Discovering Additive Structure in Black Box Functions

Giles Hooker
Department of Statistics
Stanford University
Stanford, CA, USA
gilesh@stanford.edu

ABSTRACT

Many automated learning procedures lack interpretability, operating effectively as a black box: providing a prediction tool but no explanation of the underlying dynamics that drive it. A common approach to interpretation is to plot the dependence of a learned function on one or two predictors. We present a method that seeks not to display the behavior of a function, but to evaluate the importance of non-additive interactions within any set of variables. Should the function be close to a sum of low dimensional components, these components can be viewed and even modeled parametrically. Alternatively, the work here provides an indication of where intrinsically high-dimensional behavior takes place.

The calculations used in this paper correspond closely with the functional ANOVA decomposition; a well-developed construction in Statistics. In particular, the proposed score of interaction importance measures the loss associated with the projection of the prediction function onto a space of additive models. The algorithm runs in linear time and we present displays of the output as a graphical model of the function for interpretation purposes.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Nonparametric statistics; I.5.2 [**Pattern Recognition**]: Design Methodology—*Feature Evaluation and Selection*; I.6.5 [**Simulation and Modeling**]: Model Development—*Modeling Methodologies*

General Terms

Algorithms, Measurement, Design, Verification

Keywords

Visualization, Diagnostics, Functional ANOVA, Additive models, Graphical models, Interpretation, Feature Selection

KDD'04, August 22–25, 2004, Seattle, Washington, USA.
Copyright 2004 ACM 1-58113-888-1/04/0008 ...$5.00.

1. INTRODUCTION

Many procedures in Data Mining produce prediction functions that act effectively as a "black box": predicting a certain quantity given particular inputs without providing insight into the underlying dynamics that would allow a better understanding of the system. This is particularly the case for neural networks, support vector machines and other kernel methods, and also for many ensemble methods such as boosting and bagging. Indeed, trees – the most obviously interpretable Machine Learning models – have proved to be highly unstable (c.f. [2]) and using them in ensemble methods destroys that interpretability.

This paper initiates a computationally feasible investigation into the structure of such black boxes. In particular, we are interested in measuring the importance of variables in determining the output of the function and in finding underlying additive, or approximately additive, interactions between subsets of variables. Doing this will allow us to decompose a response function into additive parts[1]. If each of these components is low dimensional, then they may be individually viewed, interpreted, and possibly even modeled parametrically. Additionally, this allows us to produce graphical models of function interactions to better see the important components in the structure of a potential functional ANOVA decomposition.

The diagnostic tools presented here also have a limiting function. It is common, when attempting to interpret black box predictions, to produce a matrix of bivariate plots, each describing aggregate behavior on two of the predictors. These are often complemented with plots of conditional variance: how much functional variation is not being accounted for at each value of those two predictors. [13] and [9] advocate this approach, and an example is given in Figure 1. If the function in question is additive up to first order interactions, then such a plot-matrix exactly captures its dynamics: one only needs to sum up the values of the plots, a relatively simple cognitive procedure. Plots of bivariate conditional variances may indicate where something intrinsically higher-dimensional is going on, but they do not indicate what variables to look for as additional interactions, or indeed how many variables might be included in this interaction. The work that we present here solves this problem.

Most of the approaches mentioned above also involve an evaluation of the function based on a uniform distribution.

[1] It is important to emphasize the distinction here between additivity of functional effects and independence of the underlying predictors. That we can write $F(x, y) = f(x)+g(y)$ does not make x and y independent.

This introduces the problem of extrapolation. Prediction functions are rarely learned on data that is even close to uniform, so that Monte Carlo evaluation on uniformly sampled data will inevitably lead to evaluating the function at points of extrapolation. Indeed, such a uniform sample may contain more points of extrapolation than points close to the learning sample. The approach presented below mitigates this problem.

