Desiderata for a Predictive Theory of Statistics
B. Clarke
Department of Medicine, Department of Epidemiology and Public Health, and Center for Computational Sciences
University of Miami

June 2009

SUMMARY

In many contexts the search for good models and validation of them is as important as standard notions of credibility or confidence. This is particularly so in fields involving complex or high dimensional data. This paper suggests a unified treatment for such data based on six ‘desiderata’ that are an effort to list and explain criteria that a good predictive theory of statistics should satisfy.

Key words: online prediction, bias variance analysis, overall sensitivity, residual analysis.
1 Introduction.

The conventional model for inference rests on using outcomes of random variables such as $X : (\Omega, \mathcal{F}, P) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to select from a class of objects say $\mathcal{P}$. Often $\mathcal{P}$ is a parametric family, but may be a nonparametric class of densities, distributions, or functions or a class of decisions, among other possibilities. A probability measure $P_X$ can be induced on the image of $X$ in which case the range measure space is written $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$; it is the “visible” part of the random variable. That is, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ is the part of a phenomenon that we model; the underlying triple, $(\Omega, \mathcal{F}, P)$, is essentially masked from us and not modeled. The idea is that the macroscopic world of real data that we analyze arises from an unseen world and we use our data analysis and $\mathcal{P}$ in an effort to uncover $P$. From this point, mainstream statistical analysis roughly separates into two philosophies, Bayes and Frequentist.

The conventional model, while amazingly successful, is not immune from criticism. These criticisms come from two directions. The first direction is that it is not at all clear how well the conventional model scales up to high dimensional and complex data. There is evidence that as the overall complexity increases, the search for good models becomes more important than formal inference about good models. Indeed, it is not necessarily clear that the idea of a true model is even very useful in settings with complex and high dimensional data. Wrong but simple models that are good approximations to a complex true model may give better results than the true model for realistic sample sizes and collections of covariates that are feasible to measure. In addition, there may be settings in which few realistic approximations, if any, are satisfactory.

In the Bayesian context, the problem with scaling up to high dimensions is seen in the phenomenon of dilution – a sufficient number of wrong but good models may split the posterior probability so finely that a genuinely poor model may appear better, see Chipman et al. (2001). An extension of this idea is that the posterior may in fact converge to zero if the model list is permitted to grow too quickly and includes enough good models that they all dilute. Steps can be taken to correct this, see George (2001), but they rest on ‘uniformizing’ the prior over models on neighborhoods. Essentially this limits the choice of prior and appears to rule out priors that have a claim to objectivity in the complex data setting. (The only two well-known proposals for objective priors in high dimensional settings are the Dirichlet process prior with a uniform base measure and the Jeffreys prior). Importantly, prior selection and posterior exploration in high dimensional contexts is hard because it is difficult to form an intuition for the local properties of such large spaces, especially as it relates to data collection.

Current efforts to scale up linear regression models to high dimensions via shrinkage criteria include SCAD, Adaptive Elastic Net, ALASSO, and ACOSSO among others. These are Bayesian because the penalty term is equivalent to a prior so maximizing the objective function amounts to finding the posterior mode. It is unclear how susceptible these methods are to dilution when $p >> n$, however, this class of techniques derives its justification from the oracle property, see Fan and Li (2001). That is, the oracle property requires that the number of covariates $p$ grows slowly enough with the sample size $n$. In practice, it is unclear
in general what this growth rate means, but for the Adaptive elastic net, \( p = n^\gamma \) for \( \gamma < 1 \), see Zou and Zhang (2008) and for the SCAD penalty, Huang and Xie (2008) require \( p = o(\sqrt{n}) \). Commonly this is violated by -omics data where \( p \approx 20,000 \) and \( n \approx 50 \). Moreover, it is unclear whether these methods reduce to Hodges’ super-efficiency, see Leeb and Potscher (2001), and it is seen in practice that they can give poor results in the presence of dependence among the covariates.

There are other efforts besides shrinkage methods to achieve variable selection or dimension reduction. While these methods are helpful and often give improved performance, most have limitations. One of the most important and most recent techniques in this class is called sure independence screening (SIS), see Fan and Lv (2008). While SIS does scale up to the case that \( p \) is exponential as a function of \( n \), it rests on correlation and so is essentially a linear, marginal procedure subject to the usual problems of such methods. (For instance, if \((Y, X_1, X_2) = (3, 3, 1), (4, 4, 1), (5, 5, 1), (3, 8, 2), (4, 9, 2)\) and \((5, 10, 2)\) both \(X_1\) and \(X_2\) have correlation zero with \(Y\) even though \(Y\) clearly depends strongly on both.) It is not clear yet how well this method performs on complex data even when marginal methods will suffice. For instance, Webster (2008) used SIS and permutation methods to reduce one million dimensional SNP data to around 250,000 dimensions. Even with this reduction, typical shrinkage methods will break down for realistic sample sizes.

From another perspective, classical and Bayesian nonparametrics are also efforts to deal with complex and high dimensional data. The usual problem with classical nonparametrics is that it suffers the Curse of Dimensionality. There are regression techniques such as neural networks and projection pursuit that evade the Curse, see Barron (1993) and Zhao and Atkeson (1993). However, neural networks are notoriously unstable (even with regularization) and projection pursuit does not seem to have been investigated as well as it appears to deserve. Here, neural networks and projection pursuit are regarded as techniques for constructing black-box predictors. The contribution of this paper is to suggest how analyses based on them may be improved.

Bayes nonparametrics, by contrast, is promising, but to a large extent is still under development and can be difficult to implement in general. Gaussian process priors and Dirichlet process priors seem to be the main choices that have been used. Thus, prior selection in general remains very much an open question for good performance and interpretation and easy implementation.

The second direction of criticism of conventional models is that there is a strong tendency for them to fail to validate for complex data. Ransohoff (2004, 2005) documents many examples from the cancer literature where the results of studies failed to stand up under repeated testing, despite the conclusions having seemed persuasive. Ransohoff (2004) identifies overfitting as the most obvious shortcoming of otherwise plausible analyses and Ransohoff (2005) states that bias is such a pervasive and hard-to-detect problem that “results are guilty of bias until proven innocent”. In part, Ransohoff’s examples are the logical outcome of failing to account adequately for all the uncertainties in modeling, see Draper (1995), which are especially severe in the ‘-omics’ fields that Ransohoff surveys. Overall, Ransohoff’s criticisms of existing approaches lead him to emphasize extensive validation of
results before they should be considered established. It would be naive to think that it is only in the -omics fields that validation often fails.

Overall, inference techniques derived under the conventional model are well suited to models with a relatively small number of coefficients in a relatively simple model with a relatively simple data set and decent sample size i.e., when a sparse model is “true” or at least best. In some of these cases, the population for which a sparse model exists will be somewhat narrow or artificial so that actual use is open to question. Moreover, in many contexts, it is not reasonable to assume sparsity. The typical case encountered in complex and high dimensional data types, such as in the -omics world, is often that the observed phenomenon is the result of many small contributions that may themselves be complex. Moreover, in the -omics world, sparsity is often interpreted as a proportion of terms selected from a very large set of variables rather than as an absolute number. Thus, a sparse model may have 10,000 variables out of a million measured, with a sample size of 100.

Emerging data types such as Next Generation Sequencing, in which literally millions of variables are measured, promises to challenge both statistical methods based on sparsity and their applications even further. The severity of the challenge is even greater when sparsity does not hold. For sparse models in enormous model classes, or for settings in which sparsity does not hold, the dimension and complexity is so great that computing is a challenge, let alone inference method formulation and implementation. As noted above, the search for good models often is the focus rather than inference per se.

Taken together, these two points – difficulty finding reliable methods in complex data settings and ensuring that they work well in reality – suggests there may be room for a reformulation of the statistical paradigm, away from the conventional model and, to an extent, away from both the orthodox Bayesian and Frequentist philosophies. That is, it may be that difficulty of scaling up and validation may not be with the methods per se so much as the limitations of the conventional model for inference.

An alternative to the conventional model is provided by the Prequential approach, see Dawid (1984). The Prequential Principle comes in three forms, weak, strong and super-strong. The weak prequential principle – that is assumed unless otherwise specified – requires that any criterion of agreement between a forecaster and a data generator (DG) should depend only on the actual observed sequences of forecasts and outcomes, and not on any of the strategies which might have been used to produce the forecasts or outcomes. The strong and super-strong prequential principles add conditions to the weak prequential principle, see David and Vovk (1999). One of the undernoticed implications of the Prequential principle is that it can be interpreted to be “strictly forward” in the sense that retrodiction is ruled out. Consider one-day-ahead predictions of rain over \(n\) days. Then, using a construct such as the probability assigned to rain for 14 July on 13 July given that it did rain on July 15 would be disallowed.

One of the benefits of the Prequential approach is that the centrality of comparing forecasts and outcomes emphasizes the role of validation. Indeed, the Prequential principle only requires forecasts and outcomes to be applied, thereby reducing the importance of the conventional model. This is partially because good prediction is a more general task than
model identification but also because the DG is ruled out as a factor in evaluating forecasts. Thus, Prequentiality is more closely related than conventional modeling is to Popperian ideas about the falsification of hypotheses on which science rests more generally. Note that the Prequential approach is also parsimonious in that it does not invoke any unseen world. Indeed, the conventional model is only brought into the Prequential approach as a method for analyzing longrun performance, as is typical of other techniques such as support vector machines.

One of the criticisms of the Prequential approach is that because of its sequentiality, it is eminently suited to sequential data but provides little guidance for analyzing a batch of data that have been collected. With batch data, often the order of collection has been lost or, since independence from sample to sample is assumed, many would argue that any ordering is artificial. The obvious answer to this is to ‘batchify’ the Prequential approach by choosing a number of random permutations of the data, do the analysis sequentially for each, and then average over the results to make predictions and quantify aggregate behavior like the sequence of predictive errors. This sort of procedure – essentially a sort of bootstrapping – is used in an example in Section 4.

Separate from averaging over permutations, there are three reasons why looking at batch data sequentially is useful. First, looking for good predictors is more general problem than finding a good model and so is a weaker criterion that has a better chance of being achievable. Roughly, every model has an associated predictor, but it is not clear that every predictor corresponds to a model. Searching more broadly for predictors may result in obtaining better predictive properties than a model would have. The cost would be not having a model, but many of the interpretive properties one can derive from a model can also be derived from a predictor. For instance, the importance of a covariate in linear regression can be evaluated by the effect of leaving it out not just by comparing its coefficient to the coefficients of other covariates. In fact, the focus on prediction may not be too costly from a modeling standpoint. Often, it will be possible to identify a model whose natural predictor will approximate a good predictor and give satisfactory predictions, if not as quite as good as the predictor itself. Thus, if a useful model really is true, it will convert to a good predictor and conversely. Moreover, in this fortunate case, the near equivalence may carry over to hypothesis tests. For instance, parameter estimation and testing is de facto equivalent in the presence of a pivotal quantity.

Second, there are insights into the data that emerge from looking at it sequentially that are not likely to be found from a batch analysis approach. First, it is only a sequential approach that reveals some predictors routinely outperform Bayes predictors, see Wong and Clarke (2004). This only happens in finite samples, of course, because Bayes predictors are asymptotically optimal. Likewise, Clarke (2003) has shown that Bayes predictors are often outperformed, again in small samples, when the model list does not contain the true model; i.e., in the presence of bias. These examples are related to the fact that prediction and Bayesian analysis are subtly different. Roughly, by focussing attention on the support of the prior, the Bayesian works with a “closed mind” and can never discover that a model list is wrong without having a clear alternative, see Dawid (1992) who argues that falsification of
hypotheses (in the sense of Popper) must be included in Bayesian analysis. Essentially, looking at the data sequentially permits one to ask the question whether the observed outcomes have diverged far enough from the predictions that a change in model space is necessary.

