Vol. TBA (2021) 1

Testing for the Important Components of Predictive Variance

Dean Dustin[1], **Bertrand Clarke**[2]

[1] *Charles Schwab, Denver, CO e-mail:* ddustin8@huskers.unl.edu

[2] *Department of Statistics, University of Nebraska-Lincoln, NE, USA, 68583-0963 e-mail:* bclarke3@unl.edu

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 identify which of the discrete random variables i.e., aspects of modeling, are most important to prediction intervals. 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 subject classifications: Primary 62F15; secondary 62J10.
Keywords and phrases: prediction interval, predictive variance, law of total variance, stacking, ANOVA, bootstrap testing, Cochran's theorem.

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.

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 therefore possibly reduce the number of levels in a hierarchical model by removing those that contribute too little to be important.

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

arXiv: 2209.00636

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 components of modeling rather than the data.

Our decomposition is based on iterating an empirical version of the the law of total variance for future outcomes given $\mathcal{D}_n$. Recall that for a single random V variable we have

$$\text{Var}(Y_{n+1};\mathcal{D}_n) = E(\text{Var}(Y_{n+1};V,\mathcal{D}_n)) + \text{Var}(E(Y_{n+1};V,\mathcal{D}_n)). \tag{1.1}$$

In our examples, V will be discrete although continuous V's satisfy (1.1) as well. The semicolon means that we are conditioning on V as a random variable but may allow a more general functional dependence on $\mathcal{D}_n$. The first term on the right can be interpreted as the average location of the variance taking into account the variability of V. The second term is the variability contributed by V to the location of the predictive distribution. If the second term 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. On the other hand, if the first term is small, then the contribution of V to the variance of $(Y_{n+1};V,\mathcal{D}_n)$ as a random variable equipped with an estimated distribution may be ignored. The difference between these two terms is in how much V affects the variability in location versus the variability in variance.

COMMENT: From referee 1:
the multiple-K part theory is a selling point of this paper, especially because the multi-level total variance is likely not well understood in the past.

Intuitively, the values of V represent some feature 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. Trivially, knowing the true model would correspond to $K=1$ and V_1 equal a constant and the second term in (1.1) would be zero. More realistically, as seen in Sec. 3, V may represent the choice of penalty in penalized linear regression. The penalty corresponds to a prior, so our method includes a technique for assessing variability due to prior selection within a class, where the assessment uses post-experimental weights on the priors. In another example here, V represents a model type itself. More generally, V may represent a link function in generalized linear models, a nonlinear regression technique, a selection of variables, etc.; see Dustin and Clarke (2022).

We can also consider multidimensional $V = V_K = (V_1,\ldots,V_k,\ldots,V_K)$ and apply (1.1) iteratively to itself, generating one new term for each V_k at each iteration. For $K=2$ we have

$$\begin{aligned} Var(Y_{n+1};\mathcal{D}_n) = {} & E_{V_1,V_2}Var(Y_{n+1};V_1,V_2,\mathcal{D}_n) \\ & + E_{V_1}Var_{V_2}E(Y_{n+1};V_1,V_2,\mathcal{D}_n) \\ & + Var_{V_1}E(Y_{n+1};V_1,\mathcal{D}_n); \end{aligned} \tag{1.2}$$

for general K see Prop. 2.1. In general, this gives $K+1$ terms that can be interpreted in terms of means and variances. Thus, we must choose a K and we can regard each V_k as an aspect of a modeling strategy. For instance, if $K=2$, V_1 may be a 'scenario' and V_2 may be a 'model' in the sense of Draper (1995).

To see one way these decompositions may be useful, consider the last terms in either of (1.1) and (1.2). Regardless of the distribution used to take the variance, there are two basic ways we can get $\mathrm{Var}(E(Y_{n+1};V,\mathcal{D}_n)=0$. First, the distribution of $V=V_1$ concentrates at a single value $V=v_1$. Second, the models i.e., values of V that get non-zero weights, give the same predictions given $\mathcal{D}$ [1]. That is,

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

for any v_1 and v_2 getting positive weight. Solving for a set like

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

amounts to inverting an integral operator a problem which is known to be intractable. However, by carefully selecting the models $V=v$ to ensure they are meaningfully different 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. Indeed, as data accumulate, it is harder and harder for two different m odels to be accidentally predictively equivalent. Thus, on pragmatic grounds, with some foresight, 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 level of modeling drops out. In the case of (1.2) 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 (1.1).

It will rarely be the case that $\mathrm{Var}(E(Y_{n+1};V,\mathcal{D}_n))=0$, however, there will be many cases when $\mathrm{Var}(E(Y_{n+1};V,\mathcal{D}_n)\leq\gamma$ for suitable choices of γ – small enough that the term can reasonably be neglected but large enough that it can be detected in well-chosen models with large enough n. Essentially, we test for this event in a relative sense after explaining how we use stacking weights in place of posterior weights. The benefit of this procedure is that we may be able to eliminate entries in V_K for use in prediction – or be sure that they are important to include. This thinking parallels ANOVA where we try to determine which factors can be reduced.

To describe our method heuristically, note that in (1.1) and (1.2), the dependence on the V's is by conditioning since we assume we have a likelihood function for V. The dependence on the data may also be through conditioning however we wish to allow more general forms of mathematical dependence. Inded, there are many instantiations of expressions like (1.1) and (1.2) because their weights may be chosen in many ways that yield a valid assessment of post-data predictive variability. For instance, the data dependent weights may come

[1] A slight variant on this is dilution where there is a small region of models that 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.

from any model averaging strategy provided they are non-negative and sum to one: Simply use such weights in place of posterior model weights for values of the V. If the actual posterior weights were used, the predictive density would correspond to the Bayesian model average.

Here, for a variety of reasons we choose the stacking average, see Wolpert (1992), to get expressions for the terms in decompositions such as (1.1). Stacking weights are based on a cross-validation criterion, are often used for model averaging, and can be regarded as summaries of model uncertainty. In many cases, stacking weights are easier to compute than posterior probabilities. Zhang and Liu (2022) show that, like posterior probabilities, when the true model is on the model list, its stacking weight is asymptotically one and otherwise the stacking average converges to the model on the list that is predictively optimal. Lastly, Yao et al. (2018) and Clarke (2003) argue that stacking distributions and means, respectively, often outperform those from Bayesian model averaging.

Given that the components in V correspond to modeling choices, knowing the true model would correspond to $K = 1$ and V_1 equal a constant. Usually this is unrealistic. So, we choose the components of V to represent components of modeling that are uncertain. The intuition is that we should choose K and V to obtain optimal predictions given the modeling assumptins we think are true. Different modeling strategies will use different V's and our work here is a general methodology to analyze the relative importance of the V_k's in a V in terms of their contributions to the predictive variance.