2. THE FUNCTIONAL ANOVA

Visualization procedures for the dynamics of a function or model often rely on plotting univariate or bivariate effects. A standard definition of these is found in the functional ANOVA decomposition, given for example in [11]. This construction has a long history going back to 1948 in [7]. This section will provide an introductory summary.

Let $F(x) : \mathbb{R}^k \to \mathbb{R}$ be square integrable, with

$$x = (x_1, \ldots, x_k)$$

For $u \subset \{1, \ldots, k\}$ we denote by x_u the subset of variables whose indexes are in u. Similarly x_{-u} will indicate the variables with indexes not in u. We can write $f(x)$ uniquely as

$$F(x) = \sum_{u \subseteq \{1,\ldots,k\}} f_u(x)$$

with f_u depending only on x_u and defined by

$$f_u(x) = \int_{x_{-u}} \left(F(x) - \sum_{v \subset u} f_v(x) \right) dx_{-u}$$

In more concrete terms, F is represented as a constant ($u = \phi$), plus terms in one variable ($u = \{i\}$), plus terms in two variables and so forth. Each term is calculated as the projection of F onto a particular subset of the predictors, taking out the lower-order effects which have already been accounted for.

It can be shown that the f_u are orthogonal and that the functional variance $\sigma^2(f) = \int f^2 dx$ may be decomposed as

$$\sigma^2(F) = \sum_{u \subseteq \{1,\ldots,k\}} \sigma_u^2(f_u). \tag{1}$$

Note that this definition can be generalized trivially to any product measure without compromising orthogonality. The Functional ANOVA decomposition was used in [9] and [13] as the basis for a representation of functional behavior using bivariate plots. Throughout the following, we will assume that F is scaled to give $\sigma^2(F) = 1$.

3. PARTIAL DEPENDENCE PLOTS

The interest in this work is to understand functions learned on data. Since most of the data sets are far from uniform, or any product measure, creating a Monte Carlo sample from one of these will distort our picture of the dynamics of the function; placing greater emphasis on areas of low probability mass. [5] defines the *partial dependence* of a function $F(x)$ on x_l to be

$$F_u(x_u) = E_{x_{-u}}\left[F(x)\right] = \int F(x_u, x_{-u}) p_{-u}(x_{-u}) dx_{-u}$$

where $p_{-u}(x_{-u})$ is the marginal distribution of x_{-u}. [5] notes that taking the marginal distribution rather than a conditional distribution given x_u preserves additive structure in F in the sense that the partial dependence recovers the components of an additive function up to a constant. This comes at the expense of distorting the underlying probability space. This distortion, however, is less severe than using a uniform distribution, or than taking the product of marginal densities in all dimensions in the sense that the Kullback-Leibler distance of the underlying distribution with respect to the original data distribution is smaller.

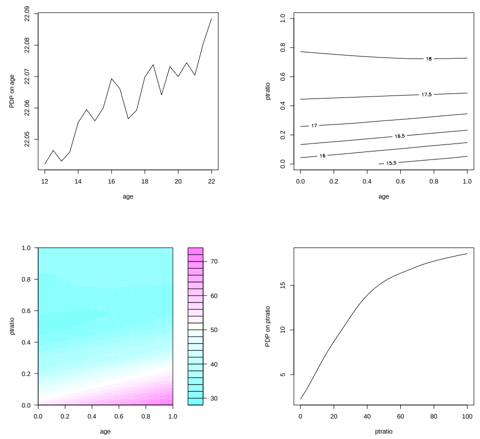

Figure 1: A typical matrix of partial dependence plots given for a neural network trained on the Boston Housing Data. The partial dependence on "age" (top left), and that on "ptratio" (bottom right). A contour plot of the partial dependence on the pair (top right) and a filled contour plot of the variance of the function given "age" and "ptratio" (bottom left).