Third, one of the strengths of the predictive approach to validation for batch data is that it helps avoid taking the batch data so seriously that one ends up fitting noise and being unable to detect it in the model. That is, because Prequential techniques develop models sequentially, they may be less prone to over- and under-fit because extraneous terms are eventually ruled out and missing terms may be found as \( n \) increases, addressing Ransohoff (2005). In addition, in Prequential modeling validation is inherent in the procedure whereas in batch modeling only fitting is inherent in the procedure; validation has to be done as a separate procedure. Essentially, this addresses Ransohoff (2004).

To see this third point in more detail, consider comparing sequential prediction to cross-validation (CV) in a regression problem. Recall that residuals are a natural assessment of fit more than validation. However, residuals can be converted into an assessment of validation by internal prediction. So, consider a fixed sample of size \( n \). Suppose \( k_1 \) data points are used to fit a model, \( k_2 \) data points are used for validation, and \( k_3 = n - k_1 - k_2 \) data points are left over. Then, to do internal validation, values of \( k_1 \) and \( k_2 \) must be chosen. It makes sense to choose more than one pair \((k_1, k_2)\) because otherwise inferences may depend on the way the data are split. So, the question is which pairs to choose. In \( K \)-fold CV, the data points are given a fixed order and partitioned into \( K \) disjoint subsets. Each subset is held out in turn for validating a model formed using the other \( K - 1 \) subsets. Thus, \( k_1 = (K - 1)n/K \), \( k_2 = n/K \), and \( k_3 = 0 \) and \( K \) different partitions of the data are used. By contrast, in Prequential validation, i.e., online prediction, \( k_1 = 0, 1, 2, \ldots, k_2 = k_1 + 1 \) and \( k_3 = n - k_2 \) and \( n - 1 \) different partitions of the data are used. Thus as \( n \) increases, the raw number of tests of a predictor increases, unlike in CV. So, Prequential validation in batch data is just another instance of internal validation. The reason to choose Prequential validation over cross-validation is that letting the size of the training set increase is a proxy for permitting models of increasing complexity as data accumulate.

The central issue in this paper is how to manage the overwhelming importance of ensuring inferences from a specific data set validate beyond a specific narrow context. The narrow context may be as narrow as the specific data set itself, but more generally is the output of a data generator, perhaps one that it quite complex. To address the validation challenges, we propose a list of set of six desiderata that may provide a framework for addressing a large class of statistical problems predictively. Another way to say this is that these desiderata may define a sort of ideal set of criteria a good analysis should satisfy. These six desiderata are not formal enough to be a paradigm or analytic framework, however, they are formal enough that their prescriptions can be meaningfully interpreted in some cases. For the sake of giving it a name, this set of desiderata is called a Coordinating Theory in the hope that it may inter-relate many of the foundational ideas in statistics. It will be seen that the approach advocated here is not purely Bayesian, but is much closer to Bayesian than it is any other existing school of statistical thought.

The six desiderata themselves are an effort to complete the Prequential approach to an
effective alternative to the main philosophies such as Bayes, Frequentist, Likelihood and Conditional Likelihood that are derived from the conventional model described at the beginning. Note that use of the word “complete” is meant to convey the belief that, while many properties of the Prequential approach are known, there are statistical issues conceptually disjoint from online prediction (narrowly defined) that are important and should be integrated into Prequentialism to help make it a full prescription for statistical analysis. The benefits of proposing an ideal analysis include identifying what features of modeling are generically most important to address and identifying what features of a statistical problem are most important to be used in further study. The hope is that this Coordinating Theory will be well suited to complex data for which the conventional model does not scale up well.

Briefly, the six desiderata are (i) predictive optimality subject to the Prequential principle, (ii) the use of prediction errors to update the prediction scheme, (iii) a proper treatment of all sources of variance and bias, (iv) the complexity of the prediction procedure should be related to the difficulty of approximating the true model, (v) a complete robustness analysis, and (vi) a sanity check that the procedure behaves properly in limiting senses. It is easy to imagine alternative desiderata – many have argued that Savage’s axioms for Bayesian inference, von Neumann’s axioms for decision theory or Rostek’s axioms for percentile based inference are quite satisfactory and little more is needed. Also, it is not hard to imagine revamping the set of six desiderata by using ideas from sequential hypothesis testing, data compression and transmission, or even an analysis of the dependency structures in a data set, in place of those for prediction. However, the argument that the six desiderata presented here are appropriate rests on the notion that prediction is the central optimality property to seek and that there are subsidiary criteria that we want an ideal prediction scheme to satisfy. Thus, the Coordinating Theory perspective on batch analysis is not that it is wrong or bad but, rather, if batch analysis succeeds, it does so by achieving good prediction, or, in effect, by satisfying the six desiderata.

The three most novel features of the six desiderata are that they introduce the concept of going outside the Bayesian paradigm to rechoose a decision problem in (ii), a comprehensive variance bias decomposition in (iii) and a complete robustness analysis in (v), distinguishing between local and global perturbations. It must be reiterated that the point of the desiderata is to provide structure for a generic problem, not to criticize existing methods per se. The criticism of methods above based on validation probably arises from using those methods under criteria based on fit rather than validation. Thus, the intent is here is to situate existing methods in a predictive structure to compare their performance.

It will be seen that the relationship between a Coordinating Theory and Bayes or Frequentist philosophies is intricate. From the Coordinating Theory perspective, other schools of thought such as Bayes, Frequentist, likelihood and conditional inference, are techniques by which to construct predictors with certain properties. The Coordinating Theory would then compare the predictors on the basis of the six desiderata. The setting of desideratum (i) is decision-theoretic and hence more Bayesian than Frequentist. Desideratum (ii) is, at root, a Frequentist criterion meant to quantify bias and variance over repeated sampling. Desideratum (iii) recognizes that Bayes is optimal in a given decision problem but that choosing
which decision problem to solve to get a good predictor is usually not Bayesian. Desideratum (iv) is neither Bayes nor Frequentist; it is imported from the minimum description length philosophy of statistics. Desideratum (v) invokes a sort of objective Bayes robustness. And desideratum (vi) is again a Frequentist property from repeated sampling. Taken together, the Desiderata provide much more structure for inference than the usual Machine Learning view which rests on predictive performance almost exclusively.

The structure of this paper is as follows. In the next Section the six desiderata of the Coordinating Theory are stated and explained. Section 3 then discusses the relationship of the Coordinating Theory to existing schools of thought in Statistics, focusing on Bayes and Frequentist thinking. In Section 4, several aspects of the first four desiderata are explored in a computed example. Section 5 discusses several of the implications of the perspective developed here.

2 Six Desiderata for Predictive Analysis

Four typical inference problems in statistics are classification, model identification, decision making and prediction. It is easy to see that classification and model identification can be expressed in terms of prediction: A good classifier or model based on a sample of size \( n \) should perform well on future outcomes compared to other classifiers or models based on the same data. Also, decision making can be evaluated predictively: Rather than using a decision rule to compare two hypotheses, for instance, convert both hypotheses into predictors and evaluate their performance. The hypothesis with the lower prediction error is selected as true. (For \( H_1 : \theta \in \mathcal{R}_1 \) vs. \( H_2 : \theta \in \mathcal{R}_2 \), given data \( D_n = \{(Y_i, X_i) : i = 1, ..., n\} \) where the \( X_i \)'s are covariates, one can convert hypotheses to predictors \( \hat{F}_{j,i} \) for \( j = 1, 2 \) by writing \( \hat{F}_{j,i} = E(Y_{i+1}|R_j(\theta)|D_i) \), i.e., mixing over the sets in the hypotheses.) More generally, the decision with lower predictive error is preferred. Thus, it may be reasonable to use a generic prediction problem as the central setting of statistics instead of measure-theoretic probability as for conventional models.

So, suppose the task is to predict the next outcome \( Y_{n+1} \) using \( X_{n+1} \) and to determine the predictor \( \hat{F}_n \) given a stream of data \( D_n \). Here, for each time step \( i \), a value of \( X = (X_1, ..., X_p) \) is a set of explanatory variables for \( Y_i \). Thus, we write the prediction for time \( n + 1 \) as \( \hat{Y}_{n+1} = \hat{F}_n(X_{n+1}) \). Now, we can state the six desiderata.

Desideratum #1: Evaluate a predictive scheme by its online cumulative predictive error subject to the Prequential principle.

The intuition behind this desideratum is that we must have some way to evaluate how well a predictive scheme (our modeling) has performed relative to reality and that the evaluation must be fair. The fairness is imposed by the Prequential principle: When comparing two predictors it is fair if (i) we only look at how each predictor makes use of the data and (ii) we ensure that no predictor uses information about the model that is unavailable to other predictors.

The structure of this paper is as follows. In the next Section the six desiderata of the Coordinating Theory are stated and explained. Section 3 then discusses the relationship of the Coordinating Theory to existing schools of thought in Statistics, focusing on Bayes and Frequentist thinking. In Section 4, several aspects of the first four desiderata are explored in a computed example. Section 5 discusses several of the implications of the perspective developed here.

2 Six Desiderata for Predictive Analysis

Four typical inference problems in statistics are classification, model identification, decision making and prediction. It is easy to see that classification and model identification can be expressed in terms of prediction: A good classifier or model based on a sample of size \( n \) should perform well on future outcomes compared to other classifiers or models based on the same data. Also, decision making can be evaluated predictively: Rather than using a decision rule to compare two hypotheses, for instance, convert both hypotheses into predictors and evaluate their performance. The hypothesis with the lower prediction error is selected as true. (For \( H_1 : \theta \in \mathcal{R}_1 \) vs. \( H_2 : \theta \in \mathcal{R}_2 \), given data \( D_n = \{(Y_i, X_i) : i = 1, ..., n\} \) where the \( X_i \)'s are covariates, one can convert hypotheses to predictors \( \hat{F}_{j,i} \) for \( j = 1, 2 \) by writing \( \hat{F}_{j,i} = E(Y_{i+1}|R_j(\theta)|D_i) \), i.e., mixing over the sets in the hypotheses.) More generally, the decision with lower predictive error is preferred. Thus, it may be reasonable to use a generic prediction problem as the central setting of statistics instead of measure-theoretic probability as for conventional models.

So, suppose the task is to predict the next outcome \( Y_{n+1} \) using \( X_{n+1} \) and to determine the predictor \( \hat{F}_n \) given a stream of data \( D_n \). Here, for each time step \( i \), a value of \( X = (X_1, ..., X_p) \) is a set of explanatory variables for \( Y_i \). Thus, we write the prediction for time \( n + 1 \) as \( \hat{Y}_{n+1} = \hat{F}_n(X_{n+1}) \). Now, we can state the six desiderata.

Desideratum #1: Evaluate a predictive scheme by its online cumulative predictive error subject to the Prequential principle.

The intuition behind this desideratum is that we must have some way to evaluate how well a predictive scheme (our modeling) has performed relative to reality and that the evaluation must be fair. The fairness is imposed by the Prequential principle: When comparing two predictors it is fair if (i) we only look at how each predictor makes use of the data and (ii) we ensure that no predictor uses information about the model that is unavailable to other predictors.
There are two ways this desideratum can be interpreted. The first is for stochastic data, the second is for data types that cannot plausibly be regarded as the outcome of random sampling.

Since it is more familiar, we begin with the stochastic case. For stochastic data, the CPE can be written as

\[ CPE(n) = \frac{1}{n} \sum_{i=1}^{n} L(\hat{Y}_i, Y_i), \]  

for a loss function \( L \), where it is understood that the data points are ordered so that only points up to stage \( i \) are used to predict \( Y_{i+1} \). One instantiation of this is to use the mean squared predictive error (MSPE) in which \( L \) is taken to be squared error loss. The CPE is calculated at each time step \( i \) and represents the sum of the sequence of residuals from the use of a sequence of predictors \( \hat{F}_i \) for \( i = 1, ..., n \) from a specified predictive process. Since \( e_i = Y_i - \hat{Y}_i \) is the residual from a predictor, \( e_i \) is often called a predictual. More generally, the error \( L(\hat{Y}_i, Y_i) \) is also called a predictual. Note that using the CPE is equivalent to online prediction: (1) can be found from the one step errors from strictly online prediction and the one step errors from online prediction can be recovered from the sequence \( CPE(1), ..., CPE(n) \).

