Statistical Analysis and Data Mining: An ASA Data Science Journal

WILEY

RESEARCH ARTICLE

Testing for the Important Components of Predictive Variance

Dean Dustin[1] | Souparno Ghosh[2] | Bertrand Clarke[2]

[1]First Citizens' Bank, Raleigh, North Carolina, USA | [2]Statistics Department, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

Correspondence: Bertrand Clarke (bclarke3@unl.edu)

Received: 16 August 2023 | **Revised:** 9 May 2025 | **Accepted:** 19 May 2025

Funding: This work was supported by Leidos, National Cancer Institute, Nebraska Program of Excellence in Computational Science and National Science Foundation (2007418).

Keywords: ANOVA | bootstrap testing | Cochran's theorem | law of total variance | prediction interval | predictive variance | stacking | variance decomposition

ABSTRACT

We give a decomposition of the predictive variance based on the law of total variance by making the response variable dependent on a finite dimensional discrete random variable representing our modeling assumptions. Then, we test which terms in this decomposition are small enough to ignore. This allows us to identify which of the discrete random variables, that is, aspects of modeling, are most important to prediction variance. The terms in the decomposition admit interpretations based on conditional means and variances and are analogous to the terms in a Cochran's theorem decomposition of squared error often used in analysis of variance. Thus, the modeling features are treated as factors in completely randomized design.

MSC2020 Classification: Primary 62F15, Secondary 62J10

1 | Introduction

The goal of this paper is to present an additive decomposition for $\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)$, the variance of a future outcome Y_{n+1} as a function of the data available, $\mathcal{D}_n$, before the next outcome Y_{n+1} is revealed. The data set $\mathcal{D}_n$ contains y_i for $i = 1, \ldots, n$ and may also contain values of explanatory variables X_i. We assume the y_i's are independent, but not necessarily identically distributed. We write the density used to define $\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)$, as $p(Y_{n+1}; \mathcal{D}_n)$ to indicate dependence on the data. The dependence is not in general through conditioning.

An additive decomposition is important because $\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)$ controls the length of prediction intervals (PI's) for Y_{n+1}. The idea is that by examining the terms we can tell which ones contribute most to the width of PI's and which ones can be neglected. That is, we can identify which features of modeling are most important for controlling variance and which aspects can be neglected so as to simplify models.

Our desired additive decomposition has three key properties: (i) The terms are individually interpretable as a sort of variability intrinsic to Y_{n+1}; (ii) Each term can be tested to see if it is small enough relative to the other terms that it can be neglected, and (iii) The terms in the decomposition of $\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)$ are analogous to the terms in Cochran's theorem including allowing flexibility as to how many terms are included. These components of the predictive variance can be examined to determine what they say about the various ingredients used to formulate the model. That is, for a given modeling scheme with multiple components we can test to see which are most important. Essentially, we put an ANOVA-like structure on the model features rather than on the data because, eventually, we want to use multiple decompositions for the same problem to assess a modeling strategy.

[Correction added on August 14, 2025, after first online publication: The copyright line was changed.]

Our decomposition is based on iterating an empirical version of the law of total variance for future outcomes given $\mathcal{D}_n$. Recall that in the posterior distribution, the law of total variance (LTV) for a single random variable V is

$$\mathrm{Var}(Y_{n+1}|\mathcal{D}_n) = E(\mathrm{Var}(Y_{n+1}|V,\mathcal{D}_n)) + \mathrm{Var}(E(Y_{n+1}|V,\mathcal{D}_n)) \quad (1)$$

Suppose V assumes finitely many values. The variance in the first term on the right is conditioned on $V = v$ (as well as the data) which is then integrated out with respect to the posterior as indicated by the expectation. So, from a Bayes standpoint we can regard the values of V as aspects of the model for the data. For instance, $V = v$ might indicate the inclusion of a prescribed subset of explanatory variables and the posterior for $V = v$ would indicate the post-data probability of that. An analogous interpretation applies to the inner expectation on the right.

Since our goal is to decompose the predictive variance into terms that represent various features of modeling, we would like to determine if any of the terms in (1) are small enough to omit. In the context of model averaging, this means that the smaller ensemble (decomposition with omitted terms) explains nearly the same amount of predictive variability as the larger ensemble (including all the terms). Thus, we want to design hypothesis tests to make this determination. Unfortunately, we cannot propose Bayesian hypothesis tests because we do not have a likelihood given the value of either of the terms. However, we can design frequentist bootstrap tests for whether a term can realistically be taken as zero. To be consistent, we therefore study a frequentist analog of (1).

With some hindsight, we rewrite the LTV as

$$\mathrm{Var}(Y_{n+1};\mathcal{D}_n) = E(\mathrm{Var}(Y_{n+1};V,\mathcal{D}_n)) + \mathrm{Var}(E(Y_{n+1};V,\mathcal{D}_n)) \quad (2)$$

Each of the three terms in (2) is an approximation of the corresponding term in (1). The central feature of the approximation is the consistent replacement of the posterior distributions by the corresponding stacking distributions, whence the semicolon in (2) in place of the conditioning symbol in (1), cf. [1]. The replacement is done in all three terms separately so that equality results. Now, the LHS is the stacking variance of Y_{n+1} using all the models indexed by V and it is philosophically consistent to use frequentist bootstrap tests on the terms of (2).

The first term on the right in either (1) or (2) is the average location of the variance taking into account the variability of V. If it is small, the variance as a function of the modeling features is small, perhaps indicating they make little difference predictively. The second term on the right is the variability contributed by V to the location of the predictive distribution. If it is small, then we know that $E(Y_{n+1};V,\mathcal{D}_n)$ is not affected much by the variability of V so it may make sense to ignore this term. The conceptual difference between these two terms is in how much V affects the variability in variance versus the variability in location. A caveat to this interpretation is that Y_{n+1} and V can be dependent even when the second term is zero. Indeed, suppose $Y_i|V$ is normally distributed and V has a distribution with positive support. Then, $\mathrm{Var}_V[E(Y|V)] = 0 \Rightarrow E(Y|V) = \text{Constant}$. So, we can choose $Y_i|V \sim \text{Normal}(\text{Constant}, V)$. Now, $\mathrm{Corr}(Y,V) = 0$, but obviously $Y \not\perp V$. Loosely, if dependence amongst the first m moments is ruled out, it is possible that dependence remains in moments at or above the $m+1$ moment.

To extend this variance decomposition, note we can apply (1) to itself in either term of (2). For instance, if we write $V = V_1$, introduce a second random variable V_2 also taking finitely many values, and apply (1) to the "E-Var" term we get,

$$\begin{aligned}\mathrm{Var}(Y_{n+1}|\mathcal{D}_n) &= E_{V_1,V_2}\mathrm{Var}(Y_{n+1}|V_1,V_2,\mathcal{D}_n)\\ &+ E_{V_1}\mathrm{Var}_{V_2}E(Y_{n+1}|V_1,V_2,\mathcal{D}_n) + \mathrm{Var}_{V_1}E(Y_{n+1}|V_1,\mathcal{D}_n) \quad (3)\end{aligned}$$

Using the stacked densities, now over the values of both V_1 and V_2, we can rewrite (3) as

$$\begin{aligned}\mathrm{Var}(Y_{n+1};\mathcal{D}_n) &= E_{V_1,V_2}\mathrm{Var}(Y_{n+1};V_1,V_2,\mathcal{D}_n)\\ &+ E_{V_1}\mathrm{Var}_{V_2}E(Y_{n+1};V_1,V_2,\mathcal{D}_n) + \mathrm{Var}_{V_1}E(Y_{n+1};V_1,\mathcal{D}_n) \quad (4)\end{aligned}$$

Again, ";" in (4) means we are replacing the posterior distribution in (3) by the stacking distribution in all four terms.

It is easy to extend this variance decomposition by including a random variable V_3. Indeed, in general, we can consider a multidimensional random variable $V = V_K = (V_1, \ldots, V_k, \ldots, V_K)$, apply (1) to itself $K-1$ times, and obtain stacking-based approximations analogous to (2) or (4) generating one new term for each V_k at each iteration. That is, for V_K we get a $K+1$ term decomposition that can be interpreted in terms of means and variances, see Proposition 1. For $K = 2$, there are multiple versions of (4). The left hand side is a fixed number given K, V_K and the data but it is easy to see that even with those quantities fixed there are several version of the terms on the right, that is, several versions of the variance decomposition. This means that our decompositions reflect a conservation of variance law. The number of possible decompositions increases with K and we regard the $K \geq 2$ cases as an important aspect of our proposed methodology since multi-level variances are not well understood.

The values of V can be used to represent features of the modeling strategy for $\mathcal{D}_n = \{(x_1 y_1), \ldots, (x_n y_n)\}$ where the x_i's are p-dimensional explanatory variables giving response y_i under some error structure. For instance, trivially, knowing the true model would correspond to $K = 1$ and V_1 equal a constant and the second term in (2) would be zero. As a first nontrivial example, we use our variance decomposition for $K = 1$ in Section 2 to quantify the effect of penalty selection in shrinkage methods on predictive variance. We find, via simulations, that if the penalty varies over a class of penalties the variability cannot be ignored. That is, not knowing the correct penalty to use and representing penalty selection as a hyperparameter substantially affects the predictive variance. This finding is counterintuitive.

When $K = 2$ an early variant of our technique was used for uncertainty quantification. Roughly [2] called V_1 a "scenario" and V_2 may be a "model." This was done in a Bayesian context and our methods can be used to extend Draper's example, see [3].

Here, we will focus on using the LTV in the "E-Var" terms so that there will be a single "Var-E" term (on the right) depending explicitly only on V_1. The idea is that this term—and perhaps V_1—can be omitted if it doesn't affect the predictive variance very

much. So, consider the last term in (2) or (4). Regardless of the distribution used to take the variance, there are two basic ways we can get $\mathrm{Var}\left(E\left(Y_{n+1};V,\mathcal{D}_n\right)\right)=0$. First, the distribution of $V=V_1$ concentrates at a single value $V=v_1$. Second, the models, that is, values of V that get non-zero weights, have the same predictions given $\mathcal{D}$[1]. That is,

$$E\left(Y_{n+1};V=v_1,\mathcal{D}\right)=E\left(Y_{n+1};V=v_2,\mathcal{D}\right)$$

at least approximately, for any v_1 and v_2 getting positive weight. Solving for a set like

$$I_{n+1}(\mathcal{D},c)=\left\{v\mid E\left(Y_{n+1};V=v,\mathcal{D}\right)=c\right\}$$

amounts to inverting an integral operator which is an intractable problem. However, by carefully selecting the models $V=v$ to ensure they are meaningfully different, (e.g., the models are parameterized so that the posterior means are unique), and having a large enough n, the chance of I_{n+1} being both nonvoid and larger than a singleton set will be vanishingly small. Thus, on pragmatic grounds, if the last terms that explicitly depend only on a single component of V are small, we can simply set V_1 to be a constant meaning that the level of modeling it represents drops out. In the case of (4) we would be left with only the first two terms on the right hand side that depend on V_2 in which V_1 was a constant. The resulting expression reduces to (2).