To be more explicit, for $K = 2$, the stacking predictive distribution is

$$p(Y_{n+1}) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \hat{w}(v_{1i}, v_{2j}) p(Y_{n+1}|v_{1i}, v_{2j}), \tag{1.3}$$

where the levels of V_1 and V_2 are indicated by v_{1i} and v_{2j} respectively. For ease of notation, dependence of the summands on $\mathcal{D}_n$ is suppressed. The bar $|$ in (1.3) is the model for Y_{n+1} conditional on the modeling components. Writing an explicit likelihood is the only place that conditioning per se, as opposed to more general data dependence, is used. Often, the p in (1.3) has parameters that are estimated in which case we write $\hat{p}$. Throughout, we reserve the hat-notation when we have a specific estimator in mind. More generally, we use the generic form of of $\mathcal{D}$, as in $E(\cdot; \mathcal{D})$, to indicate functional dependence.

The stacking weights, denoted $\hat{w}(\cdot)$, are found from a quadratic optimization analogous to cross-validation, and can be taken positive and summing to one. The stacking weights do not necessarily factor like posterior probabilities. However, if the outer weights for the models are specified as $w(v_{2i})$ and are positive and sum to one then the inner weights of the form $w(v_{1i}; v_{2j})$ can be found analogously to the overall weights in (1.3).

If we write $\mathrm{Var}_{V_K}(Y_{n+1}; \mathcal{D}_n)$ 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 $\mathrm{Var}_{V_K}(Y_{n+1}; \mathcal{D}_n) \neq \mathrm{Var}_{V'_{K'}}(Y_{n+1}; \mathcal{D}_n)$. On the other hand, the relative sizes of terms in decompositions of the form (1.1) depend delicately on the choice of K and V_K. Fortunately, in practice, we usually only have one V_K

that we most want to consider, but the order of the V_k's may matter and it is partially a matter of statistical judgement how big V_K should be and what components it should have. We regard the selection of V in general as an aspect of modeling and so beyond our present scope.

Given that we are using stacking, once K and V_K have been chosen, the predictive variance decomposition based on V_K can be generated e.g., as in (1.2), and its terms examined to see which are large enough to be important. We do this using a bootstrap testing procedure. We are forced to use bootstrapping because the terms we want to test for proximity to zero are latent quantities, i.e., they do not directly depend on the data, and hence do not have an accessible likelihood. Moreover, using frequentist bootstrap tests is philosophically consistent with using stacking, a frequentist model average. A Bayesian treatment of our methodology is outlined in Dustin and Clarke (2022).

We regard our bootstrap tests as an approximation to the usual F-tests that are used in ANOVA and come from a Cochran's theorem decomposition. Indeed, our general predictive decomposition resembles the Cochran's theorem decomposition of the squared error into a sum of quadratic forms; see Subsec. 2.2. In fact, our tests resemble ratios of χ-squared distributions but we cannot ensure the independence or determine the degrees of freedom explicitly. The overall procedure involves several steps that are readily computable. Here, we do this for $K = 1$, but it is clear how to handle $K \geq 2$ and (Bayesian) computed examples are in Dustin and Clarke (2022).

The structure of this paper is as follows. Sec. 2 presents our full method with justifications. There are subsections to explain the variance decomposition in terms of quadratic forms and the testing procedure for terms in a variance decomposition. In Sec.3 we give an example with simulated data of how our methodology can be used to determine which shrinkage method within a finite collection of shrinkage methods is best in the sense of minimizing the predictive variance. For contrast, since shrinkage methods correspond to priors, in Sec. 4 we give a real data example where the levels of V correspond to possible likelihoods. In Sec. 5, we discuss our overall contribution.

2. Decomposing the Predictive Variance

In this section we give our variance decomposition in full generality, indicate how to choose amongst candidate variance decompositions, and explain our testing procedure for the terms in a given variance decomposition. We will see that our decomposition of the predictive variance is analogous to the Cochran's theorem decomposition of the squared error into quadratic forms.

2.1. The Effect of the Model List on Overall Variance

We can enlarge a model list simply by including more plausible models. However, this may lead to problems such as dilution; see George (2010). So, we want to assess the effect of a model list on the variance of predictions. Consider a model

list $\mathcal{M}$ and suppose we don't believe it adequately captures the uncertainty (including mis-specification) of the the predictive problem. We can expand the list by including other competing models 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 Draper (1995).

In the simplest case, once a new model list $\mathcal{M}'$ is constructed, if it contains new models with positive probability, the predictive distribution $p(Y_{n+1}; \mathcal{D}_n)$ using $\mathcal{M}'$ will be different from $p(Y_{n+1}; \mathcal{D}_n)$ resulting from $\mathcal{M}$. Recalling that we are using the stacking model average we denote dependence on $\mathcal{M}$ by

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

In our variance decomposition below, V includes dependence on the model list. This dependence is by conditioning (and so we should use | to indic te it) but we continue to use ; because the dependence on the data is more general.

2.1.1. Predictive Variance Decomposition "P-ANOVA"

To quantify the uncertainty of our subjective choices, recall $V = (V_1, \ldots, V_K)$, where V_k represents the values of the k-th potential choice 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 define the *levels* of V_k to be $v_{k1}, \ldots, v_{km_k}$. That is, $v_{k\ell}$ is a specific value V_k may assume. Thus, V is discrete and has probability mass function $W(v) = W(V_1 = v_1 \ldots, V_K = v_K)$. The V_k's are not in general independent and W corresponds to a prior on V. Define the model list by

$$\mathcal{V}^K = \{v_{11}, \ldots, v_{1m_1}\} \cup \ldots \cup \{v_{K1}, \ldots, v_{Km_K}\}.$$

There are $m_1 \times \cdots \times m_K$ distinct models in $\mathcal{V}^K$ and they may or may not have a hierarchical structure. Our first result gives a decomposition of the predictive variance by conditioning on V.

Proposition 2.1. *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

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

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

$$\begin{aligned} Var(Y_{n+1}; \mathcal{D}_n)(\mathcal{V}^K) = {} & E_{(V_1,\ldots,V_K)} Var(Y_{n+1} | V_1, \ldots, V_K, \mathcal{D}_n) \\ & + \sum_{k=2}^{K} E_{(V_1,\ldots,V_{k-1})} Var_{V_k} E(Y_{n+1}; V_1, \ldots, V_k, \mathcal{D}_n) \\ & + Var_{V_1} E(Y_{n+1}; V_1, \mathcal{D}_n), \end{aligned} \tag{2.1}$$

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* $Var(Y_{n+1}; \mathcal{D}_n)(\mathcal{V}^K)$ *can be condensed into a two term decomposition:*

$$\begin{aligned} Var(Y_{n+1}; \mathcal{D}_n)(\mathcal{V}^k) = E_{(V_1,\ldots,V_K)} Var(Y_{n+1}; V_1, \ldots, V_K, \mathcal{D}_n) \\ + Var_{(V_1,\ldots,V_K)} E(Y_{n+1}; V_1, \ldots, V_K, \mathcal{D}_n). \end{aligned} \tag{2.2}$$

The proof of Clause (i) is inductive: The case $K = 1$ is (1.1) The case $K = 2$ results from one iteration with the law of total variance as in (1.2). 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 vector V_K.