The CPE converges to a limit \( E(CPE) \) by the strong law of large numbers. So, the limiting quantity may be minimized. However, for finite \( n \) minimization of \( CPE(n) \) over a large enough class of predictors will often yield an interpolant for \( D_n \) which will have poor predictive properties. Thus, to avoid getting a sequence of interpolants from minimization in (1) the CPE is only used to evaluate predictors. Optimizing \( E(CPE) \) is disallowed by the Prequential principle as well because it depends on the model under which \( E \) is taken.

It is easy to imagine variations on Desideratum #1. After all, the intent is only to compare a sequential vector of predictions to a vector of outcomes fairly. Thus, a median predictive error such as

\[ MedPE = \text{Median}_{i=1,...,n} L(\hat{Y}_i, Y_i), \]

could be used in place of (1). (Again, it is understood that each prediction \( \hat{Y}_i \) depends only on the previous data.) Other operations to be performed on the predictuals are also possible.

Another type of variation on Desideratum #1 is to relax the weak Prequential principle to a sort of weaker Prequential principle. For instance, as noted in the Introduction, CV and online prediction can be regarded as members of the class of internal prediction criteria for batch data. So, if the key point is validation, then a weaker Prequential principle permitting some retrodiction may be enough. Consider the setting that \( M \) predictors are available and suppose each is derived from a specific model for the data. Then, if the data is IID or stationary more generally, it may make sense to ignore the ordering up to and including the \( i \)th time step when making a prediction for time \( i + 1 \). Thus, the sequentiality of the prediction is maintained for present-time use but all the accumulated data can be used in a less restricted fashion. Under this weaker Prequential principle, it would be permissible to
use K-fold CV with all the data at time \( i \) to select one of the \( M \) models to make a prediction at time \( i + 1 \).

Now consider the case that we have a nonstochastic data type that cannot plausibly be regarded as the outcome of random sampling. In these cases, the data often is a long vector \((x_1, \ldots, x_p)\) with statistical regularities so a sequential approach may be useful. However, neither a mean nor a variance is a reasonable summary. The sequence of As, Ts, Gs and Cs that define a chromosome in an individual would be an instance of this and it is possible that patterns of base pairs throughout the chromosomes would be of interest. For these cases, a variant on (1) would be better.

Start by regarding online prediction as a sequential game between Nature, \( N \), and a Forecaster, \( F \), permitting the Forecaster access to a collection \( \mathcal{E} \) of experts indexed by \( \theta \) where \( \theta \in \mathcal{E} \), see Shtarkov (1988), Haussler and Barron (1992), and Haussler, Kvinien, and Warmuth (1998). Each round of the game is organized by a Referee who sequentially gets the opinions of the experts, receives \( F \)'s density from which predictions will be made, receives \( N \)'s choice of outcome \( x \), and then calculates how much \( F \) must pay to \( N \) or the reverse. At each step each player knows the history of the entire game, so the data accumulate.

Suppose each round of the game uses a log scoring rule. Then, to start the game, each expert \( \theta \) announces a density \( p_\theta(\cdot) \) for \( x_1 \). The Forecaster sees this and tries to match the performance of the best expert in \( \mathcal{E} \) by choosing the density \( q(\cdot) \) from which to make predictions for \( x_1 \). Then, \( N \) chooses the actual value of \( x_1 \) arbitrarily. The Referee then makes \( F \) pay \( \ln \frac{1}{q(x_1)} \) to \( N \). So, the question is how much \( F \) must pay to \( N \) or the reverse. At each step each player knows the history of the entire game, so the data accumulate.

Thus, \( F \) should choose \( q \) to minimize the difference

\[
\ln \frac{1}{q(x_1)} - \inf_\theta \ln \frac{1}{p_\theta(x_1)} = \sup_\theta \ln \frac{p_\theta(x_1)}{q(x_1)}. \tag{2}
\]

Expression (2) is called the regret. Clearly, the worst \( N \) could do to \( F \) would be to choose \( x_1 \) to maximize (2). So, \( F \) might be led to choose \( q \) to achieve the minimax value

\[
\inf_{q \in \mathcal{P}} \left( \sup_{x, \theta} \log \frac{p_\theta(x)}{q(x)} \right). \tag{3}
\]

Shtarkov (1988) proved that the optimal \( q \) is

\[
q_{\text{opt}}(x) = \arg_{q \in \mathcal{P}} \left[ \inf_{q \in \mathcal{P}} \left( \sup_{x, \theta} \log \frac{p_\theta(x)}{q(x)} \right) \right] = \frac{p(x_1|\hat{\theta})}{\int p(x|\hat{\theta}(x))dx}, \tag{4}
\]

the normalized, maximized likelihood. When (4) exists, the value of (3) is the log of the normalizing constant in \( q_{\text{opt}} \) and when \( x \) is a vector of length \( n \), say \( x \) is replaced by the vector \((x_1, \ldots, x_n) = x^n\), it has an asymptotic form (in \( n \)) given by Rissanen (1996, eq. 6). The Shtarkov solution \( q_{\text{opt}} \) is an average of the models over the sample space, but the \( n \)-th solution (for \( x^n \)) is not the marginal from the \( n+1 \) solution (for \( x^{n+1} \)). Xie and Barron (1997,
2000) give a complete minimax analysis of this case for discrete \( x \), including determining the least favorable priors.

In the Bayesian version of this game, the experts are taken as subjectively weighted by a prior \( w(\theta) \) based on their reliability. So, (4) becomes

\[
\arg \left[ \inf_{q \in \mathcal{P}} \left\{ \sup_{x,\theta} \log \frac{w(\theta)p_{\theta}(x)}{q(x)} \right\} \right] = \frac{w(\tilde{\theta})p(\tilde{\theta})}{\int w(\theta)p(\theta|\theta(x))\,d\theta},
\]

(5)
a variant on Shtarkov (1988) and Rissanen (1996), where \( \tilde{\theta} \) is the posterior mode rather than the MLE. The asymptotics for (5) when \( x \) is replaced by \( x^n \) can be found, see Clarke (2007).

To put this in the setting of Desideratum #1, imagine F plays \( n \) rounds of the Bayesian game (5); this will reduce to the Frequentist game if \( w \equiv 1 \). So, use \( x = x^n \) but examine the sequence of \( n \) univariate games for each \( x_i \), \( i = 1, ..., n \) assuming that at each time step all the players have access to the previous moves. Omitting subscripts on the \( q \)s and sequentializing the Bayesian version of the Shtarkov game means finding

\[
\arg \left[ \inf_{q \in \mathcal{P}} \left\{ \sup_{x,\theta} \log \frac{w(\theta|x^{i-1})p_{\theta}(x_i|x^{i-1})}{q(x_i|x^{i-1})} \right\} \right].
\]

(6)

So, set

\[
\tilde{\theta}_i = \arg \max_{\theta} p_{\theta}(x|\theta|x^{i-1})
\]
to see that the optimal density for the \( i \)-th round is now

\[
q_{opt,i}(x_i|x^{i-1}) = \frac{w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})}{\int w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})\,dx_i}.
\]

(7)

The value of (6) is now

\[
\ln \int w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})\,dx_i.
\]

The cumulative regret is now the CPE for the non-stochastic case, parallel to (1) for the stochastic case. For the Shtarkov predictor, write the CPE as

\[
C\text{Regret}(n, x^n) = \sum_{i=1}^{n} \ln \frac{w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})}{q_{opt,i}(x_i|x^{i-1})} = \sum_{i=1}^{n} \ln \int w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})\,dx_i.
\]

Note that predictive performance requires specification of a class of predictors and a criterion for good prediction both at time \( n \), but there is no necessity that the class or criterion must be constant as a function of \( n \). In some cases, this means that the Frequentist principle can lead to unexpected results. Wong and Clarke (2004) developed “mongrel” predictors based on re-selection of the error criterion used to form them and showed they
outperformed Bayes methods in small samples in a Prequential sense in the stochastic case. This verifies that while Bayes methods can be asymptotically optimal they needn’t be finite sample optimal; this holds for the Shtarkov setting as well. In essence, standard Bayes methods can be outperformed predictively when the decision problem that a predictor solves at time \( n + 1 \) is allowed to be different from the decision problem it solves at time \( n \). That is, the criterion and action space used to choose a predictor is updated from time \( n \) to time \( n + 1 \), not just the the predictor itself. The interpretation of this kind of extensive updating is that over time, we refine our idea of what problem to solve to find good predictors. This point motivates the second desideratum.

**Desideratum #2: Use Prediction Errors to Update the Prediction Problem.**

In the sequential prediction problem, a prediction must be made at each time step based on the accumulated data. So, imagine fitting a linear regression model using a data set and making a prediction from it. When the next data point is revealed, there are two possibilities. First, the prediction is good in the sense of (1) say and the model is validated. Then, it is enough to update the parameter estimates with the new data point and make another prediction. Second, the prediction is deemed inadequate: The new data point gives an excessive value for (1), is outside the 95\% prediction interval, leads to too many predictions that are systematically higher than their observed \( y_i \)'s (although \( L(y_i, \hat{y}_i) \) is small), or leads to some other undesirable feature. In this case, we may be led to reformulate the problem before estimating any parameters.

Apart from the predictive scenario, this is standard thinking when using residual analysis to improve the model of a data generator until only noise is left. Indeed, Desideratum #2 is little more than automated residual analysis on a larger scale because it permits whole classes of predictors to be rechosen as well as the technique for choosing a member of such a class, not just rechoosing a predictor from a class by the same technique.

To see this more formally, suppose reality chooses \( F \) to be true and we first choose a space of functions, \( F \), to help form \( \hat{F} \). Then, we might set up a decision problem for predicting the next outcome as follows. Fix an initial list \( M_0 \subset F \) with \( M_0 \) items, where \( M_0 = \{ f_{0,1}, f_{0,2}, \ldots, f_{0,M_0} \} \) is the action space \( A_0 \) and the loss \( L \) is squared error. Put a prior \( W \) on the list \( M_0 \), and assign priors for any parameters within each model \( f_{0,k} \). Since the Bayes model average (BMA) minimizes the posterior risk, it is the Bayes action from which to make predictions at \( t = n + 1 \).

Now, suppose that as an experimenter accumulates data, the CPE grows so large that for \( n > n_0 \) the adequacy of the modeling strategy is doubtful. Perhaps the list \( M_0 \) was wrong (too large, too small, poorly located etc.) thereby giving a poor action space \( A_0 \) with high CPE. Perhaps the loss function or the priors, or even \( F \) is found to be wrong. Then, the experimenter might reformulate the prediction problem with new choices of the model list, optimality criterion, or action space. After problem reselection, the CPE based on all the accumulated data would still be used provided all the observations were from the same data generator.
An instance of reselecting the loss function would be starting with say $L^1$ loss giving the median of the posterior as the optimal action, but switching to the $L^2$ loss and using the posterior mean as data accumulate. The median would give robustness for small sample sizes but convert to a mean when enough data was available that robustness was not so important. Another instance of changing the action space would be using stacking initially until a model list with low enough bias had been found so that BMA (under $L^2$) would be appropriate. An example of rechoosing the model list is given in Section 4.

Repeating these reselections, in principle for each time step, one version of the general problem may be stated as, for each $n$, find $M_n \subset \mathcal{F}$ with $M_n = \{f_{n,1}, \ldots, f_{n,M_n}\}$, finitely many finitely parametrized models to be used at time $n$ and find a criterion under which to choose $\hat{F}_n$ as a model average from $\mathcal{M}_n$, with weights determined by the data up to time $n$.

Note that updates of the prediction problem can have new optimal actions that might or might not be Bayes. Indeed, it is likely that this reselection of the prediction problem to solve cannot be done in a Bayesian fashion. However, between times at which the prediction problem is reformulated, the Bayesian version would be either correct or correct for a long enough run of predictions. Because the prediction problem is reformulated from time to time, the prediction scheme effectively searches over prediction problems to find the right one to solve to obtain good prediction. That is, the early prediction problems and their solutions are estimates of the correct prediction problem and its solution.