If we write $\mathrm{Var}_{V_K}\left(Y_{n+1};\mathcal{D}_n\right)$ to mean the predictive variance using a specific choice of V_K, it is easy to see, in general, that for another choice, say, $V'_{K'}$, we will usually find $\mathrm{Var}_{V_K}\left(Y_{n+1};\mathcal{D}_n\right)\neq\mathrm{Var}_{V'_{K'}}\left(Y_{n+1};\mathcal{D}_n\right)$. On the other hand, the relative sizes of terms in decompositions of the form (2) depend delicately on the choice of K and V_K and the order in which each successive V_k is introduced. Consequently, while the most important test is the fraction of the total predictive variance represented by the last term, that is, for some preselected $\tau>0$

$$H_0:E\left(\frac{\mathrm{Var}_{V_1}\left(E\left(Y_{n+1};\mathcal{D}_n,V_1\right)\right)}{\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)}\right)\geq\tau$$

our bootstrap testing procedure applies to any ratio of terms. Indeed, there is a parallel between our variance decompositions using a V_K and Cochran's theorem in ANOVA using K factors. Therefore, our bootstrap tests can be seen as a variation the standard F-tests, see Section 3.2, in that they are ratios of terms that look like squared errors, even though the hypotheses are quite different. On the other hand, our tests include ratios of "between group variance" and "total variance" which are not independent and don't have an explicit degrees of freedom whereas F-tests are a ratio of "between group variance" and "within group variance" that are independent and do have an explicit degrees of freedom.

The structure of this paper is as follows. We begin in Section 2 with a $K=1$ example to show how our methodology assesses the contribution to predictive variance from penalty selection in shrinkage methods. Penalty selection is mathematically equivalent to prior selection so our example amounts to assessing the predictive effect of a discrete uniform hyperprior. Section 3 presents our full method with justifications. One subsection explains our predictive variance decomposition in the context of Cochran's theorem and another subsection gives our testing procedure for the terms in our variance decomposition. In Section 4, we give details on two implementations of stacking in the context of a real data example. The first, used in Section 2, is a stabilized "full" stacking method that we advocate for small sample sizes. The second, "iterative" stacking, is a block coordinate descent method that does not require a stabilization step and can be directly applied in a $K\geq 2$ decomposition. We use the stabilized "full" stacking method for a $K=1$ decomposition where V_1 represents model choice and but then only iterative stacking with a $K=2$ decomposition where V_1 represents variable choice. Our examples here are limited to $K=2$ problems where V_1 is binary. This is only due to the complexity of coding not anything conceptual. In Section 5, we discuss the implications of our overall contribution.

2 | A Simulated Example

An example will show the importance of including the last term in (2).

There has been much discussion about when different shrinkage methods are appropriate, see [5] for instance. The consensus from simulations and applications seems to be that for easy, general use LASSO or Elastic Net (EN, a generalization of LASSO) are usually best when there is enough sparsity in the data and multicollinearity is not a problem; see [6]. Otherwise, when sparsity is low or multicollinearity is a problem, ridge regression is often preferable. In this section, we show that our variance decomposition provides a more formal basis for this intuition.

The question is whether we should choose a single shrinkage method for predictive purposes or use several shrinkage methods and combine their results. Combining multiple shrinkage methods effectively retains model variability, which may be desirable to ensure the nominal coverage from a prediction scheme equals the actual coverage. Otherwise put, is retaining the extra variability from using multiple shrinkage techniques predictively useful compared to selecting a single one?

Let's generate data as follows. Set $n=50$ and $p=100$ and write the linear model

$$Y_i=X_i^T\beta+\varepsilon_i \tag{5}$$

for $i=1,\ldots,n$ where X_i is a vector of explanatory variables with $\dim\left(X_i\right)=\dim(\beta)=p<n$ and $\varepsilon_i\sim N(0,1)$ IID. Take 95 of the β_j coefficients to be zero and five to be generated independently from a $N\left(5,(1.5)^2\right)$. Next, let V be a uniform random variable taking values $m_1,\ldots,m_5$ corresponding to five penalized methods, namely LASSO, Ridge Regression (RR), Adaptive LASSO (ALASSO), EN, and Adaptive EN (AEN), respectively.

Let us apply the two term variance decomposition in (2) and consider the following reasoning. For the sparse data we generated, $\mathrm{Var}_V E\left(Y_{n+1};\mathcal{D}_n,V\right)$ should be small relative to $\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)$ because $P(V=EN;\mathcal{D})$ should be near one and the probabilities of other values of V should be near zero. The reason is that (i) adaptive methods have so many parameters they often perform poorly, (ii) RR is usually only good for non-sparse problems whereas here we have sparsity, and (iii) EN has only one

more parameter than LASSO and includes an extra sense of error, L^2, and so EN should be preferred—the cost of one parameter is small so EN should give results better than (or at least no worse than) LASSO. Pre-data, therefore, our intuition is that testing

$$H_0 : E\left(\frac{\mathrm{Var}_V E(Y_{n+1}; \mathcal{D}_n, V)}{\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)}\right) \geq 0.05$$

versus

$$H_1 : E\left(\frac{\mathrm{Var}_V E(Y_{n+1}; \mathcal{D}_n, V)}{\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)}\right) < 0.05$$

will reject the null, meaning we can drop the second term in (2) at the 0.05 level and simply use EN.

As will be described in Section 3.3, this test can be performed by bootstrapping the argument of the expectation in the null hypothesis. We are effectively forced to this sort of test because we do not have a likelihood for expectation in H_0 given the data. Heuristically, for normal error, the distributions of the numerator and the denominator can be regarded as, approximately, convex combinations of χ^2 distributions, see Deriving a χ^2 distribution for $K = 2$. So, their ratio is expected to behave like an F distribution. However, even though the convex combinations can be precisely defined they are generally numerically inaccessible. Nevertheless, our testing procedure can be regarded as a pragmatic nonparametric alternative to standard normal theory.

Let's use the first 49 data points to form a predictive distribution for the 50th data point for each of the five methods and for the stacking average of the five methods. To obtain the stacking weights $\hat{w}_1, \ldots, \hat{w}_5$ under the non-negativity and sum-to-one constraint, we use the methodology described in pp. 283–284 of [7]. Briefly, [7] obtains stacking weights by minimizing the cross-validation error. Hence, for $i = 1, \ldots, n$, we internally predict Y_i using

$$\tilde{Y}_i = \sum_{j=1}^{M} w_j \tilde{Y}_{j,i}$$

where w_j is the stacking weight corresponding to the j-th model, $\tilde{Y}_{j,i}$ is the prediction generated from the j-th model for the i-th test data point, $j = 1, 2, \ldots, M$. Following [7] write

$$\boldsymbol{e}_{(j)}^{1\times n} = \left(e_{ij}\right)_{i=1}^{n} = \left(Y_i - \tilde{Y}_{j,i}\right)_{i=1}^{n}$$

as the vector of cross-validation errors produced by the j-th model and $\tilde{\boldsymbol{e}}^{n\times M} = \left(\boldsymbol{e}_{(1)}, \boldsymbol{e}_{(2)}, \ldots, \boldsymbol{e}_{(M)}\right)$ as the collection of cross validation errors from all M models. Then, denoting $\boldsymbol{w} = \left(w_1, w_2, \ldots, w_M\right)$, the optimization problem becomes

$$\begin{aligned} &\underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{n}\boldsymbol{w}^T \tilde{\boldsymbol{e}}^T \tilde{\boldsymbol{e}} \boldsymbol{w} \\ \text{subject to } & w_j \geq 0, j = 1, 2, \ldots, M \\ & \boldsymbol{w}^T \mathbf{1} = 1 \end{aligned} \tag{6}$$

In theory, the positive-semidefiniteness of $\boldsymbol{e}^T\boldsymbol{e}$ guarantees the convexity of the loss function in (6). However, to achieve numerical stability, particularly when $\boldsymbol{e}^T\boldsymbol{e}$ becomes ill-conditioned, we have to project this Gram matrix to the nearest positive definite matrix [8]. We refer to this optimization as "stabilized full" stacking as against the "iterated stacking" procedure that we introduce in Section 4.2 which uses component-wise gradient descent algorithm and does not require projection of the analog of $\boldsymbol{e}^T\boldsymbol{e}$ to its nearest PD matrix.

Since the `glmnet` package is easy to use and computationally fast, obtaining the stacking coefficients in this example is straightforward. Generically, write the stacking model average as

$$\sum_{j=1}^{5} \hat{w}_j(\mathcal{D}_{49})\hat{p}(Y_{50}; X_{50}, m_j) \tag{7}$$

where the $\hat{w}_j$'s are the stacking weights and the dependence of the $\hat{\beta}_j$'s in the linear model is indicated by $\hat{p}$. More explicitly,

$$\hat{p}(Y_{50}; X_{50}, m_j) = N\left(X_{50}\hat{\beta}_{m_j}, \hat{\sigma}^2_{m_j} + \widehat{\mathrm{Var}}\left(X_{50}\hat{\beta}_{m_j}\right)\right) \tag{8}$$

The use of normality in (8) comes from the normality in the errors in (5). As can be seen, we are neglecting the variability in the parameter estimators. We think we have enough data that assuming the predictive distribution is normal will not be too far wrong, for example, if it is a t-distribution the degrees of freedom will be large enough that it is close enough to normal that the difference can be neglected, at least when compared with other sources of error. In (8), apart from RR, (about which shortly) the $\hat{\beta}$'s are the "usual" shrinkage estimators where the estimation of the decay parameters λ_j is suppressed in the m_j's. Thus, to find $\widehat{\mathrm{Var}}\left(X_{50}\hat{\beta}_{m_j}\right)$ we use the bootstrapped variance estimator from the boot package in R:

$$\widehat{\mathrm{Var}}\left(X_{50}\hat{\beta}_{m_j}\right) = \frac{1}{B}\sum_{i=1}^{B}\left(Y_i - \tilde{Y}_i\right)^2$$

More generally, bootstrapping can be used even when normality is violated. Also, in (8), again except for RR, the $\hat{\sigma}^2_{m_j}$'s used are from the standard OLS estimator of σ^2 based on the variables selected by m_j. We justify this by citing [9] who showed that doing this would be consistent for LASSO and we also observe that the proof can be extended to EN and, we think, to any shrinkage method with the oracle property (e.g., AEN and ALASSO). Returning to RR, which is not a sparsity criterion, we used $\hat{\beta}_{RR}$ and $\hat{\sigma}^2_{RR}$ from regressing on the variables selected by EN on the grounds that EN uses a combination of the L^1 and L^2 penalties and we simply think it will be close enough that the results won't be too far wrong.

The first two rows of Table 1 give the stacking weights and predictive variances for the five penalized regression models. In addition, the first entry in the second row is the predictive variance of the stacking average computed from the stacking weights and predictive variances for the models in the following way:

$$\begin{aligned} 3.01 = \widehat{\mathrm{Var}}(Y_{50}; \mathcal{D}_{49}) &= \widehat{E_V\left(\mathrm{Var}(Y_{50}; V, \mathcal{D}_{49})\right)} \\ &+ \widehat{\mathrm{Var}_V\left(E(Y_{50}; V, \mathcal{D}_{49}); \mathcal{D}_n\right)} = 2.43 + 0.58 \end{aligned} \tag{9}$$

where the first term on the RHS $\left(\widehat{E_V(\mathrm{Var}(.;.))}\right)$ is given by $\sum_{j=1}^{5}\hat{w}_j\widehat{\mathrm{Var}}\left(X_{50}\hat{\beta}_{m_j}\right)$, and the second term on RHS $\left(\widehat{\mathrm{Var}_V(E(.;.))}\right)$

TABLE 1 | Stacking shrinkage methods: This table gives the stacking weights and the variances of the predictive distributions for the five shrinkage methods and their stacking average.