Note that on the right we have used ';' because the dependence on the data is not via conditioning even though we are conditioning on a V_k.

We summarize the decomposition in (2.1) using what we call "P-ANOVA", or *predictive analysis of variance*. In Table 1, 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

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

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

$$Var_{V_j} E(Y_{n+1}; = V_i = v_1, V_2 = v_2, \ldots, V_{j-1} = v_{j-1}, V_j, \mathcal{D}_n).$$

Using the Bayes model average – or any other model averaging procedure – in place of stacking leads to a P-ANOVA table analogous to Table 1.

TABLE 1

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. All terms are dependent on* $\mathcal{D}_n$*, but not necessarily in a conditional sense*

Source	Interpretation	Variance
V_1	Between V_1 variance	$Var_{V_1} E(Y_{n+1}; \mathcal{D}_n, V_1)$
V_2	Expected between V_2 across V_1	$E_{V_1} Var_{V_2} E(Y_{n+1}; \mathcal{D}_n, V_1, V_2)$
$\vdots$	$\vdots$	$\vdots$
V_K	Expected between V_K across $V_{K-1} \ldots V_1$	$E_{V_1} \ldots E_{V_{K-1}} Var_{V_K} E(Y_{n+1}; \mathcal{D}_n, V_1, V_2, \ldots, V_K)$
Predictions	Expected variance across $V_1 \ldots V_k$	$E_{V_1} \ldots E_{V_K} Var(Y_{n+1}; \mathcal{D}_n, V_1, V_2, \ldots, V_K)$
Total	Posterior predictive variance	$Var(Y_{n+1}; \mathcal{D}_n)$

2.2. 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^TY" 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 Scheffé (1959) Appendix VI.

Theorem 2.1. *(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*,*

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

Next, we a discuss 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 Subsec. 2.3, we describe a bootstrap based testing procedure for the individual terms in our expansion so as to determine if the means within the V_k's with strictly positive stacking weights are different enough that they contribute substantially to the overall predictive variance. Our results are fundamentally different from Gustafson and Clarke (2004) 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 Draper (1995), 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. Now, Prop. 2.1 gives

$$
\begin{aligned}
Var(Y_{n+1};\mathcal{D}_n) &= E_{V_1}E_{V_2}Var(Y_{n+1};\mathcal{D}_n,V_1,V_2) + E_{V_1}Var_{V_2}E(Y_{n+1};\mathcal{D}_n,V_1,V_2) \\
&\quad + Var_{V_1}E(Y_{n+1};\mathcal{D}_n,V_1) \\
&= \sum_{i=1}^{I} p(s_i;\mathcal{D}_n) \sum_{j=1}^{J} p(m_{ij};\mathcal{D}_n,s_i) Var(Y_{n+1};\mathcal{D}_n,s_i,m_{ij}) \\
&\quad + \sum_{i=1}^{I} p(s_i;\mathcal{D}_n) \sum_{j=1}^{J} p(m_{ij};\mathcal{D}_n,s_i) \left[E(Y_{n+1};\mathcal{D}_n,m_{ij},s_i) - E(Y_{n+1};\mathcal{D}_n,s_i)\right]^2 \\
&\quad + \sum_{i=1}^{I} p(s_i;\mathcal{D}_n) \left[E(Y_{n+1};\mathcal{D}_n,s_i) - E(Y_{n+1};\mathcal{D}_n)\right]^2 . \qquad (2.3)
\end{aligned}
$$

For ease of notation, let

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

Now we can restate (2.3) as

$$
\begin{aligned}
Var(Y_{n+1};\mathcal{D}_n) &= \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} Var(Y_{n+1};\mathcal{D}_n,m_{ij},s_i) \qquad (2.4) \\
&\quad + \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} \left(\hat{y}_{ij} - \bar{y}_{i\cdot}\right)^2 \qquad (2.5) \\
&\quad + \sum_{i=1}^{I} \xi_i \left(\bar{y}_{i\cdot} - \bar{y}\right)^2 . \qquad (2.6)
\end{aligned}
$$

Our strategy is to express each term in $Var(Y_{n+1};\mathcal{D}_n)$ in vector notation so we can recognize quadratic forms. These quadratic forms can be shown to have distributions that parallel the distributional statements in Cochran's Theorem. This is seen in detail in Appendices A.1 and A.2.

We emphasize that the analogy is conceptually useful if incomplete. In classical Cochran's Theorem settings, the χ^2 distributional results are used to form F-tests. Here,this is not feasible because the quadratic forms are not independent, the matrices in them are not idempotent, and some distributions have to be assumed to be χ^2 directly. We think this is reasonable but it only shows an analogy between the Cochran's Theorem decomposition and our predictive variance decomposition, not a result that can be used directly. In practice, instead of F tests, our decomposition leads to bootstrap tests that we present in the next subsection.

Indeed, a full analogy to Cochran's Theorem would lead to testing ratios of one 'Q_j' to another '$Q_{j'}$'. In practice, when we did this computationally i.e., tried the analog of F-tests, we found they always rejected or never rejected and hence were badly calibrated. Developing bootstrap tests seemed easier than trying to calibrate them correctly. More to the point, our bootstrap procedure tests the relative size of a term on the right in (2.1) to the posterior variance on its left hand side, not to another term on the right. This seemed more appropriate for predictions because the posterior predictive variance controls the width of PI's.

2.3. 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. Here, 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 $Var(Y_{n+1}; \mathcal{D}_n)$. Now, we want to test hypotheses of the form

$$H_0 : E\left(\frac{Var_{V_1}(Y_{n+1}; \mathcal{D}_n, V_1)}{Var(Y_{n+1}; \mathcal{D}_n)}\right) \geq \tau$$
$$H_1 : E\left(\frac{Var_{V_1}(Y_{n+1}; \mathcal{D}_n, V_1)}{Var(Y_{n+1}; \mathcal{D}_n)}\right) < \tau.$$

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.

Assuming that the data is representative of of the DG, we use bootstrapping on the argument of the expectation in H_0. The result is a data set of the form

$$Z_b = \frac{Var_{V_1} E(Y_{n+1}; \mathcal{D}_n^b, V_1)}{Var(Y_{n+1}; \mathcal{D}_n^b)} = \frac{\sum_{v_1=1}^{m_1} \hat{w}_{v_1}^b (\hat{y}_{v_1}^b - \sum_{v_1=1}^{m_1} \hat{y}_{v_1}^b)^2}{\sum_{v_1=1}^{m_1} \hat{w}_{v_1}^b (\hat{y}_{v_1}^b - \sum_{v_=1}^{m_1} \hat{y}_{v_1}^b)^2 + \sum_{v_1=1}^{m_1} \hat{\sigma}_{v_1}^2(b)},$$

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

$$t = \frac{\bar{z} - \tau}{SE(\bar{z})}.$$

We use τ in this expression because it corresponds loosely to seeking the uniformly most powerful test for H_0, a one sided hypothesis. Note that $\bar{z}$ 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 a important point but we do not wish to clutter the notation.