Desideratum #3: Generate a Unified Bias-Variance Analysis.

The intuition is that the nature of the uncertainty associated to each input to a predictive scheme should be examined. Variance and bias are two established assessments of uncertainty, but, as seen below data compression formulations may also be relevant in the non-stochastic data case. It is hard to state clearly in generality how Desideratum #3 can be implemented so consider the following examples, the first of which are stochastic. The last two are nonstochastic.

Begin with a very simple predictive system using one model that is known up to finitely many parameters denoted $\beta = (\beta_1, \ldots, \beta_p)$. The only inference problem is to estimate $\beta$. Provided there is a value $\beta_T$ that makes the model correct, the predictor based on the plug in estimate $\hat{\beta}$ should give better and better predictions as it converges to $\beta_T$. That is, parameter consistency gives predictive optimality asymptotically. In this case, there is only one bias-variance decomposition based on the mean squared error

$$MSE(\beta_T, \hat{\beta}) = \sum_{j=1}^{p} \text{bias}(\hat{\beta}_j, \beta_j, T)^2 + \sum_{j=1}^{p} \text{Var}(\hat{\beta}_j),$$

in which $\text{bias}(\hat{\beta}_j, \beta_j, T) = E\hat{\beta}_j - \beta_j, T$ and the expectations and variance are taken in the distribution indexed by $\beta_T$. Note that in this case there is an estimate $\hat{\beta}$ using the data and there is a true value $\beta$ so that both terms are well defined.

As an extension of (8) to function estimation, Domingos (2000) established a decomposition of the prediction risk into three terms, valid for a class of loss functions that includes...
zero-one and squared error. Two of the terms are recognizable as variance and bias while the third represents the irreducible noise of the data generator. This analysis rests on the regression model \( Y = F(X) + \epsilon \) in which \( F \) is estimated by \( \hat{F}(\cdot) = \hat{F}(\cdot; D_n) \) where \( D = D_n \) is the sample of size \( n \) accumulated up to time \( n + 1 \). The Domingos (2000) analysis would apply to a nonparametric predictor that had no inputs. This might correspond to, say, kernel regression in which the tuning parameter were chosen as a fixed value independent of the data so its value could be treated as a robustness issue under Desideratum #5. (If the smoothing parameter \( h \) were chosen by CV for instance, or estimated as a function \( h(x|D) \), then a separate bias-variance decomposition would be required for \( h \).) Here, it is seen that \( \hat{F} \) plays the same role in this example as \( \hat{\beta} \) does in the last case; however, the analog of the mean \( E\hat{\beta} \) is the main prediction \( y_m \) for \( Y \) at \( x \) defined by

\[
y_m(x) = \arg \min_{y'} E_D L(\hat{Y}(x), y'),
\]

in which the expectation is taken with respect to the joint distribution for \( D \) and \( L \) is a loss function. Thus, the variance of the predictor \( \hat{Y}(x_{n+1}) = \hat{F}(x_{n+1}) \) is

\[
\text{Var}(x_{n+1}) = E_D L(y_m, \hat{Y}(x_{n+1})).
\]

Note that (9) and (10) are Case I in James and Hastie (1997) who treated the absolute error and zero-one loss cases and this is consistent with Heskes (1998) who developed similar expressions for the relative entropy loss.

To define a general concept of bias, let \( y^* = y^*(x) \) be the ideal predictor for \( Y \) defined by

\[
y^*(x) = \arg \min_{\hat{y}} E_{\text{noise}} L(Y(x), \hat{y}),
\]

in which the expectation is taken in the noise model. Note that, with respect to noise, \( y_m \) is the best predictor under the joint distribution for \( D \) and \( y^* \) is the best predictor under noise as in (8). Then, Domingos (2000) defines

\[
\text{Bias}(x) = L(y^*, y_m).
\]

Thus, \( y^* \) and \( y_m \) play roles analogous to \( \beta_T \) and \( \hat{\beta} \) in (8).

Given (10) and (12), the proof of Theorem 1 in Domingos (2000) establishes that an analog to (8) for the function estimation case is

\[
E_D L(y^*(x), \hat{Y}(x)) = L(y^*(x), y_m(x)) + E_D L(y_m(x), \hat{Y}(x)),
\]

thereby giving a bias-variance decomposition for a variety of choices of \( L \). More generally, Domingos (2000) establishes

\[
E_{D,\text{noise}} L(Y(x), \hat{Y}(x)) = E_{\text{noise}} L(Y(x), y^*(x)) + L(y^*(x), y_m(x)) + E_D L(y_m(x), \hat{Y}(x))
= N(x) + \text{Bias}(x) + \text{Var}(x),
\]

14
for a variety of loss functions. The term \( N(x) \) in (14) is the inherent, irreducible variability of the data generator independent of and \( \hat{Y} \). In a sense, it is a variance for \( Y \), rather than for \( \hat{Y} \).

Note that (8) and (13) each involve essentially one variance around the true value of the parameter and around the main prediction, respectively. Likewise, (8) and (13) each involve essentially one bias away from the true value of the parameter and away from the ideal predictor, respectively. That is, in these two examples, there was only a single unknown input to the prediction scheme, the unknown parameter or the unknown function, and so only a single variance and bias to examine. The term unified bias-variance analysis here refers to using several versions of (8) or (13) to analyze, in principle, all inputs to the predictor. First, there are three kinds of input, apart from the data. Some inputs are knowable in the sense that they can be estimated from the data. Some inputs are not estimable – a prior density would be the paradigm case. Some inputs are taken as known, e.g., it might be known that the best way to analyze a signal is in its Fourier basis. The third class of input will be examined in Desideratum 5 since if an input is known it doesn’t have a meaningful bias or variance apart from zero.

The idea is that each input that is estimable from the data, such as \( \beta \) or \( F \) in the last two cases, requires a variance-bias analysis. By contrast, non-estimable inputs require a variance analysis only since they must be chosen and this will have uncertainty but they do not have a true value that can be used to form a notion of bias.

To see how a unified bias-variance analysis can be done for more than one estimable input, consider a simple case of Frequentist model averaging. Suppose \( Y = F(x) + \epsilon \) in which \( F \in \mathcal{F} \), a space of functions, and let the class of terms that can be used in a linear model predictor for \( F \) be \( \mathcal{E} \). So, the full list of models is \( \mathcal{M} = \mathcal{M}(\mathcal{E}) \) with cardinality \( \text{card}(\mathcal{M}) = m \). A typical element of \( \mathcal{M} \) is \( f(x|\beta) = x^T \beta_f \) where \( x_f \) is the vector of terms from \( \mathcal{E} \) defining \( f \) and \( \beta_f \) is the corresponding vector of coefficients.

Given data \( D \), the stacking average, see Wolpert (1992), is

\[
\hat{F}_{\text{stack}}(x_{n+1}) = \sum_{j=1}^{m_n} \hat{w}_{n,j} \hat{f}_j(x_{n+1}),
\]  

(15)

where \( \hat{f}_j(x) = f_j(x|\beta_f) \) for some estimator and \( m_n = \text{card}\mathcal{M}_n \) is the cardinality of the set \( \mathcal{M}_n \) of models used to form the average for time \( n + 1 \). The weights \( w_{n,j} \) are found by a CV-type procedure. Let \( f_j^{(-u)}(x) \) be the prediction at any \( x \) using model \( j \), as estimated from training data with the \( u \)-th observation removed. Then the estimated weight vector \( \hat{w}_n = (\hat{w}_{n,1}, \ldots, \hat{w}_{n,m_n}) \) solves

\[
\hat{w}_n = \text{argmin}_w \sum_{u=1}^{m_n} \left[ y_u - \sum_{j=1}^{m_n} w_j \hat{f}_j^{(-u)}(x_u) \right]^2.
\]  

(16)

This puts low weight on models that have poor leave-one-out accuracy; five or ten-fold cross-validation would be better in practice but is not needed for the present argument. The class

\[ \mathcal{C} \]
over which \( w_n \) is optimized can make a difference; natural choices include no constraints, all for all \( j \) \( w_{n,j} \geq 0 \), and \( \sum_j w_{n,j} = 1 \). (Equation (16) does not satisfy the first desideratum; however, it can be modified so that the only data used to estimate \( \hat{f}_j^{(-u)} \) precedes \( u \); this may require a burn-in period.)

Now, suppose that under some optimality criterion, the best predictor of \( Y_{n+1} \) using a linear combination of models in \( \mathcal{M} \) based on \( D \) can be defined. For instance, parallel to (11), this might be given by the list, weights, and parameter values

\[
(\mathcal{M}_T, (w_1^T, ..., w_m^T), (\beta_1^T, ..., \beta_m^T)) = \arg \min_{A,w_1,\ldots,w_{\text{card}(A)},\beta_1,\ldots,\beta_{\text{card}(A)}} E\| F(x) - \sum_{j \in A} w_j f_j(x|\beta_j) \|, (17)
\]

where \( A \subset \mathcal{M} \), \( m_T = \# \mathcal{M}_T \), \( \| \cdot \| \) is the norm based on the Euclidean inner product, and the \( \beta_j \)s are to be estimated, for instance by MLEs from a sample of size \( n \).

To obtain a unified bias-variance analysis for the present case, note there are three estimable inputs, namely, the parameters \( \theta_f \), the weights \( \hat{w}_n \), and the model list \( \mathcal{M}_n \). Using (17) in place of (11), an analog to (8) can be constructed for each of the estimable inputs. First, the “true” model list \( \mathcal{M}_T \) can be regarded as a vector \( e = (e_1, ..., e_{\text{card}(\mathcal{M})}) \) of zeros and ones and the estimated model list can be regarded as \( a = (a_1, ..., a_{\text{card}(\mathcal{M})}) \) where \( a_j = 1 \) if model \( j \) is in \( \mathcal{M}_n \) and zero otherwise. In \( a \), exactly \( m \) of the entries are ones and \( a_j = a_j(D) \) where \( j \) indexes the \( j \)-th \( f \) in \( \mathcal{M} \). Now, parallel to (8), the bias-variance expression for the model list is

\[
\text{MSE}(\mathcal{M}_T, \mathcal{M}_n) = \| \mathcal{M}_T - Ea \|^2 + \sum_{j=1}^{\text{card}(\mathcal{M})} \text{Var}(a_j). \tag{18}
\]

The use of \( Ea \) as the location in (18) in place of the \( w_j \)s is justified for \( n \) large enough if the estimators \( a_j \) are consistent for their means; similar remarks apply to the cases below.

Second, generic model weights may be regarded as a vector \( (w_1, ..., w_{\text{card}(\mathcal{M})}) \) with one entry for each model in \( \mathcal{M} \). So, the model weights for \( \mathcal{M}_T \), namely \( (w_1^T, ..., w_{m_T}^T) \), can be embedded into the weight vector \( \mathbf{w}^T = (w_1^T, ..., w_{\text{card}(\mathcal{M})}^T) \) in terms of \( \mathcal{M} \). Each element \( w_j^T \) of \( \mathbf{w}^T \) is now either the corresponding element in the vector for \( \mathcal{M}_T \) or zero. Also, the estimated model weights \( \hat{w}_n \) for \( \mathcal{M}_n \) can be embedded into a vector \( \hat{\mathbf{w}} \) of length \( \text{card}(\mathcal{M}) \). Again, \( \hat{w}_j \) in \( \hat{\mathbf{w}} \) is either the corresponding estimated weight \( \hat{w}_j = \hat{w}_j(D) \) or is zero. Now, the variance-bias expression for the model weights is,

\[
\text{MSE}(\mathbf{w}_T, \hat{\mathbf{w}}) = \| \mathbf{w}_T - E\hat{\mathbf{w}} \|^2 + \sum_{j=1}^{\text{card}(\mathcal{M})} \text{Var}(\hat{w}_j). \tag{19}
\]

Third, like the model list and the weights, the parameters can also be embedded in a higher order space. Write \( \mathbf{B} = (\beta_1, ..., \beta_{\text{card}(\mathcal{M})}) \) to mean the collection of parameters from the models in \( \mathcal{M} \) and let \( (\beta_1, ..., \beta_{\text{card}(\mathcal{M})}) \) be the parameters from the models in \( \mathcal{M}_T \). Then we can define \( \mathbf{B}_T \) to represent the parameters from the models in \( \mathcal{M}_T \); i.e., an entry \( \beta_j^T \) in
\( B_T \) is \( \beta_f^T \) if \( f \in \mathcal{M}_T \) and is \( \beta_f = 0 \) for \( f \in \mathcal{M} \setminus \mathcal{M}_T \). Similarly, \( \hat{B} \) is the vector of estimated parameters with entries \( \hat{\beta}_f \) for \( f \in \mathcal{M}_n \) and \( \hat{\beta}_f = 0 \) for \( f \in \mathcal{M} \setminus \mathcal{M}_n \). Now,

\[
\text{MSE}(B_T, \hat{B}) = \sum_{f \in \mathcal{M}} \text{bias}(\hat{\beta}_f, \beta_f^T) + \sum_{f \in \mathcal{M}} \text{Cov}(\hat{\beta}_f). \tag{20}
\]

Taken together, the three equations (18), (19) and (20) are unified bias-variance analysis for the stacking predictor.

