# Modeling with Gamuts

William D Heavlin, Google, Inc., P O Box 2846, El Granada, CA 94018

October 1, 2016

**Abstract**

Consider predicting a response using an additive model. A gamut is an additional, usually continuous variable that can segment data and any associated estimates; the underlying model varies smoothly. Examples of gamuts: (a) for disease, ages of subjects, their initial severity status, and/or cumulative exposure doses; (b) for learning, measures of cumulative experience and/or engagement; and (c) for economic activity, levels of income and/or spending. In previous work, gamuts have helped identify metric changes, detect coefficient shifts, and formulate statistical narratives.

Gamuts can be classified into four types: (1) gamuts exogenously specified and known a priori; (2) those endogenously constructed and therefore latent; (3) gamuts derived from an auxiliary model's predictions; (4) gamuts chosen to optimize a predictive model. Here we use gamuts of type (3) to parametrize model coefficients. By construction, the in-sample goodness-of-fit is always improved, so we focus on out-of-sample cross-validating methods. We also address computational issues.

*key words:* deconstruction, lowess, nonparametric regression, regression diagnostics, statistical narrative, varying-coefficient models.

## 1 Gamuts

The term *gamut* derives from Middle English. The medieval hexachordal music system has three octaves and assigns unique case-sensitive letter names to each. The lowest note of the lowest octave was called *Γ ut*. By elision, this name collapsed to *gamut*. Through synecdoche, its meaning shifted from referring to this particular lowest note to encompass the entire three-octave range, just as today we refer to the alphabet as our ABCs.

In modern usage, a gamut now implies an entire range or spectrum. The present paper uses the term gamut in this sense, as the name of a one-dimensional spectrum, the meaning of which can change from one application to another. We consider here models whose coefficients vary as a function of the gamut.

### 1.1 The Predictor-Response Model with Gamuts

Consider this $n$-observation dataset with response, $J$ predictors, and a gamut dimension: $\{(y_i,\, x_i,\, g_i) :\ i = 1, ..., n\}$. The values $y_i$ and $g_i$ are scalars, while $x_i$ is a $J$-vector. For standard reasons, our error structure assumes observations $i$ and $j$ are uncorrelated when $i \neq j : \mathbb{C}\{y_i,\, y_j | x_i,\, x_j\} = 0$. The gamut-based model asserts this semi-linear model:

$$\mathbb{E}\{y_i | x_i\} = x_i^T \beta(g_i) \tag{1}$$

In this paper, the gamut $g_i = g(x_i)$, that is, $g_i$ is the evaluation of the auxiliary function $g(\cdot)$. Alternative formulations are conceivable: Conventionally, equation (1) is modeled by including interaction terms, either among the $x$ variables or between the $x$ variables and the measured gamut $g$.

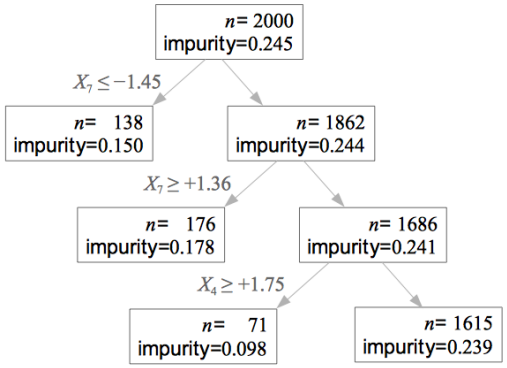


Figure 1: Example of a CART-derived tree, from Zola (2015).

## 1.2 Sources of Gamuts

By their origin, four types of gamuts seem available:

*a priori:* Some applications have a readily available dimension that functions as a gamut. For example, in medical applications, one can use treatment subjects' before-treatment disease severity. For college achievement, one can use students' pre-college SAT scores.

*exogenous composite:* Rankings of colleges, for instance, have multiple measures that are combined by suitable weights. Formally, $g(x) = x^T w$, where the weights $w$ are selected by some rationale. In the case of US News college rankings, the weights are hand-curated. In other applications, the choice of the weights $w$ more resemble the first eigenvector. Gamuts of this form are quite interesting, and subject to a companion paper.

*endogenous composite:* Such gamuts have the same formal structure as exogenous ones, in that $g(x) = x^T w$. The difference is that the weights of the endogenous form are derived directly from the dataset $\{(y_i, x_i) : i = 1, ..., n\}$. Gamuts of this type are the focus of this paper.