	STK avg	LASSO	RR	ALASSO	EN	AEN
Stacking weights		0.74	0.00	0.00	0.25	0.00
Pred. variance	3.01	1.02	6.71	0.99	6.73	6.70

is given by $\sum_{j=1}^{5}\widehat{w}_j\left(X_{50}\widehat{\beta}_{m_j}-\sum_{j=1}^{5}\widehat{w}_j X_{50}\widehat{\beta}_{m_j}\right)^2$. In effect, we are treating $\mathrm{Var}(E(\cdot\cdot);\cdot)$ and $E(\mathrm{Var}(\cdot\cdot);\cdot)$ as single operations. Hence we see the ratio of the between-models variance to total variance is

$$\frac{\mathrm{Var}_V\,\widehat{E}\left(Y_{50};V,\mathcal{D}_n\right)}{\mathrm{Var}\left(Y_{50};\mathcal{D}_n\right)}=\frac{0.58}{3.01}=0.19$$

Informally, the ratio is high enough that it suggests there is too much between-models variance to ignore when making predictions. More formally, we can use the foregoing test of hypotheses to arrive at a decision rule. We resort to a bootstrap procedure to generate the null distribution of

$$\frac{\mathrm{Var}_V\,\widehat{E}\left(Y_{50};V,\mathcal{D}_n\right)}{\mathrm{Var}\left(Y_{50};\mathcal{D}_n\right)}$$

the details on enforcing the null hypothesis are in Section 3.3. Once samples from the null distribution are obtained, we compute a bootstrapped p-value, commonly called the achieved significance level (ASL). The null hypothesis is rejected when the ASL is smaller than the specified level of significance.

Using our bootstrap-based test, we obtain an $\widehat{\mathrm{ASL}}=0.99$ meaning we cannot reject the null at any reasonable level. This leads us to conclude that the second term on the LHS of (2) contributes more than 5% of the total predictive variance. Consequently, we should account for penalty uncertainty when making predictions. This confirms our initial intuition.

Going beyond the information provided by a single use of our test we ask if we allowed ourselves to ignore a larger proportion of variance—that is, increase the threshold τ in H_0—at what threshold could we reject H_0? We observe that, if we change the RHS of H_0 and H_1 to 0.09 instead of 0.05, our test gives an $\widehat{\mathrm{ASL}}=0.0095$. Hence, we would conclude that 9% is roughly the smallest percentage at which we could ignore the contribution of the between-models variance to the overall variance. We emphasize that when we calculate the components in (9), that is, computing stacking weights and prediction variances, we use the entire training data set. However, when we perform the test, we use bootstrapping and we recompute stacking weights and predictive variances for each bootstrap replicate.

3 | Decomposing the Predictive Variance

In this section, we give our general variance decomposition, indicate how to choose amongst different candidate variance decompositions, and explain our testing procedure for the relative size of their terms. We will see that our decomposition of the predictive variance parallels Cochran's theorem decomposition of the squared error into quadratic forms.

3.1 | The Effect of the Model List on Overall Variance

Consider a model list $\mathcal{M}$ and suppose we don't believe it adequately captures the uncertainty (including mis-specification) of the predictive problem. This may lead us to expand $\mathcal{M}$ and this can be done by adding more models to it or by embedding the models on the list in various "scenarios" as is done in [2]. Expanding the list simply by including more plausible models may lead to problems such as dilution; see [4]. So, we are led preferentially to Draper's approach. Moreover, we want to assess the effect of a model list on the variance of predictions.

In the simplest case, expanding $\mathcal{M}$ to $\mathcal{M}'$ where $\mathcal{M}\subset\mathcal{M}'$, where $\mathcal{M}'$ has models with positive probability that are not in $\mathcal{M}$ the predictive distribution $p\left(Y_{n+1};\mathcal{D}_n\right)$ using $\mathcal{M}'$ will be different from the predictive distribution using $\mathcal{M}$. Recalling that we are using the stacking model average we denote dependence on a model list $\mathcal{M}$, when we need it, by

$$p\left(Y_{n+1};\mathcal{D}_n\right)=p\left(Y_{n+1};\mathcal{D}_n\right)(\mathcal{M})$$

In this notation, we think of $\mathcal{M}=\left\{m_1,\ldots,m_M\right\}$, each model having a positive weight w_m so that the sum over $m=1,\ldots,M$ gives one. In this case, we have $\dim(V)=1$ and treating the model list as one factor with M levels is fine. However, if we want to expand a given $\mathcal{M}$ by including more scenarios, we are led to choosing a V with $\dim(V)\geq 2$ and regarding $\mathcal{M}$ as having a corresponding multivariate structure. In either case, in our variance decomposition below, V encapsulates dependence on the model list. Ideally, this dependence is by conditioning (and so we should use | to indicate it). However, we continue to use ; because we are taking expectations in the stacking distribution for Y_{n+1}.

3.1.1 | Predictive Variance Decomposition "P-ANOVA"

To quantify the uncertainty of our chosen features, treat V as multivariate and recall $V=\left(V_1,\ldots,V_K\right)$, where V_k represents the values of the k-th potential feature that must be made to specify a predictor. Analogous to terminology in ANOVA, we call V_k a *factor* in the prediction scheme, and we call the $v_{k1},\ldots,v_{km_k}$ the *levels* of V_k. That is, $v_{k\ell}$ is a specific value that a specific V_k may assume. Thus, V is discrete and has probability mass function $W(v)=W\left(V_1=v_1\ldots,V_K=v_K\right)$. The V_k's are not in general independent and W corresponds to a prior on V here given by the stacking weights.

Perhaps the most natural way to represent $\mathcal{M}$ is to replace it by

$$\mathcal{V}^K=\left\{v_{11},\ldots,v_{1I}\right\}\times\ldots\times\left\{v_{K1},\ldots,v_{KI}\right\}\qquad(10)$$

where it is understood that a model is uniquely identified by a vector in $\mathcal{V}^K$. There are now $M=IK$ distinct models in $\mathcal{V}^K$ and they have a hierarchical structure along with a joint prior. This representation of our model list makes it easy to interpret

expansions and contractions of $\mathcal{M}$ in terms of the features of the modeling.

Our first result gives a decomposition of the predictive variance by conditioning on V.

Proposition 1. *We have the following two expressions for the stacking predictive variance.*

Clause (i): For $K = 1$, the stacking predictive variance for Y_{n+1} is

$$\mathrm{Var}\left(Y_{n+1}; \mathcal{D}_n\right)\left(\mathcal{V}^K\right) = E_{V_1}\left(\mathrm{Var}\left(Y_{n+1}; V_1, \mathcal{D}_n\right) + \mathrm{Var}_{V_1} E\left(Y_{n+1}; V_1, \mathcal{D}_n\right)\right)$$

and for $K \geq 2$, the stacking predictive variance for Y_{n+1} as function of the K factors defining our predictive scheme is given by

$$\begin{aligned}
&\mathrm{Var}\left(Y_{n+1}; \mathcal{D}_n\right)\left(\mathcal{V}^K\right) \\
&= E_{(V_1,\ldots,V_K)}\mathrm{Var}\left(Y_{n+1}; V_1, \ldots, V_K, \mathcal{D}_n\right) \\
&\quad + \sum_{k=2}^{K} E_{(V_1,\ldots,V_{k-1})}\mathrm{Var}_{V_k} E\left(Y_{n+1}; V_1, \ldots, V_k, \mathcal{D}_n\right) \\
&\quad + \mathrm{Var}_{V_1} E\left(Y_{n+1}; V_1, \mathcal{D}_n\right) \qquad (11)
\end{aligned}$$

where the distribution of $V = (V_1, \ldots, V_K)$ is defined by the stacking weights.

Clause (ii): For any K, the stacking predictive variance $\mathrm{Var}\left(Y_{n+1}; \mathcal{D}_n\right)\left(\mathcal{V}^K\right)$ can be condensed into a two term decomposition:

$$\mathrm{Var}\left(Y_{n+1}; \mathcal{D}_n\right)\left(\mathcal{V}^k\right) = E_{(V_1,\ldots,V_K)}\mathrm{Var}\left(Y_{n+1}; V_1, \ldots, V_K, \mathcal{D}_n\right) + \mathrm{Var}_{(V_1,\ldots,V_K)} E\left(Y_{n+1}; V_1, \ldots, V_K, \mathcal{D}_n\right) \qquad (12)$$

Proof. Clause (i) follows by induction: The case $K = 1$ is (2) The case $K = 2$ results from one iteration with the law of total variance as in (4). Then for any given value of K repeat this $K - 2$ times and replace the posterior weights with stacking weights. Obviously, the corresponding result holds if the posterior weights are kept. For Clause (ii), simply use the law of total variance on the whole range of the vector V_K. □

In this result, we have only used the LTV in one sequence of K iterations. In fact, the LTV can be used in any term (provided a V_k exists for it) and the V_k's can be introduced in any sequence. Moreover, we may use the entire V_K in the left hand side but leave some of the V_k's as "latent," that is, not explicitly appearing, on the right side.

We summarize the decomposition in (11) using what we call "P-ANOVA," or *predictive analysis of variance*. In Table 2, each row corresponds to a different source of variability associated with the factors in V. Note that the interpretation "Expected between V_j across $V_{j-1}, \ldots, V_1$" for the term

$$E_{V_1} \ldots E_{V_{j-1}} \mathrm{Var}_{V_j} E\left(Y_{n+1}; V_1, V_2, \ldots, V_j, \mathcal{D}_n\right)$$

means we have averaged the variance due to V_j across all the values

$$\mathrm{Var}_{V_j} E\left(Y_{n+1}; V_1 = v_1, V_2 = v_2, \ldots, V_{j-1} = v_{j-1}, V_j, \mathcal{D}_n\right)$$

Using the Bayes model average—or any other model averaging procedure—in place of stacking leads to a P-ANOVA table analogous to Table 2.

3.2 | Analogy to Cochran's Theorem

Cochran's theorem is used in standard ANOVA problems to identify hypothesis tests that determine whether a factor or its levels should be dropped as having little effect on the observed variability. Informally, the theorem states that, under various regularity conditions, the corrected sum of squares from an ANOVA problem can be written as a sum of independent quadratic forms each of which is distributed as a χ^2 random variable with a degrees of freedom specified by the statement of the problem. Equivalently, the sum of squares "$Y^T Y$" can be written as a sum of scaled χ^2_1 random variables, where the scaling constants are eigenvalues from the corresponding quadratic form. More formally, we have the following distilled from [10] appendix VI.

Theorem 1 (*Cochran's Theorem*). *Let $Y_i \sim N(\mu, 1)$ for $i = 1, \ldots, n$ be independent. Suppose $Q_1, \ldots, Q_s$ are quadratic forms of rank $n_1, \ldots, n_s$ respectively in variables $Y_1, \ldots, Y_n$ and $\sum_{i=1}^{n} y_i^2 = Q_1 + \ldots + Q_s$. Then, $n_1 + \ldots + n_s = n$ if and only if $Q_1 + \ldots + Q_s$ are independent $\chi^2_{n_j}(\Delta_j)$ where the noncentrality parameters in the χ^2's are $\Delta_j^2 = Q_j(EY_1, \ldots, EY_n)$ for $j = 1, \ldots, s$. Then, if $Z \sim \chi^2_\nu$ is independent of Q_j,*

$$F_j = \frac{\nu}{n_j}\frac{Q_j}{Z} \sim F_{n_j,\nu}$$

TABLE 2 | Sources of predictive variation for $K \geq 3$. We have listed the generic terms in our decomposition of the predictive variance together with their interpretations. Following the conventions of ANOVA, we have also listed the source of the variability. The Source labeled Predictions is analogous to the correction term in the corrected sum of squares. All terms are dependent on $\mathcal{D}_n$, but not necessarily in a conditional sense.