To see how a unified bias-variance analysis can include inputs that are not estimable, consider a new predictor, the BMA. Given a prior \( W \) across the models in \( \mathcal{M} \) and a set of priors \( \{W_f \mid f \in \mathcal{M}\} \) for the parameters \( \beta_f \), the BMA can be written

\[
\hat{Y}_{n+1} = \sum_{f \in \mathcal{M}} W(f \mid D) X_f E_{W_f}(\beta_f \mid D). \tag{21}
\]

It is well-known that BMA is optimal in an \( L^2 \) sense. Now, the estimable inputs are the model list \( \mathcal{M}_n \), the model weights \( W(f \mid D) \), and the parameter estimates, here taken to be posterior means, \( E(\beta_f \mid D) \). The non-estimable inputs are the within model priors and the across model priors. Since it is obvious how to adapt (18), (19) and (20) to the BMA setting, it is enough to give the extra expressions for the variances of the two kinds of priors used to form the BMA. Note that priors do not have a bias because they are not estimable.

To give a variance for a single \( W_f \) on \( \beta_f \) recall Gustafson and Clarke (2004) and regard \( W_f \) as an element of a class of priors indexed by \( \lambda = \lambda_f \in S_f \). That is, there is some \( \lambda_0 \in S_f \) so that \( W_f = W_{\lambda_0} \). Since \( \lambda_f \) is a hyperparameter, write its distribution as \( \Pi_f \).

Now, by the conditional covariance identity,

\[
\text{Cov}(\beta_f \mid D) = E_{\Pi_f} \text{Cov}_{W_f}(\beta_f \mid D, \Lambda_f) + \text{Cov}_{\Pi_f}(E_{W_f}(\beta_f \mid D, \Lambda_f)), \tag{22}
\]

where \( \Lambda_f \) is the random variable with distribution \( \Pi_f \) taking values \( \lambda_f \in S_f \). The first term in (22) is the usual parameter uncertainty averaged with respect to \( \Pi_f \) across the possible priors, in principle it is already included in the variance expression for the parameters as in (20). The second term, however is new. It is the uncertainty about the estimator \( E(\beta_f \mid D, \lambda) \) of the parameter from the across-prior variation due to \( \Pi_f \). This second term therefore represents the variance of the prior, relative to \( \Lambda_f \).

An analogous technique can be used to give a variance for \( W \). First, regard \( W \) as an element of a class of priors \( \mathcal{P} \) indexed by \( \psi \in S \) with associated random variable \( \Psi \) having distribution \( Q \). Let the vector of model weights be

\[
w = (W(1 \mid D), \ldots, W(\text{card} \mathcal{M} \mid D)), \tag{23}
\]

where \( i = 1, \ldots, \text{card} \mathcal{M} \) indexes the \( f \)’s in \( \mathcal{M} \). Extending (23) to include the hyperparameter \( \psi \) leads to

\[
w_\psi = (W_\psi(1 \mid D), \ldots, W_\psi(\text{card} \mathcal{M} \mid D)). \tag{24}
\]
Now, let us use $w_\psi$ in place of $w$ to evaluate variability. Let $\text{Dir}(w_\psi)$ denote the Dirichlet distribution on the $\text{card} \mathcal{M} - 1$ dimensional simplex of weights $(w_1, \ldots, w_{\text{card} \mathcal{M}})$ on the models in the BMA. Again, the conditional covariance identity can be applied. It gives

$$
\text{Cov}(W|\mathcal{D}) = E_Q \text{Cov}_{\text{Dir}}(w|\mathcal{D}, \Psi) + \text{Cov}_Q E_{\text{Dir}}(w|\mathcal{D}, \Psi).
$$

(25)

The first term in (25), like the first term in (22), represents the uncertainty in the model weights as parameters that is accounted for in (19) and the second term in (25), like the second term in (22) is the uncertainty about the estimator $E_{\text{Dir}}(w|\mathcal{D}, \psi)$ of the model weights from the across-model prior variation due to $Q$. Clarke and Gustafson (2004) show how this reasoning can be extended to assign a variance to the choice of space of functions used in models; this would give a third term in (26) below.

Now, the extra variation due to the inputs that cannot be estimated is from (22) and (25), given by

$$
\sum_{f \in \mathcal{M}} \text{trace} (\text{Cov}(E(\beta_f|\mathcal{D}, \Lambda_f)) + \text{Cov}(E(w|\mathcal{D}, Q)).
$$

(26)

So the unified variance bias decomposition for BMA is given by the analogs of (18), (19), (20) and (26).

It is easy to imagine further examples. For instance, developing a variance bias analysis for neural net regression or projection pursuit regression would be useful. In addition to the number of terms to include in each case, neural nets would have variance and bias terms for the number of nodes, the connections between the nodes, the within-node weights and the across node weights. Projection pursuit would have variance and bias terms for the number of terms and the direction of projection.

In some simple cases, an elaborate bias-variance analysis is not explicitly necessary. Consider a binary classification problem with many explanatory variables under zero-one loss. The model list would be very large including some simple models and some very complex models. Let the data be partitioned as $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ with $\mathcal{D}_1 \cap \mathcal{D}_2 = \phi$ to represent training and testing. Suppose a simple model built on $\mathcal{D}_1$ was found to perform quite well on $\mathcal{D}_2$. It might be decided, for instance, that this sort of verification were enough if squared-error CV were used since it seeks it a good tradeoff between bias and variance. That is, under enough assumptions (light tails in the error, the validity of a parsimonious model etc) squared error CV might be deemed de facto equivalent to a bias-variance analysis concluding that all the terms were satisfactorily small. Nevertheless, the point remains that a detailed examination of bias and variance is important to confirm validity. Confirmation is important because it is not clear a priori whether such limited validation is enough. For instance, a detailed bias-variance analysis may reveal how weak predictors can be improved. A specific example is given in Ransohoff et al. (2008) who tries to correct excessive variance and bias in earlier studies developing molecular markers for colorectal cancer. This paper develops a classifier with a sensitivity of 78 %, a specificity of 53 %, and an accuracy of 63% – better than chance and enough to motivate further work but not really good enough for clinical use.
As a fifth example, consider the case that the data type is a long nonstochastic vector and recall the use of the log-scoring rule at the end of the discussion of Desideratum #1. In this setting, the fact that the data is nonstochastic means that variances of “estimators” are not reasonable. However, there will be a distance between a prediction given by an action and an outcome so that a bias expression may be reasonable.

For the Shtarkov solution (4) in the non-Bayes setting, it is enough to assign a bias to (7). Doing so necessitates a concept of the “true” action, a density \( q_T(x_i) \). One way to do this is to enlarge the problem by considering a large set \( \Gamma \) of densities that contains the parametric family \( p(\cdot|\theta) \) and to seek the minimal complexity density estimator, see Barron and Cover (1990) for a given set of data. Let \( L(\cdot) \) be a fixed coding scheme for the elements of \( \Gamma \) and recall that the Shannon codelength for \( x^n \) with respect to \( q \in \Gamma \) is \( \ln(1/q(x^n)) \). Now, the two stage codelength for \( x^n \) is

\[
\ell_\Gamma(x^n, q) = \ln L(q) + \ln \frac{1}{q(x^n)}. \tag{27}
\]

Using \( \Gamma \) enlarges the problem so that the performance of \( q_{\text{opt}}(\cdot) = q_{\text{opt}}(\cdot) \) can be assessed relative to \( q_T(x^n) = \arg \min_{q \in \Gamma} \ell_\Gamma(x^n, q) \).

Since \( q_T \) is a minimum over a larger set relative to \( q_{\text{opt}}(\cdot) \), the difference in data compression properties between \( q_{\text{opt}}(\cdot) \) and \( q_T(\cdot) \) is

\[
\ln \frac{1}{q_T(x^n)} - \ln \frac{1}{q_{\text{opt}}(x^n)} = \ln \frac{q_T(x^n)}{q_{\text{opt}}(x^n)}. \tag{28}
\]

Taking the expectation in the true density gives the relative entropy \( D(q_T \| q_{\text{opt}}) \geq 0 \). If \( \Gamma \) is convex then \( q_T \in \Gamma \) and the information projection \( q_{\text{proj}} \) of \( q_{\text{opt}} \) satisfies

\[
D(q_T \| q_{\text{opt}}) = D(q_T \| q_{\text{proj}}) + D(q_{\text{proj}} \| q_{\text{opt}}) \tag{29}
\]

where

\[
D(q_{\text{proj}} \| q_{\text{opt}}) = \min_{p \in \Gamma} D(p \| q_{\text{opt}}),
\]

see Csiszar (1975, eq. 1.5,7). The first term in (29) measures how close the minimum complexity density estimator is to the best approximation within \( \Gamma \). The second term represents how close \( \Gamma \) can approximate the Shtarkov action \( q_{\text{opt}} \). In this case, the variance-bias treatment consists of two biases and no variance.

The Bayesian version of the prediction along a string of data introduces a prior \( w \) and the Shtarkov solution is (5). So, a variance term for \( w \) must be given. As in the BMA case above, a variance for \( w \) can be obtained from the trace of the diagonal of the second term in (22). So, in this case, the variance-bias treatment consists of two biases and a variance.
No variance-bias decomposition is needed for $\tilde{\theta}$ because the data are not stochastic. Instead, the issue is the robustness of the solution to $\tilde{\theta}$ as in Desideratum #5 below.

**Desideratum #4:** Every prediction procedure should have an associated complexity with a meaning formally related to the difficulty of approximating the true model.

Intuitively, the complexity of a prediction task i.e., the difficulty approximating a DG, and the complexity of predictors are both related to predictive errors. More complex predictors are often better at approximating a wider range of DGs than less complex predictors and so give better predictive performance when the DG is complex or little is known about it. Less complex predictors are often better at giving serviceable approximations to a DG when a simple approximation is good. In both cases, a good approximation means small prediction errors and this is dependent on the complexity of the DG. Note that the converse often holds as well. That is, if the complexity of a predictor is allowed to be too large relative to the approximation task, then ridiculous predictors with seemingly small prediction errors can be found while if the complexity of a predictor is too small then the best approximation may still be quite poor. Although these two cases would likely be ruled out by their excessive CPEs under Desideratum #1 as more data accumulated or by their excessive variance or bias respectively under Desideratum #3, the main deficiency is inappropriate complexity.

The complexities of predictors and prediction problems are important because, under the Prequential principle alone, two predictors giving the same sequence of predictions are indistinguishable even though one must be chosen. It can be seen that bias-variance by itself is not enough because an approximation task does not intrinsically have a meaningful bias or variance although it does have a complexity; this can be measured in a variety of ways including the number of terms required to give an approximation of prespecified exactitude on a domain. The implication is that it is often good strategy to seek a predictor with a complexity appropriate to the prediction problem. The belief undergirding Desideratum #4 is that when a predictor of the appropriate complexity is used, its complexity will match the complexity of the approximation problem, yield the optimal predictive error, and give the best unified bias-variance analysis.