Let $J > B$. In a second layer of bootstrapping, draw J samples of size B from $z_1, \ldots, z_B$, with replacement. Denote these by $z'_1, \ldots, z'_J$ where each z'_j has B entries. To get a distribution for $T = t$ as a random variable under the null, we generate the vectors

$$\tilde{z'}_j = z'_j - (\bar{z'}_j - \tau)\mathbf{1}_B = z'_j - \left(\frac{1}{B}\sum_{b=1}^{B} z'_{j,b} - \tau\right)\mathbf{1}_B$$

where $\mathbf{1}_B = (1, \ldots, 1)$ is B-dimensional. Now, we have J different samples for which the mean is τ. From the samples corrected by their means and τ so they satisfy the null, we form the t-statistics

$$\tilde{t}_j = \frac{\bar{\tilde{z}}'_j - \tau}{SE(\bar{\tilde{z}}'_j)}$$

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

$$\widehat{ASL} = \frac{1}{J}\sum I(\tilde{t}_j \leq t).$$

When the $\widehat{ASL}$ is small, we reject H_0 and this tells us that $Var_{V_1}E(Y_{n+1}; \mathcal{D}_n, V_1) \approx 0$ suggesting that $E(Y_{n+1}; \mathcal{D}_n, V_1)$ is constant in V_1. Therefore, omitting this term in forming the PI for Y_{n+1} does not affect the width. Here, when we do this testing, we default to a threshold of $\alpha = 0.05$ for the ASL for convenience. Note that this threshold is different from τ – α is the significance level and τ is a parameter of interest.

Below we have used normality in some of our computational work because it was justified by auxilliary 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 Subsec. 2.2 and in Prop. A.2 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 Bonferoni correction or other simple 'fix' may be practical. However, for large K, we may have to use some sort of Westfall-Young correction 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 due to small sample sizes.

3. A Simulated Example

An example will make the point regarding the importance of the last term in (1.1). There has been much discussion about when different shrinkage methods are appropriate, see Wang et al. (2020) 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 Dustin et al. (2024). Otherwise, when sparsity is low or multicollinearity is a problem ridge regression is preferable. In this section, we see 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 for accurate prediction. Otherwise put, is retaining the extra variability from using multiple shrinkage techniques useful compared to selecting a single one?

Let us compare five penalized methods, namely LASSO, Ridge Regression (RR), Adaptive LASSO (ALASSO), EN, and Adaptive EN (AEN) as applied to a linear model

$$Y_i = X_i^T \beta + \epsilon_i$$

for $i = 1, \ldots, n$ where X_i is a vector of explanatory variables with $\dim(X_i) = \dim(\beta) = p < n$ and $\epsilon_i \sim N(0, 1)$ IID. Write $m_1, \ldots, m_5$ to mean the five penalty functions. Write V to be the discrete random variable assuming values over the five penalties.

Let us apply the two term variance decomposition in (1.1) using V. We suspect that the second term on the right is small relative to the left hand side because we think the models from the five methods will be very similar, i.e., they will have similar locations even if their variances are not identical. That is, we suspect the test

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

versus

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

will end up rejecting the null, meaning we can drop the second term in (1.1) at the .05 level. As noted, we are using a typical frequentist test because we do not have a likelihood for V given the data. Indeed, since we are de facto forced to use a test statistic based on bootstrapping, the frequentist definition of probability and framework for testing may be appropriate.

To investigate the behavior of the terms in the predictive variance decomposition we generate data as follows. Let $n = 50$ and $p = 100$ and take 95 of the β_j coefficients to be zero and five to be generated independently from a $N(5, (1.5)^2)$.

As described in Subsec. 2.3, this test can be performed by bootstrapping the argument of the expectation in the null hypothesis. In fact, for normal error, the distributions of the numerator and the denominator can be regarded as, approximately, convex combinations of χ^2 distributions, see Appendix A.1. So their ratio is expected to behave like an F distribution. The convex combinations can be precisely defined but are generally numerically inaccessible. Nevertheless, our testing procedure can be regarded as a simple nonparamertic approximation to standard normal theory.

Let's use the first 49 data points to form predictive distributions for each of the five methods as well as for the stacking average (based on five-fold cross-validation) of the five methods. To obtain the stacking weights $\hat{w}_1, \ldots, \hat{w}_5$, we use the methodology in Zhang and Liu (2022); see steps 1-4 in Sec. 3 for full details. Since the glmnet package is easy to use and computationally fast, obtaining the stacking coefficients in this example is easy. Now, write the stacking model average as

$$\sum_{j=1}^{5} \hat{w}_j(\mathcal{D}_{49})\hat{p}(Y_{50}; X_{50}, m_j) \tag{3.1}$$

where we have indicated the dependence of the $\hat{\beta}_j$'s in the model by writing $\hat{p}$. More explicitly,

$$\hat{p}(Y_{50}; X_{50}, m_j) = N(X_{50}\hat{\beta}_{m_j}, \hat{\sigma}^2_{m_j} + \widehat{Var}(X_{50}\hat{\beta}_{m_j})) \tag{3.2}$$

where the estimation of the decay parameters λ_j is suppressed in the m_j's and $\hat{\sigma}^2_{m_j}$ is the standard OLS estimator of σ^2 using only the variables selected by m_j – except for RR where we use the $\hat{\sigma}$ from EN since it is a combination of the L^1 and L^2 penalties. We justify this by citing Zhao et al. (2021) who showed that this procedure is consistent for LASSO. 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). To find $\widehat{Var}(X_{50}\hat{\beta}_{m_j})$ we use the bootstrapped variance estimator from the boot package in R. The use of normality in (3.2) is consistent with the fact that the shrinkage methods we have used implicitly assume normality because they are effectively based on penalized squared error. However, our bootstrapping can be used even when normality is violated.

Now, we draw another $100,000$ data points from each of the five models. Then we sample $n_j = 100,000\hat{w}_j$ from each model $\hat{p}(Y_{50}; X_{50}, m_j)$, for $j = 1, \ldots, 5$. This gives us 100,000 data points from the stacking mixture (3.1). We use these data points to assess coverage of the PI's from the five shrinkage methods and their stacking average. The PI for stacking is of the form $PI_{stack}(.05) = [q_{.025}, q_{.975}]$ where the q's are the quantiles from (3.2). Similarly, we have $PI_{m_j}(.05) = [q^{m_j}_{.025}, q^{m_j}_{.975}]$. This gives us 6 PI's.