[5] provides an efficient algorithm for producing plots of these functions for ensembles of trees. A data-driven approximation can be produced for any black box function F by calculating

$$\hat{F}_u(x_u) = \frac{1}{N} \sum_{i=1}^{N} F(x_u, x_{i,-u}) \tag{2}$$

where $\{x_i\}_{i=1}^N$ is the sample used to learn F.

As an example, we trained a 13-26-1 neural network on the Boston Housing Data [6] using the default settings in the `R` package, `nnet`[12]. A typical display might look something like Figure 1: a matrix of plots containing the partial dependence of a prediction function on individual variables on the diagonal. The upper diagonal elements then contain contours of the partial dependence of the function on pairs of variables. Below the diagonal is the variance of the function at each value of a pair of variables. We have presented the two variables "age" and "ptratio", but more would typically be given. A constant variance would indicate that the function has no interaction between the pair of variables specified and the other predictors.

4. TESTS OF ANOVA STRUCTURES

Throughout this section we will assume that the predictor variables are drawn from a product distribution in order to retain the interpretational properties of the Functional ANOVA. In §5 we will give a data-driven approximation that is similar to partial dependence plots (2).

We say that a function f has ANOVA structure described by a collection $\mathbf{U}$ of subsets of $\{1,\dots,k\}$ if $f_u(x) = 0$ for all u that are proper supersets of some element of $\mathbf{U}$. Expressing the subsets of $\{1,\dots,k\}$ as a lattice space ordered by the subset operator, this is equivalent to $\mathbf{U}$ being a least upper bound on the set of elements u of the lattice with non-zero $\sigma_u^2(f_u)$. In concrete terms, a function of the form $f_1(x_1) + f_2(x_2, x_3)$ would be said to have structure $\{\{1\},\{2,3\}\}$, describing the terms that are necessary to recover the function.

The $\mathcal{L}^2$ projection of F onto the set of functions with ANOVA structure described by $\mathbf{U}$ is given by:

$$G_{\mathbf{U}}(x) = \sum_{i=0}^{|\mathbf{U}|}(-1)^{|\mathbf{U}|-i}\sum_{v\in\cap_i\mathbf{U}} E_{-v}F(x) \qquad (3)$$

where $|\mathbf{U}|$ is the cardinality of $\mathbf{U}$, $\cap_i\mathbf{U}$ represents the collection of i-way intersections among the elements of $\mathbf{U}$, and $E_{-v}F(x)$ represents expectation conditioned on x_v. The natural measure of goodness of fit for this projection is therefore:

$$E(F(x) - G_{\mathbf{U}}(x))^2$$

with the expectation taken over the underlying product measure.

Consider a 3 dimensional function, $f(x_1, x_2, x_3)$, with underlying uniform measure. Let us project this function onto the set of functions with ANOVA structure $\{\{1\},\{2\}\}$: additive in x_1 and x_2 and constant in x_3. The projection is given by:

$$\begin{aligned}E(f|x_1) + E(F|x_2) - E(f) &= \int f(x_1,x_2,x_3)dx_2dx_3\\ &+ \int f(x_1,x_2,x_3)dx_1dx_3\\ &- \int f(x_1,x_2,x_3)dx_1dx_2dx_3\\ &= f_1(x_1) + f_2(x_2) + f_0\end{aligned}$$

the corresponding two first effects plus the mean. In general (3) is equal to the sum of the effects indexed by subsets of the elements of $\mathbf{U}$.

