wines with low alcohol content, the illustration suggests a nonlinear and cumulatively *negative* association. For these wines, the predicted preference score is actually negatively associated with pH for low values of pH and then begins to increase — a severe departure from what the PDP suggests. However, the area of increase contains no data points, signifying that the increase is merely an extrapolation likely driven by the positive trend of the high alcohol wines. Overall, the ICE plots indicate that for more alcoholic wines, the predicted score is increasing in pH while the opposite is true for wines with low alcohol content. Also, the difference in cumulative effect is meaningful; when varied from the minimum to maximum values of pH, white wine scores vary by roughly 40% of the range

<sup>2</sup>Note that NN models are highly sensitive to the number of hidden units and weight decay parameter. We therefore offer the following results as merely representative of the type of plots which NN models can generate.



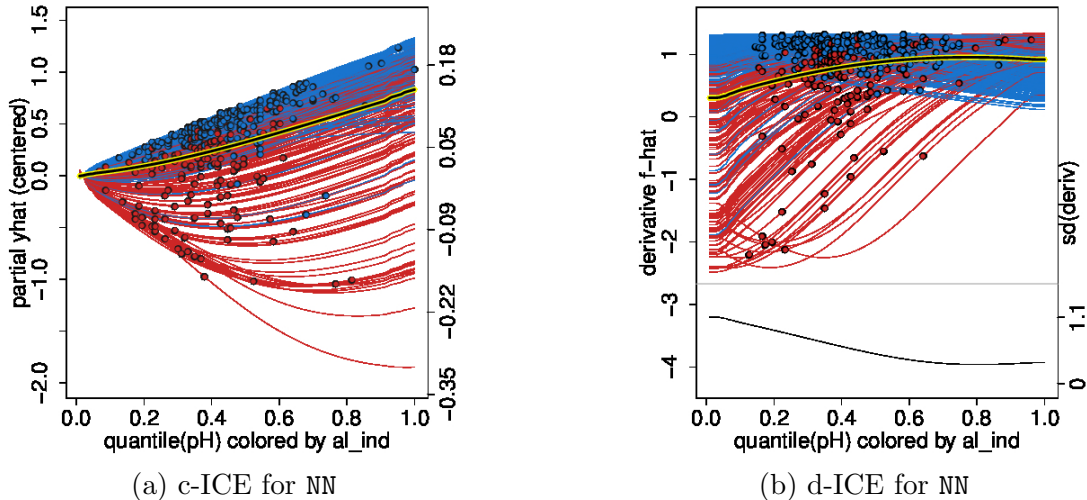


Figure 11: ICE plots of NN model for wine ratings versus pH of white wine colored by whether the alcohol content is high (blue) or low (red). To prevent cluttering, only a fraction of the 5,000 observations are plotted.

of the response variable.

Examining the derivative plot of Figure 11b confirms the observations made above. The NN model suggests interactions exist for lower values of pH in particular. Wines with high alcohol content have mostly positive derivatives while those with low alcohol content have mostly negative derivatives. As pH increases, the standard deviation of the derivatives decreases, suggesting that interactions are less prevalent at higher levels of pH.

### 5.3 Diabetes Classification in Pima Indians

The last dataset consists of 332 Pima Indians (Smith and Everhart, 1988) obtained from the R library MASS. Of the 332 subjects, 109 were diagnosed with diabetes, the binary response variable which was fit using seven predictors (with body metrics such as blood pressure, glucose concentration, etc.). We model the data using a RF and achieve an out-of-bag misclassification rate of 22%.

Once again, ICE plots offer the practitioner a more comprehensive view of the output of the black box. For example, the covariate `skin` thickness about the triceps is plotted as a c-ICE in Figure 12a. The PDP clearly shows an increase in the predicted centered log odds of contracting diabetes. This is expected given that `skin` is a proxy for obesity, a major risk factor for diabetes. However, the ICE plot illustrates a more elaborate model fit. Many subjects with high `skin` have a flat risk of diabetes according to  $\hat{f}$ ; others with comparable thickness exhibit a much larger centered log-odds increase.<sup>3</sup> Figure 12b shows that the RF model fits interactions across the range of `skin` with the largest heterogeneity in effect occurring when `skin` is slightly above 30. This can be seen in the standard deviation of the derivative in the bottom pane of Figure 12b.