Source	Interpretation	Variance
V_1, m_1 levels	Between V_1 variance	$\mathrm{Var}_{V_1} E\left(Y_{n+1}; \mathcal{D}_n, V_1\right)$
V_2, m_2 levels	Expected between V_2 across V_1	$E_{V_1}\mathrm{Var}_{V_2} E\left(Y_{n+1}; \mathcal{D}_n, V_1, V_2\right)$
⋮	⋮	⋮
V_K, m_K levels	Expected between V_K across $V_{K-1} \ldots V_1$	$E_{V_1} \ldots E_{V_{K-1}}\mathrm{Var}_{V_K} E\left(Y_{n+1}; \mathcal{D}_n, V_1, V_2, \ldots, V_K\right)$
Predictions	Expected variance across $V_1 \ldots V_k$	$E_{V_1} \ldots E_{V_K}\mathrm{Var}\left(Y_{n+1}; \mathcal{D}_n, V_1, V_2, \ldots, V_K\right)$
Total	Posterior predictive variance	$\mathrm{Var}\left(Y_{n+1}; \mathcal{D}_n\right)$

Next, we argue that our decomposition, see (11), is a predictive analog to Cochran's Theorem. In our analog, we expand the predictive variance into a sum of quadratic forms that have χ^2 distributions, as shown in Appendix A. However, we do not obtain the analogous statements about degrees of freedom or independence. Nor do we obtain F-tests. However, in Section 3.3, we describe a bootstrap based testing procedure for the individual terms in our expansions so as to determine if they contribute substantially to the overall predictive variance. Our results are fundamentally different from [11] who gave an "ANOVA" like decomposition of the posterior variance for estimation because we have used the ANOVA framework in a predictive setting and proposed hypothesis tests.

As an illustration of how our variance decomposition resembles Cochran's Theorem, we explicitly convert the terms in a three term decomposition to a convex combination of quadratic forms. Consistent with the notation of [2], we write s_i to represent "scenarios" $i = 1, \ldots, I$ and m_{ij} to represent models within scenarios, $j = 1, \ldots, J$. Now, the s_i's correspond to the values of V_1 and the m_{ij}'s correspond to values of V_2 nested within V_1. Our strategy is to express each term in $\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)$ in vector notation so we can recognize analogous quadratic forms. Now, Proposition 1 gives

$$\begin{aligned}
\mathrm{Var}(Y_{n+1}; \mathcal{D}_n) &= E_{V_1} E_{V_2} \mathrm{Var}(Y_{n+1}; \mathcal{D}_n, V_1, V_2) \\
&\quad + E_{V_1} \mathrm{Var}_{V_2} E(Y_{n+1}; \mathcal{D}_n, V_1, V_2) \\
&\quad + \mathrm{Var}_{V_1} E(Y_{n+1}; \mathcal{D}_n, V_1) \\
&= \sum_{i=1}^{I} p(s_i; \mathcal{D}_n) \sum_{j=1}^{J} p(m_{ij}; \mathcal{D}_n, s_i) \\
&\quad \times \mathrm{Var}(Y_{n+1}; \mathcal{D}_n, s_i, m_{ij}) \\
&\quad + \sum_{i=1}^{I} p(s_i; \mathcal{D}_n) \sum_{j=1}^{J} p(m_{ij}; \mathcal{D}_n, s_i) \\
&\quad \times \left[E(Y_{n+1}; \mathcal{D}_n, m_{ij}, s_i) - E(Y_{n+1}; \mathcal{D}_n, s_i) \right]^2 \\
&\quad + \sum_{i=1}^{I} p(s_i; \mathcal{D}_n) \\
&\quad \times \left[E(Y_{n+1}; \mathcal{D}_n, s_i) - E(Y_{n+1}; \mathcal{D}_n) \right]^2 \qquad (13)
\end{aligned}$$

For ease of notation, let

- $p(s_i; \mathcal{D}_n) = \xi_i$
- $p(m_{ij}; \mathcal{D}_n, s_i) = \omega_{ij}$
- $E(Y_{n+1}; \mathcal{D}_n) = \overline{y}$
- $E(Y_{n+1}; \mathcal{D}_n, s_i) = \overline{y}_{i\cdot}$
- $E(Y_{n+1}; \mathcal{D}_n, m_{ij}, s_i) = \hat{y}_{ij}$.

Now we can restate (13) as

$$\mathrm{Var}(Y_{n+1}; \mathcal{D}_n) = \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} \mathrm{Var}(Y_{n+1}; \mathcal{D}_n, m_{ij}, s_i) \qquad (14)$$

$$+ \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} (\hat{y}_{ij} - \overline{y}_{i\cdot})^2 \qquad (15)$$

$$+ \sum_{i=1}^{I} \xi_i (\overline{y}_{i\cdot} - \overline{y})^2 \qquad (16)$$

The quadratic forms in (14–16) can be shown to have distributions that parallel the distributional statements in Cochran's Theorem. This is given in detail in Deriving a χ^2 distribution for $K = 2$ and General K.

3.3 | Testing

In the ANOVA context, it is common to test the equality of levels of a factor. Here, the corresponding null hypothesis would be the equality of expectations of the predictive distributions within a factor or the model weight being close to one for a single level within a factor. So, we rephrase these tests as a way to determine the relative importance of terms in our decomposition. Specifically, we want to test whether a term in the variance decomposition is a substantial fraction of the overall variance.

Consider the case $K = 1$ that gives a two-term decomposition for $\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)$. Now, we want to test hypotheses of the form

$$H_0 : E\left(\frac{\mathrm{Var}_{V_1}(E(Y_{n+1}; \mathcal{D}_n, V_1))}{\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)} \right) \geq \tau$$

$$H_1 : E\left(\frac{\mathrm{Var}_{V_1}(E(Y_{n+1}; \mathcal{D}_n, V_1))}{\mathrm{Var}(Y_{n+1}; \mathcal{D}_n)} \right) < \tau$$

for some pre-selected value of $\tau > 0$. Since we do not have a likelihood for the argument of the expectation in H_0, we are led to a nonparametric test based on bootstrapping. Our bootstrapping procedure to enforce the null is as follows.

First, we generate bootstrap replicates of the argument of the expectation in H_0. This results in a set of Z_b given by

$$\begin{aligned}
Z_b &= \frac{\mathrm{Var}_{V_1} E(Y_{n+1}; \mathcal{D}_n^b, V_1)}{\mathrm{Var}(Y_{n+1}; \mathcal{D}_n^b)} \\
&= \frac{\sum_{v_1=1}^{m_1} \hat{w}_{v_1}^b \left(\hat{y}_{v_1}^b - \sum_{v_1=1}^{m_1} \hat{y}_{v_1}^b \right)^2}{\sum_{v_1=1}^{m_1} \hat{w}_{v_1}^b \left(\hat{y}_{v_1}^b - \sum_{v_1=1}^{m_1} \hat{y}_{v_1}^b \right)^2 + \sum_{v_1=1}^{m_1} \hat{\sigma}_{v_1}^2(b)}
\end{aligned} \qquad (17)$$

for $b = 1, \ldots, B$. Each Z_b can be regarded as a random variable representing $\frac{\mathrm{Var}_{V_1}(Y_{n+1}|\mathcal{D}_n, V_1)}{\mathrm{Var}(Y_{n+1}|\mathcal{D}_n)}$. We note that none of the quantities in this formula rely on a specific distribution. The estimates $\hat{w}_j^b$ are numerically obtained as a solution to (6), $\hat{y}^b$ takes the form of the predictor from the specific j-th model, and $\hat{\sigma}_j^2(b)$ is the estimated predictive variance from the j-th model. These quantities do not have specific formulas because they depend on the model being used. Writing $\overline{z}$ and $\mathrm{SE}(\overline{z})$ for the mean and its standard error for the Z_b's we form

$$t = \frac{\overline{z} - \tau}{\mathrm{SE}(\overline{z})} \qquad (18)$$

Note that $\overline{z}$ is (mild) abuse of notation. In fact, we should write the Z_b's with "hats" over the variances and expectations since we are bootstrapping. This is an important point but we do not wish to clutter the notation.

Next, we must generate samples from the null distribution of the random variable T taking values t in (18). Hence, for the second

layer of bootstrapping, draw C samples of size B, with $C > B$, from the set of realizations $z_1, \ldots, z_B$, with replacement. Denote these by $z'_1, \ldots, z'_C$ where each z'_c has B entries. To get a distribution for the random variable T under the null, we generate vectors

$$\tilde{z}'_c = z'_c - \left(\frac{1}{B}\sum_{b=1}^{B} z_b - \tau\right)1_B$$

where $1_B = (1, \ldots, 1)$ is the usual B-dimensional vector of 1's. Now, we have C different samples of the vector $\tilde{z}'_c$ with mean τ. Since these realizations of $\tilde{z}'_c$ are corrected by their means and τ, so they satisfy the null. Hence, realizations from the null distribution of the test statistic described in (18) are obtained as

$$\tilde{t}_c = \frac{\overline{\tilde{z}}'_c - \tau}{\mathrm{SE}\left(\overline{\tilde{z}}'_c\right)}$$

for $j = 1, \ldots, J$ and we calculate the estimated achieved significance level

$$\widehat{\mathrm{ASL}} = \frac{1}{C}\sum I\left(\tilde{t}_c \le t\right)$$

When the $\widehat{\mathrm{ASL}}$ is small, we reject H_0 and this tells us that $\mathrm{Var}_{V_1} E\left(Y_{n+1}; \mathcal{D}_n, V_1\right) \approx 0$ suggesting that $E\left(Y_{n+1}; \mathcal{D}_n, V_1\right)$ is constant in V_1. Here, when we do this testing, we default to a threshold of $\alpha = 0.05$ for the ASL for convenience. Note that this threshold α for testing is compared with the usual p-value and so is different from the threshold τ for the expected ratio of variances. As a generality, both α and τ should be chosen pre-experimentally.

Below we have used normality in some of our computational work because it was justified by auxiliary reasoning. However, when the normal assumption fails, we would use parameter estimators based on the actual family if it were known, defaulting to standard estimators for variance, for instance, in the hope they would be effective. Otherwise, our bootstrapping approach allows us to move beyond the assumption that the predictions follow a normal distribution as used in the discussion at the end of Section 3.2 and in Proposition 3 in General K because we can generate the bootstrap sampling distribution for any parameter estimator.

As a final point about the testing, we comment on multiple comparison issues. Here we have shown the $K = 1$ case for simplicity, but the testing procedure can be used for general K to test if each term in the variance is important. Hence, we may be interested in $K + 1$ tests. For small K, a Bonferroni correction or other simple "fix" may be practical. However, for large K, we may have to use some sort of Westfall-Young correction, see [12], since our testing procedure is in the same spirit as permutation tests. On the other hand, because we interpret the components of V as components of modeling, large K's will be uncommon with small sample sizes.

4 | A Real Data Example

In this section, we analyze the `Superconductivity` data presented in [13]. This data set has 81 regressors of a physical or chemical nature to explain a response Y representing temperature measurements (in °K) for when a compound begins to exhibit superconductivity. The full data set has $n = 21263$, and we assume the relationship between Y and the regressors (X) follows a signal plus noise structure, that is,

$$Y_i = f\left(X_i\right) + \varepsilon_i$$

for $i = 1, \ldots, n$ and where $\varepsilon_i \sim N\left(0, \sigma^2\right)$. Hamidieh [13] used a linear model (LM) as a "benchmark model" and then improved on it by developing an XGBoosting model—a boosted, penalized tree model. The goal in their paper was to minimize predictive error on a hold out set. So, they did not consider the variance of predictive distributions.