Generally, we are interested in the significance of an interaction u, indexed by a subset of $\{1,\dots,k\}$. Formally, we are asking: $\exists v \in \mathbf{U} : u \subseteq v$: *Is there a non-zero functional ANOVA component v that contains u?* To measure this, we project onto $\sum_{(v\not\subseteq u)} f_v(x_v)$: the set of all functions with no interaction in u. The equivalent ANOVA structure is $\mathbf{U} = \{\{1,\dots,k\}\backslash\{i\}\}_{i\in u}$. This is an upper bound for the set of ANOVA structures not containing u. In this case (3) simplifies to:

$$G_u(x) = \sum_{v\subseteq u}(-1)^{k-|v|-1}E_vF(x). \qquad (4)$$

Here the quantity of interest, $E(F(x) - G_u(x))^2$, corresponds, in functional ANOVA terms, to the measure

$$\bar{\sigma}_u^2 = \sum_{v\supseteq u}\sigma_v^2(f_v). \qquad (5)$$

In the three dimensional example above, $\bar{\sigma}_1^2$ measures the error associated with approximating $f(x_1, x_2, x_3)$ with a function of the form $g(x_2, x_3)$, leaving out x_1. We could test the interaction $\{x_2, x_3\}$ by projecting onto the set of functions of the form $g_1(x_1, x_2) + g_2(x_1, x_3)$.

This measure is of interest in the statistical quadrature literature [10] although it differs from the measures of subset importance given by [14]. It can be viewed as the $\mathcal{L}^2$ cost of excluding the interaction u from the model and will alternatively be labeled the $\mathcal{L}^2$ *Cost of Exclusion* or $\mathcal{L}^2$CoE.

5. EMPIRICAL ESTIMATES

In order to estimate the quantities (5) empirically we will perform a simple Monte Carlo integration using our training data, in the vein of §3. Here,

$$\hat{E}_{-v}F(x) = \frac{1}{N}\sum_{i=1}^{N} F(x_v, x_{i,-v})$$

is the empirical partial dependence function of F on v. We use this to construct the empirical projection $\hat{G}_{\mathbf{U}}(x)$ from (3) and measure

$$\frac{1}{N}\sum_{i=1}^{N}(F(x_i) - \hat{G}_{\mathbf{U}}(x_i))^2. \qquad (6)$$

Although we assumed that the data are drawn from a product distribution in §4, this is not strictly necessary. The inequalities below will still hold using partial dependence operators. Particularly, a function which exactly fits a given ANOVA structure will be exactly recoverable with these estimates. However, we are no longer producing the $\mathcal{L}^2$ projection of the prediction function under a known measure onto a space of functions defined by a given ANOVA structure. In this sense, the exact interpretation of the empirical $\mathcal{L}^2$CoE is less clear.

Note that there is a variance associated with the estimate of both the $G_u(x_i)$ and the outer sum of (6). The computational cost of (6) is $O(N^2)$. However, if for each i we estimate $\hat{G}_{\mathbf{U}}(x_i)$ using a randomly drawn subsample of size N_1, we can express

$$\hat{G}_{\mathbf{U}}(x_i) = G_{\mathbf{U}}(x_i) + \epsilon_i$$

where ϵ_i has mean zero and variance $\frac{1}{N_1}E(F(x) - G_{\mathbf{U}}(x))^2$. This gives

$$E\frac{1}{N}\sum_{i=1}^{N}(F(x_i) - \hat{G}_{\mathbf{U}}(x_i))^2 = \left(1 + \frac{1}{N_1}\right)E(F(x) - G_{\mathbf{U}}(x))^2$$

with a variance of $O\left(\frac{1}{N} + \frac{1}{NN_1}\right)$. If we can set $N_1 = 1$ to provide an $O(N)$ estimation scheme without sacrificing variance. At this point, the estimate of $\bar{\sigma}_u^2$ becomes

$$\hat{\bar{\sigma}}_u^2 = \frac{1}{2N}\sum_{i=1}^{N}\left(F(x_i) - \sum_{i=0}^{|\mathbf{U}|}(-1)^{|\mathbf{U}|-i}\sum_{v\in\cap_i\mathbf{U}} F(x_{i,v}, x_{r(i),-v})\right)^2$$

where $r(i)$ is an integer randomly chosen in $\{1,\dots,N\}$. Some algebra shows that if $\bar{\sigma}_u^2 = 0$ (the function does not have an interaction in u) then the estimate also returns zero. This estimate is very similar to the Monte Carlo estimates employed by [10]; although that paper only examines a uniform distribution.