<sup>3</sup>The curves at the top of the figure mainly correspond to younger people. Their estimated effect of high thickness is seen to be an extrapolation.

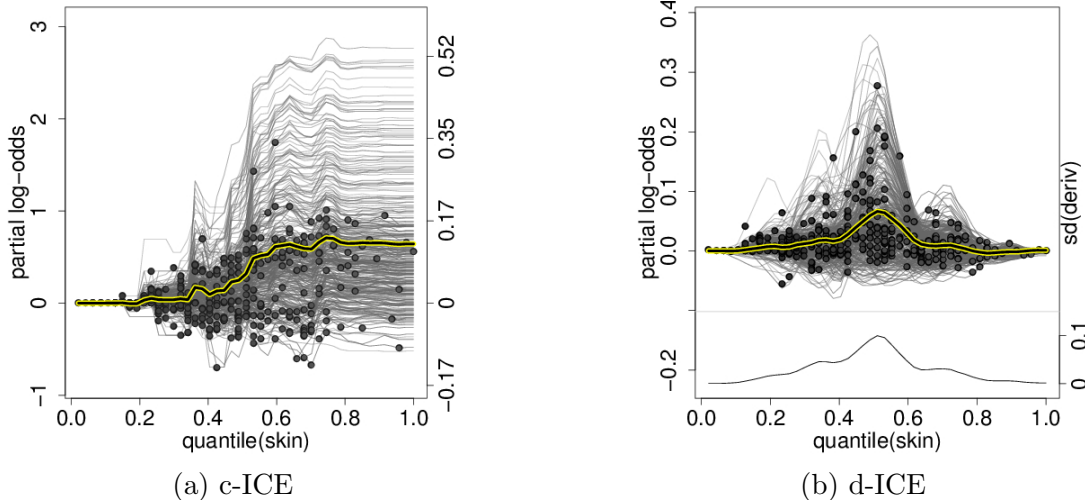


Figure 12: ICE plots of a RF model for estimated centered logit of the probability of contracting diabetes versus `skin` colored by subject `age`.

## 6 A Visual Test for Additivity

Thus far we have used the ICE toolbox to explore the output of black box models. We have explored whether  $\hat{f}$  has additive structure or if interactions exist, and also examined  $\hat{f}$ 's extrapolations in  $\mathcal{X}$ -space. To better visualize interactions, we plotted individual curves in colors according to the value of a second predictor  $x_k$ . We have *not* asked whether these findings are reflective of phenomena in any underlying model.

When heterogeneity in ICE plots is observed, the researcher can adopt two mindsets. When one considers  $\hat{f}$  to be the fitted model used for subsequent predictions, the heterogeneity is of interest because it determines future fitted values. This is the mindset we have considered thus far. Separately, it might be interesting to ascertain whether interactions between  $x_S$  and  $x_C$  exist in the data generating model, denoted  $f$ . This question exists for other discoveries made using ICE plots, but we focus here on interactions.

The problem of assessing the statistical validity of discoveries made by examining plots is addressed in Buja et al. (2009) and Wickham et al. (2010). The central idea in these papers is to insert the observed plot randomly into a lineup of null plots generated from data sampled under a null distribution. If the single real plot is correctly identified amongst 19 null plots, for example, then “the discovery can be assigned a  $p$ -value of 0.05” (Buja et al., 2009). A benefit of this approach is that the procedure is valid despite the fact that we have not specified the form of the alternative distribution — the simple instruction “find the plot that appears different” is sufficient.