To estimate the empirical coverage of the six PI's, we use the bootstrap again now on the entire procedure up to this point. We choose $B = 1000$. Letting $j = 1, \ldots, 6$ index the predictive distributions ($j = 6$ corresponds to the

	STK avg	LASSO	RR	ALASSO	EN	AEN
Stacking weights (raw data)		0.74	0.00	0.00	0.25	0.00
Pred. Variance (raw data)	2.97	1.02	6.71	0.99	6.73	6.70
Coverage (B-strap)	0.97	0.98	0.43	0.12	0.94	0.25

TABLE 2

***Stacking shrinkage methods**: This table gives the stacking weights, the variances of the predictive distributions, and the coverage of the PI's for five shrinkage methods and their stacking average.*

stacking average) we compute

$$\widehat{\mathsf{Coverage}}_j = \frac{1}{B}\sum_{b=1}^{B} \chi_{\{Y_b \in PI_{j,b}\}}.$$

We also have the bootstrapped variance from the j-th predictive distribution from the RHS of (3.2). This procedure bootstraps the three terms in (1.1). The details on enforcing the null hypothesis are in Subsec. 2.3. Essentially, we get a bootstrapped p-value, commonly called the achieved significance level (ASL), and reject when the ASL is too small. Our results are in Table 2.

In these computations, some values are bootstrapped and some are not as indicated in Table 2. Specifically, the stacking weights and the predcitive variances are calculated from the original data only. In this case, the weights are 'static' and we are using stacking weights analogously to how one would use posterior quantities. By contrast, when we find the coverage or perform tests we use bootstrapping and we did recalculate the weights. That is, model weights for the predictive variance decomposition are calculated once using the whole data set, however, when we find the bootstrap estimate of the stacked predictive distributions for the entries of a V_k or perform tests we recompute the model weights for each bootstrap sample. For instance, whenever we find Z_b's, the weights are not static. This holds for the results in Tables 3 and 4, also.

We see that only LASSO and EN have positive stacking weights. LASSO achieves greater than the nominal 95% coverage while EN is slightly less at 94% despite having a much larger predictive variance than LASSO. The stacked predictive distribution has an estimated variance of 2.97 and decomposes as

$$\widehat{\mathrm{Var}}(Y_{50};\mathcal{D}_n) = \widehat{E_V(\mathrm{Var}}(Y_{50};V,\mathcal{D}_n)) + \widehat{\mathrm{Var}_V}(E(Y_{50};V,\mathcal{D}_n);\mathcal{D}_n) = 2.39 + 0.58,$$

where we have indicated both the data and the estimation explicitly for clarity. In effect, we are treating $Var(E(\cdot;\cdot);\cdot)$ as a single operation. Hence we see the ratio of the between-models variance to total variance is

$$\frac{\mathrm{Var}_V \hat{E}(Y_{50};V,\mathcal{D}_n)}{\mathrm{Var}(Y_{50};\mathcal{D}_n)} = \frac{0.58}{2.97} = 0.195.$$

Informally, this suggests that there is too much between-models variance to ignore when making predictions. More formally, using our test, we obtain an $\widehat{ASL} = 0.99$ meaning we cannot reject the null. This leads us to conclude that

the second term on the LHS of (1.1) contributes more than 5% of the total predictive variance. Consequently, we should account for penalty uncertainty when making predictions.

Going beyond the information provided by a single use of our test we see that despite both LASSO and EN having good coverage, the small size of n relative to p leads us to ask what level of between-models variance would lead to rejection. In other words, if we allowed our selves to ignore a larger proportion of variance – i.e. increase the threshold 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 $\widehat{ASL} = .0095$. Hence, we would conclude that 9% is the smallest percentage at which we could ignore the contribution of the between-models variance to the overall variance. Note that we are considering several values for the proportion of variance we are willing to ignore which is reasonable depending on specific predictive setting. This is different from changing the threshold for which we compare $\widehat{ASL}$ to decide if we reject the test.

To conclude this example, we can go beyond testing, look at Table 2 and reason as follows. Since we want the correct nominal predictive coverage with the smallest K and V, we note that LASSO has smaller or equivalent variance to the other methods and at least the desired coverage. We can rule out ALASSO, AEN, and RR on the basis of poor coverage and zero stacking weight. Thus, if we choose, say, 10% (or any number bigger than 9%) as our threshold, we are led to use PI's from LASSO only. That is, V reduces to a single level. However, below this number, we are better to use LASSO and EN. If the sample size were to change e.g., we used $n = 75$ rather than $n = 50$, we would have got different values from those in Table 2 and a lower threshold than 9% – and our reasoning would likely have been different. We provide more discussion on choosing K and V in Sec. 5.

4. A Real Data Example

In this section, we analyze the data set Superconductivity presented in Hamidieh (2018). This data set has 81 explanatory variables of a physical or chemical nature to explain a response Y representing temperature measurements (in degrees 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 explanatory variables follows a signal plus noise structure, i.e.

$$Y_i = f(X_i) + \epsilon_i$$

for $i = 1, \ldots, n$ and where $\epsilon_i \sim N(0, \sigma^2)$. Hamidieh (2018) 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.

Here we use 5 common predictive models; m_1 = LM , m_2 = neural nets (NN), m_3 = projection pursuit regression (PPR), m_4 = support vector machine

with a radial kernel (SVM), and m_5 = XGBoosting (XGB). Hence we write $V = (m_1, \ldots, m_5)$. We note that these models do not have a unique error structure. However, upon examining the residuals from the other fitted models, we confirmed that the residuals were normally distributed. So, we use a normal to form a predictive distribution 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 $n+1$ observation to predict Y_{n+1}.

Let the predictor from model k be $\hat{f}_k$, $k = 1, ..., 5$. Then the next outcome is normally distributed, centered at the point predictor $\hat{f}_k(X_{n+1})$ with estimated variance

$$\widehat{Var}\left(Y_{n+1} - \hat{f}_k(X_{n+1})\right) = \widehat{Var}(\hat{f}_k(X_{n+1})) + \widehat{Var}(\hat{\epsilon}_k).$$

We calculated $\widehat{Var}(\hat{f}_k(X_{n+1}))$ by bootstrapping. That is, we found a bootstrap distribution for it and then took its variance. For $\widehat{Var}(\hat{\epsilon}_k)$, we found the variance of the residuals from the fitted model.

Formally, the predictive distribution for each model is given by

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

Since these models are implemented in a frequentist sense and we used stacking (as described in Zhang and Liu (2022)) to average over the models based on the cross-validated predictive performance, the stacked predictive distribution for Y_{n+1} is

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

Now we present two cases, one where we randomly sample 500 of the data points to form the predictive distributions and test whether the between-models variance is important, and another where we use the whole data set 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%.

First we took a random sample of 500 observations from the whole data set. We set $B = 200$ and $J = 10000$. The results are given in Table 3. Overall the results are unsatisfactory: While stacking did put a lot of weight on XGB, the procedure advocated by Hamidieh (2018), its predictive coverage is weak. On the other hand, SVM, which performed better in terms of coverage got a low stacking weight. This is likely due to the difference between coverage (what proportion of new data points are in a PI) and minimzing L^2 predictive error.