To implement this, observe that in the simplest case, the prediction problem devolves to estimating a real parameter and the complexity of estimating a real parameter can be assessed in several ways, the MSE being perhaps the most common. Thus, the MSE can be regarded as an overall complexity of approximating the true model. Separately, a predictor has a complexity and the most common way to express this is by its bias and variance. Then, the usual bias-variance decomposition (8) is the relationship between the complexity of the predictor and the complexity of the prediction problem. The optimal complexity of the prediction problem is the minimal MSE which corresponds to the optimal complexity of the predictor in terms of bias and variance. More generally, Desideratum #4 requires an evaluation of how complex or difficult the prediction problem is, how complex the predictor is (expressed perhaps as a sum of difficulty terms involving biases and variances), and a relationship between them.
There are numerous ways to define the complexity of a prediction task and of a predictor. Although not entirely satisfactory, the most obvious possibility is based on a unified bias-variance analysis. The idea is to generalize from the bias-variance decomposition in (8) by merely defining the complexity of the predictor to be the sum of its squared bias and variance terms and then taking the sum as the definition of the complexity of the prediction task.

To see this explicitly, observe that the unified bias-variance analyses for Desideratum # 3 leads to four expressions for the complexity of the prediction problem, one for each choice of predictor, namely, stacking, BMA, the Shtarkov solution, and the Bayes Shtarkov solution. For the stacking predictor, the unified variance-bias analysis leads to a complexity given by the sum of (18), (19) and (20). So, the complexity of the prediction problem i.e, of the function approximation, may be taken as

\[
C(\text{stacking}) = \text{MSE}(\mathcal{M}_T, \mathbf{w}_T, \mathbf{B}_T; \mathcal{M}_n, \hat{\mathbf{w}}, \hat{\mathbf{B}})
= \text{MSE}(\mathcal{M}_T, \mathcal{M}_n) + \text{MSE}(\mathbf{w}_T, \hat{\mathbf{w}}) + \text{MSE}(\mathbf{B}_T, \hat{\mathbf{B}}).
\]

In the case of BMA, two extra terms are added from (26) so the complexity of the prediction task is

\[
C(\text{BMA}) = \text{MSE}(\mathcal{M}_T, \mathbf{w}_T, \mathbf{B}_T; \mathcal{M}_n, \hat{\mathbf{w}}, \hat{\mathbf{B}}) + \sum_{f \in \mathcal{M}} \text{trace}(\text{Cov}(E(\beta_f | \mathcal{D}, \Lambda_f)))
+ \text{Cov}(E(\mathbf{w} | \mathcal{D}, Q)),
\]

in which \(\mathbf{w}\) corresponds to the Bayes weights not the stacking weights. In the nonstochastic case, the complexity expression for the Shtarkov solution has the two biases from (29) and leads to regarding the complexity of the prediction problem as

\[
C(\text{Shtarkov}) = D(q_T || q_{\text{proj}}) + D(q_{\text{proj}} || q_{\text{opt}}).
\]

Likewise, in the Bayesian version, the complexity of the Shtarkov solution has an extra variance term. This gives

\[
C(\text{Bayes Shtarkov}) = D(q_T || q_{\text{proj}}) + D(q_{\text{proj}} || q_{\text{opt}}) + \text{trace}(\text{Cov}(E(\Theta | \mathcal{D}, \Lambda))),
\]

where the third term is of the form of the first term in (26) and \(\Lambda\) represents the hyperparameter introduced for the prior \(w\) on \(\theta\).

All four of these complexity expressions represent the difficulty of the prediction problem in terms of a sum of biases and variances associated to a predictor. However, this approach has the deficiency that a single prediction task now has two different complexities. Stacking and BMA can both be used to make predictions for the same problem, but treating the sum of their bias and variance terms as a complexity leads to \(C(\text{stacking})\) and \(C(\text{BMA})\) which are not equal. The expressions for Shtarkov and Bayes Shtarkov have the same problem even though they can be related to complexity in the description length sense. This limitation merely means that the complexity of the prediction task has been artificially backformed from the bias-variance decomposition rather than found independently.
Despite this, it can be helpful to look at the sum of biases and variances and treat it as a complexity even though its relation to the actual complexity of the prediction problem is not as tight as desired. For instance, if a predictor has inadequate performance, and the four expressions above are sufficiently accurate representations of the complexity of the prediction task, the terms can be examined to see which aspect of the modeling is most important to improve to get a better approximation of the DG. For instance, if the prior covariance terms in $C(BMA)$ are quite large compared to the MSE terms, one might be led to infer that the prior specification was inadequate. On the other hand, if $MSE(M_T, M_n)$ were very large it would reveal that model list uncertainty was the greatest problem. Examination of the bias and variance contributions in $MSE(M_T, M_n)$ would reveal whether it was the specific models on the list or the selection from the list that was a problem. In this light, looking at the complexity of the prediction task as a function of the complexity of the predictor formalizes the standard intuition that one should reduce the variances or biases that will give the most benefit for one’s effort.

Note that treating the sum of a predictor’s variances and squared biases as a complexity is just one possibility. It is easy to intuit alternatives but they are as yet undeveloped. For instance, in Section 4 an example is given in which the complexity of the predictor is defined by the extent of the search over models that it uses and the complexity of the problem is assessed by how large a class of functions is needed to approximate the DG well in terms of CPE. There it will be seen that Desideratum #4 is consistent with a sort of “Principle of Matching” between the complexity of the approximation and the complexity of the predictor. The best predictor may have a complexity level matched to the complexity level of the DG. Thus, simple predictors may typically do best for simple DG’s and complex predictors may do best for complex DG’s. The intuition is that when a predictor is more complex than the DG, the predictor searches through so many functions that its performance is worse than a simple predictor which does a smaller search. Likewise, when a predictor is simpler than the DG, the predictor must accommodate the complexity of the DG in order to predict well so that its performance is worse than a more complex predictor that tracks the DG more readily. An example of this is seen in Section 4.

Desideratum #5: Generate a Comprehensive Robustness Analysis.

Roughly, bias-variance analysis is a way to quantify how well a prediction strategy performs in a given setting. By contrast, robustness asks that a prediction strategy perform well even when the setting is changed. For instance, robustness asks that a predictor’s performance not be affected too much when the data is bad (outliers, missing data) or when the data generator is different from anything the predictive scheme in intended to model. Thus, varying the true model for which a predictor is found would be a matter of robustness, while comparing different predictors for the same data generator would typically be within a bias-variance analysis.

Both bias-variance analysis and robustness analysis can be used to compare predictors. An important operational difference is that ideally, both bias and variance should be as small as possible whereas robustness should be at the “right” level: Too much robustness
leads to posterior insensitivity and too little leads to posterior instability. It is not generally obvious how much robustness is optimal. Nevertheless, predictors that have good robustness properties have a greater claim to plausibility than those that don’t.

A comprehensive robustness analysis parallels the unified bias-variance analysis in Desiderata # 3 and used in # 4. The idea is that each aspect of the setting in which a predictor scheme is to be used must be varied to ensure the right degree of robustness and this must be done for each “output” of the predictor scheme. The outputs include the CPE, measures of fit, various features of the predictor itself, and possibly others. The analogs of bias and variance are played by distinguishing two sorts of robustness, local and global, depending on which aspect of the setting is being varied. First, local robustness concerns aspects of the setting that are objective; they are fixed and deterministic, usually arising from modeling assumptions believed to be true. The data are in this category, too. The natural way to evaluate robustness in these cases is through local perturbations: Choose a neighborhood around each of the nominal inputs and ensure that the consequent predictions do not vary overmuch. The local perturbations should capture the small deviations from the model that one would expect the experimental set up to have. At root, this is little different from using an $\epsilon$-$\delta$ argument to find the modulus of continuity for a choice of neighborhoods.

By contrast, global robustness concerns aspects of the setting which are subjective, not justified by any physical understanding of the DG. There is no good reason to choose one value for the inputs over another, apart from convenience. Thus, it makes sense to model these perturbations as stochastic, invoking the Principle of Insufficient Reason see Kass and Wasserman (1996). These perturbations are, therefore, not generally local. Compared to the ranges used in local perturbations for deterministic inputs, these are large scale perturbations. They range over possibilities that might reasonably give moderately different answers. The stochasticity is a proxy for the arbitrariness of the selection. In this case, one would be led to choose a hyperprior that was objective in some sense, possibly as close to uniform as possible, at least over the range of priors that were not unreasonable to entertain.

Since this is rather abstract, consider the following instances. First, an objective prior $W$ might arise from an optimality principle that represented modeling information believed to be true. As a consequence, $W$ would be the only reasonable prior to use, unless the modeling information were called into question. This implies that varying $W$ makes sense as a way to test the stability of the CPE against small perturbations in the efficacy of the assumed information. The perturbation of $W$ would be over a neighborhood, a test of local robustness. By contrast, if no objective information went into the choice of $W$ it would make sense to exhibit $W$ as say $W_{\lambda_0}$, a prior in a class defined by the hyperparameter $\lambda$ to which a distribution $\pi(\lambda)$ was assigned. This would be one way to test $W$ by mimicking its subjective selection as a random process. That is, it might make sense to regard $W = W_{\lambda_0}$ as an outcome of $\Lambda$ and so consider the marginal distribution from integrating over $\lambda$. The distribution assigned to $\Lambda$ would be as dispersed as reasonably possible and again the effect on CPE would be evaluated.

As a second example, consider the model space in a regression problem and suppose the regression function is expressed in a basis. The setting is defined by the notion that a true
model is most parsimoniously expressed in a certain function class. If it were believed that the function really was a waveform then a Fourier analysis would be appropriate for physical reasons. In this case, it would make sense to evaluate the effect on CPE of varying the Fourier basis locally. For instance, one could vary the Fourier basis by replacing \(\sin(nx)\) with \(\sin(nx + \delta(x))\) for some small function \(\delta\) or \(\sin^{1+\alpha}(nx)\). These would be quite difficult, so it would probably be enough for evaluating local robustness to vary the Fourier coefficients over a small neighborhood as a proxy for varying the Fourier basis. By contrast, if there were no extra modeling assumptions that were reasonable, a large scale perturbation or global robustness would be appropriate. For instance, re-estimating the regression function using Legendre polynomials or wavelet thresholding would assess the robustness of the predictions to the choice of basis. Naturally, the results of this large scale perturbation would be averaged, treating each choice of basis as equivalent to the others. That is, in effect, a uniform prior would be assigned for basis selection among a class of available bases.

As a third example, note that in some settings there are several ways to summarize a data set. Suppose there are two treatment groups and within each group there are two classes of subjects. Then, one could use a two dimensional summary statistic, perhaps a mean for each treatment group, or a four dimensional summary statistic including means for the two classes of subject. If it were believed that only the two treatment groups were important, then a local perturbation would be enough: The use of the mean could be varied locally. This could be interpreted as some level of trimming on each side, possibly using some power \(1 + \alpha\) of the data, or modifying the data by adding small perturbations to it. However, if the four dimensional summary statistic were plausible as well, it would make sense to consider a larger variety of summarizations and treat them stochastically.

Just as there is a desire that all the biases and variances be combined in a single notion of problem difficulty or complexity, so it is important to consider the relative contributions of the different components of robustness to an overall robustness. In other words, convert the global robustness problem over a range of possibilities to the local robustness of the prior over the range of possibilities. That is, regard the nearly uniform prior assigned to a range of possibilities in the stochastic robustness case as if it were a deterministic aspect of modeling. Then, the various aspects of modeling – now all regarded as deterministic – can be combined as follows into an overall sensitivity.