*jointly optimized composite:* In this formulation, two equations,

$$g(x) = x^T \gamma, \tag{2}$$

$$\mathbb{E}\{y_i | x_i\} = x_i^T \beta(g(x_i)) \tag{3}$$

are posed to hold simultaneously, and (2) and (3) are jointly optimized with respect to $(\gamma, \beta(\cdot))$ — subject to the usual regularity and smoothness requirements. For example, in predicting first-year college performance, one might take into account high school GPA, SAT scores, and financial aid package. Further, one can imagine the coefficients for GPA to be somewhat different for high-SAT students than for low-SAT students: for low-SAT students, the GPA may discriminate more strongly, because the grades might track more strongly with an underlying non-intelligence-related latent variable such as work ethic ("grit").

These four gamut types themselves comprise a spectrum. At one end, a priori gamuts are conceptually simple and easily interpreted. At the other end, jointly optimized gamuts have the best empirical fits — at the expense of both simplicity and interpretation ease. From an application point of view, the exogenous and endogenous composites strike an interesting compromise: improving model accuracy while maintaining plausible narratives.

Here we focus on the case of the endogenous composite; the exogenous case merits a separate treatment. Like its exogenous cousin, both in applications and in statistical theory the gamut approach gives rise to interesting analyses. As we shall see below, working out the details for the endogenous case has enough depth that such focus is merited.

## 1.3 Motivating Gamuts

From at least two points of view, one can intuit from statistical practice the underlying concept of gamuts.

(a) Consider a tree-based classification model such as that from Zola (2015) presented in Figure 1. The first split finds a single variable that sheers off a relatively small subset of high purity, and this is repeated with two more small cuts from the remaining majority branch. As one proceeds into the body of the data, ever more features are used for partitioning. This says that the edges of the predictive range may require or support fewer features and simpler models, while the middle tiers of the predictive range, with a greater mix of categories, may require more features and richer models.

(b) Consider the case of physical constants, such as the speed of light and Avogadro's number. These constants correspond to coefficients of particular proportional systems. It is a statement of profound physics that these coefficients are constant, that is, invariant across a wide range of experimental conditions. So if in the physical sciences the strongest and deepest statements assert constant model coefficients, then it follows that, scientifically speaking, the corresponding scientific null hypothesis should assert that coefficients vary from context to context.

Of course, argument (b) conflicts with Occam's razor, which uniformly favors simple models over more complex ones. For our purposes, it is sufficient to note the tension between these two points of view. As will unfold below, gamuts constitute a simple, one-dimensional formulation in which this tension can be contemplated without premature commitment to a constant-coefficient model.

## 1.4 Examples of Gamuts

Figure 2 (a) presents a numerator-denominator plot from Heavlin and Koslov (2008). The quantity $b_{y \cdot x} = \sum y_i x_i / \sum x_i^2$ estimates the regression coefficient of a zero-intercept model. In Figure 2 (a), the y-axis plots $\texttt{cumsum}\, y_i x_i$ versus $\texttt{cumsum}\, x_i^2$. As a consequence, the slope of the line between the origin and the last point $(\sum y_i x_i, \sum x_i^2)$ is $b_{y \cdot x}$, the overall simple regression coefficient. Were the summands of Figure 2 (a) sorted at random, the slope of the cumsum-cumsum curve would also be linear, on average, with slope approximately $b_{y \cdot x}$. However, the actual order of the summands in Figure 2 (a) sorts by $x_i$, so $x_i \leq x_{i+1}$, that is, the gamut is $g(x) \equiv x$. As a result, locally, the slope between any two points with gamut labels $g_1$ and $g_2$ on the curve in Figure 2 (a) corresponds to an estimate of $b_{y \cdot x}$ specific to that gamut interval: $\sum y_i x_i / \sum x_i^2$, where the summation takes place over the set $\{i : g_1 < g_i \leq g_2\}$. Observe that the slope of Figure 2 (a)'s curve is largest for the lower values of $x$. Heavlin and Koslov report that $y$ consists of a hard-to-measure ground truth, while $x$ is a proposed clickstream-based calculation of quality. From Figure 2 (a), we conclude that low values of $x$ strongly imply non-quality, and that going from low $x$-values to medium $x$-values corresponds to a greater improvements in quality than going from medium $x$-values to high ones.

