THE VIRTUE OF TRANSPARENCY:

HOW TO MAXIMIZE THE UTILITY OF DATA WITHOUT OVERFITTING

THIS VERSION: July 22, 2024

Megan Czasonis, Mark Kritzman, and David Turkington

Megan Czasonis is a managing director at State Street Associates in Cambridge, MA.
mczasonis@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Mark Kritzman is the chief executive officer at Windham Capital Management in Boston, MA, and a senior lecturer at the MIT Sloan School of Management in Cambridge, MA.
kritzman@mit.edu
100 Main Street, Cambridge MA, 02142

David Turkington is senior managing director and head of State Street Associates in Cambridge, MA.
dturkington@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Key Takeaways

- The key challenge to prediction is to maximize the utility of available data without detrimentally overfitting the data.
- One approach for doing so is to expand the prediction model by applying non-linear functions to transform the original predictive variables into a much larger set of variables, and then to reduce the dimensionality of the expanded set of variables by applying regularized regression.
- An alternative solution for maximizing the utility of data is to curate the observations and predictive variables for each individual prediction task using a model-free technique called relevance-based prediction.

Abstract

The key challenge to prediction is to maximize the utility of data without overfitting the observed data to the detriment of future predictions. The authors describe two techniques for addressing this challenge. The first technique, proposed by Kelly, Malamud, and Zhou (2024), is referred to as a high-complexity model (HCM). This approach applies non-linear functions to transform the original predictive variables into a much larger set of variables. It then applies regularized regression to guard against overfitting. The second technique, called relevance-based prediction (RBP), described by Czasonis, Kritzman, and Turkington (2020, 2022a, 2022b, and 2023), is a model-free approach that identifies the optimal combination of observations and predictive variables for each individual prediction task. Whereas a single high complexity model uses many variables to address all prediction tasks, relevance-based prediction relies on relatively few variables but considers many combinations of observations and variables to address all prediction tasks. The authors compare these alternative techniques.

THE VIRTUE OF TRANSPARENCY:

HOW TO MAXIMIZE THE UTILITY OF DATA WITHOUT OVERFITTING

Formal data-based prediction originated circa 1795 when Carl Friedrich Gauss introduced linear regression analysis to predict astronomical motion. This technique survives today as the most widely used approach for data-based prediction. However, there are many prediction tasks that involve complicated relationships between the predictive variables and outcomes which demand more sophisticated methods than linear regression analysis. These complicated relationships are driven by conditionalities in which the nature of the relationship shifts as conditions change. The key challenge to prediction is to extract as much information as possible from a sample of data that potentially has many conditional relationships in a way that does not overfit the observed data and thereby harm the effectiveness of future predictions.

It is generally assumed from principles of classical statistics that overfitting arises when the number of predictive variables is too large relative to the number of observations. Kelly, Malamud, and Zhou (2024) describe and illustrate a technique for maximizing the utility of data that persuasively belies the common view that proliferation of predictive variables leads to overfitting. Their approach, referred to as a high-complexity model, hereafter referred to as HCM, uses nonlinear functions to transform the original predictive variables into a much larger set of variables and then applies regularized regression to guard against overfitting.[1] The premise of this approach is twofold: if enough transformations are included, the expanded model will capture every conditional relationship in the data sample; and by applying regularized regression, the model will only rely on a smaller set of useful transformations.

Czasonis, Kritzman, and Turkington (2020, 2022a, 2022b, and 2023) propose an alternative solution for maximizing the utility of data without overfitting. Rather than build a single comprehensive model of many variables to address all prediction tasks, their approach, called relevance-based prediction and hereafter referred to as RBP, curates the observations and predictive variables for each individual prediction task. RBP therefore uses many different model-free prediction routines all composed from relatively few variables in contrast to Kelly, Malamud, and Zhou who use a single model composed of many variables.

We proceed as follows. We first provide detailed descriptions of the HCM and RBP approaches, offering a synthesized view of multiple articles on both topics along with further extensions and perspectives. Then we construct a simple illustration using a contrived set of HCM nonlinear transformations to show in a transparent way how these different techniques render similar predictions by applying similar weights to observations. We then extend this illustration to demonstrate that the near equivalence of these results holds when we expand the number of predictive variables to a much larger set of randomly produced nonlinear variable transformations. We then conduct a simulation to expand our understanding of these different techniques. We conclude with a summary.

High-Complexity Models

As described by Kelly, Malamud, and Zhou (2024), HCMs use non-linear functions to transform a model's original set of predictive variables into a much larger set to extract as much information as possible from the available data. HCMs then apply linear regression analysis along with

regularization techniques to reduce the model's dimensionality to guard against overfitting. Let us investigate this process in more detail within the context of linear regression analysis.

Suppose we wish to predict an unknown value of y_t using any given values of an M-dimensional row vector x_t. To inform our predictions, we observe N historical outcomes y_i which we stack into a column vector, Y, and we observe N corresponding row vectors of x_i, consisting of M predictive variables which we stack into a matrix X. Without any loss of generality, let us assume that all variables have been recentered to have average values of zero: $1'_N Y = 0$ and $1'_N X = 0$.

A classical linear regression model using ordinary least squares gives the following prediction.

$$\hat{y}_{t,linear} = x_t (X'X)^{-1} X'Y \quad (1)$$

In Equation 1, the operator $'$ denotes matrix inverse. The solution is identical if we transform X into normalized z-scores by dividing each variable by its standard deviation: $Z = X\Omega_{diag}^{-1/2}$ and $z_t = x_t \Omega_{diag}^{-1/2}$, where $\Omega = X'X(N-1)^{-1}$ is the covariance matrix of X and Ω_{diag} contains the diagonal elements of Ω with zeros elsewhere.

$$\hat{y}_{t,linear} = z_t (Z'Z)^{-1} Z'Y \quad (2)$$

We can write the same expression in terms of the correlation matrix $P = Z'Z(N-1)^{-1}$ in which case the entire expression is divided by $N-1$.

$$\hat{y}_{t,linear} = z_t P^{-1} Z'Y (N-1)^{-1} \quad (3)$$

Further, because correlations are symmetric, we can carry out principal components analysis to decompose P into eigenvectors V (columns) and eigenvalues D (diagonal entries in a square matrix).

$$P = VDV' \tag{4}$$

Expanding the inverse P^{-1}, which only requires inverting the diagonal matrix D, we have the following equivalent expression.

$$\hat{y}_{t,linear} = z_t(VD^{-1}V')Z'Y(N-1)^{-1} \tag{5}$$

These transformations are equivalent to performing linear regression analysis on the principal component transformations of Z: $\Pi = ZV$ and $\pi_t = z_t V$. We obtain the same linear regression solution because principal components are linear transformations that retain all the linear information in the data. The covariance matrix of the predictive variables Π is equal to D, a diagonal matrix of the variances (eigenvalues) of the uncorrelated principal components.

$$\hat{y}_{t,linear} = \pi_t D^{-1}\Pi'Y(N-1)^{-1} \tag{6}$$

To summarize, we perform two linear multiplicative transformations on X, first to normalize the variables, and second to project them as values on principal component vectors. Therefore, instead of performing linear regression analysis on X, we perform it on a new set of the same number of variables $\Pi = X\Omega_{diag}^{-1/2}V$ and $\pi_t = x_t\Omega_{diag}^{-1/2}V$. Linear regression predictions $\hat{y}_{t,linear}$ are invariant to this transformation.

Representing observations in terms of principal component transformations Π has useful properties:

1. It retains all the linear information in Z in as many or fewer variables because linear regression predictions are invariant to principal component transformation.
2. The principal component variables do not share any linear information because they are uncorrelated.
3. The principal components can be organized in order of their linear information content.