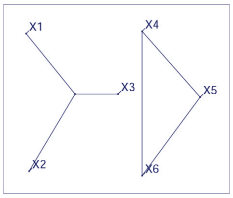

Figure 2: An example of the difference in the VIN graph for functions of the form $f(x_1, x_2, x_3)$ and $g_1(x_4, x_5) + g_2(x_4, x_6) + g_3(x_5, x_6)$.

6. GRAPHICAL DISPLAYS

In order to make use of the cost measures developed above as a diagnostic tool, it is helpful to have a graphical representation of the ANOVA structure of a function. In a learned-function context, very rarely will the importance score for any particular interaction be identically zero. Therefore, a representation that conveys which interactions are deemed significant, as well as the overall importance of interactions, is needed.

For first-order interactions we can represent the set of significant interactions as edges in a graph that uses predictor variables as nodes. In this case, the graph has a similar interpretation to a Bayesian Network. That is: F is additive in x_i and x_j given the collection of variables x_u if any path from x_i to x_j crosses x_u. In fact, a Bayesian Network represents exactly this structure for the log density of the variables. An equivalent functional interpretation is that F is additive in x_i and x_j if for any fixed x_u, $F(x_i, x_j; x_u) = f(x_i) + g(x_j)$ for some functions f and g. This implies the statement

$$\int F(x_i, x_j, x_u) dx_u = f(x_i) + g(x_j).$$

We have labeled such a representation a Variable Interaction Network (VIN).

The inverse of the $\mathcal{L}^2$CoE measure can be taken as a distance between nodes, so that the graph representation may be incorporated into a multi-dimensional scaling routine. The package XGvis [3] has been used to produce the graphs in Figures 2, 3, and 4, although the nodes have been positioned by hand. We have found that while dynamic views of the scaled graphs – having a plot rotate with a three dimensional representation – are informative about interaction strengths, static two-dimensional plots have not provided a good representation of interaction strengths.

A representation limited to edges in a graph with variables as nodes still does not distinguish higher-order interactions, except in so far as they must appear as cliques in the graph: a fact that will appear below. We will therefore represent these higher interactions by a "cartwheel" to distinguish them from a set of additive first order terms. The VIN network we now produce can be interpreted as a hypergraph, with cartwheels representing multi-dimensional edges. We demonstrate this difference in Figure 2.

As an example, we consider the function

$$F(x) = \pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7. \qquad (7)$$

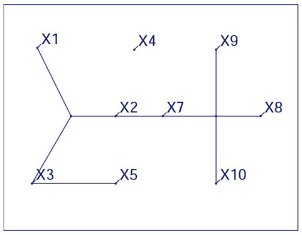

Figure 3: Variable Interaction Network for the function (7) showing non-additive structure.

The true VIN components for F are

$$\mathbf{U} = \{\{1, 2, 3\}, \{2, 7\}, \{4\}, \{3, 5\}, \{7, 8, 9, 10\}\} \qquad (8)$$

which induces the plot in Figure 3. Here edges $\{1, 2\}$, $\{1, 3\}$ and $\{2, 3\}$ would normally form a clique if only first-order interactions were considered. A similar clique occurs for $\{x_7, x_8, x_9, x_{10}\}$.

Note that these cartwheels do not alter the graphical interpretation if we allow as paths any route through an "intersection".

These plots may also be thought of as a graphical version of the representation for hierarchical log-linear models given in [4]. There, a log-linear model is specified on categorical data with the highest order interaction terms listed as in (8). In that case, the absence of an interaction represents independence between categorical variables instead of additivity of a response in either categorical or continuous predictors.