The analysis of Figure 2 (a) embodies two concepts: (a) that a coefficient can change over the range of the feature space, and (b) that one-dimensional curves can describe these changes. Heavlin and Koslov (2008) describe generalizations of ND plots to various ratio statistics. Heavlin (2012) connects Hill's (1965) causality criteria to particular gamut choices.

Figure 2 (b), from Heavlin (2014), presents another ND plot, now corresponding to a one coefficient $b_{yx}$ in a regression model with other predictors in addition to $x$. Here, the data are stratified by the ads exposure volume; the strata are indexed by $g$. A set of coefficients is calculated for each stratum separately; in stratum $g$ denote the coefficient of interest by $b_{yx}[g]$. Plotted are the weighted cumsums of $b_{yx}[g]$, $\texttt{cumsum}\, b_{yx}[g]w[g]$; the weights, $w[g]$, which sum to 1, are proportional to reciprocals of squared standard errors. The last plotted point, with value of 0.89, is the (weighted) average coefficient, averaged over all exposure doses. By construction, the slope from the origin to the last plotted point is also 0.89.

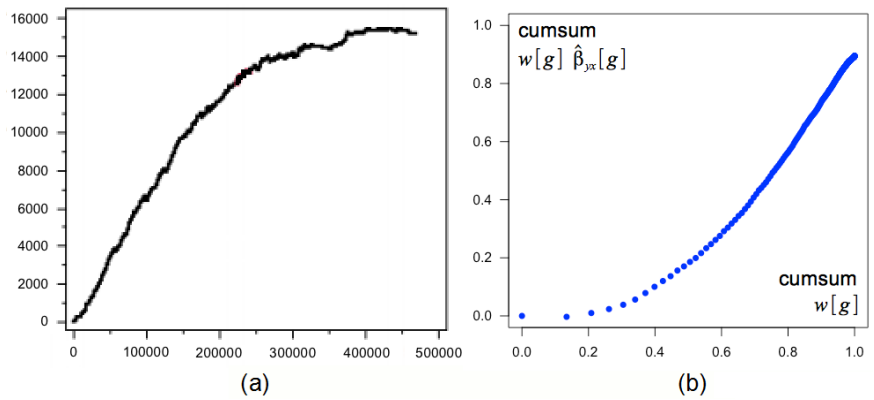


Figure 2: (a) ND plot of a simple regression coefficient, from Heavlin and Koslov (2008). The gamut consists of the regressor $x_i$. (b) ND plot of a stratified multiple regression coefficient, from Heavlin (2014). The gamut consists of the tertiary variable, exposure dose.

The J-shape in Figure 2 (b) tells us that for low values of $g$, the coefficients $b_{yx}[g]$ are near zero, while for higher values of $g$, the slopes run somewhat above 1.0. Observe the change in structure at about 12 to 17 points from the left: this corresponds to exposure levels above which users become progressively more sensitive to their quality experience. Ads exposure therefore acts as a dose, and after a certain amount of this dose, users become sensitized to the quality of their ads experience. In the terminology of the present paper, ads exposure defines a gamut, and we identify an important effect as a function of this gamut. (Note also that the gamut of exposure dose essentially counts users' learning opportunities.)

For each point plotted in Figure 2 (b), the data volume is sufficient to calculate a full multiple regression with each unique value of ads exposure dose. In that sense, the ND plot merely displays many separately calculated regression coefficients while respecting their relative precisions. A natural question arises of how to handle the case of a smaller amount of data. A related question is how to proceed when the gamut is measured on a continuous scale, and we have, unlike the integer gamut values in Figure 2 (b), no natural discrete strata. Section 2.2 offers a concrete proposal to this question: generalizing gamut-dependent model coefficients to finer-grained, continuous-scale gamuts and/or to smaller data volumes.

# 2 Primary Example

## 2.1 Prices of 505 California Homes

Our dataset consists of the selling prices of 505 homes around San Luis Obispo, California, sold in 2009 (Statcrunch, 2016, dataid=1758729). As extracted, we have each home's 2009 sales price, its community, square footage, and numbers of bed- and bathrooms. To this we supplement the 2016 median sales prices for each corresponding community by searching on zillow.com in early 2016; this allows us to measure on a continuous scale the otherwise categorical variable designating community; in what follows, we refer to the terms making use of median home price as *location*.