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

$$\begin{aligned}\widehat{Var}(Y_{500}; \mathcal{D}_{499}) &= E_V\widehat{Var}(Y_{500}; V, \mathcal{D}_{499}) + Var_V\widehat{E}((Y_{500}; V, \mathcal{D}_{499}); \mathcal{D}_{499}) \\ &= 135.85 + 262.23 \\ &= 398.08.\end{aligned}$$

	STK avg	LM	NN	PPR	SVM	XGB
Stacking weights (raw data)		0.10	0.26	0.12	0.01	0.51
Pred. Variance (raw data)	398.08	237.33	260.46	57.06	172.11	69.39
Coverage (B-strap)	1.00	0.87	0.79	0.26	1.00	0.79

TABLE 3

Small sample results for the Superconductivity data*: The only model with reasonable coverage, SVM, has a low stacking weight. Also, the stacking average while giving superb coverage, does so at the cost of high variance. – larger than the variance of any single model. This is consistent with high between-models variance.*

	STK avg	LM	NN	PPR	SVM	XGB
Stacking weights (raw data)		0.01	0.26	0.21	0.01	0.52
Pred. Variance (raw data)	173.73	308.60	315.28	184.14	155.32	78.71
Coverage (B-strap)	1.00	1.00	1.00	1.00	1.00	1.00

TABLE 4

Re-analyzing with all available data*: The predictive variances in this table are bigger than in Table 3 but the overall stacking variance is less than half of the earlier value. This suggests the between models variance is less important than with* $n = 500$.

Now, to test whether the between-models variance term matters, we have the hypotheses

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

versus

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

and the test statistic $\bar{z} = \frac{262.23}{398.08} = 0.66$. For $\tau = 0.05$ we obtain $\widehat{ASL} = 1$ and cannot reject the null. In this case, we cannot reject the null for any reasonable value of τ. This confirms what Table 3 showed, namely that the between-models variance is much bigger than the between-predictions within-models variance.

For contrast we redo the analysis using all the available data. Here we let $B = 50$, and $J = 5000$. Note that here we only used 50 bootstrap samples due to computational burden. The results are given in Table 4. With the larger sample size we find that all coverages are one and superficially if we had to choose one method it would be XGB.

Now the variance decompositions is

$$\begin{aligned}\widehat{\mathrm{Var}}(Y_{21263}; \mathcal{D}_{21262}) &= \widehat{E_V \mathrm{Var}}(Y_{21263}; V, \mathcal{D}_{21262}) + \widehat{\mathrm{Var}_V E}((Y_{21263}; V, \mathcal{D}_{21262}); \mathcal{D}_{21262}) \\ &= 166.57 + 7.16 \\ &= 173.73. \end{aligned} \tag{4.1}$$

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

$$H_0 : E\left(\frac{Var_V(E(Y_{21263}; V, \mathcal{D}_{21262}); \mathcal{D}_{21262})}{Var(Y_{21263}; \mathcal{D}_{21262})}\right) \geq \tau$$

versus

$$H_1 : E\left(\frac{Var_V(E(Y_{21263}; V, \mathcal{D}_{21262}); \mathcal{D}_{21262})}{Var(Y_{21263}; \mathcal{D}_{21262}))}\right) < \tau,$$

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

TABLE 5

$\widehat{ASL}$ for different choices of τ: *The reliability of the entries is potentially limited because B is low.*

and the test statistic is $\bar{z} = \frac{7.16}{173.73} = 0.041$. We used different choices of τ and observed the results in Table 5. It is seen that for $\tau = 0.05$ there is not enough evidence to say T 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 five and six percent. We suggest that if a larger value of B could have been used, the threshold for rejecting the null would likely decrease to around $\tau = .05$.

Thus, 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 null at τ around 6%. In this latter case, we are left with only the first term in (4.1) 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 modeling down to a single model. Alternatively, if we take 6% as our threshold and invoking the conclusion from our test, we can reason further from examining the entries in Table 4. We are then led to choose the method with the desired coverage and smallest predcitive variance taking into consideration the results for the stacking average. Doing this, we confirm that the preferred method of Hamidieh (2018) is well-justified. It gives high coverage and the smallest variance among the alternatives and we only need the first term in our decomposition. Moreover, XGB received the highest stacking weight, presumably because it had the smallest cross-validated error. Another way to say this is that if we retaining the variability over methods is important, 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.

5. Discussion

Here we have proposed a decomposition of the stacking predictive variance. The decompositions are based on representing modeling choices by a discrete random variable $V = V_K$ and then iterating the law of total variance for each component of V. The predictive variances control the width of prediction intervals so our decomposition lets us assess the contribution of each source of variability in V to the overall variance. We proposed a testing procdure to assess the relative contributions of the terms in the decomposition so that we can, in principle, eliminate some components of V thereby simplifying the resulting prediction intervals where possible. We show how our analysis proceeds in a series of examples and verified that our methods give 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 χ^2 random variables. We note that 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 Toutenberg and Shalabh (2009) Chap. 6.3 for some details. However, we have not developed this here.

A recurrent theme in our findings is the discrepancy between the relative contribution of a variance term to the total variance and its absolute level. The relative importance of a term depends on the sample size differently from the total variance. In particular, if the absolute level of variance is small enough, then it is not important how much each term in the decomposition contributes.

Another theme 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 crrespond to factors. Here, there is a correspondence but it is weaker and we are not sure how dropping a term related to dropping a factor.

We conclude with the observation that there may be two different choices of V that an analyst may want to consider. This leads to the question as to how to choose one over the other. In Sec. 3 and in Sec. 4 we faced a special case of this problem when we reduced a one dimensional V to a single model. Our approach can be formalized as the following empirical optimization. Recall that in expanding our model list, we want to ensure we have close to the proper coverage and the smallest variance possible among model lists with good coverage. This leads us to choose K and V in the following manner. First calculate estimated coverage using g-fold cross-validation or g bootstrapping samples and define the estimated coverage to be

$$\hat{C}(\mathcal{V}^{(K)}) = \frac{1}{g}\sum_{i=1}^{g} I_{\{y_{i,new}\in PI(\mathcal{V}^{(k)})\}},$$

where $\mathcal{V}^K$ is the model list corresponding to V. Then for given $\alpha, \delta > 0$, we choose

$$\hat{K} = \arg \min_{k\in\{k|\hat{C}(\mathcal{V}^{(k)})\in(1-\alpha-\delta,1-\alpha+\delta)\}} Var(Y_{n+1}|\mathcal{D}_n)(\mathcal{V}^{(k)}).$$

That is, we choose the value of K and the corresponding V to minimize the variance among all model lists that have estimated coverage δ-close to the nominal $1-\alpha$ coverage. (An easy example of this is to consider two model lists, one based on a fixed number of terms in a Fourier basis and another based on a set of feedforward neural nets with a fixed number of nodes; our example in Sec. 4 is a special case of this.) Despite this data-driven proposal, the problem of model list selection remains both difficult and open.

Appendix A: Cochran's Theorem

Here we continue the derivation from Subsec. 2.2 showing hown the decomposition in Clause (i) of Prop. 2.1 is analogous to Cochran's Theorem.