7. THE VIN ALGORITHM

While the evaluation of the strength of a particular interaction can be made in $O(N)$ functional evaluations, there are still 2^k interactions to be evaluated. Denoting by $|u|$ the size of an interaction, the complexity of evaluation also scales as $2^{|u|}$. This becomes prohibitive as k and $|u|$ increases. However, for many functions, the strength of interactions drops off quickly with their size, enabling a very aggressive search strategy.

We will make use of the following monotonicity property:

$$u \subset v \Rightarrow \bar{\sigma}_u^2 \geq \bar{\sigma}_v^2.$$

That this holds is clear from equation (5). This allows us to observe that a d-way interaction can only be considered significant if all its $(d-1)$-way interactions are significant[2]. Thus we can begin by considering main effects - removing one variable - giving a measure of variable importance. We then proceed to first order interactions whose components are all in the significant list. Second order interactions are only considered if all the first order interactions that they contain are included, and so forth. The algorithm here bears strong resemblance to the *Apriori* algorithm for association rules [1]. Both of these rely on a monotonicity property to dramatically reduce the search space as we increase the complexity of interactions (or item-sets) that we are considering.

Suppose we have a threshold $\epsilon \geq 0$, and we wish to find $u : \bar{\sigma}_u^2 \leq \epsilon$. The following algorithm provides a least upper bound for this set.

[2] We will define significance by whether its $\mathcal{L}^2$CoE exceeds some threshold ϵ.

```
$i \ = \ 1$
$\mathbf{U} \ = \ \phi$
Loop:
  $K = \{u \in 1, \ldots, n : |u| = i, \forall v \subset u, v \in \mathbf{U}, |v| = i - 1\}$
  For Each $u \ \in \ K$:
      Calculate $P_F(u)$
      If $P_F(u) \ \leq \ \epsilon$
           $\mathbf{U} \ \leftarrow \ \mathbf{U} \ \cup \ u$
           $\mathbf{U} \ \leftarrow \ \mathbf{U} \ \setminus \ \{v \subset u\}$
      End If
  End For
$i \ \leftarrow \ i \ + \ 1$
End Loop ($K = \phi$ or $i > D$)
```

So long as the function in question does not exhibit very high-dimensional behavior, this algorithm will examine only a very small subset of interactions, which themselves will be of low order. We believe that it is unlikely that we will find functions that are intrinsically high dimensional. [11] explores many apparently high dimensional functions and finds that many are very close to additive. In the event of interactions exceeding some given order (say 6), the algorithm can be curtailed as an indication that interpretation will become very difficult.

8. UPPER BOUNDS ON LATTICE SPACES

In §7, we produced a search algorithm designed to find $u : \bar{\sigma}_u^2 \leq \epsilon$. The natural measure to concern ourselves with is the overall error resulting from a choice of interaction terms. We would like to find a minimal $\mathbf{U}$ to give

$$E(F(x) - G_{\mathbf{U}}(x))^2 < \epsilon$$

for $G_{\mathbf{U}}$ defined in 3: a representation that explains almost all of the function. Here, minimality is taken to be a least upper bound on a lattice. Instead of the set of elements with non-zero score $\sigma_u^2(f_u)$ given in §4, we are interested in an upper bound $\mathbf{U}$ that gives[3]

$$\sum_{u \prec \mathbf{U}} \sigma_u^2(f_u) \geq 1 - \epsilon. \tag{9}$$

Unfortunately, this problem is ill-defined and many such $\mathbf{U}$ may exist.

8.1 Breadth and Depth Searches

In order to specify the problem, importance can be given either to having low-order interactions, or a small number of variables. For the former, a breadth-first search can be performed, including all low-order interactions until the fitting requirement (9) is met. In this case we successively include the term with highest $\mathcal{L}^2$CoE among those candidates of lowest order. In doing this we maintain the hierarchical requirement that possible candidates must already have all of their subsets included.