Our analysis of the predictive variance rests on computing variance decompositions for specific choices of V_K for $K = 1$ and $K = 2$ using Proposition 1. For a two term decomposition, that is, $K = 1$, we use the optimization (6) as discussed in Section 2 to estimate the stacking weights along with the test procedure outlined in Section 3.3. It is important to note that with $K = 1$, we are estimating a singly indexed set of stacking weights w_j so that solving (6) is possible and will give the optimal solution.

By contrast, for a three term decomposition, that is, $K = 2$, (6) cannot be immediately used as written because the optimization problem is to find two sets of stacking weights—the ξ_i's and the ω_{ij}'s in (14–16). To obtain the estimates of ξ_i and ω_{ij} directly, we introduce "iterative" stacking, as an extension of (6), to obtain test results in three term or higher cases. This is in contrast to the "stabilized full" method stacking (6) used in Section 2.

4.1 | Two Term Decompositions, $K = 1$

Here, we choose a single random variable V and generate a two-term decomposition of predictive variance. So, let V take one of five values with equal probability, namely five common predictive models: (a) linear model (LM) denoted by m_1, (b) neural nets (NN) denoted by m_2, (c) projection pursuit regression (PPR), denoted by m_3, (d) support vector machine with a radial kernel (SVM), denoted by m_4, and (e) XGBoosting (XGB), denoted by m_5. Upon examining the residuals from the individual fitted models, we confirmed that the residuals were normally distributed. So, for convenience, we use a normal density to form predictive distributions for each of the models. Moreover, to form the predictive distribution for each model, we fit the model using n data points and used the $n + 1$ value if the explanatory variables to predict Y_{n+1}.

Denote the predictor from model k by $\hat{f}_k$, $k = 1, \ldots, 5$. Then, under the signal plus noise model, the next outcome is normally distributed, centered at the point predictor $\hat{f}_k\left(X_{n+1}\right)$ with estimated variance

$$\widehat{\mathrm{Var}}\left(Y_{n+1} - \hat{f}_k\left(X_{n+1}\right)\right) = \widehat{\mathrm{Var}}\left(\hat{f}_k\left(X_{n+1}\right)\right) + \widehat{\mathrm{Var}}\left(\hat{\varepsilon}_k\right) \tag{19}$$

We calculated $\widehat{\mathrm{Var}}\left(\hat{f}_k\left(X_{n+1}\right)\right)$ by bootstrapping. That is, we obtained a bootstrap distribution for it and then computed its variance. For $\widehat{\mathrm{Var}}\left(\hat{\varepsilon}_k\right)$, we simply calculated the variance of the residuals from the fitted model. Now, formally, the predictive distribution for a model m_k is

$$\hat{p}\left(Y_{n+1}; m_k\right) = N\left(\hat{f}_k\left(X_{n+1}\right), \widehat{\mathrm{Var}}\left(\hat{f}_k\left(X_{n+1}\right)\right) + \widehat{\mathrm{Var}}\left(\hat{\varepsilon}_k\right)\right)$$

Since these models are implemented in a frequentist sense and we used full stacking (as described in (6), see [7]) to average over the models based on the cross-validated predictive performance, the stacked predictive distribution for Y_{n+1} is

$$Y_{n+1} \sim \sum_{k=1}^{5} \hat{w}_k(\mathcal{D}_n)\hat{p}(Y_{n+1}; m_k)$$

Next, we present two cases, one where we randomly sample 500 data points from the `Superconductivity` dataset and test whether the between-models variance is important, and another where we use the whole `Superconductivity` dataset to perform the same test. We will see that with the smaller sample size, the between-models variance term in the decomposition using V contributes about two-thirds of the total predictive variance. However, when the full data set is used, the estimated contribution from the between-models term drops to about 4%.

4.1.1 | Testing Terms When $n = 500$

First, we drew a random sample of 500 observations from the whole data set. Then, we followed the procedure from Section 3.3 using $B = 200$ samples for the inner bootstrap and $C = 10{,}000$ samples for the outer bootstrap. The results are in Table 3.

Using only $n = 499$, the stacking predictive variance decompositions is

$$\begin{aligned} 397.64 = \widehat{\mathrm{Var}}(Y_{500}; \mathcal{D}_{499}) &= \widehat{E_V \mathrm{Var}}(Y_{500}; V, \mathcal{D}_{499}) \\ &+ \widehat{\mathrm{Var}_V E}\big((Y_{500}; V, \mathcal{D}_{499}); \mathcal{D}_{499}\big) \\ &= 135.41 + 262.23 \end{aligned}$$

The terms on the RHS of the above expression are computed using the same technique as described in Section 2. Now, to test whether the between-models variance term matters (the second term in the RHS) we have the hypotheses

$$H_0 : E\left(\frac{\mathrm{Var}_V\left(E\left(Y_{500}; V, \mathcal{D}_{499}\right); \mathcal{D}_{499}\right)}{\mathrm{Var}\left(Y_{500}; \mathcal{D}_{499}\right)}\right) \geq \tau$$

versus

$$H_1 : E\left(\frac{\mathrm{Var}_V\left(E\left(Y_{500}; V, \mathcal{D}_{499}\right); \mathcal{D}_{499}\right)}{\mathrm{Var}\left(Y_{500}; \mathcal{D}_{499}\right)}\right) < \tau$$

and the test statistic $\overline{z} = \frac{262.23}{397.64} = 0.66$. To find a p-value, we use the empirical distribution from the set of bootstrapped values z_b for $b = 1, \ldots, B$. In this case, for $\tau = 0.05$ we obtain $\widehat{\mathrm{ASL}} = 1$ and cannot reject the null. Indeed, in this case, we cannot reject the null for any reasonable value of τ. This confirms what Table 3 showed: V must be included, that is, the between-models variance is too large to ignore.

TABLE 3 | Small sample results for `Superconductivity`: The top row of numbers are the stacking weights that solve (6). The second row shows the predictive variances from each model/column individually. The overall predictive error for the stacking average is larger than the predictive variances for the individual models. We interpret this to mean that the point predictions from the five models have a large spread in addition to their individual variances.

	STK avg	LM	NN	PPR	SVM	XGB
Stacking weights		0.10	0.26	0.12	0.01	0.51
Pred. variance	397.64	237.33	260.46	57.06	172.11	69.39

4.1.2 | Testing Terms When n Large

For contrast we redo the analysis using all the available data. Although we recommend using larger values for B and C, we only used $B = 50$ inner bootstrap samples and $C = 5000$ due to computational burden. The results are given in Table 4.

Now the variance decomposition is

$$\begin{aligned} 173.37 = \widehat{\mathrm{Var}}(Y_{21,263}; \mathcal{D}_{21,262}) &= \widehat{E_V \mathrm{Var}}(Y_{21,263}; V, \mathcal{D}_{21,262}) \\ &+ \widehat{\mathrm{Var}_V}\left(E\left(Y_{21,263}; V, \mathcal{D}_{21,262}\right); \mathcal{D}_{21,262}\right) \\ &= 166.21 + 7.16 \end{aligned} \tag{20}$$

Again, we wish to test if the between models term is a substantial portion of the total predictive variance. The hypotheses are

$$H_0 : E\left(\frac{\mathrm{Var}_V\left(E\left(Y_{21,263}; V, \mathcal{D}_{21,262}\right); \mathcal{D}_{21,262}\right)}{\mathrm{Var}\left(Y_{21,263}; \mathcal{D}_{21,262}\right)}\right) \geq \tau$$

versus

$$H_1 : E\left(\frac{\mathrm{Var}_V\left(E\left(Y_{21,263}; V, \mathcal{D}_{21,262}\right); \mathcal{D}_{21,262}\right)}{\mathrm{Var}\left(Y_{21,263}; \mathcal{D}_{21,262}\right)}\right) < \tau$$

and the test statistic is $\overline{z} = \frac{7.16}{173.37} = 0.04$. To investigate the dependence of the conclusions on τ, we used several values and generated Table 5. For $\tau = 0.05$ there is not enough evidence to say the expected ratio is statistically less than τ, but for $\tau \geq 0.06$ the test rejects the null. That is, the relative contribution of the between-models variance to the total stacking predictive variance is roughly between 5% and 6%. We suggest that if a larger value of B were used, the threshold for rejecting the null would likely decrease to around $\tau = 0.05$.

Overall, with $n = 500$, we could not reject the null at any reasonable value of τ however with the full data set we could reject the

TABLE 4 | Re-analyzing `Superconductivity` with all available data: The predictive variances for individual models are larger than in Table 3 but the overall stacking variance is less than half of the earlier value. This suggests that including the factor V, that is, the between models variance, is less important than with $n = 500$. We interpret this to mean that the spread of the point predictions from individual models using all the data is much less than their spread when $n = 500$ data points are used. As n increases, it seems that even as the individual models have a larger uncertainty, the variability shift from between models to within models enabling us to reject.

	STK avg	LM	NN	PPR	SVM	XGB
Stacking weights		0.01	0.26	0.21	0.01	0.52
Pred. variance	173.37	308.60	315.28	184.14	155.32	78.71

TABLE 5 | $\widehat{\text{ASL}}$ for different choices of τ: For the usual cutoff of 0.05, we can start to reject when $\tau \gtrsim 0.055$. The reliability of the entries is potentially limited because B is low. This is simply to demonstrate that the test will reject H_0 for relatively large value of τ indicating all the model components should be retained if we want to capture significant portion of predictive variance of the stacked model. As is customary in hypothesis testing problems τ should be pre-selected (see Section 1) and should not be altered post hoc.

τ	0.05	0.06	0.07	0.08	0.09	0.10
$\widehat{\text{ASL}}$ (B-strap)	0.16	0.03	0.003	0.0003	0	0

null with τ around 6%. In this latter case, we are left with only the first term on the right side in (20) when we want to form PI's. If model identification were our goal, we might be able to argue further that only one value of V is important and collapse our predictive modeling down to a single model. Alternatively, if we take 6% as our threshold and invoke the conclusion from our test, we can reason further from examining the entries in Table 4. That is, we may be led to choose the method with the smallest predictive variance taking into consideration the results for the stacking average. Doing this, we confirm that the preferred method XGB of [13] is well-justified and we would only need the first term in our decomposition. Moreover, XGB received the highest stacking weight, presumably because it had the smallest cross-validated error. Alternatively, if retaining the variability over methods is important, which it might well be because XGB only got weight 0.52, we must choose a threshold below 6% and then the table leads us to use at most XGB, NN, and PPR when we use both terms in the decomposition.

4.2 | Three Term Decomposition; $K = 2$

Next, to show the generality of our approach we study a three-term decomposition of predictive variance, that is, we take $K = 2$. We do this by extending our analysis from Section 4.1. So, let the factor V_2 be uniformly distributed over the same five models, that is, levels as for $K = 1$. We set the other factor, V_1, to be the binary variable taking values p_1^* and p_2^* where these denote two choices for the number of pre-selected regressors. The motivation for this choice of V_1 is that in filter feature selection protocols (for example: the Relief algorithm [14]; and its extensions [15, 16]; use the principles of instance-based learning to generate a score for each regressor. This score attempts to capture the relevance of the corresponding regressor to the response variable), the decision to use the *top-r* regressors, for some $r \in \{1, \ldots, p\}$ say, for downstream prediction, is subjective. That is, there is no proper inferential technique to ascertain the adequacy of the *top-r* pre-selected regressors.