The 2009 sales prices, the square footages, and locations are all transformed to log scale. Table 1 presents the coefficients of an ordinary least squares fit of sales price, on the four terms of square footage, location, number of bathrooms, and number of bedrooms. The coefficients of square-footage and location, with values of 1.01 and 0.70, can be interpreted as elasticities. The out-of-sample $R^2$, resulting from a cross-validation scheme detailed in section 3.3, is 0.52.

These California home prices range from a low of $27,000 to a high of $2.8 million; the first quartile is just below $200,000, the third quartile somewhat above $400,000. One can

Table 1: Coefficients of ordinary least squares fit.

| term | estimate | std error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | -3.52 | 0.991 | −3.55 | 0.0004 |
| ln square footage | 1.01 | 0.098 | 10.32 | $< 2 \times 10^{-16}$ |
| ln median home price | 0.70 | 0.069 | 10.13 | $< 2 \times 10^{-16}$ |
| number bedrooms | −0.11 | 0.037 | −2.89 | 0.0040 |
| number bathrooms | 0.03 | 0.044 | 0.748 | 0.4558 |

reasonably ask how constant are the elasticities with respect to square footage and location. The implication can be quite practical: In California, homes constitute the primary family investment: a home remodel that enlarges the square footage by 20 percent increases a home's value by 20 percent when the elasticity of price to square footage is 1.0, but by only 13 percent when the elasticity is 0.7. The difference between a 20 percent return and one of 13 percent can be substantial, as the higher rate of return may justify financing the improvement while the lower one may not.

## 2.2 Lowess-based Gamut Model

Consider now this gamut-labeled dataset: $\{(y_i, x_i, g_i) : i = 1, ..., n\}$; the $y_i$ are all scalar responses, the $x_i$ are $J$-vectors, and the $g_i$ are the gamut scalars sorted in ascending order: that is, $g_i \leq g_{i+1}$ for all indices $i$.

For a given scalar value for the gamut $g_0$, let us form the $n$-element weight vector $w_0$ based on the closeness of $g_j$ to $g_0$. Following Cleveland (1979), consider the tricube function

$$T(x) \equiv (1 - |x|^3)^3 \times 1\{|x| < 1\}. \tag{4}$$

We define the $j$-th element of $w_0$, $w_{j0}$, by the relation $w_{j0} = T((g_j - g_0)/\tau_0)$, for some positive scalar $\tau_0$. In this paper, we take $\tau_0 = \tau(g_0)$ to be $2\times$median$\{|g_j - g_0|\}$, the median being calculated over all indices $j$. (A cross-validation-based exercise indicates that a multiplier of 1.13 is slightly better than a multiplier of 2.0.) Given $n$ such weights, the weighted least squares estimate of the coefficients solves these linear equations with respect to $b_0$ :

$$X^T D_{w_0} X \, b_0 = X^T D_{w_0} y, \tag{5}$$

where $X$ consists of $n$ rows, each of the form $(1, x_i)$, $y$ is the $n$-vector $(y_1, y_2, \ldots, y_n)$, and $D_{w_0}$ =diag$(w_{10}, w_{20}, \ldots, w_{n0})$ are the weights around gamut value $g_0$. Because the coefficients $b_0$ depend ultimately on $g_0$, we can shift the notation somewhat and denote $b_0 = b(g_0)$. Further, one can sweep the gamut scalar $g_0$ over all values $g_i$, $i = 1, 2, \ldots, n$, and obtain coefficient vectors $b(g_i)$, $i = 1, 2, \ldots, n$. $b(g_0)$ is essentially a smoothed estimate of the coefficients using the neighborhood of observations near $g_0$; this neighborhood is defined implicitly, using the lowess weight system defined by the functions $(T, \tau)$.

To summarize, once the gamut values $\{g_i : i = 1, 2, \ldots, n\}$ are set, one can calculate coefficients $b(g_i)$ for all values $g_i$. In the present work, we take $g_i$ to be the fitted values from the Table 1 model.

Figure 3 draws the corresponding coefficient estimates $b(g_i)$ on the y-axis versus gamut values $g_i$ on the x-axis. The most visible feature is the shift in structure about 40 percent across the gamut range: The elasticity with respect to square footage rises above 1.0, while that with respect to location falls from about 0.7 to about 0.6. The coefficients with respect to bath- and bedroom counts similarly shift: for the higher gamut values, bathroom count offer some incremental value, while enclosing additional space into bedrooms imposes a slightly negative effect on price.