This algorithm is given formally in the pseudo-code below.

```
$S \ = \ 0$
$\mathbf{U} \ = \ \phi$
Loop:
  $K_1 = \{u \in 1, \ldots, n : \forall v \subset u, |v| = i\}$
  $K_2 = \{u \in K_1 : \ |u| = \min_{v \in K_1} |v|\}$
  For Each $u \in K_2$:
      Calculate $P_F(u)$
  $v \ \leftarrow \ \mathtt{argmax}_{u \in K} \ P_F(u)$
  $\mathbf{U} \ \leftarrow \ \mathbf{U} \ \cup \ v$
  $S \ \leftarrow \ S \ + \ \sigma^2(v)$
End Loop $S \ > \ 1 \ - \ \epsilon$
```

We can estimate

$$\hat{\sigma}_u^2(u) = \sum_{v \subseteq u} (-1)^{|u|-|v|} \hat{\bar{\sigma}}^2(v).$$

The effect of this is to place a penalty on the maximum size of an interaction, only including higher-order terms after all the lower-order have been entered. Here we are looking for a graph that minimizes the size of the greatest “cartwheel” in favor of many lower-order edges.

The alternative is a depth-first search: including the term with highest $\mathcal{L}^2$CoE among those candidates of highest order. The pseudo-code is identical to that above, replacing the definition of K_2 with

$$K_2 = \{u \in K : |u| = \max_{v \in K} |v|\}.$$

This will do the exact converse to the breadth-first search and include a new predictor variable only after all interactions in the current set of predictors have been allowed. The graph from this is expected to have large “cartwheels” but a smaller number of nodes.

8.2 Diagnostics and Greatest Lower Bounds

Given that the problem (9) is poorly specified, we believe that the algorithm as originally stated provides a reasonable set of interactions. The interactions resulting from it can be interpreted as a *greatest lower bound* on the sets $\mathbf{U}$ that satisfy (9).

THEOREM 8.1. *The collection $V = \{u : \bar{\sigma}_u^2 \geq \epsilon\}$ represents a lattice greatest lower bound on collections $\mathbf{U}$ that satisfy (9).*

PROOF. Suppose that $\bar{\sigma}_u^2 < \epsilon$, then the collection $\mathbf{U} = \{v : (v \not\supset u)\}$ has

$$\begin{aligned}\sum_{v \prec \mathbf{U}} \sigma_u^2(f_u) &= 1 - \sum_{w \in V} \sigma_w^2(f_w) \\ &> 1 - \epsilon.\end{aligned}$$

Conversely, for $\bar{\sigma}_u^2 \geq \epsilon$, any $\mathbf{U}$ with $u \not\prec \mathbf{U}$ has

$$\begin{aligned}\sum_{v \prec \mathbf{U}} \sigma_u^2(f_u) &= 1 - \sum_{w \in V} \sigma_w^2(f_w) \\ &\leq 1 - \epsilon.\end{aligned}$$

□

This is satisfying in providing a collection of subsets that must be included to provide a good fit. Further, it aids interpretation significantly by providing a smaller collection of interactions.

There remains the question of choosing ϵ. In §9 we have employed a criteria of explaining 99% of the variance of the function on the empirical distribution of the training data. An alternative is to consider an ordered plot of scores for main effects - these typically die off exponentially - and choose a cutoff manually using a natural break in the scores. This cutoff could be employed throughout, or rechosen at every interaction size. It may well be advantageous to be more or less aggressive in choosing how many terms to include when the relative scores are seen.

[3] $u \prec \mathbf{U}$ is here taken to indicate $\exists v \in \mathbf{U} : u \subset v$.

9. EXAMPLE: BOSTON HOUSING DATA NEURAL NETWORK

We examine the ANOVA structure of the 13-16-1 neural network from §3. Here we picked $\epsilon = 0.7$ which corresponds to 1% of the over-all variance of the function, calculated from the predictions given at the original data. At this cut-off, we include all but three variables: “chas,” “indus” and “zn.” The remaining variables all have interactions with “lstat” apart from “b.” These are very similar findings to those of [13]. There is only one second-order interaction, in “age”-”tax”-”b”. This gives us the final VIN plot in Figure 4. An additive model fit using these interactions and cubic splines improved test error by 12% indicating that this structure is already sufficiently flexible to model the data well.