Thus, in context, our procedure provides a test for the adequacy of a set of pre-selected variables. Let p_1^* and p_2^* be the objects that identify the top-p_1 and top-p_2 regressors in, for instance, the `RReliefF` algorithm [16] with $p_1 < p_2$. Then the variables in p_1^* are a proper subset of the variables in p_2^*. So, if we fail to reject the null hypothesis $H_0 : E\left(\frac{\text{Var}_{V_1}(E(Y_{n+1};D_n,V_1))}{\text{Var}(Y_{n+1};D_n)}\right) \geq \tau$, we can infer that the between-models variance in V_1, averaged over the elements of V_2, contributes significantly to the total predictive variance of Y_{n+1}. Thus, we need to retain both the elements of V_1. On the other hand, if we can reject H_0, it implies that between-models variance in V_1, averaged over the elements of V_2, contributes insignificantly to the total predictive variance of Y_{n+1}. Thus, here, we may infer that the levels of V_1 can be collapsed and additional criteria could be introduced to identify which level of V_1 should be retained. We emphasize that rejection of the null hypothesis does not offer any information regarding which element of V_1 should be retained.

As indicated in Section 3.2, see (14–16), we must estimate the marginal stacking weights, the ξ_i's, associated with the elements in V_1 and the conditional stacking weights, the ω_{ik}'s, associated with the elements in V_2 for each element in V_1. Here, $i = 1, 2$ and $k = 1,2,3,4,5$. Therefore, we propose the following "iterative stacking" procedure. Minimize

$$\text{Loss}(\boldsymbol{\xi}, \boldsymbol{\omega}) = \frac{1}{n}\sum_{l=1}^{n}\left(Y_l - \sum_{i=1}^{I}\xi_i\left(\sum_{k=1}^{K}\omega_{ik}\hat{Y}_{-l,ik}\right)\right)^2$$

$$\text{subject to : } \sum_i \xi_i = 1, \text{ and} \sum_j \omega_{ik} = 1 \text{ for each } i = 1, 2, \ldots, I \text{ and}$$

$$\xi_i \geq 0, \omega_{ik} \geq 0, \text{for } i = 1, 2, .., I, ; k = 1, 2, \ldots, K \qquad (21)$$

with respect to $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_I)$ and $\boldsymbol{\omega} = (\boldsymbol{\omega}_{1.}, \ldots \boldsymbol{\omega}_{I.})$, where $\boldsymbol{\omega}_{i.} = (\omega_{i1}, \ldots, \omega_{iK})$.

To see the equivalence of (6) and (21) relabel the w_m's as α_{ik}'s with $i = 1, \ldots, I$ and $k = 1, \ldots, K$ where $M = IK$, i is the level of V_1, and k is level of V_2. Write $\xi_i = \sum_k \alpha_{ik}$ and $\alpha_{k|i} = \alpha_{ik}/\sum_k \alpha_{ik}$. Now, the transformation

$$T\left((\alpha_{ik})_{i=1,k=1}^{i=I,k=K}\right) = \left(\xi_1, \ldots, \xi_I; (\alpha_{k|i})_{i=1,k=1}^{I,K}\right)$$

is a homeomorphism from the interior of its domain to its range. So, optimizing in (6) over the w_m's is equivalent to iteratively optimizing over the ξ_i's and the $\omega_{k|i}$'s in (21). Essentially, optimizing (6) will produce estimates of the product of the marginal and conditional stacking weights.

In the present setting we use a block coordinate descent algorithm to solve (21) and directly obtain estimates of the marginal and conditional stacking weights. Observe that if $\boldsymbol{\omega}$ is fixed then the optimization of (21) becomes a standard constrained least square optimization problem in $\boldsymbol{\xi}$. Similarly, if we fix $\boldsymbol{\xi}$ and $(\boldsymbol{\omega}_{1.}, \ldots, \boldsymbol{\omega}_{i-1.}, \boldsymbol{\omega}_{i+1.} \ldots, \boldsymbol{\omega}_{I.})$ then (21) becomes constrained least square optimization problem in $\boldsymbol{\omega}_{i.}$. Thus, in each block a component-wise gradient descent algorithm could be used to update the target optimization variable in that block. Choosing the step size in an adaptive fashion ensures non-negativity of the optimization target. Chen et al. [17] provide the updating equations for such simplex constrained least squares optimization along with the conditions for choosing the step-size. This algorithm does not require stabilization via projection to the nearest PD matrix as required by the (6) algorithm. The adaptive step size selection stabilizes the algorithm. The algorithm proceeds as follows:

- Initialize $\boldsymbol{\xi}^{(0)} \geq 0$, $\boldsymbol{\omega}_{1.}^{(0)} \geq 0, \ldots, \boldsymbol{\omega}_{I.}^{(0)} \geq 0$ such that $\left\|\boldsymbol{\xi}^{(0)}\right\|_1 = 1$, and $\left\|\boldsymbol{\omega}_{i.}^{(0)}\right\|_1 = 1$ for $i = 1, 2, \ldots, I$.

- At tth iteration update:
- $\boldsymbol{\xi}^{(t)} = \underset{\boldsymbol{\xi}}{\text{argmin}}\ Loss\left(\boldsymbol{\xi}, \boldsymbol{\omega}_{1.}^{(t-1)}, \ldots, \boldsymbol{\omega}_{I.}^{(t-1)}\right)$, subject to $\boldsymbol{\xi} \geq 0$ and $\|\boldsymbol{\xi}\|_1 = 1$
- $\boldsymbol{\omega}_{1.}^{(t)} = \underset{\boldsymbol{\omega}_{1.}}{\text{argmin}}\ Loss\left(\boldsymbol{\xi}^{(t)}, \boldsymbol{\omega}_{1.}, \boldsymbol{\omega}_{2.}^{(t-1)} \ldots, \boldsymbol{\omega}_{I.}^{(t-1)}\right)$, subject to $\boldsymbol{\omega}_{1,.} \geq 0$ and $\|\boldsymbol{\omega}_{1.}\|_1 = 1$.
- $\boldsymbol{\omega}_{2.}^{(t)} = \underset{\boldsymbol{\omega}_{2.}}{\text{argmin}}\ Loss\left(\boldsymbol{\xi}^{(t)}, \boldsymbol{\omega}_{1.}^{(t)}, \boldsymbol{\omega}_{2.} \ldots, \boldsymbol{\omega}_{I.}^{(t-1)}\right)$, subject to $\boldsymbol{\omega}_{2,.} \geq 0$ and $\|\boldsymbol{\omega}_{2.}\|_1 = 1$.
- ...
- $\boldsymbol{\omega}_{I.}^{(t)} = \underset{\boldsymbol{\omega}_{I.}}{\text{argmin}}\ Loss\left(\boldsymbol{\xi}^{(t)}, \boldsymbol{\omega}_{1.}^{(t)}, \boldsymbol{\omega}_{2.}^{(t)} \ldots, \boldsymbol{\omega}_{I.}\right)$, subject to $\boldsymbol{\omega}_{I,.} \geq 0$ and $\|\boldsymbol{\omega}_{I.}\|_1 = 1$.
- Repeat until convergence

In principle, this algorithm can be used for V_k's with any number of values. Here, for ease of exposition, we have designed an example with $I = 2$ and $K = 5$.

To illustrate the application of our iterated stacking algorithm and inferential framework, we choose $n = 500$ samples from `Superconductivity` dataset and use the `RreliefF` algorithm to rank the 81 regressors in decreasing order of importance. We consider two cases. In the first case both elements of V_1 correspond to very sparse model. In the second case both elements of V_1 correspond to relatively rich models. We show below that in the first case our test does not reject H_0 indicating that factor V_1 must be retained, that is, both levels of V_1 must be retained to capture at least τ proportion of the total predictive variance produced by the stacked model. In the second case, our test rejects H_0 indicating V_1 can be collapsed to a singleton set, that is, it disappears. This is in accord with our intuition but now we can formally verify it.

Case I. The two values of V_1 are p_1 which denotes the top-5 regressor model and p_2 which denotes the top-10 regressor model. The hypothesis test enables us to ask whether marginally collapsing V_1, that is, using only the top-5 regressor model, is adequate for predictive purposes, or if we should use a richer set of regressors. Our computed results are in Table 6.

Recall the three term decomposition for $K = 2$ given by (13) and its expression as the sum of (14–16). From Table 6, we see that $\widehat{\xi}_1 = 0$ so the contributions of the five models with five predictors to the total variance is zero. Also, the $\widehat{\omega}'_{2,k}s$ are 0 for $k = 1,2,3,4$. So, only the $(i,k) = (2,5)$ term is nonzero. Now, the variance decomposition is degenerate and gives

$$\widehat{\text{Var}}\left(Y_{500}; \mathcal{D}_{499}\right) = 13.72$$

Thus, the observed value of test statistic is 0 and $\widehat{\text{ASL}} = 1$. We cannot reject the null hypothesis for any reasonable value of τ. This implies that the between-model variances in V_1 completely capture the total predictive variance because all the variability is coming from one model. From a practical standpoint, the non-rejection of H_0 implies that a 5-regressor model is too reductive for this dataset and the predictive capacity of 10-regressor model is overwhelmingly large as compared to its 5-regressor counterpart.

Case II. We now start with a rich set of explanatory variables and ask whether adding more regressors substantially decreases total prediction variance. So, let $V_1 = p_1^*$ denote the top-45 regressor model and $V_1 = p_2^*$ denote the top-50 regressor model. The sample size, n, remains the same at 500. This generates Table 7 which is analogous to Table 6, but now for the top 45- and top 50-regressor models.

When we plug the estimates of the stacking weights and predictive variances from Table 7 into expression (13), we get:

$$\begin{aligned}\widehat{\text{Var}}\left(Y_{500}, \mathcal{D}_{499}\right) &= E_{V_1}\widehat{E_{V_2}}\text{Var}\left(Y_{500}; \mathcal{D}_{499}, V_1, V_2\right) \\ &+ E_{V_1}\widehat{\text{Var}_{V_2}}E\left(Y_{500}; \mathcal{D}_{499}, V_1, V_2\right) \\ &+ \widehat{\text{Var}_{V_1}}E\left(Y_{500}; \mathcal{D}_{499}, V_1\right) \\ 16.41 &= 16.15 + 0 + 0.26\end{aligned}$$

The observed value of the test statistic is $0.26/16.41 = 0.016$. We reject the null hypothesis for $\tau = 0.05$ with $\widehat{\text{ASL}} < 0.0001$. This

TABLE 6 | Marginal and conditional stacking weights for a $K = 2$ model with V_1 representing the top-5 and top-10 regressors and V_2 representing the model class. The third and fifth rows give the predictive variances for $V_1 = v_1, V_2 = v_2$.

Values for $V_1, \downarrow$	Marginal stacking weights	Values for $V_2, \rightarrow$	LM	NN	PPR	SVM	XGB
Top-5 regressor	$\widehat{\xi}_1 = 0$	$\widehat{\omega}_{1k}$	0.24	0.19	0	0.25	0.32
		Pred. variance	633.40	538.86	581.40	480.25	67.00
Top-10 regressor	$\widehat{\xi}_2 = 1$	$\widehat{\omega}_{2k}$	0	0	0	0	1
		Pred. variance	556.23	403.76	455.50	299.60	13.72

TABLE 7 | Marginal and conditional stacking weight for $K = 2$ model with V_1 representing the top-45 and top-50 regressors and V_2 representing the model class. The third and fifth rows give the predictive variances for $V_1 = v_1, V_2 = v_2$.

Values for $V_1, \downarrow$	Marginal stacking weights	Values for $V_2, \rightarrow$	LM	NN	PPR	SVM	XGB
Top-45 regressor	$\xi_1 = 0.122$	ω_{1k}	0	0	0	0	1
		Pred. variance	366.85	158.11	242.17	219.23	20.05
Top-50 regressor	$\xi_2 = 0.878$	ω_{2k}	0	0	0	0	1
		Pred. variance	379.42	178.91	223.75	222.53	15.61

implies that the factor between-V_1, that is, the within models variance, fails to capture even 5% of the total predictive variance. In other words, V_1 can be collapsed to a single level or element. From a practical standpoint, the rejection of H_0 implies that the predictive capacity of 45-regressor model is comparable to that of 50-regressor model. This implies that the smaller model is just as good as the larger one.