To see this, suppose there are two aspects of the setting in which a predictor is to be used, say \(a_1, a_2\) and one output \(o\), CPE for instance, with baseline inputs \(\tilde{a}_1, \tilde{a}_2\) and corresponding baseline output \(\tilde{o}\). Then an overall sensitivity can be evaluated as follows. Choose distances \(d_j\) on \(i_j, j = 1, 2\) and then set

\[d_i((a_1, a_2), (\tilde{a}_1, \tilde{a}_2)) = d_1(a_1, \tilde{a}_1) + d_2(a_2, \tilde{a}_2).\]  

(30)

Also, choose a distance on the output to be \(d_o(o, \tilde{o})\). Solving the optimization problem

\[\max d_o(o, \tilde{o}) \text{ subject to } d_i((a_1, a_2), (\tilde{a}_1, \tilde{a}_2)) < \epsilon\]

(31)

for \(\epsilon \to 0^+\) can be done through second order matrix approximations in many cases. If there are three or more aspects, then there are three or more terms in (30), respectively.
Clarke and Gustafson (1998) use the relative entropy to evaluate the overall sensitivity of the posterior to three aspects, namely the prior, likelihood, and data. The result is a vector corresponding to the maximal eigenvalue indicating the direction among the collection of inputs along which the deviation of \( \tilde{o} \) from the baseline \( \tilde{o} \) is fastest. The consequence of this is that the influence of each input to the output can be evaluated relative to each other, given the choice of \( d_i \)'s.

Desideratum #6: Determine the limiting properties of the predictors, any estimates they use, and any functions of them that are of interest and verify they do not contradict the context of the problem; the limiting properties may be for any output of the predictive scheme over any input to the predictive scheme.

The content of Axiom 6 is a logical consistency requirement: You don’t want to be able to derive something under reasonable conditions that is different from what would be observed.

3 The Desiderata and Other Theories

In this section, we examine the interplay between Coordinating Theory, Bayes theory and Frequentist theory. The three approaches have some common features, but Coordinating Theory is closer to Bayes. It tries to incorporate some Frequentist (and coding) ideas to effect a serviceable compromise between Bayes and Frequentist.

Before going into these details, note that the desiderata are mostly a formalized way to evaluate a predictor class. That is, one uses the desiderata to determine systematically how good a predictor class is for a given DG. To apply the desiderata, the predictor must be exhibited as an output of its inputs and then each desideratum considered in turn. For instance, if the predictor for stochastic data is the predictive density \( m(x_{n+1}|x^n) \), the inputs are the prior, the likelihood, and the data in which the prior is subjective and the likelihood is given by modeling assumptions. Then, looking at predictive performance and updating (here trivially since the inputs are fixed) satisfies #1 and #2. None of the inputs are estimable, however, a variance can be assigned to the prior as in #3 and a bias can be assigned to the likelihood, possibly as in the non-stochastic case. (If the likelihood were not based on modeling assumptions, then it would be assigned a variance possibly using the conditional covariance approach of situating it in a larger family of likelihoods.) The sum of these terms measures the complexity of the prediction problem as in #4. A local perturbation of the likelihood and data would be combined with a stochastic perturbation of the prior with respect to a near uniform hyperprior and the three would be combined to give a comprehensive robustness analysis of CPE or of the predictor itself for #5. Examining the asymptotics would satisfy #6. Notice that the parameter does not appear explicitly on this list. However, if the predictor were formed using an estimate of \( \hat{\theta} \), say \( \hat{p}_\theta \), then the variance and bias of \( \hat{\theta} \) would be included under desideratum #2.
3.1 Coordinating Theory and Bayes

The Bayesian approach can be regarded as one rational way to construct a predictor. The construction rests on a collection of axioms such as those in Bernardo and Smith (2000) leading to the use of the posterior distribution or on Savages postulates characterizing choice as maximizing expected posterior utility, see Savage (1954). The independence of the prior from the data is optimal under a coherence criterion as in Freedman and Purves (1969). An important sense in which Bayes methods are best is provided by the Complete Class Theorem which states (roughly) that the procedures which are Bayes for a given problem form a complete class, i.e., a class that contains all admissible procedures, see Brown (1981). Take together, this provides a unified framework for data analysis, hypothesis testing, and prediction.

There is nothing wrong with the Bayesian framework. However, it makes two simplifying assumptions that are not reasonable in general. First, it is static. That is, once the problem has been specified in terms of loss function, prior, likelihood and so forth, the problem is not updated. It is straightforward, in principle, to use the Bayes approach to get solutions but it is not straightforward to reformulate the Bayes problem e.g., to rechoose the inputs such as a model list in response to lack of fit or poor prediction.

Second, related to the first point, Bayes procedures do not have an in built way to detect bias. That is, Bayes inference is limited to the support of the prior and there is no way to decide (within the Bayes framework) that the prior was incorrectly chosen and fails to put mass on a neighborhood of the true model.

Despite these limitations, Bayes prediction is asymptotically as good as optimal prediction in many cases. Indeed, Dawid and Skouras (1998, 1999) establish Bayesian prediction systems are efficient asymptotically and and Clarke (2004) found that optimizing a conditional risk, chosen by using the accumulated data, to predict a future outcome outperformed standard Bayes techniques in small sample sizes, asymptoting to the performance of Bayesian predictors. In the presence of bias, Clarke (2003) has also shown that prediction based on stacking can predictively outperform Bayes model averaging. Roughly, to outperform Bayes predictively on stochastic data it is enough to rechoose the decision problem from time step to time step so that a stronger optimization is done, essentially tracking the data closer than a single Bayes predictor would.

Essentially, Coordinating Theory enlarges the Bayes predictive problem with sequential thinking by replacing the notion of credible set with predictive accuracy and invoking Desiderata #2 and #3 to structure it. That is, periodic re-selection of the model list or other elements and going outside the Bayes framework to include bias (and variances) permits the Bayesian to be a better Bayesian by finding the right decision problem to solve and evaluating how well the solution performs. Note that the other desiderata do not particularly conflict with established Bayesian thinking although they may go beyond common practice. For instance, Desideratum #5 is well recognized when applied to the prior but not so well recognized when applied to the data or likelihood.

To see how the desiderata can be applied to BMA, for instance, list the inputs to the predictive procedure. Suppose sequential outcomes $Y_i$ from $Y = F(X) + \epsilon$ with $F$ unknown
but an IID error structure are available where \( \mathbf{X} \) is a collection of \( p \) explanatory variables.

Then, for time step \( n \), we have a collection \( S_n \) of models. The models in \( S_n \) may have parameters and be written \( F_j(\mathbf{X}|\theta_j) \). To form the average we use a prior \( w_n(j) \) on the models in \( S_n \), which is updated to form \( w(j|D_n) \), where \( D = \{ (\mathbf{X}_i, Y_i) : i \leq n \} \) and equip parameter \( \theta_i \) with prior \( w(\theta_i) \). So, there are six inputs: (1) the \( \theta_j \)'s within the models \( F_j \), (2) the priors on the \( \theta_j \)'s within each \( F_j \), (3) the model lists for each \( n \), (4) the error term, (5) the prior on the models, and (6) the space \( S_n \) is drawn from. Note that including the individual models \( F_j \) on this list would be redundant from a bias-variance standpoint because varying the prior (5) on the models on the list effectively varies the models conditional on the list while varying the model list (6) necessitates varying the \( F_j \)'s. From the sensitivity standpoint, local variation of the models would be reasonable only if the models on the list were regarded as true in some sense, as in the Fourier basis earlier and then there would be no point in varying the model list stochastically for bias-variance.

Now, we go through the desiderata in turn. Desideratum #1 is satisfied if the MSPE satisfies the prequential principle and this makes sense since the BMA is optimal under squared error loss. Desideratum #2 is satisfied if the model lists \( S_n \) are reselected as a function of the predictuals. For instance, one may start with linear functions of the variables in \( \mathbf{X} \), find that there are patterns in the residuals, and be led to expand the collection of regression functions to include square and rectangular terms in the \( \mathbf{X}_i \)’s. Doing this may require changing the prior on \( S_k \) and assigning priors to the parameters in the extra models included in \( S_{k+1} \). More general cases are readily imagined.

Desideratum #3 requires variances and biases. The bias-variance decomposition has already been given in Section 2 and is the sum of (18), (19), (20), and (26), the first three suitably modified. Note that of the six overall inputs, the only two left out are the error term and the model space.

Desideratum #4 requires looking at the sum of the terms in the bias-variance decomposition to identify which terms are largest. This would necessitate calculating all the terms; this can be done, but is the subject of a future paper. However, some heuristics can be easily given. First, if the true model is linear and is on the list \( S_n \) and \( \dim(\mathbf{x}) \) is not large, then the complexity of the prediction task is probably small because the true model is easily approximated, even identified. This is the \( M \)-closed case of Bernardo and Smith (1994). Accordingly, in this case, the biases and variances should all be small. Second, in the \( M \)-complete case the complexity is probably in some mid-range under any reasonable calibration because the true model is not in any \( S_n \) on account of it being so difficult to approximate. So, the biases should be small at the cost of higher variances or the variances should be small at the cost of higher biases. In the \( M \)-open case, the situation is more difficult because the true model is not in \( S_n \) and is not really approximable by the available models. This last would be a higher bias and higher variance situation. It is in these contexts that a Principle of Matching may be reasonable: For the \( M \)-closed case low complexity predictors would likely be best; for the \( M \)-complete case medium complexity predictors would likely be best, and, for the \( M \)-open case, high complexity predictors would likely be best.

Desideratum #5 would apply to the data, the error term and the model space itself,
the latter being summarized by a prior over different representations of a model space in terms of basis elements, for instance. In some cases, the formulation in (30) and (31) is amenable to standard quadratic maximization problems, as used in Clarke and Gustafson (1998). Essentially, a quadratic approximation can be given for each term locally so that maximizing a quadratic objective function subject to quadratic constraints becomes feasible. Note that the quantities subject to the sensitivity analysis are not subject to a bias-variance analysis and conversely, but each input to the predictor is examined under Desideratum #3 or #5. There is no prohibition on looking at the sensitivity of CPE, say, to the prior over models. However, it should be enough to look at its variance. If the variance is too small, then there may be too much sensitivity and if the variance is too large there may be too little sensitivity. Thus, a proper bias-variance analysis should reflect the sensitivity of outputs like CPE to their inputs.

Finally, Desideratum #6 means that once we have formalized our procedure, we want to be sure that its limiting properties are reasonable. For BMA, we want to be sure that if there is a true model in $S$ that in the limit $S_n$ converges to a list that includes the true model and that the weight in the BMA on the true model converges to one. If the true model is not in $S$, then we want to know that the BMA converges to the model in $S$ closest to the true model. Moreover, we expect that the estimates of the parameters in the true model will converge with the usual asymptotic normality and we expect the MPSE to have a limiting form with rate $nI(\theta_T)$, in regular cases.

### 3.2 Coordinating Theory and Frequentism

The Frequentist approach can be regarded as another way to construct a predictor. Unlike Bayes, Frequentism does not have a comprehensive foundational construction. The analog to Savage’s postulates is roughly provided by the von Neuman-Morganstern expected utility theorem for representing choices as expectation maximization and Frequentist decision theory is correspondingly well-developed. However, there seems to be little justification for the use of a sampling distribution for estimation apart from the repeated sampling interpretation from probability theory.

Coordinating Theory replaces the concept of confidence with predictive accuracy and only uses Frequentist concepts based on the sampling distribution for aggregate characterizations as in Desideratum # 3, with variances taken over the sample space or in Desideratum #6 where asymptotics are used as a sort of sanity check. One could sidestep Frequentism further by using posterior variances in Desiderata #3 and # 4. However, the Frequentist variance is a broader summary of variability since it involves integration over the sample space not just dependence on a string of data. Moreover, Coordinating Theory allows Frequentist thinking between time steps in the respecification of the decision problem leading to a predictor, in Desideratum #2. The reselection of a decision problem is non-Bayesian because it rests on correcting bias, a problem that is more readily examined within the Frequentist paradigm than in the Bayesian.

For contrast with BMA, consider the stacking model average as in (15) and (16). Instead
of the six inputs, there are now only five since the two sorts of priors are not used but the model weights are still required. The five inputs are (1) the $\theta_j$’s within the models $F_j$, (2) the model weights, (3) the model lists for each $n$, (4) the error term, and (5) the space $S_n$ is drawn from.