In cases with more variables than observations ($M > N$), the principal components compress all the linear information in centered zero-average variables Z into a maximum of $N - 1$ variables, which is the maximum amount of linear information that N centered observations can contain.[2] The eigenvalues of any additional principal components are equal to zero. The transformation is now $\Pi_{sub} = X\Omega_{diag}^{-1/2}V_{sub}$ and $\pi_{sub,t} = x_t\Omega_{diag}^{-1/2}V_{sub}$ where V_{sub} includes only the eigenvectors corresponding to nonzero eigenvalues. We can express this solution in terms of the truncated diagonal matrix of eigenvalues, D_{sub}.

$$\hat{y}_{t,linear} = z_t V_{sub} D_{sub}^{-1} V_{sub}' Z'Y \tag{7}$$

The same expression can be written in terms of the full eigenvector transformation but with the Moore-Penrose pseudo inverse D^+ in place of the traditional matrix inverse D^{-1}. The Moore-Penrose pseudo inverse in this setting simply computes the reciprocal of the nonzero diagonal eigenvalues but leaves zero values as zero, achieving the same effect as truncation.

$$\hat{y}_{t,linear} = z_t V D^+ V' Z'Y \tag{8}$$

This information compression is useful because we can perform standard linear regression analysis on the compressed subset of transformed Π variables, which is not possible on the original set of more than N variables. The linear regression solution obtained this way

provides equivalent predictions to the method of ridgeless regression (see Hastie, Montenari, Rosset, and Tibshirani, 2022). Ridgeless regression performs the same information compression for cases where $M > N$ by invoking ridge regression, which adds a penalty $\zeta\|\beta\|_2^2 = \zeta\beta'\beta$ to the squared error loss function of OLS corresponding to the solution $\hat{y}_{t,ridge}(\zeta) = z_t(Z'Z + \zeta I)^{-1}Z'Y$, and uses it to approach the unpenalized solution, which does not inherently have a unique solution, by taking the limit as $\zeta \to 0$ from above. We can think of this solution as retaining all the linear information in Z.

Now consider a case in which Z contains significantly fewer variables than observations ($M \ll N$). Linear regression analysis is limited in its capacity to explain Y. Instead of using the variables in Z, HCMs generate nonlinear transformations of them for use in a new expanded linear regression. Each transformation of the observation z_i forms a new variable $s_{ik} = f_k(z_i)$. The transformation functions f_k could take any form, mapping z_i to a real number. Though there are an infinite number of possible functions, there is a limit to how much information they can collectively convey about Z, because Z does not contain infinite information. To put it differently, many or most of the f_k transformations will contain overlapping information about the data. HCMs could use principal component transformations, as discussed earlier, to condense the linear information in the transformed variables S into a subset of principal components which we call Q. No matter how many nonlinear transformations we use, there is a maximum of $N - 1$ condensed Q variables. For intuition on the overlapping information, consider an extreme case with just two observations. Every function will either generate a higher value for observation 1, a higher value for observation 2, or the same value for both. Regardless of what the values are, any functions that have higher values for observation 1 are

linearly redundant because they are perfectly correlated. There are only a few distinct ways to differentiate between two observations. The same intuition applies to larger N.

To summarize the foregoing discussion, HCMs expand a prediction model with many nonlinear transformations of the original predictive variables to extract as much information as possible from a sample of data that may contain many conditional relationships. HCMs then reduce the dimensionality of the covariance matrix by applying ridgeless regression or equivalently by applying principal components analysis to guard against overfitting the data. Kelly, Malamud, and Zhou also consider the use of ridge regression on S with a nonzero regularization parameter ζ to further mitigate the risk of overfitting, which is similar but not identical to retaining a subset Q_{sub} of principal components of S. In addition to the method and degree of regularization, HCM predictions also depend on the method of nonlinear transformation and the number of random transformations. In practice the optimal calibration choices are not known in advance but may be chosen using cross-validation routines that evaluate average efficacy on holdout samples synthesized from the observed data.

Next, we describe RBP, which also seeks to maximize the utility of data without overfitting but in a very different way.

Relevance-Based Prediction

As described by Czasonis, Kritzman, and Turkington (2020, 2022a, 2022b, and 2023), RBP is a model-free prediction technique that forms a prediction as a relevance-weighted average of observed outcomes in which relevance has a precise statistical meaning.[3] Although RBP gives

the same prediction as linear regression analysis if it is applied across all observations, it usually gives a more reliable prediction if it is applied to a subset of relevant observations. When RBP is applied to a subset of relevant observations, it is called partial sample regression. RBP also depends crucially on fit, which measures the average alignment of relevance and outcomes across all pairs of observations that go into a prediction task. Fit assesses the expected reliability of individual predictions before they are rendered. The final feature of RBP is grid prediction which forms a composite prediction as a reliability-weighted average of many predictions given by different combinations of observations and predictive variables.

Relevance

Relevance is a statistical measure of the importance of an observation to forming a prediction given a chosen set of predictive variables. It is composed of two components, similarity and informativeness.

$$r_{it} = sim(x_i, x_t) + \frac{1}{2}\big(info(x_i, \bar{x}) + info(x_t, \bar{x})\big) \quad (9)$$

If a prediction is formed from a single predictive X variable, which we may call A, similarity and informativeness are measured as squared z-scores.

$$sim(x_{iA}, x_{tA}) = -\frac{1}{2}(x_{iA} - x_{tA})^2/\sigma_{x_A}^2 \quad (10)$$

$$info(x_{iA}, \bar{x}_A) = (x_{iA} - \bar{x}_A)^2/\sigma_{x_A}^2 \quad (11)$$

$$info(x_{tA}, \bar{x}_A) = (x_{tA} - \bar{x}_A)^2/\sigma_{x_A}^2 \quad (12)$$

In these equations, x_{iA} is the value of the predictive variable A for observation i, x_{tA} is the value of the variable for a chosen prediction circumstance, $\bar{x}_A$ is the average of all the observations of variable A, and σ_{x_A} is the standard deviation of all the observations of A.

If we instead form a prediction from more than a single predictive variable, we must use the more general Mahalanobis distance[4] to measure multivariate similarity and informativeness.

$$sim(x_i, x_t) = -\frac{1}{2}(x_i - x_t)\Omega^{-1}(x_i - x_t)' \quad (13)$$

$$info(x_i, \bar{x}) = (x_i - \bar{x})\Omega^{-1}(x_i - \bar{x})' \quad (14)$$

$$info(x_t, \bar{x}) = (x_t - \bar{x})\Omega^{-1}(x_t - \bar{x})' \quad (15)$$

In these equations, x_i is a vector of the values of M predictive variables for a prior observation, x_t is a vector of the values of the predictive variables for a specific prediction task, $\bar{x} = 1_N 1_N' X N^{-1}$ is the average of the predictive variables across all observations, and Ω^{-1} is the inverse covariance matrix of all the observations of the variables. The vector $(x_i - x_t)$ measures how distant the observations are independently from the circumstances of the prediction task. By multiplying this vector by the inverse of the covariance matrix, we capture the interaction of the predictive variables, and at the same time we standardize the distances by dividing by variance. By multiplying this product by the transpose of the vector $(x_i - x_t)$ we consolidate the outcome into a single number.

Notice that for our measure of similarity we multiply by negative $1/2$. The negative sign converts a measure of distance into a measure of similarity. We multiply by $1/2$ because the average distance between observations is twice as large as the observations' distances from the

average of all observations, which means that pairwise comparisons, including the comparison of a prior observation to the prediction circumstance, are measured in units that are twice as large as distances from average. When we measure informativeness, we retain its positive sign, and we have no need to multiply by $1/2$. By measuring informativeness as a difference from average, we are claiming that unusual observations contain more information than common observations, which follows from Claude Shannon's information theory.[5] Finally, note that we also measure the unusualness of the current observation. We do so to center our measure of relevance on zero, $\sum_{i=1}^{N} r_{it} = 0$. All else being equal, observations that are like current circumstances but different from average circumstances are more relevant than those that are not.

This definition of relevance is not arbitrary. We know from the Central Limit Theorem that the relative likelihood of an observation from a multivariate normal distribution is proportional to the exponential of a Mahalanobis distance. We also know from information theory that the information contained in an observation is the negative logarithm of its likelihood. Therefore, the information contained in a point on a univariate or multivariate normal distribution is proportional to a Mahalanobis distance.

We can also justify the non-arbitrariness of relevance by considering a limiting case of the predictions it yields. RBP forms a prediction as a weighted average of prior outcomes for Y.

$$\hat{y}_t = \sum_{i=1}^{N} w_{it} y_i \qquad (16)$$

If we define weights in terms of relevance as follows, which admits the relevance-weighted average of every prior outcome in the observed data sample, the result is precisely equivalent to the prediction that results from linear regression analysis.[6]

$$w_{it,linear} = \frac{1}{N} + \frac{1}{N-1} r_{it} \tag{17}$$

Owing to this equivalence, the theoretical justification given by Gauss for linear regression analysis applies as well to RBP. In most cases, however, we can produce a more reliable prediction by taking a relevance-weighted average of a subset of relevant observations, which is called partial sample regression.

Partial Sample Regression

Partial sample regression censors the influence of observations that are less relevant than a chosen threshold, which leads to the following definition of prediction weights.

$$w_{it,psr} = \frac{1}{N} + \frac{\lambda^2}{n-1} (\delta(r_{it}) r_{it} - \varphi \bar{r}_{sub}) \tag{18}$$

$$\delta(r_{it}) = \begin{cases} 1 & if\ r_{it} \geq r^* \\ 0 & if\ r_{it} < r^* \end{cases} \tag{19}$$

$$\lambda^2 = \frac{\sigma^2_{r,full}}{\sigma^2_{r,partial}} = \frac{\frac{1}{N-1}\sum_{i=1}^{N} r_{it}^2}{\frac{1}{n-1}\sum_{i=1}^{N} \delta(r_{it}) r_{it}^2} \tag{20}$$

In Equations 18 through 20, $n = \sum_{i=1}^{N} \delta(r_{it})$ is the number of observations that are fully retained (not censored), $\varphi = n/N$ is the fraction of observations in the retained sample, and $\bar{r}_{sub} = \frac{1}{n}\sum_{i=1}^{N} \delta(r_{it}) r_{it}$ is the average relevance value among the retained sample. It is important to note that $w_{it,psr}$ depends crucially on the prediction circumstances x_t. Relevance is reassessed for each prediction circumstance which further affects the identification of the

retained subsample and introduces nonlinear conditional dependence of the prediction $\hat{y}_t$ on the prediction circumstances x_t. The scaling factor λ^2 compensates for a bias that would otherwise result from relying on a small subsample of highly relevant observations. In the case of linear regression analysis, where $n = N$, we have $\lambda^2 = 1$. Lastly, note that partial sample regression weights always sum to 1.[7] The question that now emerges is how to select the censoring threshold r^*, which leads to fit.

Fit

Fit is a crucial component of RBP. It reveals how much confidence we should have in a specific prediction task, separately from the confidence we have in the overall prediction system. In addition, it provides a principled way to evaluate the relative merits of alternative calibrations for each prediction task.

Consider, for example, a pair of observations that are used, in part, to form a prediction. Each observation has a weight and an outcome. We are interested in the alignment of the weights of the two observations with their outcomes. We must first standardize them by subtracting the average value and dividing this difference by standard deviation – in essence, converting them to z-scores. We then measure their alignment by taking the product of these standardized values. If the product is positive, their relevance is aligned with their outcomes, and the larger the product, the stronger the alignment. We perform this calculation for every pair of observations in our sample. We should also note that all the formulas we have thus far considered for weights rely only on relevance, which in turn relies only on the x_is, the x_t, and the $\bar{x}$. They do not make use of any of the information from observed outcomes. To determine fit, however, we must consider outcomes (the y_is).

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j z_{w_{it}} z_{w_{jt}} z_{y_i} z_{y_j} \quad (21)$$

Equation 22 intuitively describes fit as the squared correlation of relevance weights and outcomes, which conceptually matches the notion of the conventional R-squared statistic. As we soon show, this connection of fit to R-squared is critically important.

$$fit_t = \rho(w_t, y)^2 \quad (22)$$

Although we compute fit from the full sample of observations, the weights that determine fit vary with the threshold we choose to define the relevant subsample. As we focus the subsample on observations that are more relevant, we should expect the fit of the subsample to increase, but we should also expect more noise as we shrink the number of observations. The fit across pairs of all observations in the full sample implicitly captures this tradeoff between subsample fit and noise by overweighting observations that are more relevant and underweighting observations that are less relevant accordingly.

Like relevance, fit is not arbitrary. In the case of linear regression with $n = N$, the informativeness-weighted average fit across all prediction tasks in the observed sample equals the classical R-squared statistic.[8]

$$R^2 = \frac{1}{N-1} \sum_{t=1}^{N} info(x_t, \bar{x}) fit_t \quad (23)$$

This convergence of fit to R-squared reveals an intriguing insight. R-squared is the result of some good predictions, some average predictions, and some bad predictions; that is, some predictions with high fit, some with average fit, and some with low fit. R-squared reveals the average reliability of a prediction model. It reveals much less about the reliability of specific

prediction tasks, which can vary substantially. Fit is much more nuanced. It gauges the reliability of a specific prediction task in a non-arbitrary way, as demonstrated by its convergence to R-squared. Fit is the fundamental building block of R-squared. To compute fit, we must know the weight of each observation in a prediction. These weights are inherent to RBP, but they are not available in model-based prediction algorithms which rely exclusively on calibrated parameters rather than weighted observations to form predictions.

Grid Prediction

We have thus far shown how to form a prediction as a relevance-weighted average of outcomes (y_is). And we have shown how we can use fit to measure the reliability of a specific prediction task. But we have left unanswered the question of how to determine the threshold for the subsample of relevant observations. We have only noted that a partial sample regression prediction depends on the choice of a parameter, r^*, which is the censoring threshold for relevance. In addition, we have implicitly assumed up to this point that the full menu of predictive variables is used to measure relevance and form a partial sample prediction. However, it is possible that a subset of the predictive variables will render a better prediction for a specific task. The efficacy of observations for a given prediction task depends on the predictive variables, and the efficacy of the predictive variables depends on the observations. These choices are codependent. We, therefore, turn to the last key feature of RBP, which is grid prediction. But before we show how to form predictions that consider a range of alternative calibrations, we must first describe an enhanced version of fit called adjusted fit.

Partial sample regression using relevance is more effective to the extent there is strong alignment between relevance and outcomes, as measured by fit. It is also more effective to the

extent there is asymmetry between the fit of the weights formed from the retained subsample of observations and the fit of the weights formed from the complementary set of censored observations. In the presence of asymmetry, we trust the more relevant sample based on principle. In the absence of asymmetry, the full sample relationship is linear, and linear regression analysis, which is a special case of RBP, will suffice. Therefore, to compare properly the efficacy of two predictions formed from different values of r^*, we need a way to measure not only fit but asymmetry.

We measure asymmetry between the fit of the retained and censored subsamples as shown by Equation 24. The $(+)$ superscript designates weights formed from the retained observations while the $(-)$ superscript designates weights formed from the censored observations. Asymmetry recognizes the benefit of censoring non-relevant observations that contradict the predictive relationships that exist among the relevant observations. This assessment also inherently considers the relative sample sizes of the complementary groups.

$$asymmetry_t = \frac{1}{2}\left(\rho\left(w_t^{(+)}, y\right) - \rho\left(w_t^{(-)}, y\right)\right)^2 \tag{24}$$

To calculate adjusted fit, we add asymmetry to fit and multiply this sum by K, the number of predictive variables included in the prediction, as shown by Equation 25. Multiplication by the number of predictive variables allows us to compare predictions based on different numbers of predictive variables. It corrects a bias that would otherwise occur, whereby adding a pure noise variable decreases fit in proportion to the increase in the number of variables, even if the predictions themselves do not change (consider, for example, the case of a full sample linear regression analysis with a large sample of observations). Another way to

view the intuition for K is that we are more likely to observe a spurious relationship from prediction weights based on any one variable in isolation than we are based on a collection of many variables.

$$adjusted\ fit_t = K(fit_t + asymmetry_t) \tag{25}$$

We now return to the question of how to form a prediction given uncertainty in the calibration of r^* and variable selection, which are codependent choices. To address this issue, we could consider every possible calibration that combines a choice of r^* with a choice of a subset of variables and select the prediction with the greatest reliability as measured by adjusted fit. It is important to note that the assessment of reliability using adjusted fit is made before the prediction is rendered and the subsequent outcome is known and that the assessment of reliability is specific to the prediction task.

However, instead of selecting one optimal calibration for a given prediction task, it may be more prudent to compute a composite prediction as a reliability-weighted average of the predictions from all possible calibrations. Equation 26 defines reliability weights, ψ_θ, as the adjusted fit for a parameter calibration, θ, divided by the sum of all adjusted fits across all parameter calibrations.

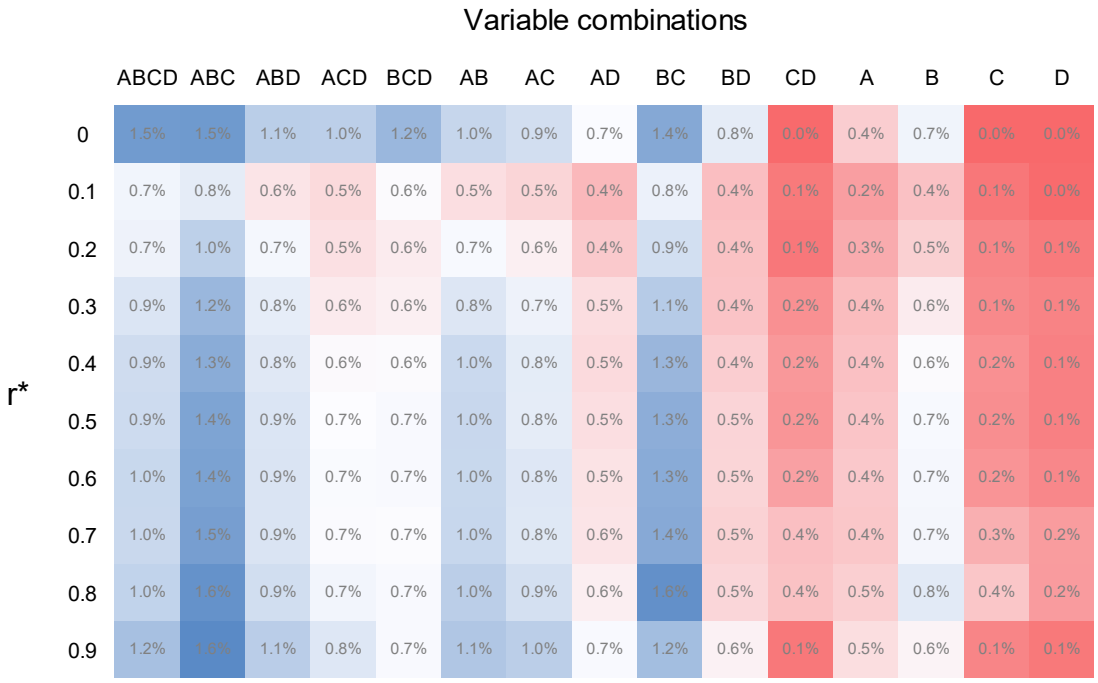
$$\psi_\theta = \frac{adjusted\ fit_\theta}{\sum_{\tilde{\theta}} adjusted\ fit_{\tilde{\theta}}} \tag{26}$$

Equation 27 describes the composite prediction.

$$\hat{y}_{t,grid} = \sum_\theta \psi_\theta \hat{y}_{t,\theta} \tag{27}$$

Exhibit 1 gives a visual representation of grid prediction, based on a contrived data set of four predictive variables and 400 randomly simulated observations. The column labels represent alternative variable subsets, and the row labels represent alternative observation subsets. Each cell represents a codependent calibration θ; that is, a unique combination of predictive variables and observations. The values in the cells are the weights (ψ_θ) we apply to the calibration-specific predictions to form the composite grid prediction. Cells that are shades of red are less important to forming the prediction while blue shaded cells are more important. The values in the grid are specific to each prediction task.

Exhibit 1: Grid Prediction – Illustrative Example



Note that each cell's prediction is a linear function of observations, and the grid prediction is a linear function of each cell's prediction. Therefore, we can express the grid prediction in terms of composite weights applied to each observation, as shown in Equation 28.

Composite weights are important because they preserve the transparency of how each observation contributes to the current prediction task, and they allow us to calculate fit from composite weights as a final gauge of the grid prediction's reliability.

$$w_{it,grid} = \sum_{\theta} \psi_{\theta} w_{it,\theta} \tag{28}$$

One final point is worth noting about grid prediction. In cases where informative (statistically unusual) observations do not extrapolate reliably to other circumstances, it may be advisable to consider subsamples of observations and predictive variables based on similarity filtering rather than relevance filtering. We need not worry whether we should use similarity or relevance to identify the optimal combination of observations and variables. We simply include multiple observation censoring rules as candidates in the grid. However, even when we censor based on similarity, we should still form the predictions as a relevance-weighted average of the retained observations.

We next present a simple illustration of HCMs and RBP to show how they form predictions differently but give similar results.

A Simple Illustration of HCMs and RBP

Exhibit 2 shows evenly spaced observations of a predictive X variable and corresponding values for Y. This example clearly is contrived to reflect a conditional relationship in which the relationship of X and Y is sometimes negative and sometimes positive.

Exhibit 2: Illustrative Data for a Conditional Relationship

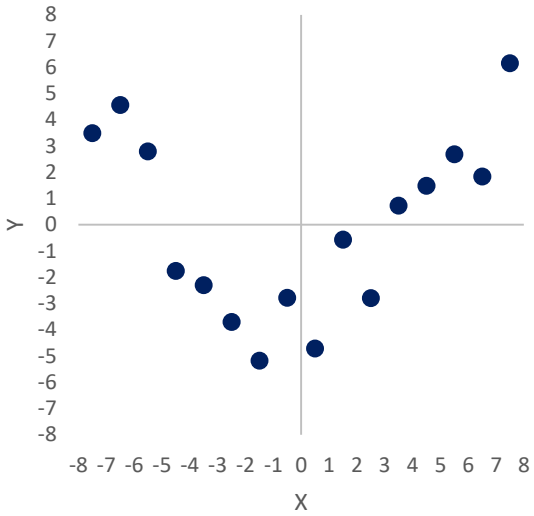


Exhibit 3 compares the prediction of Y given an x_t value of -5.50 performed two different ways. The left panel of Exhibit 3 uses an HCM to form the prediction. It creates two nonlinear transformations of X to create new predictive variables $S1$ and $S2$. The rule for $S1$ is to keep the value of X if it is negative and replace it with zero if it is positive. The rule for $S2$ is to keep the value of X if it is positive and replace it with zero if it is negative.

The right panel in Exhibit 3 uses RBP to form the prediction, though only up to partial sample regression. It does not incorporate grid prediction. Partial sample regression is configured to censor the least relevant half of the observations to generate prediction weights to apply across observations of Y. It leads to prediction weights that are very similar to the HCM.

Exhibit 3: Comparison of HCM and RBP Predictions ($x_t = -5.5$)

HCM						RBP					
Prediction for X = -5.5						Prediction for X = -5.5					
Observation	S1	S2	Relevance	Censor	Observation Weight	Observation	X	Y	Relevance	Censor	Observation Weight
1	-7.50	0.00	2.76	1	25%	1	-7.50	3.50	1.82	1	27%
2	-6.50	0.00	2.25	1	21%	2	-6.50	4.57	1.58	1	23%
3	-5.50	0.00	1.74	1	18%	3	-5.50	2.80	1.33	1	19%
4	-4.50	0.00	1.23	1	14%	4	-4.50	-1.75	1.09	1	15%
5	-3.50	0.00	0.72	1	11%	5	-3.50	-2.30	0.85	1	12%
6	-2.50	0.00	0.20	1	8%	6	-2.50	-3.70	0.61	1	8%
7	-1.50	0.00	-0.31	1	4%	7	-1.50	-5.17	0.36	1	4%
8	-0.50	0.00	-0.82	1	1%	8	-0.50	-2.79	0.12	1	1%
9	0.00	0.50	-1.06	1	-1%	9	0.50	-4.71	-0.12	0	-1%
10	0.00	1.50	-1.03	1	-1%	10	1.50	-0.56	-0.36	0	-1%
11	0.00	2.50	-1.01	1	0%	11	2.50	-2.80	-0.61	0	-1%
12	0.00	3.50	-0.98	1	0%	12	3.50	0.73	-0.85	0	-1%
13	0.00	4.50	-0.96	1	0%	13	4.50	1.48	-1.09	0	-1%
14	0.00	5.50	-0.93	1	0%	14	5.50	2.69	-1.33	0	-1%
15	0.00	6.50	-0.91	1	0%	15	6.50	1.85	-1.58	0	-1%
16	0.00	7.50	-0.88	1	0%	16	7.50	6.16	-1.82	0	-1%
Average:	-2	2				Average:	0.00	0.00			
Covariance:	7.07	4.27				Variance:	22.67	11.94			
	4.27	7.07									
Betas:	-1.19	1.40									
Intercept:	-5.19			**Prediction:**	**1.38**					**Prediction:**	**1.37**

Exhibit 4 shows the same analysis for predicting Y but in this case for an x_t value of +6.5. The two-variable HCM still does not censor any observations, but it shifts its focus across the two variables which has the same effect as changing the observation weights that inform the prediction. The partial sample regression chooses to censor a different half of the observations for this task, and the observation weights change accordingly.

Exhibits 3 and 4 show that, although the HCM and RBP form the predictions differently, they yield similar predictions and place similar importance on the observations. Given the conditionality of the contrived data, the HCM's betas and intercept are difficult to interpret. However, owing to the equivalence of linear regression analysis with full-sample RBP, we are able to recast the HCM predictions as relevance-weighted averages, which allows us to observe the similarity of the observation weights, and which underscores the transparency of RBP.

Exhibit 4: Comparison of HCM and RBP Predictions ($x_t = +6.5$)

HCM						**RBP**					
Prediction for X = 6.5						Prediction for X = 6.5					
Observation	S1	S2	Relevance	Censor	Observation Weight	Observation	X	Y	Relevance	Censor	Observation Weight
1	-7.50	0.00	-0.59	1	2%	1	-7.50	3.50	-2.15	0	-2%
2	-6.50	0.00	-0.75	1	1%	2	-6.50	4.57	-1.86	0	-2%
3	-5.50	0.00	-0.91	1	0%	3	-5.50	2.80	-1.58	0	-2%
4	-4.50	0.00	-1.07	1	-1%	4	-4.50	-1.75	-1.29	0	-2%
5	-3.50	0.00	-1.23	1	-2%	5	-3.50	-2.30	-1.00	0	-2%
6	-2.50	0.00	-1.39	1	-3%	6	-2.50	-3.70	-0.72	0	-2%
7	-1.50	0.00	-1.55	1	-4%	7	-1.50	-5.17	-0.43	0	-2%
8	-0.50	0.00	-1.71	1	-5%	8	-0.50	-2.79	-0.14	0	-2%
9	0.00	0.50	-1.42	1	-3%	9	0.50	-4.71	0.14	1	0%
10	0.00	1.50	-0.69	1	2%	10	1.50	-0.56	0.43	1	4%
11	0.00	2.50	0.05	1	7%	11	2.50	-2.80	0.72	1	8%
12	0.00	3.50	0.78	1	11%	12	3.50	0.73	1.00	1	13%
13	0.00	4.50	1.51	1	16%	13	4.50	1.48	1.29	1	17%
14	0.00	5.50	2.25	1	21%	14	5.50	2.69	1.58	1	22%
15	0.00	6.50	2.98	1	26%	15	6.50	1.85	1.86	1	26%
16	0.00	7.50	3.71	1	31%	16	7.50	6.16	2.15	1	30%
Average:	-2.00	2.00				Average:	0.00	0.00			
Covariance:	7.07	4.27				Variance:	22.67	11.94			
	4.27	7.07									
Betas:	-1.19	1.40									
Intercept:	-5.19			**Prediction:**	**3.92**					**Prediction:**	**3.15**

Next, we expand our set of predictive variables beyond just $S1$ and $S2$. We now consider 1,000 randomly generated nonlinear transformations. In the spirit of Kelly, Malamud, and Zhou, we create one new variable by multiplying all 16 values of X by a randomly drawn number from a normal distribution with unit variance centered on zero, and we shift those results by another randomly drawn number from the same distribution. We then apply the highly nonlinear sine function to the result. We repeat this process 1,000 times to generate 1,000 distinct transformed S variables. We then perform principal components analysis on the covariance matrix of the 1,000 S variables and project their values to obtain 15 uncorrelated Q variables which we sort in decreasing order of variance. Recall that the information in any number of S variables can be condensed into a maximum of 15 Q variables, which is one less than the number of observations. This is possible because the S variables contain partly

redundant information. The Q variables represent the dominant sources of variation in the randomly transformed variables.[9] As we discussed previously, regressing Y on the Q variables derived from principal components analysis is equivalent to regressing on the S variables directly using the Moore-Penrose pseudo-inverse of the covariance matrix of S, which is also equivalent to ridgeless regression.

Exhibit 5 shows the values of the Q variables for each observation. We observe, for example, that $Q1$ mainly distinguishes high values of X from low values of X, while $Q3$ distinguishes moderate values from high and low values.

Exhibit 5 also shows the observation weights, based on relevance, for predictions that pay attention to the top 5 Q variables, the top 10 Q variables, and all 15 Q variables. The observation weights based on the top 5 Q variables are broadly similar to the weights we observed in the prior examples; observations for low values of X similar to the prediction circumstances (observation 3) receive high weights, while other observations receive weights close to zero. As we include more Q variables, the weights become increasingly concentrated. If we include all 15 Q variables, the prediction places 100% of its weight on the observation that matches the prediction circumstance. It therefore results in a prediction that is precisely equal to the actual Y value for observation 3. In the Appendix we show mathematically that this result always obtains for ridgeless regression when the nonlinear transformations are extensive enough that they generate the full range of $N - 1$ principal components of the covariance structure, which means they are capable of perfectly isolating each observation in the data sample. This scenario would be deemed overfitting because there is no generalization of a relationship from multiple observations, and if there is noise involved in the process, we would

expect a new observation of Y for the same X circumstances to differ from the value of 2.80. Using multiple observations to inform a prediction, where viable, has the advantage of diversifying the noise components of each observation in the prediction. To put it differently, the prediction with 15 Q variables has not learned a generalized relationship, it has merely memorized the training data.[10]

The regression betas corresponding to each Q variable are shown at the bottom of Exhibit 5. It is difficult to interpret these regression betas, and it is not necessarily apparent from visually inspecting them that the prediction from 15 Q variables places 100% weight on a single observation. It would be even more difficult to interpret 1,000 regression betas corresponding to the S variables. However, invoking the equivalence of linear regression analysis with full-sample RBP, we can again recast the HCM predictions as relevance-weighted averages, which reveals the importance of each observation to the predictions. The larger point, though, is that the efficacy of HCMs is not dependent on a contrived two-variable example; it generalizes to applications that comprise a very large number of randomly generated predictive variables.

Exhibit 5: HCM Predictions with 15 Qs from 1,000 Randomly Generated Sine Transformations

Predictions X =	13.19	5.60	8.01	1.58	0.37	-7.62	8.98	3.86	2.45	2.17	-1.83	-1.64	-2.12	-0.51	0.31	Relevance			Observation Weights		
Observations	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	5 Qs	10Qs	15 Qs	5 Qs	10 Qs	15 Qs
1	-5.62	-9.48	4.90	8.29	-8.11	-8.81	-5.82	-3.47	4.59	-2.31	2.89	-1.04	0.79	-0.11	0.01	-0.37	-0.59	-0.94	4%	2%	0%
2	-8.97	-12.43	6.08	5.47	-6.70	0.08	-0.31	1.89	-4.53	4.30	-4.29	3.66	-0.95	1.33	0.24	-0.97	-1.28	-0.94	0%	-2%	0%
3	-11.21	-9.69	5.65	-3.64	0.26	8.30	7.04	5.21	-2.67	1.41	1.77	-2.23	1.65	-1.16	-0.30	-1.46	-0.90	-0.94	-3%	0%	0%
4	-12.07	-2.24	4.22	-9.84	7.61	2.67	6.10	-2.11	5.59	-2.20	1.93	1.92	-1.03	2.04	0.67	-1.47	-0.51	-0.94	-4%	3%	0%
5	-11.58	6.38	1.81	-6.09	10.50	-6.70	-1.51	-5.23	-0.94	2.51	-3.95	1.43	0.91	-1.34	-0.80	-1.02	-1.30	-0.94	-1%	-2%	0%
6	-9.94	12.02	-1.98	4.45	7.45	-4.58	-4.74	4.08	-5.83	1.86	2.19	-1.88	0.25	1.61	1.16	-0.59	-1.13	-0.94	2%	-1%	0%
7	-7.23	12.23	-6.63	10.73	0.19	4.60	0.09	5.86	3.60	-1.24	1.10	3.43	-0.53	-0.35	-1.22	-0.64	-0.25	-0.94	2%	5%	0%
8	-3.49	7.45	-10.95	5.62	-6.28	5.55	6.21	-4.48	3.31	2.89	-3.41	-0.73	1.47	0.31	1.45	-1.21	-1.15	-0.94	-2%	-1%	0%
9	0.77	-0.11	-13.58	-4.96	-6.68	-3.26	6.46	-4.29	-5.60	1.24	2.82	-0.28	-1.04	0.96	-1.31	-1.88	-1.51	-0.94	-6%	-4%	0%
10	4.91	-7.12	-13.15	-8.97	-0.55	-6.62	0.18	5.70	-0.85	-2.75	-0.38	3.07	0.87	-0.78	1.31	-1.94	-0.02	-0.94	-7%	6%	0%
11	8.57	-10.24	-9.04	-2.37	7.05	1.49	-5.44	3.47	4.94	4.12	-2.49	-1.79	0.58	1.68	-0.96	-0.93	-0.98	-0.94	0%	0%	0%
12	11.49	-7.82	-2.38	6.94	10.28	6.73	-2.27	-4.62	-1.90	3.27	3.31	1.70	-1.16	-1.04	0.82	0.86	-2.55	-0.94	12%	-11%	0%
13	13.18	-1.34	4.14	8.81	7.23	-0.87	7.02	-1.16	-3.64	-4.90	-1.56	1.41	2.33	1.48	-0.42	2.59	2.85	-0.94	24%	25%	0%
14	13.19	5.60	8.01	1.58	0.37	-7.62	8.98	3.86	2.45	2.17	-1.83	-1.64	-2.12	-0.51	0.31	3.50	8.82	14.06	30%	65%	100%
15	11.29	9.26	8.11	-6.93	-5.34	-0.39	-0.52	0.69	1.23	6.55	3.52	3.15	2.15	0.80	-0.01	3.31	4.07	-0.94	28%	33%	0%
16	7.59	7.86	5.13	-8.35	-6.05	8.84	-8.27	-0.97	-2.52	-4.48	-2.54	-0.74	-0.75	0.15	0.07	2.23	-3.55	-0.94	21%	-17%	0%
Covariance:	96.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	75.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	59.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	51.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	45.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	34.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	29.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.02	0.00	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.27	0.00	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.78	0.00	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.42	0.00	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.77	0.00	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.25	0.00						
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76						
Betas 5 Qs:	0.10	-0.13	0.32	-0.03	-0.18											**Prediction**					**3.36**
Intercept:	0.32																				
Betas 10 Qs:	0.10	-0.13	0.32	-0.03	-0.18	0.06	-0.10	0.01	-0.12	-0.15						**Prediction**					**1.58**
Intercept:	0.50																				
Betas 15 Qs:	0.10	-0.13	0.32	-0.03	-0.18	0.06	-0.10	0.01	-0.12	-0.15	-0.20	-0.06	-0.02	-0.49	0.72	**Prediction**					**2.69**
Intercept:	0.64																				

Simulation Experiments

To evaluate these approaches in a more intricate setting with multiple predictive variables, repeated random draws, and RBP grid prediction, we perform a regime-based simulation experiment similar to Czasonis, Kritzman, and Turkington (2023), but with the addition of random noise affecting outcomes for Y. We simulate four X variables which we refer to as

A, B, C, D, where A, B function as a group and influence Y in regime 1, and C, D function as a group and influence Y in regime 2.

$$\rho = \begin{pmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{pmatrix}$$

$$\mu_1 = (3 \quad 3 \quad 0 \quad 0); \quad \sigma_1 = (1 \quad 1 \quad 3 \quad 3); \quad \beta_1 = (1 \quad 2 \quad 0 \quad 0)$$

$$\mu_2 = (0 \quad 0 \quad 3 \quad 3); \quad \sigma_2 = (3 \quad 3 \quad 1 \quad 1); \quad \beta_2 = (0 \quad 0 \quad 1 \quad 2)$$

$$\begin{cases} X \sim N(\mu_1, \Omega_1) & if\ regime = 1 \\ X \sim N(\mu_2, \Omega_2) & if\ regime = 2 \end{cases}$$

$$\begin{cases} y_t = \beta_1 x_t' + \epsilon_t & if\ regime = 1 \\ y_t = \beta_2 x_t' + \epsilon_t & if\ regime = 2 \end{cases}$$

$$\sigma_\epsilon = 5$$

Regime 1 prevails randomly 75 percent of the time, and regime 2 prevails the other 25 percent of the time. We simulate 500 random training observations and 500 random testing observations.

We form predictions using multiple methods. First, we form predictions from traditional linear regression analysis. Second, we produce RBP grid predictions considering all combinations of the predictive variables along with observation subsamples based on censoring thresholds of 0, 0.2, 0.5, and 0.8 based on both relevance and similarity. Third, we produce predictions from multiple calibrations of nonlinear transformed predictive variables. Following Kelly, Malamud, and Zhou, we generate 5,000 (10 times the number of observations) S variables, each of which applies a set of four random multiples from a centered normal

distribution with variance of 10 (which provided more consistent and reliable results than variance of 1) to the normalized z-scores of the inputs and computes the sine function of the resulting sum. We condense the 5,000 S variables into sets of 10, 100, 250, and 499 Q variables. Recall from our earlier discussion that 499 Qs capture the entirety of the linear predictive information in the 500 training sample observations for the 5,000 S variables. The smaller sets of Q extract subsets of the highest variance information in S.

Exhibit 6 shows the average outcomes for the prediction tasks conditional on below median (low) and above median (high) predictions from each method, as well as a further decomposition by below median (low) and above median (high) fit within the low and high prediction subsamples. Fit effectively identifies in advance predictions with more extreme outcomes in almost every case except low predictions for 10 Qs, but the separation is most dramatic for RBP grid predictions.

Exhibit 6: Average Out-of-Sample Outcomes Relative to Full-Sample Average

	Linear	RBP	Random Sine Transformations			
	Regression	Grid	10 Qs	100 Qs	250 Qs	All Qs
Low predictions	-0.95	-1.44	-0.63	-0.74	-0.86	-0.86
Low predictions with low fit	-0.42	-0.33	-1.19	-0.18	-0.12	-0.06
Low predictions with high fit	-1.49	-2.55	-0.07	-1.31	-1.60	-1.65
High predictions	0.95	1.44	0.63	0.74	0.86	0.86
High predictions with low fit	0.80	0.77	-0.21	0.23	0.16	0.21
High predictions with high fit	1.10	2.11	1.46	1.25	1.55	1.51

Exhibit 7 reports the correlations of actual out-of-sample outcomes for each of the 500 prediction tasks with the predictions as well as the correlations among the various predictions. It is interesting to note that RBP grid prediction has the highest correlation of predictions with outcomes by a significant margin. Even though the highest correlation between HCM and RBP

is only 0.54, it is notable that the HCM predictions have a consistently higher correlation with RBP than with linear regression, which indicates that they identify at least some nonlinear predictive relationships in common. We will explore the overlapping nature of their information in more detail shortly.

Exhibit 7: Correlations of Outcomes and Predictions

		Linear	RBP	Random Sine Transformations			
	Actual	Regression	Grid	10 Qs	100 Qs	250 Qs	All Qs
Actual	1.00						
Linear	0.25	1.00					
RBP Grid	0.57	0.73	1.00				
Sine transformations: 10 Qs	0.25	0.33	0.46	1.00			
Sine transformations: 100 Qs	0.34	0.42	0.51	0.56	1.00		
Sine transformations: 250 Qs	0.37	0.43	0.54	0.48	0.85	1.00	
Sine transformations: All Qs	0.33	0.31	0.44	0.33	0.52	0.59	1.00

Exhibit 8 shows the average and standard deviation of actual test sample outcomes along with those of each prediction method. It also shows the root mean squared error of the predictions and repeats the correlations of predictions to outcomes from Exhibit 7. Lastly, Exhibit 8 shows the average standard deviation in observation weights that are used to form each prediction. Recall that observation weights sum to 1 for each prediction, and a prediction formed from a single observation has a variance (and standard deviation) of 1. Weights can also be negative, which can lead to a higher standard deviation. It is interesting to note that RBP grid prediction has the lowest root mean squared error. The apparent superiority of RBP grid prediction from Exhibits 7 and 8 may be specific to this simulation, though.

Exhibit 8: Prediction Statistics

	Actual	Linear Regression	RBP Grid	Random Sine Transformations			
				10 Qs	100 Qs	250 Qs	All Qs
Average	9.19	8.85	9.55	8.99	9.06	9.02	8.41
Standard deviation	3.25	1.08	0.99	0.93	1.73	2.09	3.58
Root mean squared error		3.18	2.82	3.15	3.12	3.15	4.02
Correlation to actual		0.25	0.57	0.25	0.34	0.37	0.33
Standard deviation of observation weights (average)		0.07	0.05	0.09	0.31	0.47	1.42

Exhibit 9 shows linear regression betas and t-statistics (in parentheses) for OLS regressions of actual test sample outcomes on multiple predictions. The RBP grid and sine transformation predictions subtract linear predictions to avoid excessive collinearity for the purpose of decomposition.[11] We consider the sine transformation predictions for the case of 250 Qs which has the highest correlation with realized outcomes and with RBP predictions. Exhibit 9 shows that despite having a correlation of only 0.54 with the RBP predictions, the useful predictive information contained in the 250 Qs HCM mostly overlaps with that of RBP, as evidenced by the differences in predictive R-squared across the regressions.

Exhibit 9: Regressions of Actual Outcomes on Prediction Components

	1	2	3	4
Intercept	-7.18	1.85	-7.50	2.67
	(-6.4)	(1.65)	(-6.7)	(2.3)
Linear Regression	1.64	0.82	1.67	0.74
	(13.79)	(6.56)	(14)	(5.65)
RBP Grid (above linear)	2.63		2.79	
	(14.57)		(16.5)	
Sine Transformations: 250 Qs (above linear)	0.16	0.50		
	(2.43)	(6.94)		
R-squared	0.40	0.14	0.39	0.06

Exhibits 10, 11, 12, and 13 show results from the same simulation setup, but the S variables are derived using a logistic S-shaped function instead of a sine function. This collection of simulations is based on a fresh draw of random numbers for X and Y. These exhibits show that, just as in the case of sine function transformations, RBP grid prediction has the highest correlation of predictions with outcomes as well as the lowest root mean squared error of all prediction methods, but again this result may be specific to this simulation.

The HCM predictions have slightly higher correlations with RBP predictions in this set of results. Once again, we observe that the 250 Qs HCM has the highest correlation with actual outcomes, but its predictive information is mostly subsumed by the nonlinear component of the RBP predictions, as seen in the R-squared comparisons of Exhibit 13.

Exhibit 10: Average Out-of-Sample Outcomes Relative to Full-Sample Average

	Linear	RBP	Random Logistic Transformations			
	Regression	Grid	10 Qs	100 Qs	250 Qs	All Qs
Low predictions	-0.97	-1.46	-1.21	-1.30	-1.42	-0.52
Low predictions with low fit	-0.32	-0.24	-0.48	-0.49	-0.61	0.15
Low predictions with high fit	-1.63	-2.67	-1.94	-2.12	-2.23	-1.19
High predictions	0.97	1.46	1.21	1.30	1.42	0.52
High predictions with low fit	0.21	0.47	0.41	0.58	0.77	0.35
High predictions with high fit	1.74	2.45	2.01	2.03	2.08	0.69

Exhibit 11: Correlations of Outcomes and Predictions

		Linear	RBP	Random Logistic Transformations			
	Actual	Regression	Grid	10 Qs	100 Qs	250 Qs	All Qs
Actual	1.00						
Linear Regression	0.30	1.00					
RBP Grid	0.59	0.74	1.00				
Logistic Transformations: 10 Qs	0.45	0.68	0.71	1.00			
Logistic Transformations: 100 Qs	0.50	0.43	0.65	0.70	1.00		
Logistic Transformations: 250 Qs	0.51	0.32	0.57	0.54	0.78	1.00	
Logistic Transformations: All Qs	0.18	0.15	0.23	0.27	0.26	0.29	1.00

Exhibit 12: Prediction Statistics

	Actual	Linear Regression	RBP Grid	Random Logistic Transformations			
				10 Qs	100 Qs	250 Qs	AllQs
Average	9.12	9.01	9.63	9.05	8.99	9.00	8.48
Standard deviation	3.40	1.12	1.00	1.63	2.28	3.13	8.99
Root mean squared error		3.24	2.97	3.05	3.00	3.25	9.04
Correlation to actual		0.30	0.59	0.45	0.50	0.51	0.18
Standard deviation of observation weights (average)		0.07	0.05	0.13	0.42	0.76	3.11

Exhibit 13: Regressions of Actual Outcomes on Prediction Components

	1	2	3	4
Intercept	-7.50	0.29	-9.31	0.83
	(-6.24)	(0.28)	(-7.89)	(0.71)
Linear Regression	1.70	0.98	1.86	0.92
	(13.62)	(8.49)	(15.03)	(7.11)
RBP Grid (above linear)	2.16		2.71	
	(10.55)		(15.16)	
Logistic Transformations: 250 Qs (above linear)	0.24	0.50		
	(5.19)	(11.36)		
R-squared	0.41	0.28	0.38	0.09

Conclusion and Summary

The key challenge to prediction is to extract the most information possible from a sample of data without detrimentally overfitting the data. One technique for doing so is to convert a relatively small number of predictive variables into a much larger set of variables by applying non-linear transformations to the original variables. If the number of transformed variables is sufficiently large, the variables will extract all the information from the data including

conditionalities that the original set of variables would fail to detect. It is commonly assumed, however, that as the number of variables increases relative to the number of observations the covariance matrix becomes unstable rendering out-of-sample predictions unreliable. Kelly, Malamud, and Zhou (2024) overcome this concern by effectively transforming the large number of transformed variables into linear combinations of fewer variables, which greatly diminishes the risk of overfitting. A prediction model derived in this fashion is called a high-complexity model (HCM).

An alternative technique for maximizing the utility of data without overfitting is called relevance-based prediction (RBP). It forms a prediction as a weighted average of observed outcomes in which the weights are a statistical measure called relevance. Relevance is composed of similarity and informativeness, which are both measured as squared z-scores in the case of a single predictive variable, or as Mahalanobis distances in the case of multiple predictive variables. RBP recognizes that observations and predictive variables are codependent; they should be selected jointly using principles that evaluate predictive reliability in advance, given the unique circumstances of each individual prediction task.

There are two critical differences between HCMs and RBP. HCMs rely on a large number of randomly manufactured predictive variables within a single model to extract information. RBP, by contrast, is model free. It is a prediction routine that extracts information by considering many combinations of relatively few variables and applying them in ways that are specific to each prediction task.

HCMs are opaque. They only reveal the average importance of the large set of S variables or their consolidated counterparts, Q. Also, they give no insight into the relative

importance of the observations to the prediction. Finally, they give no advance notice of an individual prediction's reliability. We must wait to see the outcome before we can judge the quality of a prediction. RBP is fully transparent. It shows precisely how each observation informs a specific prediction, and it reveals the relative importance of the predictive variables for each individual prediction task. Finally, it gives advance notice of the reliability of the individual predictions before they are rendered.

We have shown that HCMs and RBP give similar predictions and assign similar importance to the observations in contrived experiments designed to illustrate how they function. There are likely to be prediction tasks, though, in which there are subtle conditionalities that are beyond the reach of RBP, given its reliance on relatively few variables and a parsimonious set of censoring rules. In these instances, an HCM, with its ability to include as many variables as necessary, has the potential to extract more information from the data than RBP. One might therefore conclude that we are faced with a tradeoff: the superior transparency of RBP versus the superior completeness of HCMs. We may be able to have it both ways. Again, recalling the equivalence of linear regression analysis with full-sample RBP, we could recast an HCM prediction as a relevance-weighted average. This conversion would allow us to see how each observation informs the HCM prediction. And we could extend RBP's ability to capture conditionalities that are beyond the reach of the original predictive variables by including the Q variables engineered by an HCM as candidates in the RBP routines. Therefore, the best approach for maximizing the utility of data without overfitting may be to use an HCM in tandem with RBP.

Appendix: Observation weights for full-rank predictions of observed circumstances

Here we demonstrate mathematically that for prediction tasks in the observed sample, ridgeless regression always places 100% weight on the observed outcome that corresponds to the prediction circumstances, so long as the principal components consolidation of variables is full-rank with respect to the number of observations, N.

Consider any large set of predictive variables S for which the principal component transforms of S derived from $S'S$ are full rank. We denote these N principal component transforms as Q_+ to distinguish them from the earlier definition of Q which is based on the average-centered covariance matrix of S, and therefore has $N-1$ principal components in this scenario. The inverse matrix Q_+^{-1} is guaranteed to exist because Q_+ is full rank, and $Q_+Q_+^{-1} = I$. The OLS linear regression prediction of all the observed circumstances is $\hat{Y} = Q_+(Q_+'Q_+)^{-1}Q_+'Y$. Expanding out the inverse gives $\hat{Y} = Q_+Q_+^{-1}{Q_+'}^{-1}Q_+'Y = IY = Y$. The identity matrix represents the weights placed on each observation of Y for each prediction task and shows that each prediction places 100% weight on a single observation.

Notes

This material is for informational purposes only. The views expressed in this material are the views of the authors, are provided "as-is" at the time of first publication, are not intended for distribution to any person or entity in any jurisdiction where such distribution or use would be contrary to applicable law and are not an offer or solicitation to buy or sell securities or any product. The views expressed do not necessarily represent the views of Windham Capital Management, State Street Global Markets®, or State Street Corporation® and its affiliates.

References

Czasonis, Megan, Mark Kritzman, and David Turkington. 2020. "Addition by Subtraction: A Better Way to Forecast Factor Returns (And Everything Else)." *The Journal of Portfolio Management*, 46 (8).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022a. "Relevance." *The Journal of Investment Management*, 20 (1).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022b. *Prediction Revisited: The Importance of Observation*. Hoboken, New Jersey: John S. Wiley & Sons.

Czasonis, Megan, Mark Kritzman, and David Turkington. 2023. "Relevance-Based Prediction: A Transparent and Adaptive Alternative to Machine Learning." *The Journal of Financial Data Science*, 5 (1).

Hastie, Trevor, Andrea Montenari, Saharon Rosset and Ryan J. Tibshirani. 2022. "Surprises in High-Dimensional Ridgeless Least Squares Interpolation." *The Annals of Statistics*, 50 (2): April.

Kelly, Bryan T., Semyon Malamud and Kangying Zhou. 2024. "The Virtue of Complexity in Return Prediction." *Journal of Finance*, 79 (1).

Mahalanobis, Prasanta Chandra. 1936. "On the Generalised Distance in Statistics." *Proceedings of the National Institute of Sciences of India*, 2 (1): 49–55.

Shannon, Claude. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal*, 27 (July, October): 379–423, 623–656.

[1] Hastie, Montenari, Rosset, and Tibshirani (2022) explore related issues. Kernel ridge regression is also related to this approach but typically includes a highly curated set of nonlinear transformations of the original predictive variables rather than a large set of randomly generated transformations.

[2] If an N-by- N set of X variables are not centered, $X'X$ has a maximum of N nonzero principal components. The centered product $(X - 1_N 1_N' X N^{-1})(X - 1_N 1_N' X N^{-1})$, which is proportional to the covariance matrix, has a maximum of $N - 1$ nonzero principal components.

[3] The language in this section, particularly in the subsections titled Relevance, Partial Sample Regression, and Fit, closely follows that of Czasonis, Kritzman and Turkington (2020, 2022a, 2022b, and 2023) with minor modifications for clarity and context. Relevance-based prediction was first introduced, in successive variations, in these aforementioned publications. For clarity of exposition and to provide a self-contained reference for the analysis that follows, we review the essential components of RBP in the present section of this article. In the subsection titled Grid Prediction, we further extend RBP.

[4] This measure was first introduced by Mahalanobis (1936).

[5] Shannon showed that information is an inverse logarithmic function of probability, which is a key insight from his comprehensive theory of communication. See Shannon (1948).

[6] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[7] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[8] See Czasonis, Kritzman, and Turkington (2022b) for proof of this result.

[9] We can build intuition about which S variables contribute to the most prominent Q variables when ranked by variance. First, S transformations with greater variance contain more information in expectation because they differentiate well between observations on average. These S variables will tend to contribute to prominent (high variance) Q variables. Second, groups of S transformations that are highly similar and occur frequently across the random draws indicate aspects of the X variables that dominate the information they provide. These statistically common S variables will tend to contribute to prominent Q variables because they explain a large amount of the collective variance in S. These information-rich properties of high-ranking Q variables show why it is reasonable to consider censored subsets of the highest variance Q variables to avoid overfitting.

[10] We may, of course, form a prediction for circumstances that do not equal any of the circumstances in the observed data. In this case, the prediction cannot rely entirely on memorization. It is likely to choose a focused set of observations that are near neighbors to the novel prediction circumstances. Intuitively, because there are so many $(N - 1)$ dimensions to the Q variables, all but the closest observations are extremely distant from the prediction circumstances. We leave the detailed study of such predictions to future research. But regardless, it remains the case that any prediction circumstances that match those that occurred in the data will render based on 100% weight on a single observation. Any noise that affects that observation will translate to the prediction in full.

[11] This choice does not affect the key conclusions of the analysis, but it allows for a more useful view of the coefficients on linear regression predictions.