### 6.1 Procedure

We adapt this framework to the specific problem of using ICE plots to evaluate additivity in a statistically rigorous manner. For the exposition in this section, suppose that the response  $y$  is continuous, the covariates  $\mathbf{x}$  are fixed, and  $y = f(\mathbf{x}) + \mathcal{E}$ . Further assume  $\mathbb{E}[\mathcal{E}] = 0$  and



$$f(\mathbf{x}) = g(x_S) + h(\mathbf{x}_C), \quad (8)$$

meaning the true  $\mathbf{x}$ -conditional expectation of  $y$  is additive in functions of  $x_S$  and  $\mathbf{x}_C$ . Let  $F$  be the distribution of  $\hat{f}$  when Equation 8 holds and  $f$  is additive. We wish to test  $H_0: \hat{f} \sim F$  versus  $H_a: H_0$  is false.

Recall that ICE plots displaying non-parallel curves suggest that  $\hat{f}$  is not additive in functions of  $x_S$  and  $\mathbf{x}_C$ . Thus if we can correctly identify a plot displaying such features amongst  $K - 1$  null plots generated under  $F$ , the discovery is valid at  $\alpha = 1/K$ .

We sample from  $F$  by using backfitting (Breiman and Friedman, 1985) to generate  $g^*$  and  $h^*$ , estimates of  $g$  and  $h$ , and then bootstrapping the residuals. Both  $g^*$  and  $h^*$  can be obtained via any supervised learning procedures. The general procedure for  $|S| = 1$  proceeds is as follows.

- 1 Using backfitting, obtain  $g^*$  and  $h^*$ . Then compute a vector of fitted values  $\hat{\mathbf{y}}^* = g^*(x_S) + h^*(\mathbf{x}_C)$  and a vector of residuals  $\mathbf{r}^* := y - \hat{\mathbf{y}}^*$ .
- 2 Let  $\mathbf{r}_b$  be a random resampling of  $\mathbf{r}^*$ . If heteroscedasticity is of concern, one can keep  $\mathbf{r}^*$ 's absolute values fixed and let  $\mathbf{r}_b$  be a permutation of  $\mathbf{r}^*$ 's signs. Define  $\mathbf{y}_b := \hat{\mathbf{y}}^* + \mathbf{r}_b$ . Note that  $\mathbb{E}[\mathbf{y}_b \mid \mathbf{x}]$  is additive in  $g^*(x_S)$  and  $h^*(\mathbf{x}_C)$ .
- 3 Fit  $\mathbf{y}_b$  to  $\mathbf{X}$  using the same learning algorithm that generated the original ICE (c-ICE or d-ICE) plot to produce  $\hat{f}_b$ . This yields a potentially non-additive approximation to null data generated using an additive model.
- 4 Display an ICE (or c-ICE or d-ICE) plot for  $\hat{f}_b$ . Deviations from additivity observed in this plot must be due to sources other than interactions between  $x_S$  and  $\mathbf{x}_C$  in the underlying data.
- 5 Repeat steps (2) - (4)  $K - 1$  times, then randomly insert the true plot amongst these  $K - 1$  null plots.
- 6 If the viewer can correctly identify the true plot amongst all  $K$  plots, the discovery is valid for level  $\alpha = 1/K$ . Note that the discovery is conditional on the procedures for generating  $g^*$  and  $h^*$ .

## 6.2 Examples

An application of this visual test where  $g$  is taken to be the ‘‘supersmoother’’ (Friedman, 1984) and  $h$  is a BART model is illustrated using the depression data of Section 5.1. We sample  $\mathbf{r}_b$  by permuting signs. The data analyst might be curious if the ICE plot is consistent with the treatment being additive in the model. We employ the additivity lineup test in Figure 13 using 20 images. We reject the null hypothesis of additivity of the treatment effect at  $\alpha = 1/20 = 0.05$  since the true plot (row 2, column 2) is clearly identifiable. This procedure can be a useful test in clinical settings when the treatment effect is commonly considered linear and additive and can alert the practitioner that interactions should be investigated.