As one might expect, the corresponding out-of-sample $R^2$ rises from 0.52 to 0.68.

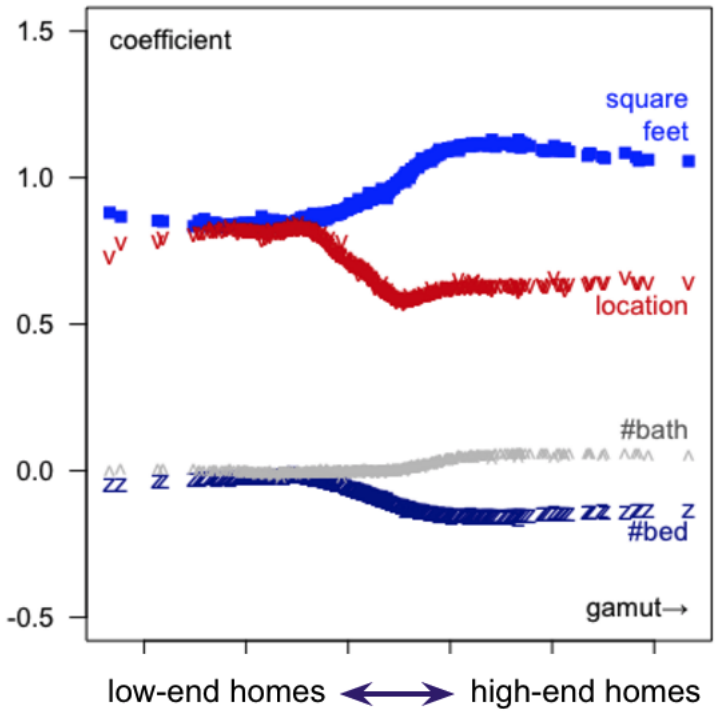


Figure 3: Gamut-based regression coefficients for the 505 California homes, as described in section 2.2.

The remainder of this paper deals with the statistical issues raised by this gamut-based calculation. Before we turn to that discussion, we note the two apparent virtues of gamuts: (a) The goodness-of-fit increases. This seems quite plausible, since gamut-based models have rather more parameters. (b) The narrative associated with the model becomes richer: For example, rather than interpret the elasticity of price with respect to square footage, Figure 3 suggests two domains, lower- and higher-end homes. The higher-end homes have a higher elasticity with respect to changes in square footage. Were this otherwise, the prospect of a constant elasticity with respect to square footage would be profound—suggesting comparisons to other coefficients observed to be constant over a wide range, analogous to the constancy of physical constants as discussed in section 1.3 (b).

# 3 Statistical Issues with Gamuts

## 3.1 Related Literature

Gamut-based methods fall under the general topic of nonparametric regression. Anscombe and Tukey (1963) plot residuals of a linear model against their corresponding fitted values. They note that visible curvature in this plot can often be mitigated by transforming the response, for instance, by taking logarithms or a related power transformation. In modern terminology, Anscombe and Tukey assess goodness of fit relative to the estimated mean function.

Cleveland (1979) introduces local weights for smoothing scatterplots. His proposed lowess weight system includes the tricube function, and the practice of smoothing with respect to robust scale estimates. Cleveland and Devlin (1988) directly generalize the lowess algorithm to the case of multiple predictors; their weight function operates directly over the covariate (feature) space in $\mathbb{R}^J$. Friedman and Stuetzle (1981) offer another generalization,

projection pursuit regression. In this scheme, a given direction in the covariate space, $X\gamma$ say, is lowess smoothed; $\gamma$ is chosen to maximize goodness of fit; the fitting process was iterated over the residuals from previous fits. Wahba (1978) establish spline-based methods. Friedman (1991) proposes an adaptive, spline-based, continuous modeling approach that dispenses with the single-direction, take-residuals-and-repeat approach of projection pursuit regression. Hastie and Tibshirani (1993) note that a wide range of predictive approaches can be expressed as varying-coefficient models. Their work relies on posing rich spaces of potential covariates and exploiting the L1-lasso technology to select that feature subsets well supported by the data. Out of the machine learning tradition, Garcia and Gupta (2009) develop elegant prediction algorithms that consist of additive, usually monotone surfaces over regular grids of spline-like knots. Their applications fit well while preserving interpretability.