A.1. Deriving a χ^2 distribution for $K = 2$

First, we see that (2.4) is an expected quadratic form, i.e.

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

For (2.5), write W_i for the column vector $W_i = (\sqrt{\omega_{i1}}, \ldots \sqrt{\omega_{iJ}})'$, and write $\tilde{Y}_i$ for the column vector $\tilde{Y}_i = (\hat{y}_{i1} - \bar{y}_{i\cdot}, \ldots, \hat{y}_{iJ} - \bar{y}_{i\cdot})'$. Now (2.5) is

$$\begin{aligned} \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} \left(\hat{y}_{ij} - \bar{y}_{i\cdot}\right)^2 &= \sum_{i=1}^{I} \xi_i W_i' \tilde{Y}_i \tilde{Y}_i' W_i \\ &= \sum_{i=1}^{I} \xi_i \tilde{Y}_i' W_i W_i' \tilde{Y}_i. \end{aligned} \quad \text{(A.2)}$$

Similarly, for term (2.6), write S for the column vector $S = (\sqrt{\xi_1}, \ldots, \sqrt{\xi_I})'$ and $\bar{Y} = (\bar{y}_{1\cdot} - \bar{y}, \ldots, \bar{y}_{I\cdot} - \bar{y})'$. Then we have that (2.6) is

$$\begin{aligned} \sum_{i=1}^{I} \xi_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 &= S' \bar{Y} \bar{Y}' S \\ &= \bar{Y}' S S' \bar{Y}. \end{aligned} \quad \text{(A.3)}$$

So, using (A.1), (A.2), and (A.3), we can rewrite (2.3) as

$$Var(Y_{n+1}; \mathcal{D}_n) = \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} E((Y_{n+1} - \hat{y}_{ij})^2; \mathcal{D}_n, m_{ij}, s_i) \quad \text{(A.4)}$$

$$+ \sum_{i=1}^{I} \xi_i \tilde{Y}_i' W_i W_i' \tilde{Y}_i \quad \text{(A.5)}$$

$$+ \bar{Y}' S S' \bar{Y}. \quad \text{(A.6)}$$

Now we see each term in the predictive variance is a quadratic form, i.e., a homogeneous polynomial of order two, even if the terms in (A.4) are (trivial) quadratic forms of dimension one.

To see how the distributional aspects of (A.4), (A.5), and (A.6) 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,

$$\begin{aligned} Var(Y_{n+1}; \mathcal{D}_n) - \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} E((Y_{n+1} - \hat{y}_{ij})^2; \mathcal{D}_n, m_{ij}, s_i) \\ = \sum_{i=1}^{I} \xi_i \tilde{Y}_i' W_i W_i' \tilde{Y}_i + \bar{Y}' S S' \bar{Y} \end{aligned} \quad \text{(A.7)}$$

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

To begin, we recall Theorem 2.1 in Box (1954) 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 $Q = X^T MX$ is any real quadratic form of rank $r \leq p$, Q is distributed like a quantity

$$\sum_{j=1}^{r} \lambda_j \chi_1^2 \tag{A.8}$$

with $r \leq p$ and λ_i the i^{th} eigenvalue of ΨM.

Now, look at the first term on the right, and let $A_i = W_i W_i'$. We know A_i is a $J \times J$, symmetric, and semi-positive definite because (A.5) 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 $\tilde{Y} \sim N(0, \Sigma^*)$ and $\sqrt{\xi_i}\tilde{Y}_i \sim N(0, \Sigma_i)$.

Now, since both terms on the right in (A.8) are quadratic forms in a normal random vector, we can apply Theorem 2.1 in Box (1954) to each of them. So, (A.8) gives

$$Var(Y_{n+1}; \mathcal{D}_n) - \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \omega_{ij} E((Y_{n+1} - \hat{y}_{ij})^2; \mathcal{D}_n, m_{ij}, s_i) \sim \sum_{i=1}^{I} \xi_i \sum_{j=1}^{J} \lambda_{ij} \chi_1^2 + \sum_{i=1}^{I} \lambda_i \chi_1^2 \tag{A.9}$$

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 (A.8) 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 $(Y_{n+1} - \hat{y}_{ij}; \mathcal{D}_n, m_{ij}, s_i) \sim N(0, \sigma_{ij}^2)$ and observe

$$\begin{aligned} E((Y_{n+1} - \hat{y}_{ij})^2; \mathcal{D}_n, m_{ij}, s_i) &= Var(Y_{n+1} - \hat{y}_{ij}; \mathcal{D}_n, m_{ij}, s_i) + E(Y_{n+1} - \hat{y}_{ij}; \mathcal{D}_n, m_{ij}, s_i)^2 \\ &= Var(Y_{n+1} - \hat{y}_{ij}; \mathcal{D}_n, m_{ij}, s_i) \\ &= \sigma_{ij}^2. \end{aligned} \tag{A.10}$$

We recognize this as equivalent to the expectation of a χ_1^2 random variable scaled by σ_{ij}^2 – i.e. $E(\sigma_{ij}^2 \chi_1^2) = \sigma_{ij}^2$. It is difficult to determine the distribution of (A.10) explicitly but because we are taking a convex combination of terms like it, computations suggest it is approximately normal.

Since all three terms in (2.3) are variances and hence corrected for their means, (A.4) is a new term that arises from trying to derive a Cochran's theorem style representation of $Var(Y_{n+1}; \mathcal{D}_n)$ 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, i.e., the number of factors, plus an extra term, (A.4), the predictive analog of the residual term.

If desired, we can approximate distributions of the right hand terms in (A.9) more compactly by using other results from Box (1954). For instance, his Theorem 2.2 gives the formula for the i^{th} cumulant of (A.8) as

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

Using this, we can approximate (A.8) 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

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

Box gives this approximation in part because it has the same first two moments as (A.8). Box also notes that when all λ_j are equal, the degrees of freedom, h, is smaller than appropriate.

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

$$g\chi_h^2 = \frac{\sum \lambda_i^2}{\sum \lambda_i}\chi^2\left(\frac{(\sum \lambda_i)^2}{\sum \lambda_i^2}\right). \tag{A.11}$$

Also, we can approximate

$$\sqrt{\xi_i}\tilde{Y}_i' A_i \sqrt{\xi_i}\tilde{Y}_i = \sqrt{\xi_i}\tilde{Y}_i' W_i W_i' \sqrt{\xi_i}\tilde{Y}_i$$

by

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

Hence, we have the approximate distribution

$$Var(Y_{n+1};\mathcal{D}_n) - \sum_{i=1}^{I}\xi_i \sum_{j=1}^{J}\omega_{ij}E((Y_{n+1}-\hat{y}_{ij})^2;\mathcal{D}_n, m_{ij}, s_i) \overset{approx}{\sim} \sum_{i=1}^{I} g_i\chi_{h_i}^2 + g\chi_h^2.$$

We emphasize that the analogy is conceptually incomplete as noted in at the end of Subsec 2.2. In addition, we do not have a definite distribution for the second term on the left in (A.8).

A.2. General K

Deriving quadratic forms and distributional expressions for $Var(Y_{n+1};\mathcal{D}_n)$ for general K is similar to the derivation of (A.8) and (A.9), respectively, seen in Subsec.2.2. 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 A.1. *For a K-factor predictive scheme, the predictive variance can be written as a sum of weighted quadratic forms as follows:*

$$\begin{aligned}
Var(Y_{n+1}; \mathcal{D}_n) = &\sum_{i_1=1}^{I_1} p\left(v_{i_1}; \mathcal{D}_n\right) \ldots \sum_{i_K=1}^{I_K} p\left(v_{i_K}; \mathcal{D}_n, v_{i_1}, \ldots, v_{i_{K-1}}\right) E\left(\left(Y_{n+1} - \tilde{y}_{v_{i_1},\ldots,v_{i_K}}\right)^2; \mathcal{D}_n, v_{i_1} \ldots, v_{i_K}\right) \\
&+ \sum_{i_1=1}^{I_1} p\left(v_{i_1}; \mathcal{D}_n\right) \ldots \sum_{i_K=1}^{I_{K-1}} p\left(v_{i_{K-1}}; \mathcal{D}_n, v_{i_1}, \ldots, v_{i_{K-2}}\right) \tilde{Y}'_{K,\ldots,1} A_{K,\ldots,1} \tilde{Y}_{K,\ldots,1} \\
&+ \sum_{i_1=1}^{I_1} p\left(v_{i_1}; \mathcal{D}_n\right) \ldots \sum_{i_{K-2}=1}^{I_{K-2}} p\left(v_{i_{K-2}}; \mathcal{D}_n, v_{i_1}, \ldots, v_{i_{K-3}}\right) \tilde{Y}'_{K-1,\ldots,1} A_{K-1,\ldots,1} \tilde{Y}_{K-1,\ldots,1} \\
&\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
&+ \sum_{i_1=1}^{I_1} p(v_{i_1}; \mathcal{D}_n) \tilde{Y}'_{2,1} A_{2,1} \tilde{Y}_{2,1} \\
&+ \tilde{Y}'_1 A_1 \tilde{Y}_1, \qquad\qquad (A.12)
\end{aligned}$$

where

$$A_{k,\ldots,1} = W_{k,\ldots,1} \left(W_{k,\ldots,1}\right)', \qquad (A.13)$$

$$W_{k,\ldots,1} = \left(\sqrt{p\left(v_{i_k=1}; \mathcal{D}_n, v_{i_1}, \ldots, v_{i_{k-1}}\right)}, \ldots, \sqrt{p\left(v_{i_k=I_k}; \mathcal{D}_n, v_{i_1}, \ldots, v_{i_{k-1}}\right)}\right),$$

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

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

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

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^2_1 random variables.

Proposition A.2. *Let $\left(Y_{n+1} - \tilde{y}_{v_{i_1},\ldots,v_{i_K}}\right) \sim N(0, \sigma^2_{i_1,\ldots,i_K})$, $\tilde{Y}_1 \sim N(0, \Sigma)$ and $\tilde{Y}_{k,\ldots,1} \sim N(0, \Sigma_{k,\ldots,1})$. Then the sum of quadratic forms in (A.12) are*

distributed like a sum of weighted χ-squared random variable as follows

$$\begin{aligned}
Var(Y_{n+1};\mathcal{D}_n) \sim & \sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_K=1}^{I_K} p\left(v_{i_K};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-1}}\right) E((Y_{n+1}-\tilde{y}_{v_{i_1},\ldots,v_{i_K}})^2;\mathcal{D}_n,v_{i_1},\ldots,v_{i_K}) \\
& + \sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_{K-1}=1}^{I_{K-1}} p\left(v_{i_{K-1}};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-2}}\right)\sum_{i_K=1}^{I_K}\lambda_{K\ldots,1}\chi_1^2 \\
& + \sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\ldots\sum_{i_{K-2}=1}^{I_{K-2}} p\left(v_{i_{K-2}};\mathcal{D}_n,v_{i_1},\ldots,v_{i_{K-3}}\right)\sum_{i_{K-1}=1}^{I_{K-1}}\lambda_{K-1\ldots,1}\chi_1^2 \\
& \qquad\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
& + \sum_{i_1=1}^{I_1} p\left(v_{i_1};\mathcal{D}_n\right)\sum_{i_2=1}^{I_2}\lambda_{2,1}\chi_1^2 \\
& + \sum_{i_1=1}^{I_1}\lambda_1\chi_1^2 \qquad\qquad (A.14)
\end{aligned}$$

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.

Acknowledgments

Dustin acknowledges funding from the University of Nebraska Program of Excellence in Computational Science. Both authors acknowledge computational support from the Holland Computing Center at the University of Nebraska.

References

Box, G. (1954). "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." *Ann. Math. Statist.*, 25: 290–302. 21, 22

Clarke, B. (2003). "Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored." *J. Machine Learning Res.*, 4: 683–712. 4

Draper, D. (1995). "Assessment and Propagation of Model Uncertainty." *J. R. S. S. B*, 57(1): 45–97. 3, 6, 8

Dustin, D. and Clarke, B. (2022). "Testing for the Important Components of Posterior Predictive Variance." arXiv:2209.00636. 2, 5

Dustin, D., Clarke, J., and Clarke, B. (2024). "Predictive Stability Criteria for penalty selection in linear models." *Comp. Stat.*. 12

George, E. (2010). "Dilution priors: Compensating for model space redundancy." In *IMS Collections Vol. 6*, 158–165. Inst. Math. Statist. 5

Gustafson, P. and Clarke, B. (2004). "Decomposing Posterior Variance." *J. Stat. Planning and Inference*, 119(2): 311–327. 8

Hamidieh, K. (2018). "A data-driven statistical model for predicting the critical temperature of a superconductor." *Comp. Materials Sci.*, 154: 346–354. 15, 16, 18

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley and Sons. 8

Toutenberg, H. and Shalabh (2009). *Statistical Analysis of Designed Experiments*. Springer, New York. 19

Wang, W., Mukherjee, S., Richardson, S., and Hill, S. (2020). "High dimensional regression in practice: An empirical study of finite-sample prediction, variable selection, and ranking." *Statistics and Computing*, 30: 697–719. 12

Wolpert, D. H. (1992). "Stacked Generalization." *Neural Networks*, 5(2): 241–259. 4

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Using Stacking to Average Bayesian Predictive Distributions." *Bayesian Analysis*, 13(3): 917–1007. 4

Zhang, X. and Liu, C.-A. (2022). "Model averaging prediction by K-fold cross-validation." *Journal of Econometrics*.
URL https://www.sciencedirect.com/science/article/pii/S0304407622000975 4, 13, 16

Zhao, S., Witten, D., and Shojaie, A. (2021). "In defense of the indefensible: A very naive approach to high dimensional inference." *Statistical Science*, 36: 562–577. 13