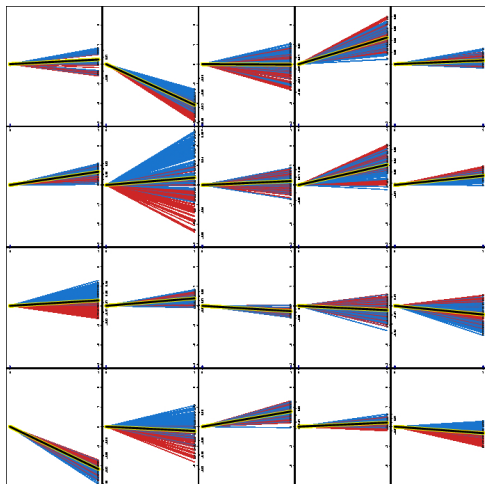


Figure 13: Additivity lineup test for the predictor `treatment` in the depression clinical trial dataset of Section 5.1.

Another application of this visual test where  $g$  is taken to be the supersmoother and  $h$  is a NN model is illustrated using the wine data of Section 5.2. Here again we sample  $\mathbf{r}_b$  by permuting signs. The data analyst may want to know if the fitted model is suggestive of interactions between pH and the remaining features in the underlying model. We employ the additivity lineup test in Figure 14, again using 20 images.

Looking closely one sees that the first and third plots in the last row have the largest range of cumulative effects and exhibit more curvature in individual curves than most of the other plots, making them the most extreme violations of the null. Readers that singled out the first plot in the last row would have a valid discovery at  $\alpha = .05$ , but clearly the evidence of non-additivity is much weaker here than in the previous example. Whereas Figure 13 suggests the real plot is identifiable amongst more than 20 images, it would be easy to confuse Figure 14's true plot with the one in row 4, column 3. Hence there is only modest evidence that pH's impact on  $\hat{f}$  is different from what a NN might generate if there were no interactions between pH and the other predictors.

## 7 Discussion

We developed a suite of tools for visualizing the fitted values generated by an arbitrary supervised learning procedure. Our work extends the classical partial dependence plot (PDP), which has rightfully become a very popular visualization tool for black-box machine learning output. The partial functional relationship, however, often varies conditionally on the values of the other variables. The PDP offers the average of these relationships and thus individual conditional relationships are consequently masked, unseen by the researcher. These individual conditional relationships can now be visualized, giving researchers additional insight into how a given black box learning algorithm makes use of covariates to generate predictions.

The ICE plot, our primary innovation, plots an entire distribution of individual conditional expectation functions for a variable  $x_S$ . Through simulations and real data examples,

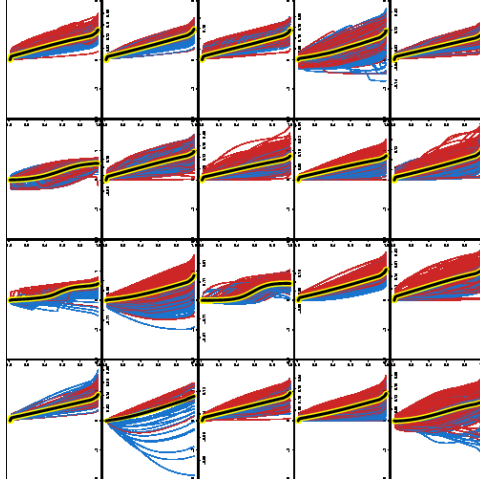


Figure 14: Additivity lineup test for the predictor pH in the white wine dataset of Section 5.2.

we illustrated much of what can be learned about the estimated model  $\hat{f}$  with the help of ICE. For instance, when the remaining features  $\mathbf{x}_C$  do not influence the association between  $x_S$  and  $\hat{f}$ , all ICE curves lie on top of another. When  $\hat{f}$  is additive in functions of  $\mathbf{x}_C$  and  $x_S$ , the curves lie parallel to each other. And when the partial effect of  $x_S$  on  $\hat{f}$  is influenced by  $\mathbf{x}_C$ , the curves will differ from each other in shape. Additionally, by marking each curve at the  $x_S$  value observed in the training data, one can better understand  $\hat{f}$ 's extrapolations. Sometimes these properties are more easily distinguished in the complementary “centered ICE” (c-ICE) and “derivative ICE” (d-ICE) plots. In sum, the suite of ICE plots provides a tool for visualizing an arbitrary fitted model’s map between predictors and predicted values.