From a diagnostic point of view, Figure 4 indicates that this prediction function can be well represented by a sum of low dimensional terms. This means that, plots of bivariate representations of the function do provide us with a close-to-comprehensive account of functional behavior. However, care should be taken when examining interactions between the “age”, “tax” and “b” predictors.

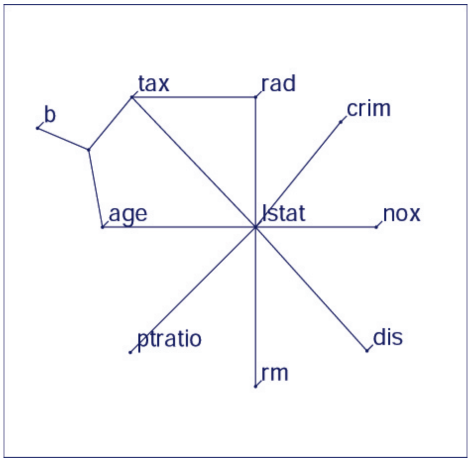

Figure 4: Variable Interaction Network for a Neural Network trained on the Boston Housing Data, with cutoff $\epsilon = 0.7$.

10. CONCLUSIONS

The work presented here can be viewed as a lattice search for a greatest lower bound on the set of hierarchical functional ANOVA components that can be used to represent a learned predictor function. This allows us to represent the complexity of a function in terms of the size of its non-additive interaction components, providing not only an indication of the intrinsic dimensionality of the system, but which predictors interact in a non-additive manner.

This algorithm employs the monotonicity of the $\mathcal{L}^2$CoE measure to provide an efficient search through this lattice. Additional sampling theory allows us to do this in $O(N)$ function evaluations. We have developed a graphical display to make the results more accessible to the user and demonstrated that it has good interpretational properties.

The empirical estimators of which we used are tied directly to the estimation of partial dependence plots. These have been chosen as providing a data-driven method which distorts the distribution of predictor variables less than the uniform distribution normally associated with the functional ANOVA. Results in [8] suggest that it is possible to estimate $\mathcal{L}^2$CoE measures without requiring any predictor variables to be independent. Where the underlying distribution is not an issue or is unknown, estimates based on a uniform distribution can be employed. The algorithms presented in this paper are compatible with any measure of interaction importance for which monotonicity holds.

Acknowledgements

The author would like to thank Jerome Friedman and Art Owen for helpful discussion.

11. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases*, 1994.

[2] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[3] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, and H. Hofmann. Xgvis: Interactive data visualization with multidimensional scaling, 2001. http://www.research.att.com/areas/stat/xgobi/index.html.

[4] S. E. Feinberg. *The Analysis of Cross-Classified Categorical Data*. MIT Press, 1980.

[5] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[6] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

[7] W. Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19:293–325, 1948.

[8] G. Hooker. Black box diagnostics and the problem of extrapolation: Extending the functional anova. Technical report, Stanford University, 2004.

[9] T. Jiang and A. B. Owen. Quasi-regression with shrinkage. *Math. Comput. Simul.*, 62(3-6):231–241, 2003.

[10] R. Liu and A. B. Owen. Estimating mean dimensionality. Technical report, Stanford University, 2003.

[11] A. B. Owen. The dimension distribution and quadrature test functions. *Statistica Sinica*, 13(1), 2003.

[12] R-project. http://www.r-project.org/.

[13] C. Roosen. *Visualization and Exploration of High-Dimensional Functions Using the Functional Anova Decomposition*. PhD thesis, Stanford University, 1995.

[14] I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 5:271–280, 2001.