Our results here can be interpreted via the notion of the marginal collapse of a factor in a model. By this we mean a factor that nominally has multiple values but can in fact be reduced or collapsed into one—irrespective of the behavior of other factors. In the context of Table 6, our test implies that we cannot marginally collapse V_1 to one value. That is, we have to retain both the top-5 and top-10 regressor models in the stacking average. If we look at the values of the $\widehat{\xi}_i$'s, we can in fact drop the top-5 regressor model. In the context of Table 7, the test implies that we can marginally collapse V_1 to one value. It is important to note that the test does not by itself tell us what value of V_1 should be retained and which others are to be discarded. To make that decision, we need additional criteria. For example, if we wish to use the marginal stacking weights ξ to determine the value of V_1, we will end up with Top-50 regressor model in this case, simply because larger models are usually better than smaller models.

5 | Discussion

Here we have used successive iterations of the law of total variance applied to itself to generate decompositions for the predictive variance. The predictive variance is important because it controls the width of prediction intervals. We have chosen our conditioning variables to be both the accumulated data and aspects of statistical modeling. In this way, we can assess the contributions of various aspects of modeling to the width of prediction intervals. The main way we assess the terms in our variance decompositions is by a bootstrap testing procedure for whether a given source of variability, that is, a factor, is small enough relative to the overall variability that it can be omitted, that is, collapsed to one level.

We approximate the conditional expectations and variances in our decompositions using weights from a stacking model average. These weights give a distribution over the features of modeling that looks a lot like a posterior because they depend on the data, are positive, and sum to one. We use these frequentist weights to be consistent with our frequentist bootstrap tests. We are forced to use bootstrap tests because, in general, we do not have a likelihood for the expected variance ratios that we use to assess contributions to the overall predictive variance.

The essence of the method is to write a predictive model in the usual way as $p(y|\upsilon)$ where υ is a K-dimensional discrete multivariate parameter that indexes a modeling strategy V_K. Then the stacking predictive variance is $\mathrm{Var}(Y_{n+1};\mathcal{D})$ where V_K only "appears" as a latent variable that is integrated it out with respect to its stacking weights. Now, the law of total variance (LTV) gives

$$\mathrm{Var}(Y_{n+1};\mathcal{D}) = E\left(\mathrm{Var}(Y_{n+1};V_K,\mathcal{D})\right) + \mathrm{Var}\left(E(Y_{n+1};V_K,\mathcal{D})\right)$$

Writing $V_K = (V_1,\ldots,V_K)$ and applying the LTV to one dimension at a time in the "E-Var" terms gives an expansion of the predictive variance into $K+1$ terms, the last one of which can be taken as $\mathrm{Var}(Y_{n+1}|\mathcal{D},V_k)$ for any k that we want. Now, for each V_K, we get a collection of decompositions depending on the order in which the individual V_k's are introduced.

Next, we convert all the conditional expectations to expectations in the stacking distribution formed from the V_k's. Denoting this by ";" rather than "|," we can use bootstrap testing to assess any term, in particular the last term of the form, say $\mathrm{Var}(Y_{n+1};\mathcal{D},V_1)$, that depends on only one component of V_K. If we decide we can drop V_1, then the decomposition for $\mathrm{Var}(Y_{n+1};\mathcal{D})$ changes and has one less term. We can then test its last term and continue until there are as few terms as possible. When the test determines we can drop a term in the decomposition, it means that a smaller ensemble of models is able to explain the same (or similar) amount of variability in the predictive distribution as a larger ensemble. A series of examples shows that our method gives intuitively plausible results for multiple choices of V.

Our analysis is analogous to the classical Cochran's Theorem decomposition of total squared error into a sum of quadratic forms with independent χ^2 distributions. We do not find as neat a distributional form, however, we show that the terms in our decomposition of the total predictive variance correspond to sums of weighted χ^2 random variables. The dependence on ordering of the V_j's here parallels the same problem in Cochran's theorem when the data is unbalanced. There is a correction for this in the classical ANOVA setting; see [18] chap. 6.3 for some details. However, we have not developed this here.

A drawback of our method is that a wicked Statistician could make a variance misleadingly small by choosing models very close together or having super-precise parameter estimates and thereby generate high bias and high mean squared predictive error. In practice, we would counter this by ensuring that the models represented by V_K, taken together, span a large enough volume in model space that some will have nontrivial coverage probabilities for future values. One way to do this would be to form a model list of representatives of the model space that are as far apart as possible from each other while ensuring that any model in the space is close enough to at least one of them that other sources of error contribute more to the variability.

Another topic that bears further work is the relationship between testing for the importance of a term in the decomposition and collapsing one of the levels V_j to a single value. In standard ANOVA, terms correspond to factors. Here, there is a correspondence but it is weaker and we do not have a formal argument relating dropping terms in a variance decomposition to dropping factors in a modeling strategy.

Author Contributions

Dean Dustin: conception and design of the project, computing, and writing. **Bertrand Clarke:** conception and design of the project, writing, result interpretation, and supervision. **Souparno Ghosh:** computing, result interpretation, and writing.

Acknowledgments

Dean Dustin acknowledges funding from the University of Nebraska Program of Excellence in Computational Science. Souparno Ghosh's research was supported in part by the National Science Foundation under grant no. 2007418 and Leidos Biomed/NCI under contract 22X049. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Leidos Biomed/NCI. All three authors acknowledge computational support from the Holland Computing Center at the University of Nebraska. All three authors are grateful to the referees who gave us constructive suggestions for how to improve our work.

Disclosure

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Endnotes

[1] A slight variant on this is dilution, see [4], where there is a small region of models that are roughly equally good and split the probabilities so finely that all the predictions are zero. We assume that V has been chosen to avoid this.

References

1. P. Smyth and D. Wolpert, "Linearly Combining Density Estiamtors via Stacking," *Machine Learning* 36 (1999): 59–83.

2. D. Draper, "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society: Series B: Methodological* 57, no. 1 (1995): 45–97.

3. D. Dustin and B. Clarke, "Testing for the Important Components of Posterior Predictive Variance," arXiv:2209.00636, 2022.

4. E. George, "Dilution Priors: Compensating for Model Space Redundancy," in *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*. IMS Collections, vol. 6, ed. J. O. Berger, T. T. Cai, and I. M. Johnstone (Institute of Mathematical Statistics, 2010), 158–165.

5. W. Wang, S. Mukherjee, S. Richardson, and S. Hill, "High Dimensional Regression in Practice: An Empirical Study of Finite-Sample Prediction, Variable Selection, and Ranking," *Statistics and Computing* 30 (2020): 697–719.

6. D. Dustin, J. Clarke, and B. Clarke, "Predictive Stability Criteria for Penalty Selection in Linear Models," *Computational Statistics* 39 (2024): 1241–1280.

7. X. Zhang and C.-A. Liu, "Model Averaging Prediction by k-Fold Cross-Validation," *Journal of Econometrics* 235, no. 1 (2023): 280–301.

8. N. J. Higham, "Computing a Nearest Symmetric Positive Semidefinite Matrix," *Linear Algebra and Its Applications* 103 (1988): 103–118.

9. S. Zhao, D. Witten, and A. Shojaie, "In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference," *Statistical Science* 36 (2021): 562–577.

10. H. Scheffé, *The Analysis of Variance* (John Wiley and Sons, 1959).

11. P. Gustafson and B. Clarke, "Decomposing Posterior Variance," *Journal of Statistical Planning and Inference* 119, no. 2 (2004): 311–327.

12. N. Meinshausen, M. Mathiuus, and P. Bühlmann, "Asymptotic Optimality of the Westfall-Young Permutation Procedure for Multiple Testing Under Dependence," *Annals of Statistics* 39 (2011): 3369–3391.

13. K. Hamidieh, "A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor," *Computational Materials Science* 154 (2018): 346–354.

14. K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in *Machine Learning Proceedings 1992*, ed. D. Sleeman and P. Edwards (Morgan Kaufmann, 1992), 249–256.

15. I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," in *European Conference on Machine Learning* (Springer, 1994), 171–182.

16. M. Robnik-Šikonja and I. Kononenko, "An Adaptation of Relief for Attribute Estimation in Regression," Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97) (Vol. 5, Citeseer, 1997), 296–304.

17. J. Chen, C. Richard, H. Lantéri, C. Theys, and P. Honeine, "A Gradient Based Method for Fully Constrained Least-Squares Unmixing of Hyperspectral Images," in *2011 IEEE Statistical Signal Processing Workshop (SSP)* (IEEE, 2011), 301–304.

18. H. Toutenberg and Shalabh, *Statistical Analysis of Designed Experiments* (Springer, 2009).

19. G. Box, "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. I. Effect of Inequality of Variance in the One Way Classification," *Annals of Mathematical Statistics* 25 (1954): 290–302.

Appendix A

Cochran's Theorem

Here we continue the derivation from Section 3.2 showing how the decomposition in Clause (i) of Proposition 1 is analogous to Cochran's Theorem.

Deriving a χ^2 Distribution for $K = 2$

First, we see that (14) is an expected quadratic form, that is,

$$\sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\omega_{ij}\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n,m_{ij},s_i\right) = \sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\omega_{ij}E\left(\left(Y_{n+1}-\hat{y}_{ij}\right)^2;\mathcal{D}_n,m_{ij},s_i\right) \tag{A1}$$

For (15), write W_i for the column vector $W_i = \left(\sqrt{\omega_{i1}},\ldots\sqrt{\omega_{iJ}}\right)'$, and write $\tilde{Y}_i$ for the column vector $\tilde{Y}_i = \left(\hat{y}_{i1}-\overline{y}_{i\cdot},\ldots,\hat{y}_{iJ}-\overline{y}_{i\cdot}\right)'$. Now (15) is

$$\sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\omega_{ij}\left(\hat{y}_{ij}-\overline{y}_{i\cdot}\right)^2 = \sum_{i=1}^{I}\xi_i W_i'\tilde{Y}_i\tilde{Y}_i'W_i = \sum_{i=1}^{I}\xi_i\tilde{Y}_i'W_iW_i'\tilde{Y}_i \tag{A2}$$

Similarly, for term (16), write S for the column vector $S = \left(\sqrt{\xi_1},\ldots,\sqrt{\xi_I}\right)'$ and $\overline{Y} = \left(\overline{y}_{1\cdot}-\overline{y},\ldots,\overline{y}_{I\cdot}-\overline{y}\right)'$. Then we have that (16) is

$$\sum_{i=1}^{I}\xi_i\left(\overline{y}_{i\cdot}-\overline{y}_{\cdot\cdot}\right)^2 = S'\overline{YY}'S = \overline{Y}'SS'\overline{Y} \tag{A3}$$

So, using (A1–A3), we can rewrite (13) as

$$\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)=\sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\omega_{ij}E\left(\left(Y_{n+1}-\hat{y}_{ij}\right)^2;\mathcal{D}_n,m_{ij},s_i\right) \tag{A4}$$

$$+\sum_{i=1}^{I}\xi_i\tilde{Y}'_iW_iW'_i\tilde{Y}_i \tag{A5}$$

$$+\overline{Y}'SS'\overline{Y} \tag{A6}$$

Now we see each term in the predictive variance is a quadratic form, that is, a homogeneous polynomial of order two, even if the terms in (A4) are (trivial) quadratic forms of dimension one.