The ICE suite has a number of possible uses that were not explored in this work. While we illustrate ICE plots using the same data as was used to fit  $\hat{f}$ , out-of-sample ICE plots could also be valuable. For instance, ICE plots generated from random vectors in  $\mathbb{R}^p$  can be used to explore other parts of  $\mathcal{X}$  space, an idea advocated by Plate et al. (2000). Further, for a single out-of-sample observation, plotting an ICE curve for each predictor can illustrate the sensitivity of the fitted value to changes in each predictor for this particular observation, which is the goal of the “contribution plots” of Strumbelj and Kononenko (2011). Additionally, investigating ICE plots from  $\hat{f}$ 's produced by multiple statistical learning algorithms can help the researcher compare models. Exploring other functionality offered by the `ICEbox` package, such as the ability to cluster ICE curves, is similarly left for subsequent research.

The tools summarized thus far pertain to *exploratory* analysis. Many times the ICE toolbox provides evidence of interactions, but how does this evidence compare to what these plots would have looked like if no interactions existed? Section 6 proposed a *testing* methodology. By generating additive models from a null distribution and introducing the actual ICE plot into the lineup, interaction effects can be distinguished from noise, providing a test at a known level of significance. Future work will extend the testing methodology to other null hypotheses of interest.

## Supplementary Materials

The procedures outlined in Section 3 are implemented in the R package `ICEbox` available on CRAN. Simulated results, tables, and figures specific to this paper can be replicated via the script included in the supplementary materials. The depression data of Section 5.1 cannot be released due to privacy concerns.

## Acknowledgements

We thank Richard Berk for insightful comments on multiple drafts and suggesting color overloading. We thank Andreas Buja for helping conceive the testing methodology. We thank Abba Krieger for his helpful suggestions. We also wish to thank Zachary Cohen for the depression data of Section 5.1 and helpful comments. Alex Goldstein acknowledges support from the Simons Foundation Autism Research Initiative. Adam Kapelner acknowledges support from the National Science Foundation’s Graduate Research Fellowship Program.

## References

- Bache, K. and Lichman, M. (2013). UCI machine learning repository.
- Berk, R. and Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology and Public Policy*, 12:513–544.
- Breiman, L. (2001a). Random Forests. *Machine learning*, pages 5–32.
- Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–231.
- Breiman, L. and Friedman, J. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391):580–598.
- Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1906):4361–83.
- Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian Additive Regressive Trees. *The Annals of Applied Statistics*, 4(1):266–298.
- DeRubeis, R., Cohen, Z., Forand, N., Fournier, J., Gelfand, L., and Lorenzo-Luaces, L. (2014). The personalized advantage index: Translating research on prediction into individual treatment recommendations a demonstration. *PloS one*, 9(1):e83875.
- Elith, J., Leathwick, J., and Hastie, T. (2008). A working guide to boosted regression trees. *The Journal of Animal Ecology*, 77(4):802–13.

- Friedman, J. (1984). A variable span smoother. Technical Report LCS-TR-5, Stanford University, Lab for Computational Statistics.
- Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.
- Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.
- Green, D. and Kern, H. (2010). Modeling heterogeneous treatment effects in large-scale experiments using Bayesian Additive Regression Trees. In *The annual summer meeting of the society of political methodology*.
- Hastie, T. (2013). *GAM: Generalized Additive Models*. R package version 1.09.
- Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Science, second edition.
- Jakulin, A., Mozina, M., and Demsar, J. (2005). Nomograms for visualizing support vector machines. In *KDD*, pages 108–117.
- Jiang, T. and Owen, A. (2002). Quasi-regression for visualization and interpretation of black box functions. Technical report, Stanford University mimeo.
- Kapelner, A. and Bleich, J. (2014). bartMachine: Machine learning with bayesian additive regression trees. *ArXiv e-prints*.
- Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.
- Plate, T., Bert, J., Grace, J., and Band, P. (2000). Visualizing the function computed by a feedforward neural network. *Neural computation*, 12(6):1337–53.
- Rao, J. and Potts, W. (1997). Visualizing Bagged Decision Trees. In *KDD*, pages 243–246.
- Ridgeway, G. (2013). *GBM: Generalized Boosted Regression Models*. R package version 2.0-8.
- Smith, J. and Everhart, J. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care*, pages 261–265. IEEE Computer Society Press.
- Strumbelj, E. and Kononenko, I. (2011). A General Method for Visualizing and Explaining Black-Box Regression Models. In Dobnikar, A., Lotric, R., and Ster, B., editors, *Adaptive and Natural Computing Algorithms, Part II*, chapter 3, pages 21–30. Springer.
- Tzeng, F. (2005). Opening the Black Box - Data Driven Visualization of Neural Networks. In *Visualization*, pages 383–390. IEEE.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010). Graphical Inference for Infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–9.