Against this background, gamut-based approaches postulate interactions of each term with the gamut direction. In the case of the endogenous gamut considered here, the gamut direction is a coarse estimate of the mean function. In doing so, it preserves the interpretability not only of the predictions, but of the coefficients themselves. In this regard, gamuts share the underlying visual esthetic of Anscombe and Tukey (1963), Cleveland (1979), and Friedman and Stuetzle (1981).

## 3.2 Regularization and Overfitting

To avoid overfitting, nonparametric regression methods have two strategies. One is to modify the objective function, so that models with more parameters are suitably penalized. The second is to divide the data into training and test sets, and evaluate goodness of fit on tests sets. Ridge regression (Hoerl and Kennard, 1970) and, later, the lasso (Tibshirani, 1996) are examples of the former strategy, while Wahba (1978), Breiman et al. (1984), and Friedman and Stuetzle (1981) use the latter extensively.

## 3.3 Confidence Intervals

For gamuts, the training-test strategy can be computationally convenient. Among other reasons, test sets are easily represented by injecting zero weights into the gamut calculations. In particular, we find the strategy of cross-validation well suited. Cross-validation methods are widely used to assess goodness of fit by calculating directly the lack of fit in the out-of-sample test set. By systematically assigning all observations to some training sets and to some test sets, the estimates of cross-validation come to resemble leave-out-$k$ jackknife samples.

Denote the set of all $n$ observations by $\mathbb{S}$, and that leaving out subsample $j$ by $\mathbb{S}_{-j}$. Further, denote the respective sizes of sets $\mathbb{S}$ and $\mathbb{S}_{-j}$ by $n$ and $n_{-j}$, respectively. Consider an estimator $G$ that operates on sets like $\mathbb{S}$ and $\mathbb{S}_{-j}$. Let us denote the variance operator by $\mathbb{V}$. By elementary methods, and assuming $\mathbb{S}_{-j}$ is the result of random without-replacement subsampling from $\mathbb{S}$, one can show that $\mathbb{V}\{G(\mathbb{S}_{-j}) - G(\mathbb{S})\} = \mathbb{V}\{G(\mathbb{S})\} \times (n - n_{-j})/n_{-j}$. Rearranging terms, we arrive at this result:

$$\mathbb{V}\{\sqrt{\frac{n_{-j}}{n - n_{-j}}} \times (G(\mathbb{S}_{-j}) - G(\mathbb{S}))\} = \mathbb{V}\{G(\mathbb{S})\}. \tag{6}$$

The righthand side of the latter expression is our quantity of interest, the squared standard error of our estimate using the whole dataset. The lefthand side is something we can calculate from our calculated set of cross-validated estimates $\{(G(\mathbb{S}_{-j}), n_{-j}) : \forall j\}$: how close are our cross-validation sample estimates to the estimate using all data $\mathbb{S}$. Note that the square-root factor tells us how to adjust the observed differences for the relative size of in-sample (training) data $\mathbb{S}_{-j}$ of cardinality $n_{-j}$ and the out-of-sample data, of cardinality $n - n_{-j}$. The case of equal-sized training and test samples can be computationally convenient, because in this case the square-root factor becomes exactly 1.

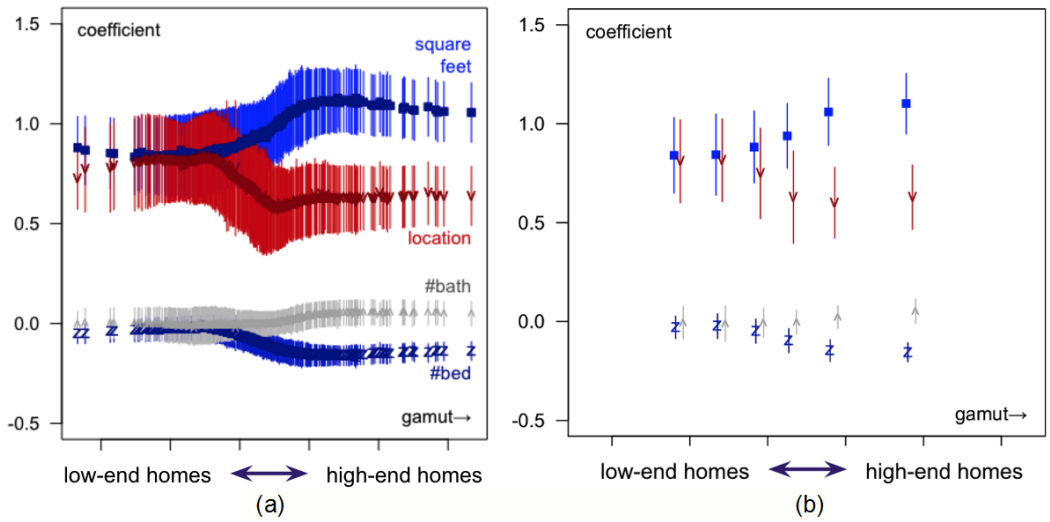


Figure 4: (a) Confidence intervals for the gamut-based coefficients, based on the 101 cross-validation subsamples, as described in section 3.3. (b) For $k = 6$ knots, the gamut coefficients and their $\pm 2$ standard error bars, as described in section 4.1.

Table 2: The efficiencies of the four coefficients relative to those of the constant-coefficient model.

| Label | relative efficiency |
|---|---|
| ln square footage | 0.68 |
| ln median house price | 0.55 |
| number bedrooms | 0.57 |
| num bathrooms | 0.84 |

In the application of 505 homes, we assign each observation to one and only one of 101 sets of 5 observations each. Each of these in turn can act as a test sample, and the corresponding training sample consists of the remaining 500 observations. For each such training sample, which we might designate as $\mathbb{S}_{-j}$, $j = 1, 2, \ldots, 101$, we calculate the gamut coefficient curves $b_{-j}(\cdot)$. These estimates are available for each observation $i$, $i = 1, 2, \ldots, 505$, and for each $i$ we have 101 coefficient estimates. We calculate the standard error of these coefficients, using the multiplier $\sqrt{500/5}$=10. Figure 4 (a) draws the $\pm 2$ standard error (and pointwise) confidence intervals. The plausibility of separate coefficients for low- and high-end homes is generally reinforced.

## 3.4 Statistical Efficiency

This cross-validation framework also allows us to compare the relative efficiency of gamut-based models to, say, a constant-coefficient model if one reduces the standard errors of the gamut coefficients to a typical value. Likewise, the same estimation pipeline that generates gamut curves can be constrained so that all weights $w_{ij}$ are equal, and the standard errors of the now-constant coefficient recorded.

The results are quantified in Table 2. The gamut model has wider standard errors by a factor of about 1.3×, or 60 percent relative efficiency. This implies that each gamut coefficient is typically estimated from the equivalent of a dataset with 303 observations.

# 4 Computational Issues with Gamuts

## 4.1 Memory Compression

Contemporary large-scale computing environments consist of a distributed file system (Ghemawat, et al, 2003) and a MapReduce programming model (Dean and Ghemawat, 2004). Compatible algorithms require less-than-linear memory scaling and a parallel computing strategy.

For gamuts, the memory requirement appears particularly pressing: at first blush, gamuts seem to require a set of coefficients for every observation, or at least a set of coefficients for each unique gamut value. However, the most straightforward of memory compression schemes works well: Suppose we constrain the space of coefficients to have not dimension $n \times J$ but rather dimension $k \times J$, where $k$ is some modest number like 5 or 6. This allows us to estimate a gamut curve that is constant below the $100/(1+k)$ percentile, constant above the $100k/(1+k)$ percentile, and by linear interpolation in between. By analogy with spline fitting, we can refer to these particular $k$ gamut values as *knots.*

Figure 4 (b) displays the estimated gamut coefficients, with standard errors, for the 6 knots. As one might expect, the smaller number of parameter estimates imparts some virtues of parsimonious models: out-of-sample $R^2$ increases slightly from 0.68 to 0.71. The coefficient-specific efficiencies relative to the constant-coefficient model also improve, from roughly 0.60 to about 0.65 — equivalent to an effective sample size $n$ of about 330 out of 505. And, of course, the number of estimated parameters is reduced from $505 \times 4 = 2020$ to $6 \times 4 = 24$.

## 4.2 Parallel Computing

Gamuts enable a particular strategy for parallel computation, based on sorting the data with respect to the gamut. Very large datasets are typically partitioned into multiple physical disk files called shards. Suppose we reshard by gamut value: we rewrite the shards to a new file system such that each shard has approximately the same gamut value. If a file system consists of $Q = 100$ shards, each shard implicitly labels a particular gamut percentile. This allows us to build a gamut-local estimate by reading one shard and, say, its two gamut neighbors. This corresponds to each shard being read by three non-dependent processes. For file systems that restrict reading to one process at the time, this implies a read throughput of around $Q/3$. In this sense, when resharding is feasible, gamuts can facilitate parallel processing. For model serving, one can calculate the equivalent of one knot for every shard, achieving a parameter count of $Q \times J$, small enough to conform to the requirements of most memory-resident models.

This strategy of sharding by gamut value is generic and applicable across predictive modeling approaches. Further, among nonparametric regression approaches, only gamuts support this property of shardability.

# 5 Summary

The principal aim of ND plots is to deconstruct ratio estimates; they use gamuts to segment the underlying data meaningfully. The endogenous gamut-based models described here preserve the rich data-driven narratives of ND plots while using their underlying principles to improve model goodness of fit. In the author's experience, goodness of fit is most improved in relatively low-dimensional, non-sparse feature spaces.

In common with most nonparametric regression methods, we rely on cross-validation methods to calibrate our model fitting. We exploit the ready availability of cross-validation's without-replacement sub-samples, and their estimates, to supply appropriate standard errors. Further, we sketch two-part computational strategy to conform with the contemporary sharded MapReduce computing environment; gamuts have some synergy with file sharding.

The example described here is compelling: gamuts identify a structural change to an economic pricing model, one quite plausible in economic theory but not easily seen by standard modeling methods. Thus, Figure 3 suggests a rather rich explanatory narrative. This narrative remains a primary virtue of gamut-based methods.

We have deliberately not addressed the case of exogenous gamuts raised in section 1.2. This case is deeply interesting, and the topic of a parallel effort.

# References


[1] Anscombe, F J and Tukey, J W (1963). "The examination and analysis of residuals," *Technometrics,* 5: 141-160.

[2] Breiman, L, Friedman, J H, Olshen, R A, and Stone, C J (1984). *Classification and Regression Trees,* Chapman and Hall, 358 pp.

[3] Cleveland, W S (1979). "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association,* 74: 829-836.

[4] Cleveland, W S and Devlin, S J (1988). "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association,* 83: 596-610.

[5] Dean, J and Ghemawat, S (2004). "MapReduce: simplified data processing on large clusters," *Sixth Symposium on Operating System Design and Implementation.*

[6] Friedman, J H and Stuetzle, W (1981). "Projection pursuit regression," *Journal of the American Statistical Association,* 76: 817-823.

[7] Friedman, J H (1991), "Multivariate adaptive regression splines," *Annals of Statistics,* 19: 1-67.

[8] Ghemawat, S, Gobioff, H, and Leung, S-T (2003). "The Google file system," *19th ACM Symposium on Operating Systems Principles.*

[9] Garcia, E and Gupta, M (2009). "Lattice regression," *Advances in Neural Information Processing Systems,* 592-602.

[10] Hastie, T and Tibshirani, R (1993). "Varying-coefficient models (with discussion)," *Journal of the Royal Statistical Society, Series B,* 55: 757-796.

[11] Hill, A B (1965). "The environment and disease: association or causation?" *Proceedings of the Royal Society of Medicine,* 58: 295-300.

[12] Heavlin, W D (2014). "Deconstruction of effects by exposure dose," *ASA Proceedings*.

[13] Heavlin, W D (2012). "Model deconstruction and Hill causality," Presentation at Joint Statistical Meetings.

[14] Heavlin, W D and Koslov, J (2008). "Numerator-denominator plots, *ASA Proceedings*.

[15] Hoerl, A E and Kennard, R W (1970). "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics,* 12: 55-67.

[16] Statcrunch (2016). "California Home Prices, 2009," https://www.statcrunch.com/app/index.php?dataid =1758729, accessed March 2016.

[17] Tibshirani, R (1996). "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B,* 58: 267-288.

[18] Wahba, G (1978). "Improper priors, spline smoothing, and the problem of guarding against model errors in regression," *Journal of the Royal Statistical Society, Series B,* 40: 364-372.

[19] Zola, M (2015). "Ensemble methods for gradient boosted trees." http://www.add-for.com/blog/ 2015/10/12/ensemble-methods-gbrt, accessed March 2016.