As in the BMA case, we go through the desiderata for the stacking model average predictor. Desideratum #1 is satisfied if the MSPE satisfies the prequential principle. Desideratum #2 is satisfied if the model lists $S_n$ are reselected as a function of the predcuiltals. The bias-variance decomposition required for Desideratum #3 has already been given in Section 2 and is the sum of (18), (19), and (20). This gives a version of the complexity for Desideratum #4.

Desideratum #5 asks for a robustness analysis. This would proceed as in the BMA case since there are three inputs that have not been assigned biases or variances, the model space and the error term, and the data that must be assessed. Finally, Desideratum #6 would be examined much as in the BMA case but from a Frequentist standpoint.

4 Aspects of Desiderata 1, 2, and 4

In this section, an instance of the kind of analysis suggested by the desiderata is presented. CPE error is calculated, and #1 and #2 are satisfied. Although neither a bias-variance analysis nor a sensitivity analysis is done, some heuristics on complexity can be given because nine different predictors based on three different function classes and three different model averaging strategies are used.

To begin, consider three model spaces: linear models (LMs), generalized additive models (GAMs), and recursive partitioning models (trees). Also, consider three model average prediction schemes: BMA, here called likelihood weighted averaging (LWA) because of the model list re-selection, stacking, and a data dependent convex combination of these here called adaptive convex average of predictors (ACAP). Because of the relative richness of the function classes they represent, it is reasonable to regard LM, GAM and trees as low, medium, and high complexity respectively. Because of the scope of the search over function spaces and model weights that they involve, it is also reasonable to regard LWA, stacking and ACAP as low, medium, and high complexity, respectively.

To specify the predictors, form the LWA for LMs as

$$\hat{Y}_{LWA,i+1} = \sum_{j \in M_{i+1}} W(j|Data) X_j \hat{\beta}_j$$

and the stacking average for LMs as

$$\hat{Y}_{CV,i+1} = \sum_{j \in M_{i+1}} \hat{w}_{i,j} X_j \hat{\beta}_j,$$

where $Data$ includes the first $i$ data points, $X_j$ means the values at time $i + 1$ of the $j$th model, and the $\hat{w}_{i,j}$’s are the stacking weights (at time $i$) parallel to the Bayes weights in
LWA (based on a uniform prior). In (33), 5-fold CV was used. The difference between (32) and (33) is seen in the weights as well as the parameter estimates because (32) used ordinary least squares estimators $\hat{\beta}_j$ and (33) used posterior means $\tilde{\beta}_j$. Now, in principle, it will be possible to compute the MSPE along a string of data so that Desideratum #1 is satisfied.

Note that $\hat{Y}_{LWA,i+1}$ and $\hat{Y}_{CV,i+1}$ may be formed from different model lists denoted $\mathcal{M}_{LWA,i+1}$ and $\mathcal{M}_{CV,i+1}$. We expect these two model lists to be different because stacking and LWA have different properties. Stacking generally has a higher variability than LWA and has slightly better performance under bias. So, when the true model is far from the models on the list stacking often does better than LWA but when the true model is on the model list LWA usually does better than stacking. Thus, the degree of model mispecification largely determines which method will do better predictively.

For comparison, consider an adaptive combination of an average of the LWA and stacking predictors, ACAPs. That is, for $\hat{Y}_{LWA,i+1}$, $\hat{Y}_{CV,i+1}$, and model lists for each, take another mixture. Let $\alpha_i \in [0, 1]$ and define a convex combination of $\hat{Y}_{LWA,i+1}$ and $\hat{Y}_{CV,i+1}$ from (32) and (33) to be

$$
\hat{Y}_{i+1, ACAP} = \alpha_i \hat{Y}_{CV,i+1} + (1 - \alpha_i) \hat{Y}_{LWA,i+1},
$$

(34)

where $\hat{\alpha}_i$ achieves

$$
\min_{\alpha} \sum_{u=i-20}^{i} \left( Y_u - [\alpha \hat{Y}_{CV,u} + (1 - \alpha) \hat{Y}_{LWA,u}] \right)^2,
$$

(35)

for each $i$. Note that (35) only uses most recent 20 data points so $\alpha_i$ won’t converge too fast relative to $\mathcal{M}_k$. When $\alpha_i < 0$, we set it to be 0 and when $\alpha_i > 1$, we set it to be 1.

An analogous procedure was used for GAMs and trees. For these two cases, however, a selection of variables is treated as a model. Given a selection of variables, a GAM can be formed using them and given several GAMs of this form a stacking average found. For the LWA case, several GAMs are formed using different selections of variables and the Bayes weights for them are found using the deviance criterion in the contributed R package `gamlss`. Likewise for trees, a stacking average is found using several selections of variables and the LWA is found using the deviance criterion in the contributed R package `tree`. In both cases, the ACAP average is found using (34).

It remains to specify how the model lists update in response to CPE so that Desideratum #2 will be satisfied. The method is described fully in Clarke and Clarke (2008). For LMs, the basic idea is as follows. Use an ensemble of terms which consists of all terms of second order or less, i.e., all $x_j$’s and all $x_jx_k$’s. A random selection of these terms is used for the initial model lists $\mathcal{M}_{LWA,0}$ and $\mathcal{M}_{CV,0}$ of fixed cardinality $K = 4$. The next step is to define a central model within $\mathcal{M}_{LWA,0}$ and $\mathcal{M}_{CV,0}$. The central model for $\mathcal{M}_{LWA,0}$ contains the terms that are in a majority of the models in $\mathcal{M}_{LWA,0}$ and the central model for $\mathcal{M}_{CV,0}$ contains the terms that are in a majority of the models in $\mathcal{M}_{CV,0}$. For both averaging procedures, new models formed by adding or deleting one term from the central model in all possible ways are considered, and the best of these is compared to the models already on the list.
The new model replaces one of the old models if it gives better CV performance. For GAM’s this procedure is modified by building models using cubic splines based on the terms in the ensemble. For trees this is modified by looking only at the terms in the selection of variables fed into the tree software, not at the terms in the actual model output by the software.

To compare the 9 methods (3 predictors, 3 model spaces) we used the benchmark data set Comp-Activ containing records of computer performance measures used to predict the fraction of time that CPUs run in user mode. This data set is considered reasonably complex. Here, we chose only 15 of the original 24 predictors. The CPE’s of the three methods for each model space are in Figs. 1, 2, and 3.

Figure 1: CPEs for LMs $n = 300$ for the Comp-Active Data. The results from fifteen random permutations of the data were averaged.

Figure 2: CPEs for GAMs, $n = 300$, for the Comp-Active Data. The results from fifteen random permutations of the data were averaged.

Note that the scales on the vertical axis are different for the 3 graphs. The first is the largest ($e^{0.5}$ means 100,000), the third is the smallest. All 3 of the smallest final CPEs occur for tree models and among these the ACAP error was less than the LWA error which was less
than the stacking error. For LM’s and GAM’s, the stacking error was less than the ACAP error which was less than the LWA error. So, the worst choices overall were ACAP’s and LWA for LM’s.

We interpret this to be broadly consistent with a sort of Principle of Matching: The complexity of the data i.e., the approximation task represented by the DG, should match the complexity of the predictor for best performance. In this case, the data set is quite complex and the complexity of the model space is the most important feature of the predictor. So, the best cases occurred for the most complex class, namely trees. The second most complex class was GAMs and it did second best while LMs, the simplest class, gave the worst performance.

Within the most complex model space, trees, the most complex method, ACAPs does best, LWA did less well and stacking did worst. The Principle of Matching would have predicted that stacking would do better than LWA because stacking is more flexible than LWA since its weights are not based on a likelihood. However this was not observed. It may be that the richness of the model space permits the higher efficiency of Bayesian methods to outperform the less efficient stacking average.

The worst cases were with LM’s, the simplest models. Among these, stacking did best. One would have expected from the Principle of Matching that the most complex predictor ACAP should have done best. However, it is possible that LWA, which converges quickly, has such a high bias, making its predictive error so bad, that it forces the ACAP to do poorly too. Further work on this example would assign a variance to the predictuals at each time step over runs for each method as a way to assess the efficiency of the nine methods.

5 Implications of the Theory

The point of this paper is to argue the merit and feasibility of expanding the statistical problem so it is phrased in terms of prediction. The 6 proposed desiderata are merely a way to structure thinking about such a general problem. They are based on the behavior of pre-
dictors rather than estimators, tests, or decision rules. Indeed, it is clear that given a model, every estimator generates a predictor. The reverse is probably not true in general. However, there may be general conditions under which a predictor generates a unique single estimator thereby giving a partial converse. Likewise, every estimator generates a test statistic but it is not true that every test statistic corresponds to an estimator. Like prediction, testing is a more general problem than estimation. However, in the presence of a pivotal quantity testing and estimation are equivalent and there may be conditions under which a three way equivalence for tests, predictors, and estimators will hold.

For the present, it is worth stating explicitly that adopting a predictive standpoint implies several immediate conclusions at odds with much common practice.

First, physical interpretation of models is downweighted in favor of actual predictive performance. Recall that, in practice, many statisticians use physical interpretations to choose among models or predictors. This is often done by preferring a model with a good physical interpretation over one with a poor physical interpretation – even when the latter performs better predictively. That is, physical interpretability is taken as a proxy for validation. Strict application of the Prequential principle will often rule out this use of physical interpretations of models on the grounds that they rest on suppositions about the true model and the other desiderata are intended to confirm whether a predictor is performing well more generally. Thus, the proper place for physical interpretations to enter a predictive formulation is in the collection of predictors entertained. Naturally, however, one wants as large a collection as can be managed with the data so that if the physical interpretation is not effective some element of the predictor class will still give good predictions.

Second, for the Bayesian, there is no prohibition on data dependent priors. Wasserman (2000) uses data dependent priors, and shows they are optimal, in a mixture setting where model uncertainty is a problem. Clarke and Yuan (2004) give general information theoretic interpretations and the sequential procedure of Wong and Clarke (2004) can be interpreted as using a prior that depends on the data to redefine its support at each time step. The origin of data independence is coherence, see Freedman and Purves (1967), but coherence is a separate criterion not directly related to the rest of the Bayes structure.

Third, the posterior variances commonly reported are conditional on the model, and much else, being correct. As soon as model uncertainty is included via updating predictors for instance, the predictive analog of confidence or credibility bands becomes much larger, see Draper (1995). It may well be that once proper modeling uncertainty is taken into account the confidence or credibility bounds become so large as to reveal that the inferences based on the data are ineffective. That is, what we commonly take as inferences are better viewed as suggestive because the actual confidence or credibility is smaller than we would have thought. This is another way to say that too many of our inferences have been based on overfitting.

Fourth, the stability or robustness of models is as important as any claim to veracity. Getting a prediction scheme whose predictive error doesn’t change overmuch under reasonable perturbations of its inputs is hard enough; insisting that the real world conform to models we can write down conveniently may be unrealistic. Indeed, the best predictors do
not usually correspond to any obvious model for the DG. Consequently, theoretical properties like consistency of estimates are desirable, however, their relevance is limited when anything like a true model is pragmatically unknowable. It may be important to back off from model identification in order to derive modeling schemes that give predictions with assessable reliability.

Fifth, and arguably most important, validation of models in the sense of doing extensive checks that the predictions they give are accurate is the most central property of any scheme. Sequential prediction is merely one natural way to do this. Pragmatically, trying to find a model that can be taken as true by surmising validity from weak validation criteria such as interpretability will often lead to overstatement of the strength of the information in complex data sets. Otherwise put, all too often the result will be models that fit adequately but fail to generalize. It should be remembered that interpretable models are a rather small class and may easily omit important variables that are not known or include variables that are only weakly relevant. In either event, they often give predictions that are not as good as desired. If a “true model” must be announced, the best way may not be to seek a model directly but instead to develop a good predictor. Then, the predictor can be used to identify a model by converting candidate true models into predictors and finding one that is not too far from the established good predictor.

6 References


George, E. (2001) see:


