

Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables

Giles Hooker

To cite this article: Giles Hooker (2007) Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables, Journal of Computational and Graphical Statistics, 16:3, 709-732, DOI: [10.1198/106186007X237892](https://doi.org/10.1198/106186007X237892)

To link to this article: <https://doi.org/10.1198/106186007X237892>



Published online: 01 Jan 2012.



Submit your article to this journal [↗](#)



Article views: 500



View related articles [↗](#)



Citing articles: 36 View citing articles [↗](#)

Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables

Giles HOOKER

This article studies the problem of providing diagnostics for high-dimensional functions when the input variables are known to be dependent. In such situations, commonly used diagnostics can place an unduly large emphasis on functional behavior that occurs in regions of very low probability. Instead, a generalized functional ANOVA decomposition provides a natural representation of the function in terms of low-order components.

This article details a weighted functional ANOVA that controls for the effect of dependence between input variables. The construction involves high-dimensional functions as nuisance parameters and suggests a novel estimation scheme for it. The methodology is demonstrated in the context of machine learning in which the possibility of poor extrapolation makes it important to restrict attention to regions of high data density.

Key Words: Extrapolation; Grid estimates; Machine learning; Variable importance.

1. INTRODUCTION

This article investigates the problem of diagnostics for high-dimensional functions. A frequent source of such functions is in the field machine learning in which functions are estimated to predict some quantity of interest from a large set of independent variables. This is the main motivating situation for this article. Another setting in which these techniques are applicable is in the analysis of computer simulations. These functions have in common that they generally do not have an easily understood algebraic formulation and may effectively be regarded as a “black box.” Once such a function, F , is produced, there are a number of questions that may be asked about it: Which variables are important? How does the function depend on them? In which variables are there strong interactions? Does the function have a close representation in terms of a sum of low-dimensional components?

Giles Hooker is Assistant Profesor, Department of Biological Statistics and Computational Biology, Cornell University Ithaca, NY (E-mail: Giles.Hooker@cornell.edu).

© 2007 American Statistical Association, Institute of Mathematical Statistics,
and Interface Foundation of North America

Journal of Computational and Graphical Statistics, Volume 16, Number 3, Pages 709–732
DOI: 10.1198/106186007X237892

Traditionally, diagnostics have been based on the *average* behavior of F as one or more variables are changed. Let x_1, \dots, x_d represent the d input variables. An *effect* can be defined for the first variable, x_1 , as being

$$f_1(x_1) = \int F(x_1, x_2, \dots, x_d) dx_2 \dots dx_d,$$

where the integral is usually taken with respect to uniform measure over the unit cube. This definition fits into the formalism of the functional ANOVA that provides an elegant theoretical framework in which to understand these diagnostics and build others. Hooker (2004b) extended these diagnostics to measures of variable and interaction importance, indicating how to represent a function using sums of low-dimensional functions.

Problems arise with this form of diagnostic, however, when strong dependencies exist between the input variables. In such situations, integration over a uniform measure can place a large amount of weight in regions of low-probability mass. In machine learning, in which prediction functions are usually chosen from a very large class of candidates, function values in these regions should be regarded as suspect. In other situations, we may not wish to place large emphasis on functional behavior that occurs in regions of low probability.

The solution proposed in this article relies on redefining the functional ANOVA in terms of *projections* of the function of interest onto spaces of additive functions. Viewed in this context, the projection operator may be defined with respect to any \mathcal{L}^2 weight function, provided certain identifiability constraints are maintained.

We use a generalized definition of the functional ANOVA that is equivalent to one proposed by Stone (1994) and further studied by Huang (1998). These articles examine the problem of function estimation from noisy data when the true function can be written as a sum of low-dimensional components. They showed that good convergence results are obtainable when the correct set of components is used. This article is somewhat different in focus; we expect that the function of interest is known, but that it cannot be recovered by low-dimensional components. Instead, we wish to find a close representation of the function in terms of such components. However, these components should not depend on specifying a particular ANOVA structure.

Finding such a representation requires the estimation of high-dimensional components for which the use of tensor-product bases examined by Stone (1994) are not computationally feasible. Instead, this article proposes a novel estimation technique that involves estimating functions on a grid of points and we show that this does provide a feasible estimate.

1.1 STRUCTURE OF THE ARTICLE

This article begins with a demonstration of the problems associated with diagnostic tools based on integral operators and lists a set of desired attributes in Section 2. It then presents a canonical description of the functional ANOVA in Section 3 and presents the proposed generalization in Section 4. We examine this construction in the light of

our desiderata in Section 5. We develop a novel estimation scheme to account for high-dimensional dependence in Section 6 and suggest possible variations in Section 7. We present an estimate of sample variance in Section 8. The method is validated in a simulated setting in Section 9. We discuss the application of this methodology in the particular setting of machine learning in Section 10 and demonstrate its practical application to the California Housing data in Section 11.

The diagnostic tools proposed here assume a known and fixed function, and a known, fixed weight function. In practice the estimation of either is difficult and may produce highly variable results. In machine learning, there are many possible estimation techniques, each of which will have its own variance properties. We therefore have not included this source of variability into our analysis. However, we shall see that known additional variability in either function can be naturally incorporated into estimates of effect variance.

2. INTEGRATION AND EXTRAPOLATION

To illustrate the difficulties associated with diagnostics based on integration, consider the function

$$F(x_1, x_2) = x_1 + x_2^2.$$

We will assume x_1 and x_2 to jointly have uniform measure $P(x_1, x_2)$ on the unit square minus the top right quadrant. A sample of 30 points drawn from this distribution is presented in Figure 1. Data distributions that are more concentrated around nonlinear relationships can be observed in many contexts. The California Housing Data example that we examine in Section 11 contains several pairs of variables with data that falls almost entirely along one axis or other leaving large empty regions of predictor space.

Now consider augmenting F with an extra term:

$$\hat{G}(x_1, x_2) = 10 * (x_1 - 1)_+^2 (x_2 - 1)_+^2,$$

which only occurs in the empty quadrant. We then have a prediction function

$$\hat{F}(x_1, x_2) = F(x_1, x_2) + \hat{G}(x_1, x_2).$$

Any data from $P(x_1, x_2)$ are unable to distinguish between F and \hat{F} . We have used the notation \hat{F} to be reminiscent of machine learning in which F will typically be “learned” from a flexible class of models. This extra term is not dissimilar to the bases used in the MARS algorithm of Friedman (1991), for example, and could plausibly occur as a result of an outlier. For our purposes, points in this quadrant will be regarded as points of extrapolation and the values of \hat{F} correspondingly suspect. In a more general context, we would not want effects such as \hat{G} to change our diagnostics when they occur in a region without probability mass.

The effect of this extrapolation can be seen in Figure 2. These plots compare $\int \hat{F}(x_1, x_2) dx_2$ with the conditional expectation $E(\hat{F}(x_1, x_2)|x_1)$ and the correct additive

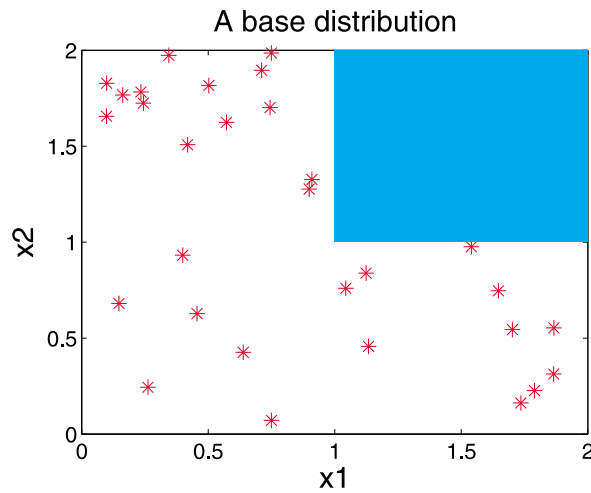


Figure 1. A set of 30 points from a potential distribution of predictor variables. A quadrant with zero probability mass is indicated by the shaded region.

effect. Corresponding plots are given for the effect for x_2 . Both diagnostics are unsatisfactory. The integrated function is clearly influenced by \hat{G} and this discrepancy can be made arbitrarily large. On the other hand, the conditional expectation fails to recover additive components and is distorted by the change in the conditional distribution. The methods proposed in this article recover the desired effects.

2.1 DESIDERATA FOR DIAGNOSTICS

Having illustrated the problems associated with diagnostic tools based on averaging operators and before embarking on developing new tools, we should ask what it is that a low-order representation of functional behavior should provide. We suggest four main properties.

2.1.1 Comprehensibility

Low-order effects should represent some understandable and appropriate quantity. Since any low-dimensional representation of a function necessarily removes information about functional behavior, in what sense such a representation captures that behavior should be well understood and appropriate.

2.1.2 Optimality

A low-dimensional representation should capture as much of the functional behavior as possible, under some well-defined metric. We would like to understand as much of the function as possible through visualizable effects. However, this can be too greedy in the sense that while individual effects may provide a close approximation to functional behavior, taken together they can be misleading.

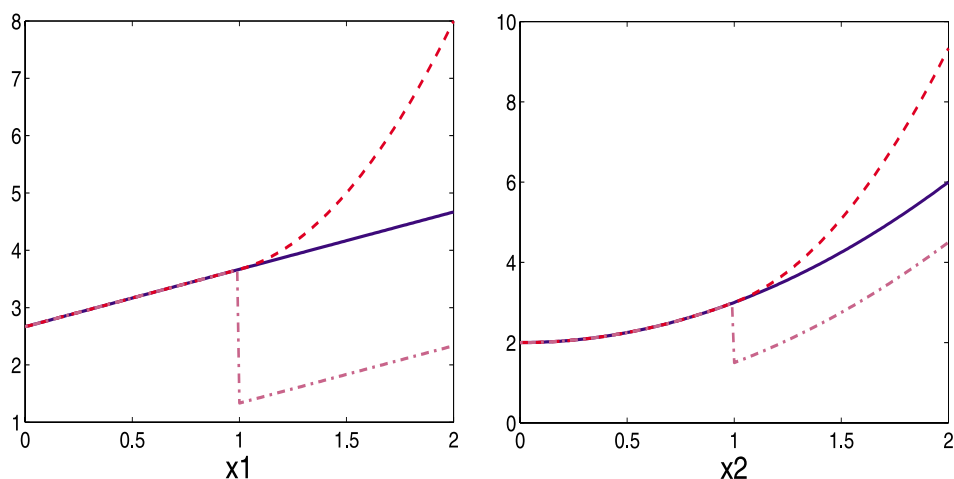


Figure 2. A comparison of Functional ANOVA effects between $F(x_1, x_2)$ (solid) and a learned approximation, $\hat{F}(x_1, x_2)$ (dashed). These two functions are indistinguishable on the data in Figure 1. The left-hand plot provides the effect for x_1 , the right for x_2 . Dotted lines provide the conditional expectations $E(\hat{F}(x_1, x_2)|x_1)$ and $E(\hat{F}(x_1, x_2)|x_2)$ with respect to the measure in Figure 1.

2.1.3 Additivity

A function that is truly additive in some set of variables on the support of their distribution should be exactly recoverable in terms of the additive components. Here, plots of conditional dependence are not satisfactory in being too greedy in the sense above.

2.1.4 Faithfulness to the Underlying Measure

As demonstrated, product measures move probability mass to regions where we do not wish to measure, or place a large emphasis on functional behavior. A good measure will put low weight on prediction values at these points.

3. THE FUNCTIONAL ANOVA

Diagnostics based on integral operators can be best understood through the machinery of the functional ANOVA. The definition of the functional ANOVA decomposition is given as follows. Let $F(x) : [0, 1]^d \rightarrow \mathbb{R}$ be \mathcal{L}^2 over $x = (x_1, \dots, x_d)$. We can then write

$$F(x) = f_0 + \sum_{i=1}^d f_i(x_i) + \sum_{i \neq j} f_{ij}(x_i, x_j) + \dots,$$

a constant mean plus first-order effects, plus second-order effects, and so on. For diagnostic purposes, we usually hope that F can be closely approximated by a few low-order terms.

For the generalization used in this article, we will need to sum over all the functional ANOVA effects. In order to make notation for that summation tractable we introduce the

following multi-index notation: let $u \subseteq \{1, \dots, k\}$, we denote by x_u the subset of variables whose indexes are in u . Similarly x_{-u} indicates the variables with indexes not in u .

We can now write $F(x)$ as

$$F(x) = \sum_{u \subseteq \{1, \dots, d\}} f_u(x_u)$$

with f_u depending only on x_u . For the sake of notational neatness, we write $u \subseteq d$ in place of $u \subseteq \{1, \dots, d\}$ throughout the rest of the article. With these conventions, each effect f_u is defined as

$$f_u(x) = \int_{x_{-u}} \left(F(x) - \sum_{v \subset u} f_v(x) \right) dx_{-u}, \quad (3.1)$$

the integral of F over the variables x_{-u} minus all the lower-order effects. We will refer to this as the “standard” functional ANOVA to distinguish it from our proposed generalization.

A list of desirable properties for the functional ANOVA can be derived sequentially:

Zero Means: $\int f_u(x_u) dx_u = 0$ for each $u \neq \emptyset$.

Orthogonality: $\int f_u(x_u) f_v(x_v) dx = 0$ for $u \neq v$.

Variance Decomposition: Let $\sigma_{(f)}^2 = \int f(x)^2 dx$ then

$$\sigma^2(F) = \sum_{u \subseteq d} \sigma_u^2(f_u). \quad (3.2)$$

We say that F is *additive* in u if we can write $F(x) = \sum_{v \subset u} g_v(x_v)$ for some $\{g_v\}_{v \subset u}$. In particular, if F is additive in u , then $f_u(x_u) = 0$ must hold. This means that F is recovered exactly using terms of the lowest possible order.

The functional ANOVA decomposition has been studied in many contexts and appears in the literature as far back as Hoeffding (1948). Modern applications have been in the study of quasi-Monte Carlo methods for integration in Owen (2003). It has been used directly in a machine learning context and studied by Stone (1994) and Huang (1998). Gu (2002) provided a comprehensive survey of this methodology and some numerical developments were given by Hegland (2002). Roosen (1995) and Jiang and Owen (2003) provided accounts of the use of the standard functional ANOVA for the visualization of high-dimensional functions. This definition should not be confused with that given by Ramsay and Silverman (2005) in which a standard ANOVA varies smoothly with time.

This standard definition for the functional ANOVA already satisfies the first three desiderata outlined in the introduction. However, when the input variables exhibit strong dependence, using this definition can place a large emphasis on regions that have low probability mass. In the next section we present a generalization that is also faithful to an underlying measure.

4. THE FUNCTIONAL ANOVA VIA PROJECTIONS

This article makes use of a generalization of the functional ANOVA proposed by Stone (1994). We replace uniform measure on the unit cube with a general measure $w(x)$. The *weighted* functional ANOVA can be defined as the projection of F onto a space of additive functions under the inner product

$$\langle f, g \rangle_w = \int_{\mathbb{R}^d} f(x)g(x)w(x)dx.$$

Explicitly, we will jointly define all the effects $\{f_u(x_u)|u \subseteq d\}$ as satisfying

$$\{f_u(x_u)|u \subseteq d\} = \underset{\{g_u \in \mathcal{L}^2(\mathbb{R}^u)\}_{u \subseteq d}}{\operatorname{argmin}} \int \left(\sum_{u \subseteq d} g_u(x_u) - F(x) \right)^2 w(x)dx \quad (4.1)$$

under the *hierarchical orthogonality conditions*

$$\forall v \subseteq u, \forall g_v : \int f_u(x_u)g_v(x_v)w(x)dx = 0. \quad (4.2)$$

In fact, (4.1) will always be minimized at zero, providing an expansion of any function in these terms. To see this, consider removing the full interaction term $f_{\{1,\dots,d\}}$ from (4.1), and defining it as the residual

$$f_{\{1,\dots,d\}}(x) = F(x) - \sum_{u \subset d} f_u(x_u).$$

This is orthogonal to the closed set of functions which are additive in u by the projection theorem (Luenberger 1969) and therefore satisfies the conditions (4.2). It will be useful, for the purposes of estimation, to retain the definition via the optimality criterion and to remove the highest-order interaction.

This article employs an equivalent definition for the functional ANOVA, replacing (4.2) with explicit conditions defined for effects in \mathcal{L}^2 . The following lemma is proved in Appendix A (p. 729):

Lemma 1. *The orthogonality conditions (4.2) are true over $\mathcal{L}^2(\mathbb{R}^d)$ if and only if the integral conditions*

$$\forall u \subseteq d, \forall i \in u \int f_u(x_u)w(x)dx_i dx_{-u} = 0. \quad (4.3)$$

hold.

That is, we require $f_u(x_u)$ to integrate to zero with respect to the *marginal* distribution of w on the variables x_u , in each coordinate direction in u .

In Appendix B (p. 730), we show that this definition leads to a unique decomposition under general conditions on w . In particular, this is true whenever the support of w is an open set. We also use these results in Section 6 when w consists of a series of point masses on a grid.

It is easy to see that the standard functional ANOVA effects are recovered when w is given by uniform measure on the unit cube. Note that the conditions (4.3) usually appear as a consequence of the definition of the functional ANOVA save that the integral is now also taken over the variables x_{-u} . This made no difference when w assumed the variables to be independent. It is required when the components are viewed as being the result of a projection.

A natural weight function to choose is the density function $p(x)$ governing the distribution of the input variables. The definition, however, is more general and there are times when we might wish to use others. For example, if $F(x)$ is regarded as being a realization of an heteroscedastic random process, we may wish $w(x)$ to include information about its local variability. In machine learning, we may feel that it is more important to exclude large empty regions of predictor space than to find an exact estimate of the density. Hooker (2004a) provided one algorithm for doing this.

5. ORTHOGONALITY, OPTIMALITY, AND ADDITIVITY

We will briefly examine the extent to which this decomposition satisfies the desiderata for functional diagnostics set out in the introduction. By defining the functional ANOVA components in terms of a projection, we have already demonstrated that the effects are a jointly optimal approximation to F . However, Equation (4.1) does not easily satisfy our need for a comprehensible quantity. Further, it seems unrealistic to attempt to estimate all 2^d effects at once. Nonetheless, from (4.2), we can write an optimization criterion for a single effect, defining $f_u(x_u)$ as the first component of the collection $g_u(x_u)$, $\{g_v(x_v) | v \subset u\}$, $g_{-u}(x_{-u})$, $\{g_{v'}(x_{v'}) | -u \subset v \subseteq -i, i \in u\}$ that minimizes:

$$\int \left(g_u(x_u) + g_{-u}(x_{-u}) + \sum_{v \subset u} g_v(x_v) + \sum_{i \in u} \sum_{-u \subset v' \subseteq -i} g_{v'}(x_{v'}) - F(x) \right)^2 w(x) dx. \quad (5.1)$$

Here g_{-u} is subject to the relaxed condition

$$\int g_{-u}(x_{-u}) w(x) dx_{-u} = 0,$$

and similarly

$$\int g_{v' \subset -j}(x_{v'}) w(x) dx_j dx_{-u} = 0$$

replaces the set of conditions given for each $v \in -j$ and each $j \in u$. Effectively, we subsume $\sum_{v \subset -u} g_v$ into g_{-u} , treating x_{-u} as a single variable, and relax our constraints accordingly.

We will make this explicit using a function $F(x_1, x_2, x_3, x_4)$ with a four-dimensional argument and underlying measure $\mu(x_1, x_2, x_3, x_4)$. Suppose that we are interested in the effect f_{12} . Then the orthogonality conditions (4.2) indicate that f_{12} is orthogonal to $g_{1234} = f_{123} + f_{124} + f_{1234}$. Further, although we need to estimate covarying effects, we

are happy to estimate $g_{234} = f_{23} + f_{24} + f_{234}$ as a single function, similarly g_{134} and g_{34} . This then requires us to estimate $f_{12}, f_1, f_2, g_{34}, g_{134}, g_{234}$ to minimize

$$\int (f_{12} + f_1 + f_2 + g_{34} + g_{134} + g_{234} - F)^2 \mu(x_1, x_2, x_3, x_4) dx_1 dx_2 dx_3 dx_4.$$

This optimization must be done subject to the constraints (4.3). First, the univariate effects integrate to zero:

$$\begin{aligned} \int f_1(x_1) \mu(x_1, x_2, x_3, x_4) dx_1 dx_2 dx_3 dx_4 &= 0 \\ \int f_2(x_2) \mu(x_1, x_2, x_3, x_4) dx_1 dx_2 dx_3 dx_4 &= 0. \end{aligned}$$

Then the effect f_{12} integrates to zero against μ for every value of x_1 held constant. It also integrates to zero for every value x_2 held constant.

$$\begin{aligned} \int f_{12}(x_1, x_2) \mu(x_1, x_2, x_3, x_4) dx_2 dx_3 dx_4 &= 0 \quad \forall x_1 \\ \int f_{12}(x_1, x_2) \mu(x_1, x_2, x_3, x_4) dx_1 dx_2 dx_4 &= 0 \quad \forall x_2. \end{aligned}$$

Finally, the controlling effects must each integrate to zero against μ while holding one of their values constant. However, since we are not interested in separating the effects of x_3 and x_4 , we only condition on both of these at once:

$$\begin{aligned} \int g_{34}(x_3, x_4) \mu(x_1, x_2, x_3, x_4) dx_1 dx_2 dx_3 dx_4 &= 0 \\ \int g_{134}(x_1, x_3, x_4) \mu(x_1, x_2, x_3, x_4) dx_2 dx_3 dx_4 &= 0 \quad \forall x_1 \\ \int g_{134}(x_1, x_3, x_4) \mu(x_1, x_2, x_3, x_4) dx_1 dx_2 &= 0 \quad \forall x_3, x_4 \\ \int g_{234}(x_2, x_3, x_4) \mu(x_1, x_2, x_3, x_4) dx_1 dx_3 dx_4 &= 0 \quad \forall x_2 \\ \int g_{234}(x_2, x_3, x_4) \mu(x_1, x_2, x_3, x_4) dx_1 dx_2 &= 0 \quad \forall x_3, x_4. \end{aligned}$$

Throughout the article we will use d' to denote the cardinality of u . The original $2^d - 1$ terms have now been reduced to $2^{d'+1} - 1$. Moreover, we can group the effects into two. First, a projection, $\sum_{v \subseteq u} f_v(x_v)$ onto x_u that we would like to visualize and second, the remaining terms which act to control for the effect of covariance in the predictor variables.

In this setting we now have a comprehensible optimality criterion: we are finding the best fit of F on the space of functions with no higher interaction than u . Alternatively, f_u is the best representation of F using only x_u *while controlling for the additive effects of other sets of variables*. Further, an examination of the form of the functions being optimized in (5.1) shows that if F really does have an additive component f_u , then we must exactly recover F , and therefore f_u . Finally, using an appropriate w allows us to concentrate the fitting criteria on regions that are regarded as important, usually through having high probability mass.

6. GRID MEASURES AND POINTWISE ESTIMATION

We have established a theoretical criterion by which we can produce a set of diagnostic tools that both preserve additivity and avoid evaluating a proposed function in regions of zero probability. In this section we propose a method to approximate that criterion by a computationally feasible scheme.

For small d' , Equation (5.1) presents a feasible number of terms to be estimated. However, we are still faced with the need to estimate $2^{d'-1}$ functions of dimension $d-1$ under integral constraints. For even moderate d , this is not feasible using the tensor-products of piecewise polynomials studied in Stone (1994). The important realization here is that we are not interested in those functions except to control for the covariance in the underlying predictor space. We will therefore confine ourselves to estimating all terms in the decomposition only at a carefully chosen set of points.

Attempting to estimate these functions at a single point is, nominally, an under-determined task—assuming the points to be distinct in all dimensions, the function can be recovered exactly at those points using only one predictor. However, following Lemma 2, if the points are placed on a grid, then under the right constraints the problem becomes a solvable linear system.

Let us start with a set of N points x_i in \mathbb{R}^d generated by a uniform distribution over the range of interest. Then we will take as our grid the set of values:

$$\{z_j\}_{j=1}^{N^{d'+1}} = \{x_{u_1, i_1}, x_{u_2, i_2}, \dots, x_{u_k, i_{d'}}, x_{-u, j}\}_{i_1, \dots, i_{d'}, j=1}^N. \quad (6.1)$$

This is formed by taking all possible combinations of values for the predictors that appear in the initial random collection. Here the vector $x_{-u, j}$ is treated as a single value and its entries are not permuted with respect to each other. We will denote by \mathcal{I}_u the union of sets of subsets $\{v \subseteq u\}$, $\{-u\}$, $\{-u \subset v' \subseteq -i \mid i \in u\}$. We will use the convention that $z_{v, k}$ denotes only those dimensions in the vector z_k which are indexed by v .

We now translate (5.1) into the problem of finding $\{f_v(z_{v, k}) \mid v \in \mathcal{I}_u, k \in 1, \dots, N^{|v|}\}$ to minimize

$$\sum_{i=1}^{N^{d'+1}} w(z_i) \left(\sum_{v \in \mathcal{I}_u} f_v(z_{v, i}) - F(z_i) \right)^2 \quad (6.2)$$

under the constraints:

$$\forall v \in \mathcal{I}_u, \forall j \in v, \forall z_{v, k} : \sum_{i=1}^N \left(\sum_{l=1}^{d'-|v|+1} w(z_{v, i}, z_{-v, l}) \right) f_v(z_{v \setminus j, k}, z_{j, i}) = 0. \quad (6.3)$$

We can regard the effect values at the data points, $f_u(z_i)$, as parameters to be estimated. Then (6.2) can be written as a linear least squares problem:

$$\text{minimize } (Xf - F)^T W(Xf - F), \quad (6.4)$$

where F is a vector whose i th component is $F(z_i)$, the vector f lists the value of $f_u(z_i)$ for each u and i and W is a diagonal matrix with $w(z_i)$ on the diagonal. This is minimized under the constraints (6.3), written as

$$Cf = 0. \quad (6.5)$$

Here both X and C perform addition on the components of f . We index the rows of X by $k \in 1 \dots N^{d'+1}$, corresponding to grid points. The columns are indexed by (v, j) , for

$v \in \mathcal{I}_u$ and $j \in 1 \dots N^{d'+1}$, corresponding to effects v evaluated at the grid points $z_{v,j}$, then

$$[X]_{k,(v,j)} = \begin{cases} 1 & \text{if } z_{v,j} = z_{v,k} \\ 0 & \text{otherwise} \end{cases}$$

This is a highly sparse system which corresponds to the matrix of effects that would be used to represent an ANOVA model for categorical predictors using all but the highest-order interaction.

Cf is the matrix representation of (6.3). Its rows are indexed by (v', j, k) corresponding to the sum over dimension j at $z_{v',k}$ of $f_{v'}(z_{v',k})$. Its columns are indexed the same way as X . We can then write the entries of C as being

$$[C]_{(v',j,k),(u,i)} = \begin{cases} \sum_{l=1}^N w(z_{-j,k}, z_{j,l}) & \text{if } u = v \text{ and } z_{v' \setminus j,k} = z_{u \setminus j,i} \\ 0 & \text{otherwise} \end{cases}$$

This can be re-expressed as

$$C = Y\tilde{W}$$

with \tilde{W} representing a matrix with $X^T w$ on the diagonal and

$$[Y]_{(v',j,k),(u,i)} = \begin{cases} 1 & \text{if } u = v \text{ and } z_{v' \setminus j,k} = z_{u \setminus j,i} \\ 0 & \text{otherwise} \end{cases}$$

Solving via the method of Lagrange multipliers results in a very large, very sparse, weighted linear system of the form

$$\begin{bmatrix} X^T W X & \tilde{W} Y^T \\ Y \tilde{W} & 0 \end{bmatrix} \begin{bmatrix} f \\ \lambda \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}. \quad (6.6)$$

where λ contains the set of Lagrange multipliers. To provide an idea of the practical size of the system, using a grid with N points in each dimensions, this system has $N^{d'+1}$ equations in

$$\sum_{j=0}^{d'} \binom{d'+1}{j} N^j = (N+1)^{d'+1} - N^{d'+1} \approx N^{d'}$$

unknowns and a further

$$\sum_{j=1}^{d'} \binom{d'+1}{j} j N^{j-1} \approx N^{d'-1}$$

constraining equations. Solving such a system is computationally feasible for small d' and moderate N . Given that we are interested in visualizing effects of order only 1 or 2, this is quite adequate.

Note that where $w(z)$ is a product of univariate functions, this system exactly reproduces the equivalent estimation of standard functional ANOVA effects, using the z_i as evaluation points. This can be seen by taking w to be a distribution of point masses on the grid of z 's.

The structure of this linear system allows for higher order effects if we employ an iterative method. For large N or d' , even storing $X^T X$ can be prohibitive. However, since both X and C effectively act as known addition operators, they need not be stored at all. In this setting, Conjugate Gradient methods, can provide a significant storage saving. See, for example, Press et al. (1997) for a description of these methods and their properties.

The full product on $N^{d'+1}$ points is also not necessary. Lemma 2 makes the observation that identifiability only requires $2^{d'+1}$ points on a product grid under the constraints (6.3). Thus we can reduce the computational complexity significantly by taking the original set of points in smaller groups and only forming products among those groups. The use of grid-based methods make the estimates of $f_u(z_k)$ highly correlated and produces smooth representations of the effect when these points are interpolated. If multiple grids are used that do not share points, these estimates will become rough, necessitating a smooth when they are plotted.

A quick and easy alternative to the numerical methods suggested earlier is to use existing software for computing a weighted standard ANOVA with weights given by the density at each evaluation point. Once a grid is formed and the prediction and density functions evaluated on it, the values of that grid in each dimension may be taken as being *categorical* values. An ANOVA, including all necessary interactions, can then be calculated and the resulting coefficients used as the vector f from (6.4). Standard ANOVA calculations do not include the constraints (6.5), but instead typically set one coefficient equal to zero. This may introduce a distortion since the effects will no longer necessarily satisfy (4.2). Experimentation with such a method using the R statistical package found that it gives very reasonable answers, but that it quickly became computationally intensive when large numbers of grid points were used. Effect values can be undefined when there is zero marginal weight where they ought to occur. R was found to helpfully flag such values as missing and to produce holes in the corresponding contour plots to indicate that this was the case.

7. NONUNIFORM SAMPLING SCHEMES

Using a base set of uniformly distributed points as the seed for the grids used above suffers from the problem of poor coverage in large dimensions. In particular, if w is zero in large parts of predictor space, we run a risk of leaving the solution to (6.2) undefined by giving weight zero to too many rows. Even when a solution can be found, by effectively removing a large number of points from the estimation, we can increase the variance associated with the estimate.

An immediate solution to this problem is to try to concentrate the points on regions of high weight and adjust $w(x)$ accordingly. Since the points must lie on a union of grids, using the original data as a seed and dividing $w(x)$ by the product of its marginals is a first solution.

There is a disadvantage to this approach in that we may desire to see the effect plotted at points more uniformly distributed than the empirical marginal of the dimensions, x_u , in question. In this case, taking a uniform sample, or uniformly spaced points in only these dimensions may be appropriate. We can then form a product distribution and divide by the

marginal on x_{-u} , treating this collection of variables as a single variate as in (5.1). This technique is employed in Section 11. Quasi Monte Carlo approaches could also be used here to get a more even distribution of approximation points.

Specific models of the distribution of the underlying training data can point to alternative, more efficient estimation schemes. Mixture of Gaussian models, or more general mixtures of products can be used to generate a grid measure for each term in the mixture. These can then be employed to provide identifiability at both lower computational cost and avoiding using rows that are given zero weight.

8. ESTIMATES OF SAMPLE VARIANCE

Variances for effect estimates are not commonly given as part of diagnostics displays. Nonetheless, it is important to have some idea of how stable the displayed effects are. Estimates of effect variance under the proposed estimation scheme are complicated by two factors: the grid-based nature of the approximation and the use of the weighting function w . The first of these will also be a concern for functional diagnostics based on the standard functional ANOVA when a set of grid points is used for evaluation. This occurs, for example, in the partial dependence plots described in Section 10. In this section, we provide an estimate of the variance of the proposed effect estimates and indicate how it may accommodate variability in F and w .

While our effect estimates have a very similar flavor to that of the standard (nonfunctional) ANOVA, the analysis of their variability is substantially different. For each point, the observation $F(z_i)$ can still be written as

$$\sum_{v \subseteq u} g_{u,z_i} + \sum_{j \in u} g_{-j,z_i} + g_{-u,z_i} + \epsilon_i, \quad (8.1)$$

where $\epsilon_i = \sum_{v \supset u} g_v(z_i)$. We have subscripted the z_i here to mimic the standard ANOVA effect. However, there may be dependence between ϵ_i and ϵ_j when it is known that $z_{v,i} = z_{v,k}$ for some v . Moreover, the least squares criterion (6.2) is weighted by $w(z_i)$ which must also be considered to be random.

We propose a δ -method approximation of effect variance using the values $w(z_i)$ and $w(z_i)F(z_i)$ as statistics. Although these statistics are likely not to have a normal distribution, observe that the estimate as defined in (6.6) may be written as

$$\hat{f} = Z(W)^{-1} X^T W F \quad (8.2)$$

where $Z(W)$ represents the left-hand side of (6.6). X^T can be regarded as an addition operator. In this case

$$X_{(u,k),i}^T G = \sum_{z_{u,j}=z_{u,k}} G(z_j);$$

summing the weighted function values that correspond to the effect $f_u(x_{u,k})$. We now note that a sum in one coordinate j :

$$\sum_{i=1}^N w(z_{j,i}, x_{-j}) F(z_{j,i}, x_{-j})$$

is a sum of independent random variables to which the central limit theorem can be applied when we condition on x_{-j} . Using these sums as statistics creates a computationally more efficient estimate than using individual $w(z_i)F(z_i)$ values. The resulting variance estimate will be the same, however, and we will employ the latter as providing a clearer exposition.

Similar calculations allow the terms $X\tilde{W}$ to be expressed as sums of the $w(x_i)$ and hence the central limit theorem can also be employed with respect to them. Unfortunately, $X^T W X$ has terms involving individual $w(x_i)$, meaning that the δ -method approach must be regarded as being heuristic. We devote the rest of this section to outlining how such an approach may be made computationally feasible.

The δ -method requires the derivatives of \hat{f} with respect to WF and W . The first of these is simply

$$Z(W)^{-1} X^T.$$

For the second, the derivative with respect to $w(z_i)$ may be given by

$$-Z(W)^{-1} \frac{dZ(W)}{dw(z_i)} Z(W)^{-1} X^T W F = -Z(W)^{-1} \frac{dZ(W)}{dw(z_i)} \hat{f}.$$

These vectors then form the columns of a large matrix which may be re-expressed as

$$-Z(W)^{-1} \begin{bmatrix} X^T \tilde{F} + \tilde{\lambda} X^T \\ Y \tilde{f} X^T \end{bmatrix} = -Z(W)^{-1} D,$$

where \tilde{F} and $\tilde{\lambda}$ are a diagonal matrices with diagonals $X\hat{f}$ and $Y^T \lambda$, respectively, and \tilde{f} has \hat{f} on the diagonal.

The covariance matrix of f and λ can now be written as

$$Z(W)^{-1} \left[X^T \Sigma_F X - X^T \Sigma_{FW} D^T - D \Sigma_{FW}^T X + D \Sigma_W D^T \right] Z(W)^{-1},$$

where Σ_F , Σ_W , and Σ_{FW} represent the covariance matrices for F , W , and between F and W , respectively. These are calculated from the data, taking into account the grid structure of the estimation. In particular, $w(z_i)$ and $w(z_j)$ are given nonzero correlation, σ_v^2 , if z_i and z_j agree on some subset v of their indexes. This is calculated by the sample correlation of all pairs $w(z_i)$ and $w(z_j)$ for which $z_{v,i} = z_{v,j}$. There are $2N^{d'}$ correlated pairs among $N^{2(d'+1)}$ entries in Σ_W , making this very sparse. Σ_F and Σ_{FW} are calculated in the same manner.

These estimates change if we fix the values of the estimation points in the dimensions u —we typically will take a fixed grid in the dimension of interest. In this case we need to calculate $\sigma_{v,i}^2$ for each point i in this grid. In this case, the matrices Σ become more sparse since the correlation between any two points is nonzero only if they share the same $z_{-u,k}$.

If F or w , or both, are regarded as random and some knowledge is assumed of their variance at each of the evaluation points, then the covariance matrices Σ may be augmented with these variances as well.

It is natural to view the proposed approximation to the effect f_u as being given by a linear interpolation of the calculated effect values. This would correspond to a finite element discretization of (5.1). Huang (1998) demonstrated that a finite-element approach provides a consistent estimate and gives rates of convergence under independent sampling. The approach in this article, however, involves terms of dimension $d - 1$ which are not amenable to a finite element approximation. If they were, the problem would reduce to straightforward numerics. More formal analysis of these techniques are beyond the scope of this article, but would require the high dimensional terms to be explicitly regarded as nuisance parameters.

9. A SIMULATED DEMONSTRATION

The goal of the article is to produce a diagnostic effect for a variable or set of variables that is capable of recovering additive functions while avoiding distortion by functional behavior in regions of low probability. This section demonstrates that the proposed weighted functional ANOVA, by taking other additive functions into account, does exactly this. This is true even if the additive components share variables.

The example that we present is defined on the unit cube. We will take as a function

$$F(x, y, z) = xy + xz + yz \quad (9.1)$$

to which we will add a spurious term,

$$5I(x > 1/2, y > 1/2, z > 1/2),$$

where $I(x > 1/2, y > 1/2, z > 1/2)$ is the indicator of all of x , y , and z being greater than $1/2$. This results in an approximation

$$\hat{F}(x, y, z) = F(x, y, z) + 5I(x > 1/2, y > 1/2, z > 1/2).$$

We have chosen an indicator function to be particularly visible in contour plots. When the term is only nonzero outside the support of the underlying measure, we do not wish it to appear in the estimated effects.

To evaluate the weighted functional ANOVA, a sample of ten uniformly distributed points was drawn and the product distribution of these points taken as in (6.1). Figure 3 presents contour plots of the second-order effects for x and y for this function defined on three different measures. The first of these is the unit cube providing the standard functional ANOVA effects and the distortion due to the spurious term is evident. The second subtracts the upper corner from the cube—exactly that part of the cube in which the spurious effect occurs. Here the desired components— $f_\phi + f_x(x) + f_y(y) + f_{xy}(x, y) = xy$ —are recovered exactly. The final distribution further removes all the upper corners from each of the faces of the cube, leaving an “L”-shaped marginal distribution in each pair of variables

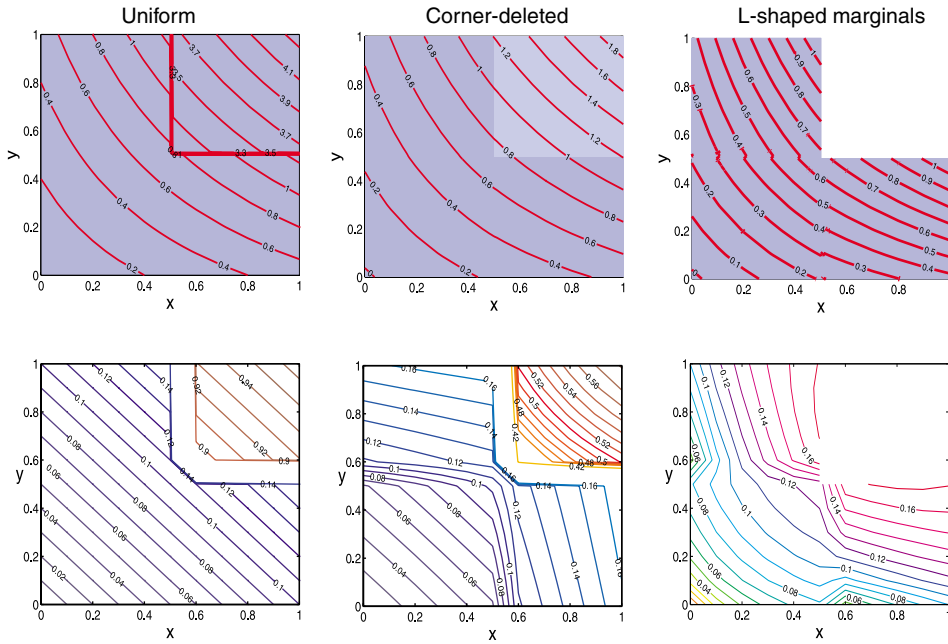


Figure 3. Bivariate effects for x and y for the function (9.1) defined on three successively more compact measures, the marginal distribution of each measure given by background shading. The bottom row gives point-wise standard errors for the effect plotted above.

such as in Figure 1 (p. 712). The bivariate effects are again recovered, apart from in the top right corner where they are left appropriately undefined.

The second row of Figure 3 plots the corresponding standard deviations of the effect estimates. Here, the standard deviation for the original functional ANOVA effect is proportional to the point-wise variability $\sqrt{\text{var}(\hat{F}(x, y, z)|x, y)}$. However, the standard deviations for the generalized ANOVAs are quite different; effects are *more* variable in regions where the density has smaller support and there are correspondingly fewer grid points, even though \hat{F} has less variation over that support.

10. DIAGNOSTICS FOR MACHINE LEARNING

Many current diagnostic tools make use of the standard functional ANOVA or some version of it. In doing this, they form product distributions at least between sets of predictors. As such, they are vulnerable to distortion by effects occurring in regions of low probability. The functional ANOVA was employed with a uniform distribution by Jiang and Owen (2003) and Roosen (1995) to provide bivariate plots of low-dimensional effects. A plot given for the effect of x_u would show

$$F_u(x_u) = \sum_{v \subseteq u} f_v(x_v),$$

the direct projection of F onto the subspace of predictors x_u . Roosen (1995) suggested

the presentation of a matrix of plots. Effects for individual predictors would be given in the diagonal elements, with elements above the diagonal providing the corresponding bivariate effect. Below the diagonals, we give the conditional variance of F at x_u . That is,

$$\int (F(x) - F_u(x_u))^2 dx_{-u}$$

which indicates the variability of F at the point x_u . This also corresponds to the sampling variability of a Monte Carlo integration and is used to create one standard error bounds around the main effect plots. We have demonstrated this presentation in Figure 5 (p. 727). We note, in particular, that if F can be expressed as $G_u(x_u) + G_{-u}(x_{-u})$, this variance should be constant. Neither constant variance, nor the correspondence between variation and sample variance are maintained for weighted functional ANOVAs. In this case it would be appropriate to display the weighted conditional variance:

$$\int w(x)(F(x) - F_u(x_u))^2 dx_{-u}.$$

in subdiagonal elements. This quantity has been employed in Figure 6.

The influence of the underlying distribution on effects has been partially recognized in machine learning diagnostics. Friedman (1999) proposed plotting the *partial dependence*, given by

$$f_u(x_u) = \int_{x_{-i}} F(x) dP_{-u}, \quad (10.1)$$

where P_{-u} is the marginal distribution of x_{-u} . P_{-u} can then be approximated by the empirical marginal, giving the data-based estimate

$$\hat{f}_u(x_u) = \frac{1}{N} \sum_{i=1}^N F(x_u, x_{i,-u}). \quad (10.2)$$

Both Friedman (1999) and Linton and Nielsen (1995) noted that this estimate recovers additive or multiplicative components up to constants.

Viewed as a projection operator, rather than as integration, this estimate corresponds to the projection with respect to the measure $P_u(x_u)P_{-u}(x_{-u})$, implicitly assuming x_u and x_{-u} to be independent. For the bivariate plots described in the introduction, the standard functional ANOVA is equivalent to the partial dependence plot, both of them being misleading. Breiman (2001) made use of a similar idea—randomly permuting the values of one of the variables—to develop measures of variable importance.

The concern of this article is that a product distribution can place potentially large probability mass in areas of extrapolation. In these areas functional behavior is dictated by the properties of the particular learning technique employed rather than the behavior of observed data. This can lead to significant distortions in the effects that are presented. Polynomial models and models using corresponding kernels can exhibit large Gibbs effects away from data. These then produce effects that exhibit larger variation than actually exists. More conservative models that tend to finite constants at infinity are also problematic

in allowing potentially interesting effects to be dampened out by large, empty regions in which the function is close to flat. Further, both Gibbs effects and some types of constant extrapolation—notably tree models—are highly variable under resampling of the data. This results in unstable and misleading plots.

11. THE CALIFORNIA HOUSING DATA

To illustrate the proposed methodology, we make use of the California Housing Data, introduced by Pace and Barry (1997) as a real-world demonstration. These data consist of the median house prices along with demographic variables in 20,640 regions of California, based on U.S. Census data. We removed 52 extreme outliers and used as attributes median income, median house age, average number of bedrooms, population, average occupancy, latitude and longitude coordinates to predict median house price given in units of \$100,000. In order to do this, we used an ϵ -support vector machine, trained on these data using the default values in the MATLAB package `libsvm` from Chang and Lin (2001). Support Vector Machines may be thought of as a kernel-regression method that optimizes a soft-thresholded absolute-value criterion. For our purposes the resulting prediction function can

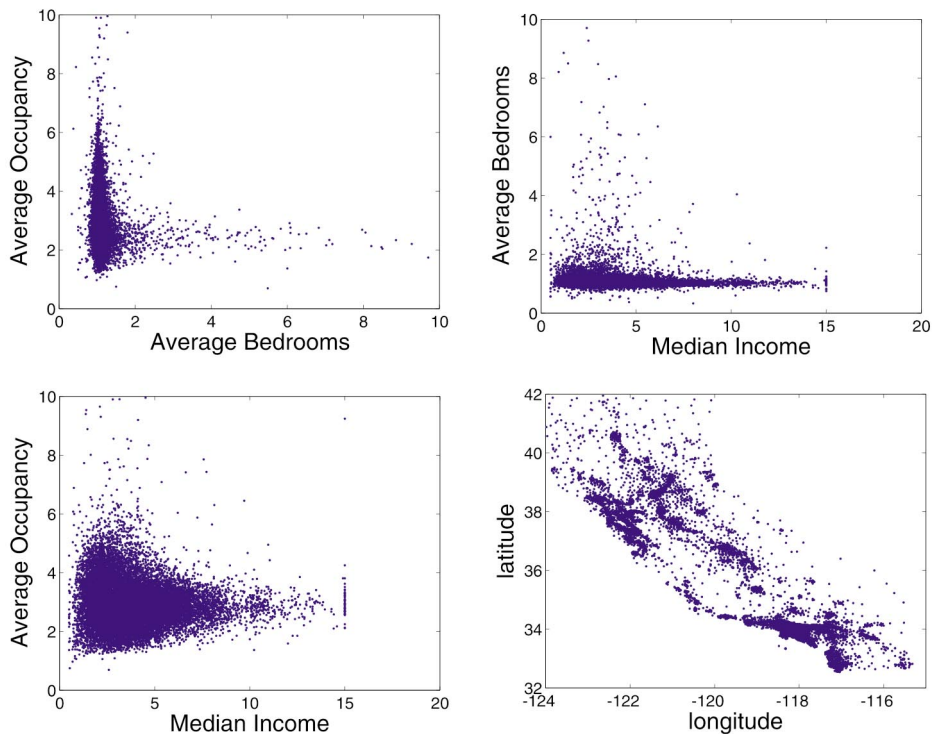


Figure 4. Bivariate distributions for the California Housing Data. Clockwise: average number of bedrooms and average occupancy, median income, and average number of bedrooms, median income, and average occupancy and latitude and longitude.

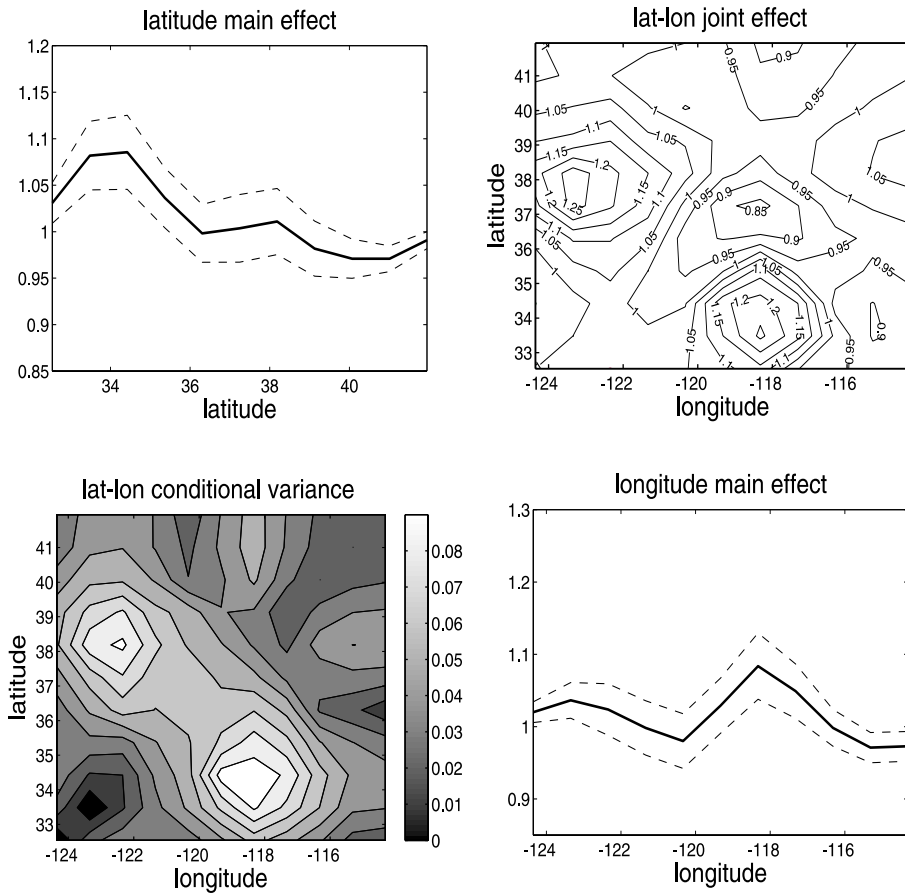


Figure 5. A matrix of plots of effects for a regression Support Vector Machine trained on the California Housing data. The effects for latitude and longitude are plotted in the diagonal elements. The upper diagonal provides the bivariate effect and the lower diagonal the functional variance conditioned on the bivariate values of the two predictors.

be thought of simply as a black box.

There are several features of the California Housing Data that make it an interesting example. The underlying distributions among some of the variables mimic the distributions used above to describe problems with the standard functional ANOVA. In particular, “L”-shaped bivariate distributions can be observed between the average number of bedrooms, median income, and average occupancy. Figure 4 demonstrates these distributions. Logically, the three variables together must have a distribution similar to that used in the final plot in Figure 3 (p. 724). We also include a scatterplot of the latitude-longitude values for these regions which we use as readily understandable predictors on which to demonstrate functional effects.

Figure 5 presents a matrix of plots created using the method of partial dependence plots described above. The Support Vector Machine used is based on a Gaussian kernel, meaning that its value tends to a constant away from data. The California Housing Data are highly

concentrated and functional ANOVA techniques based on a uniform distribution places so much emphasis on regions without data that the resulting effects are flat.

Here, both latitude and longitude show two peaks which correspond to the Los Angeles and San Francisco metropolitan areas. This is confirmed in the bivariate contour plot which also indicate a large reduction in house prices in the Central Valley. The plot of bivariate conditional variance indicates that these peaks are also the region in which the Support Vector Machine shows greatest variation. This is unsurprising; metropolitan areas have the greatest number and diversity of regions, providing more data in the remaining predictor variables for the Support Vector Machine to use.

We now employ the weighted functional ANOVA. We have used the techniques in Hooker (2004a) to provide an estimate of the density of the predictors for the California Housing Data. This method represents the density as a binary tree, taking constant values

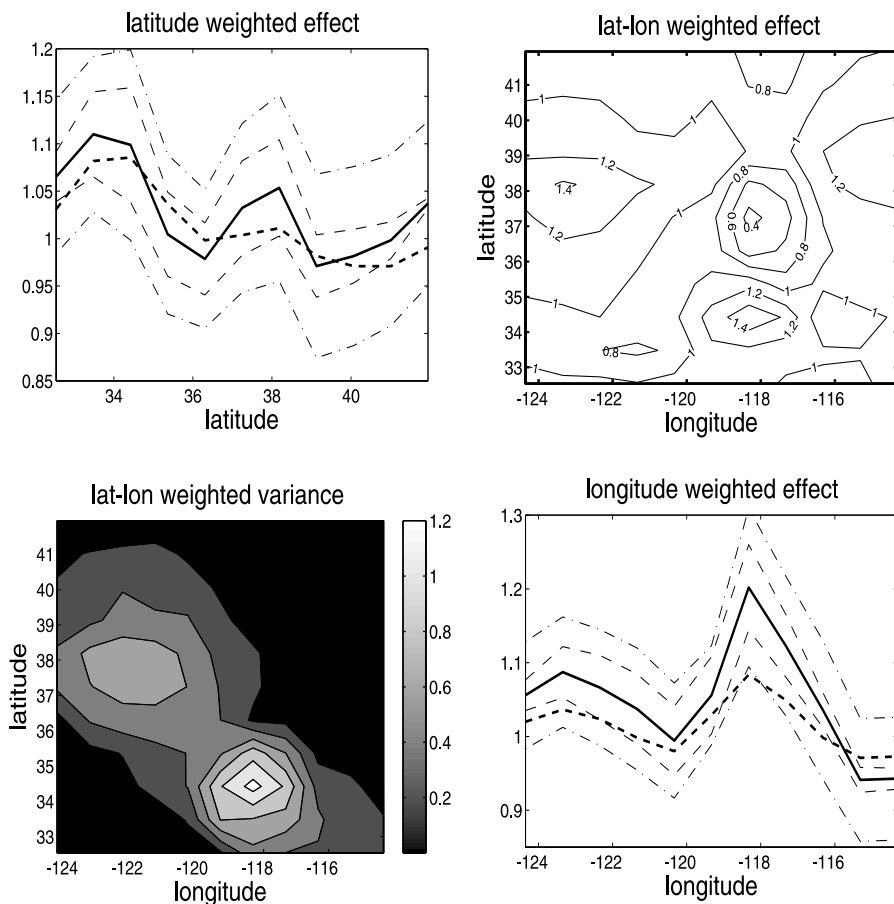


Figure 6. The effect of a regression Support Vector Machine trained on the California Housing Data for latitude and longitude. Generalized effects are given by solid lines, standard effects by dashed lines on the diagonal plots. Thin dashed lines provide bounds based on the conditional variance of the Support Vector Machine and dotted lines give one standard error bounds based on the variance of the effect estimate. The upper diagonal element presents the generalized bivariate effect and the lower the generalized conditional variance.

on nonoverlapping rectangles. It attempts to classify each point in predictor space as being either from the original data or generated from a uniform distribution. This density has then been used as a weight function in the weighted functional ANOVA and we have taken a set of 121 uniform grid points on the dimensions of interest, and 50 randomly sampled training points for the noise dimensions. The weight function has then been divided by the marginal distribution on the noise variables.

Figure 6 compares the standard and weighted effects for this function on latitude and longitude. Both effects show noticeably greater peaks and variation, suggesting that the flat regions of extrapolation that made plots based on the uniform distribution flat, have also flattened out the effect for the partial dependence plot. The plot of bivariate weighted conditional variance also indicates greater concentration of variability in regions of high data density.

12. CONCLUSION

This article has presented a new approach to diagnostics for high dimensional functions that takes into account the distribution of the input variables. This distribution can have an important influence on the presentation of low dimensional effects and scores for variable importance. In machine learning, it also allows us to avoid regions of extrapolation where learned functions are unlikely to provide reliable values.

The approach we take is developed through a weighted version of the functional ANOVA and we have proposed a novel estimation scheme to account for high-dimensional covariance among predictors. The scheme is computationally intensive, involving the solution of a large, weighted linear system. However, it can be made computationally tractable by exploiting the sparsity of the system in a straightforward manner. We have also provided approximate estimates of the variation due to sampling on a random grid. We have assumed that both the prediction and weight functions are known and fixed, but indicated ways in which known uncertainty in each may be incorporated into the estimated variability of the scheme.

A. PROOF OF LEMMA 1

Lemma 1. *The orthogonality conditions (4.2) are true over $\mathcal{L}^2(\mathbb{R}^d)$ if and only if the integral conditions*

$$\forall u \subseteq d, \forall i \in u \int f_u(x_u)w(x)dx_i dx_{-u} = 0. \quad (\text{A.1})$$

hold.

Proof: If (4.3) is true, let $i \in u \setminus v$. Then $f_v(x_v)$ is constant across x_i and x_{-u} and

$$\int f_v(x_v)f_u(x_u)w(x)dx_i dx_{-u} = f_v(x_v) \int f_u(x_u)w(x)dx_i dx_{-u} = 0.$$

Now suppose (4.3) to be false and

$$\int f_u(x_u)w(x)dx_idx_{-u} \neq 0$$

for some u, i and $x_{u \setminus i}$ on a set of positive measure. The notation $u \setminus i$ indicates $u \setminus \{i\}$ —the set u with element i removed. We can assume, moreover that (4.3) does hold for all $v \neq u$ and $j \neq i$. Then set $v = u \setminus i$ and

$$f_v = \int f_u(x_u)w(x)dx_idx_{-u}.$$

It is easy to verify that (4.3) does hold for f_v . However,

$$\langle f_u, f_v \rangle_w = \int \left(\int f_u(x_u)w(x)dx_idx_{-u} \right)^2 dx_{u \setminus i} \neq 0.$$

□

B. EXISTENCE AND UNIQUENESS

In this appendix, we consider situations under which the generalized functional ANOVA decomposition has a unique expression. Stone (1994) considered a decomposition defined for densities defined on hyper-rectangles of \mathbb{R}^d and bounded away from zero and infinity. However, somewhat less restrictive conditions may be used and will show that the estimates given by (6.2) are also unique.

For degenerate distributions—a bivariate distribution having support only on the line $x_1 = x_2$, for example—any function $F(x_1, x_2)$ can be exactly represented on the support of the distribution by $f_1(x_1) = F(x_1, x_1)$ or $f_2(x_2) = F(x_2, x_2)$. However, it is possible to show that a unique solution does exist under mild conditions on $w(x)$.

In this section we outline the conditions under which the generalization is well defined. The regularity condition that we require is the notion of grid closure:

Definition 1. A set $\Omega \subseteq \mathbb{R}^d$ is grid closed if for any $x \in \Omega$ and $u \subset d$ there exists $\{y^u \neq x\} \in \Omega$ such that $y^u_u = x_u$.

This means that for any point in Ω , we can move in each coordinate direction and find another point. For the sake of the simplicity of language, we will associate a function with its support.

Definition 2. A measure w is said to be grid closed if $\text{supp}(w)$, is grid closed.

We have used the name “grid closed” to suggest the existence of a grid in Ω for every point $x \in \Omega$. In fact, a full grid is not necessary and it is possible to define fractal sets that are grid closed but which do not contain any grid. Nonetheless, for practical purposes we will use a full grid.

Grid closure is a weak condition implied by more common regularity conditions. In particular, any open set is grid closed. This notion of directly motivates the approximation scheme in Section 6 through the following observation:

Lemma 2. Any grid or union of grids is grid closed.

Grid closure is essential to the following result

Lemma 3. *Let w be grid closed. For any $\{g_u|u \subseteq d\} \in \mathcal{L}_w^2$ with at least one $g_u \neq 0$ that satisfy the integral constraints (4.3), $\{g_u|u \subseteq d\}$ are linearly independent under the inner product defined by w .*

Proof: Suppose $\{g_u|u \subseteq d\}$ are linearly dependent and write

$$\sum_{v \subseteq d} \alpha_v g_v = 0$$

with $\alpha_v = 0$ if $g_v = 0$. Choose u such that $\alpha_u \neq 0$ and $\alpha_v = 0, \forall v \supset u$. Then $g_u = \sum_{v \not\supset u} \beta_v g_v$.

Consider any point $x = (x_u, x_{-u}) \in \Omega_w$. By grid closure, there is some other point $y = (x_u, y_{-u}) \in \Omega_w$. For any such $y \in \Omega_w$

$$g_u(x) = \sum_{v \not\supset u} \beta_v g_v(x_u, x_{-u}) = \sum_{v \not\supset u} \beta_v g_v(x_u, y_{-u}) = g_u(y)$$

and therefore for any given z_{-u} , $g_u(x_u) = \sum_{v \not\supset u} \beta_v g_v(x_u, z_{-u})$ can be written as

$$\sum_{v \not\supset u} \beta_v g_v(x_u, z_{-u}) = \sum_{v \subset u} f_v(x_v). \quad (\text{A.1})$$

for some functions $\{f_v\}_{v \subset u}$. Now, observe that (4.2) implies

$$\int g_u(x_u) f_v(x_v) w(x) dx = 0$$

for any f_v by the condition (4.3). g_u is therefore orthogonal to $\sum_{v \subset u} f_v(x_v)$ and we must have $g_u = 0$. \square

Grid closure is necessary here to ensure (A.1) holds. To demonstrate this, return to the example of a bivariate w with support only on the line $y = x$. On the support of w , $f(x) = f(y)$ for any f and we cannot insist on being able to write (A.1).

Linear independence states that each function has a unique representation in terms of lower order components. It is easy to see that the set

$$\mathcal{G} = \left\{ g : g = \sum_{v \subset d} g_v(x_v) \right\}$$

is closed in \mathcal{L}^2 under the inner product defined by w . Using the conditions (4.2), the following is now a direct corollary of the projection theorem (Luenberger 1969):

Theorem 2. *For w grid closed and $f \in \mathcal{L}_w^2$, (4.1) has a unique minimizer under the constraints (4.3).*

ACKNOWLEDGMENTS

The author would like to thank Jerry Friedman and Art Owen for thoughtful comments in the development of the material. Input from Jim Ramsay substantially improved the presentation.

[Received November 2005. Revised December 2006.]

REFERENCES

- Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.
- Chang, C.-C., and Lin, C.-J. (2001), "LIBSVM: A Library for Support Vector Machines," Software available online at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *Annals of Statistics*, 19, 1–141.
- (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232.
- Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer.
- Hegland, M. (2002), "Adaptive Sparse Grids," in *Proceedings of the 10th Biennial Computational Techniques and Applications Conference*.
- Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distributions," *Annals of Mathematical Statistics*, 19, 293–325.
- Hooker, G. (2004a), "Diagnosing Extrapolation: Tree-Based Density Estimation," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- (2004b), "Discovering ANOVA Structures in Black Box Functions," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Huang, J. Z. (1998), "Projection Estimation in Multiple Regression with Application to Functional ANOVA Models," *Annals of Statistics*, 26, 242–272.
- Jiang, T., and Owen, A. B. (2003), "Quasi-regression with shrinkage," *Mathematics and Computers in Simulation*, 62, 231–241.
- Linton, O., and Nielsen, J. P. (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, 82, 93–100.
- Luenberger, D. G. (1969), *Optimization by Vector Space Methods*, New York: Wiley.
- Pace, R. K., and Barry, R. (1997), "Sparse Spatial Autoregressions," *Statistics and Probability Letters*, 33, 291–297.
- Press, W. H., Teukolski, S. A., Vetterling, W. T., and Flannery, B. P. (1997), *Numerical Recipes in C: The Art of Scientific Computing*, New York: Cambridge University Press.
- Owen, A. B. (2003), "The Dimension Distribution and Quadrature Test Functions," *Statistica Sinica*, 13.
- Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, New York: Springer.
- Roosen, C. (1995), "Visualization and Exploration of High-Dimensional Functions Using the Functional Anova Decomposition," Ph.D. thesis, Stanford University.
- Stone, C. J. (1994), "The Use of Polynomial Splines and their Tensor Products in Multivariate Function Estimation," *Annals of Statistics*, 22, 118–171.