## A Algorithms

---

**Algorithm 1** ICE algorithm: Given  $\mathbf{X}$ , the  $N \times p$  feature matrix,  $\hat{f}$ , the fitted model,  $S \subset \{1, \dots, p\}$ , the subset of predictors for which to compute partial dependence, return  $\hat{f}_S^{(1)}, \dots, \hat{f}_S^{(N)}$ , the estimated partial dependence curves for constant values of  $\mathbf{x}_C$ .

---

```

1: function ICE( $\mathbf{X}$ ,  $\hat{f}$ ,  $S$ )
2:   for  $i \leftarrow 1 \dots N$  do
3:      $\hat{f}_S^{(i)} \leftarrow \mathbf{0}_{N \times 1}$ 
4:      $\mathbf{x}_C \leftarrow \mathbf{X}[i, C]$  ▷ fix  $\mathbf{x}_C$  at the  $i$ th observation's  $C$  columns
5:     for  $\ell \leftarrow 1 \dots N$  do
6:        $\mathbf{x}_S \leftarrow \mathbf{X}[\ell, S]$  ▷ vary  $\mathbf{x}_S$ 
7:        $\hat{f}_{S\ell}^{(i)} \leftarrow \hat{f}([\mathbf{x}_S, \mathbf{x}_C])$  ▷ the  $i$ th curve's  $\ell$ th coordinate
8:     end for
9:   end for
10:  return  $[\hat{f}_S^{(1)}, \dots, \hat{f}_S^{(N)}]$ 
11: end function

```

---

**Algorithm 2** d-ICE algorithm: Given  $\mathbf{X}$ , the  $N \times p$  feature matrix;  $\hat{f}_S^{(1)}, \dots, \hat{f}_S^{(N)}$ , the estimated partial dependence functions for subset  $S$  in the ICE plot;  $D$ , a function that computes the numerical derivative; returns  $d\hat{f}_S^{(1)}, \dots, d\hat{f}_S^{(N)}$ , the derivatives of the estimated partial dependence. In our implementation  $D$  first smooths the ICE plot using the “supersmoother” and subsequently estimates the derivative from the smoothed ICE plot.

---

```

1: function D-ICE( $\mathbf{X}$ ,  $\hat{f}_S^{(1)}, \dots, \hat{f}_S^{(N)}$ ,  $D$ )
2:   for  $i \leftarrow 1 \dots N$  do
3:      $d\hat{f}_S^{(i)} \leftarrow \mathbf{0}_{N \times 1}$ 
4:      $\mathbf{x}_C \leftarrow \mathbf{X}[i, C]$  ▷ row of the  $i$ th observation, columns corresponding to  $C$ 
5:     for  $\ell \leftarrow 1 \dots N$  do
6:        $\mathbf{x}_S \leftarrow \mathbf{X}[\ell, S]$ 
7:        $d\hat{f}_{S\ell}^{(i)} \leftarrow D \left[ \hat{f}^{(i)}(\mathbf{x}_S, \mathbf{x}_C) \right]$  ▷ numerical partial derivative at  $\hat{f}^{(i)}(\mathbf{x}_S, \mathbf{x}_C)$  w.r.t.  $\mathbf{x}_S$ 
8:     end for
9:   end for
10:  return  $[d\hat{f}_S^{(1)}, \dots, d\hat{f}_S^{(N)}]$ 
11: end function

```

---