To see how the distributional aspects of (A4–A6) parallel the distributional statements in Cochran's Theorem, we proceed as follows. Note that regarding $\mathcal{D}_n$ as a random variable rather than as observed data means that all terms in the decomposition can also be regarded as random variables. Next, assume all data are normal. Now,

$$\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)-\sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\omega_{ij}E\left(\left(Y_{n+1}-\hat{y}_{ij}\right)^2;\mathcal{D}_n,m_{ij},s_i\right)$$

$$=\sum_{i=1}^{I}\xi_i\tilde{Y}'_iW_iW'_i\tilde{Y}_i+\overline{Y}'SS'\overline{Y} \tag{A7}$$

in which each term has a distribution. We begin with the two terms on the right.

To begin, we recall Theorem 2.1 in [19] that generalizes Cochran's theorem for the distribution for quadratic forms. Namely, if $X\sim N(0,\Psi)$, with Ψ a $p\times p$ covariance matrix. Then if $Q=X^TMX$ is any real quadratic form of rank $r\le p$, Q is distributed like a quantity

$$\sum_{j=1}^{r}\lambda_j\chi_1^2 \tag{A8}$$

with $r\le p$ and λ_i the i-th eigenvalue of ΨM.

Now, look at the first term on the right, and let $A_i=W_iW'_i$. We know A_i is a $J\times J$, symmetric, and semi-positive definite because (A5) is a variance between values V_1 within V_2 and by definition variances are positive.

Next, consider the second term on the right and let $B=SS'$ which is $I\times I$, symmetric and semi-positive definite by definition of variance. Further suppose $\overline{Y}\sim N(0,\Sigma^*)$ and $\sqrt{\xi_i}\tilde{Y}_i\sim N\left(0,\Sigma_i\right)$.

Now, since both terms on the right in (A8) are quadratic forms in a normal random vector, we can apply Theorem 2.1 in [19] to each of them. So, (A8) gives

$$\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)-\sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\omega_{ij}E\left(\left(Y_{n+1}-\hat{y}_{ij}\right)^2;\mathcal{D}_n,m_{ij},s_i\right)$$

$$\sim\sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\lambda_{ij}\chi_1^2+\sum_{i=1}^{I}\lambda_i\chi_1^2 \tag{A9}$$

where λ_i is the i-th eigenvalue of $B\Sigma^*$ and λ_{ij} is the j-th eigenvalues of $A_i\Sigma_i$. That is, the two terms on the right of (A8) are convex and weighted sums, respectively, of χ_1^2 random variables.

The second term on the left is the expectation of a χ_1^2 random variable. To see this, suppose $\left(Y_{n+1}-\hat{y}_{ij};\mathcal{D}_n,m_{ij},s_i\right)\sim N\left(0,\sigma_{ij}^2\right)$ and observe

$$\begin{aligned}E\left(\left(Y_{n+1}-\hat{y}_{ij}\right)^2;\mathcal{D}_n,m_{ij},s_i\right)&=\mathrm{Var}\left(Y_{n+1}-\hat{y}_{ij};\mathcal{D}_n,m_{ij},s_i\right)\\&\quad+E\left(Y_{n+1}-\hat{y}_{ij};\mathcal{D}_n,m_{ij},s_i\right)^2\\&=\mathrm{Var}\left(Y_{n+1}-\hat{y}_{ij};\mathcal{D}_n,m_{ij},s_i\right)\\&=\sigma_{ij}^2\end{aligned} \tag{A10}$$

We recognize this as equivalent to the expectation of a χ_1^2 random variable scaled by σ_{ij}^2—that is, $E\left(\sigma_{ij}^2\chi_1^2\right)=\sigma_{ij}^2$. It is difficult to determine the distribution of (A10) explicitly but because we are taking a convex combination of terms like it, computations suggest it is approximately normal.

Since all three terms in (13) are variances and hence corrected for their means, (A4) is a new term that arises from trying to derive a Cochran's theorem style representation of $\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)$ using factors and factor level weights from stacking, Bayes model averaging, or other assessments of model uncertainty. To complete our analogy, recall Cochran's Theorem gives as many terms as there are factors plus a residual term. We get $\dim(V)$ terms, that is, the number of factors, plus an extra term, (A4), the predictive analog of the residual term.

If desired, we can approximate distributions of the right hand terms in (A9) more compactly by using other results from [19]. For instance, his Theorem 2.2 gives the formula for the i-th cumulant of (A8) as

$$Q_i=2^{i-1}(i-1)!\sum_{j=1}^{r}\lambda_j$$

Using this, we can approximate (A8) by $g\chi^2(h)$ where

$$g=\frac{1}{2}\frac{Q_1^2}{Q_2}=\frac{\sum\lambda_j^2}{\sum\nu_j\lambda_j}$$

and

$$h=\frac{2Q_1^2}{Q_2}=\frac{\left(\sum\lambda_j\right)^2}{\sum\lambda_j^2}$$

Box gives this approximation in part because it has the same first two moments as (A8). Box also notes that when all λ_j are equal, the degrees of freedom, h, is smaller than appropriate.

Using this we can approximate $\overline{Y}'B\overline{Y}=\overline{Y}'SS'\overline{Y}$ by

$$g\chi_h^2=\frac{\sum\lambda_i^2}{\sum\lambda_i}\chi^2\left(\frac{\left(\sum\lambda_i\right)^2}{\sum\lambda_i^2}\right) \tag{A11}$$

Also, we can approximate

$$\sqrt{\xi_i}\tilde{Y}'_iA_i\sqrt{\xi_i}\tilde{Y}_i=\sqrt{\xi_i}\tilde{Y}'_iW_iW'_i\sqrt{\xi_i}\tilde{Y}_i$$

by

$$g_i\chi_{h_i}^2=\frac{\sum_j\lambda_{ij}^2}{\sum_j\lambda_{ij}}\chi^2\left(\frac{\left(\sum_j\lambda_{ij}\right)^2}{\sum_j\lambda_{ij}^2}\right)$$

Hence, we have the approximate distribution

$$\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)-\sum_{i=1}^{I}\xi_i\sum_{j=1}^{J}\omega_{ij}E\left(\left(Y_{n+1}-\hat{y}_{ij}\right)^2;\mathcal{D}_n,m_{ij},s_i\right)$$

$$\overset{\text{approx}}{\sim}\sum_{i=1}^{I}g_i\chi_{h_i}^2+g\chi_h^2$$

We emphasize that the analogy is conceptually incomplete as noted at the end of Section 3.2. In addition, we do not have a definite distribution for the second term on the left in (A8).

General K

Deriving quadratic forms and distributional expressions for $\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)$ for general K is similar to the derivation of (A8) and (A9), respectively, seen in Section 3.2. For the sake of completeness, we state these two results below.

Our first result in this subsection gives the general expression for the predictive variance in terms of quadratic forms. For brevity, let

$$\tilde{y}_{v_{i_1},\ldots,v_{i_k}}=E\left(Y_{n+1};\mathcal{D}_n,v_{i_1},\ldots,v_{i_k}\right)$$

We have the following:

Proposition 2. *For a K-factor predictive scheme, the predictive variance can be written as a sum of weighted quadratic forms as follows:*

$$\begin{aligned}
\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right) = &\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_K=1}^{I_K} p\left(v_{i_K};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-1}}\right)\\
&\times E\left(\left(Y_{n+1}-\tilde{y}_{v_{i_1},\ldots,v_{i_K}}\right)^2;\mathcal{D}_n,v_{i_1}\ldots,v_{i_K}\right)\\
&+\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_K=1}^{I_{K-1}} p\left(v_{i_{K-1}};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-2}}\right)\\
&\times\tilde{Y}'_{K,\ldots,1}A_{K,\ldots,1}\tilde{Y}_{K,\ldots,1}\\
&+\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_{K-2}=1}^{I_{K-2}} p\left(v_{i_{K-2}};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-3}}\right)\\
&\times\tilde{Y}'_{K-1,\ldots,1}A_{K-1,\ldots,1}\tilde{Y}_{K-1,\ldots,1}\\
&\quad\vdots\qquad\vdots\qquad\vdots\\
&+\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\tilde{Y}'_{2,1}A_{2,1}\tilde{Y}_{2,1}\\
&+\tilde{Y}'_1A_1\tilde{Y}_1
\end{aligned}\tag{A12}$$

where

$$\begin{aligned}
&A_{k,\ldots,1}=W_{k,\ldots,1}\left(W_{k,\ldots,1}\right)',\\
&W_{k,\ldots,1}=\left(\sqrt{p\left(v_{i_k=1};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{k-1}}\right)},\ldots,\sqrt{p\left(v_{i_k=I_k};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{k-1}}\right)}\right)
\end{aligned}\tag{A13}$$

and $\tilde{Y}_{k,\ldots,1}$ is the column vector of mean adjusted predictions for factor V_k conditional on factors $V_1,\ldots V_{k-1}$. That is, we write

$$\begin{aligned}
\tilde{Y}_{k,\ldots,1}=&\left(\left(\tilde{y}_{v_{i_1},\ldots,v_{i_k=1}}-E\left(Y_{n+1};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{k-1}}\right)\right),\ldots,\right.\\
&\left.\times\left(\tilde{y}_{v_{i_1},\ldots,v_{i_k=I_k}}-E\left(Y_{n+1};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{k-1}}\right)\right)\right)'
\end{aligned}$$

where $\tilde{y}_{v_{i_1},\ldots,v_{i_k=j}}=E\left(Y_{n+1};\mathcal{D}_n,v_{i_1},\ldots,v_{i_k=j}\right)$.

Our second result gives the distributions for K of the terms in our expansion for the predictive variance. As before, we get sums of χ_1^2 random variables.

Proposition 3. *Let $\left(Y_{n+1}-\tilde{y}_{v_{i_1},\ldots,v_{i_K}}\right)\sim N\left(0,\sigma^2_{i_1,\ldots,i_K}\right)$, $\tilde{Y}_1\sim N(0,\Sigma)$ and $\tilde{Y}_{k,\ldots,1}\sim N\left(0,\Sigma_{k,\ldots,1}\right)$. Then the sum of quadratic forms in (A12) are distributed like a sum of weighted χ-squared random variable as follows*

$$\begin{aligned}
\mathrm{Var}\left(Y_{n+1};\mathcal{D}_n\right)\sim &\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_K=1}^{I_K} p\left(v_{i_K};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-1}}\right)\\
&\times E\left(\left(Y_{n+1}-\tilde{y}_{v_{i_1},\ldots,v_{i_K}}\right)^2;\mathcal{D}_n,v_{i_1},\ldots,v_{i_K}\right)\\
&+\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_{K-1}=1}^{I_{K-1}} p\left(v_{i_{K-1}};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-2}}\right)\\
&\times\sum_{i_K=1}^{I_K}\lambda_{K\ldots,1}\chi_1^2\\
&+\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_{K-2}=1}^{I_{K-2}} p\left(v_{i_{K-2}};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-3}}\right)\\
&\times\sum_{i_{K-1}=1}^{I_{K-1}}\lambda_{K-1\ldots,1}\chi_1^2\\
&\quad\vdots\qquad\vdots\qquad\vdots\\
&+\sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\sum_{i_2=1}^{I_2}\lambda_{2,1}\chi_1^2\\
&+\sum_{i_1=1}^{I_1}\lambda_1\chi_1^2
\end{aligned}\tag{A14}$$

where $\lambda_{k,\ldots,1}$ is the kth eigenvalue of $A_{k,\ldots,1}\Sigma_{k,\ldots,1}$.

Note that there are no explicit assumptions on the joint pmf for V. Our results are not asymptotic, so our results hold as long as a proper distribution is specified for V.