Combining Model Selection Procedures for Online Prediction

B. Clarke
[1]

Abstract

Here we give a technique for online prediction that uses different model selection principles (MSP's) at different times. The central idea is that each MSP in a class is associated with a collection of models for which it is best suited so that the the data can be used to choose an MSP. Then, the MSP chosen is used with the data to choose a model, and the parameters of the model are estimated so that predictions can be made.

Depending on the degree of discrepancy between the predicted values and the actual outcomes one may update the parameters within a model, reuse the MSP to rechoose the model and estimate its parameters, or start all over again rechoosing the MSP.

Our main formal result is a theorem which gives conditions under which our technique performs better than always using the same MSP. We also discuss circumstances under which dropping data points may lead to better predictions.

[1]Department of Statistics, University of British Columbia, Vancouver, B.C. Canada V6T 1Z2.
AMS 1991 Subject Classifications: Primary 62M20, secondary 62L99.
Keywords: prequential statistics, model selection, model mis-specification, model meta-selection.

1. INTRODUCTION

Although there is a vast literature on how various model selection procedures, MSP's, perform there is very little guidance about how to choose an MSP in the first place. Many people seem to advocate a specific MSP for general usage, at least in certain contexts, often for very good reasons. However, other people, with equally good reasons, may advocate a different MSP. They can't all be right. We argue that the seeming discrepancy can be cleared up by recognizing that one MSP may be better for a given class of models than another MSP is. Also, we suggest that one MSP may be better for one purpose, say prediction, than another MSP is. This means, in particular, that the different physical meanings and statistical interpretations associated to different MSP's cannot be ignored. A consequence of this point of view is that if you are unclear about which MSP to use, which class of models to search, or you are not sure about what the ultimate use of a chosen model will be, then you should keep your options open. This means that you want to search over various MSP's, and the classes of models associated to them (on which they may be presumed to perform well), and to evaluate performance by a criterion which will be generally good, independently of the purpose of the modeling.

Here, we give a technique, with justification, for how to choose an MSP from a class of MSP's for use in predicting the next outcome in a sequence. Our technique permits the use of different MSP's at different time steps depending on how well earlier outcomes were predicted. The main strength of our technique is that it makes very weak assumptions on the data generating mechanism, satisfies a weak form of the prequential principle, and performs no worse than using the 'best' MSP would, in an asymptotic, squared error sense.

To make this concrete consider two of the most popular MSP's: the Akaike Information Criterion, AIC, and the Bayesian Information Criterion, BIC. The AIC, Akaike (1977), chooses the member of a given class of parametric families having the largest value of

$$AIC = \log p(x^n|\hat{\theta}(x^n)) - d, \tag{1.1}$$

where $x^n = (x_1, ...x_n)$ is distributed according to a parametric family of the form $p_\theta(\cdot) = p(\cdot|\theta)$ and $\hat{\theta} = \hat{\theta}(x^n)$ is the maximum likelihood estimate (MLE) of the d dimensional real parameter $\theta = (\theta_1, ..., \theta_d)$. By contrast, the BIC chooses the member of a given class of parametric families having the largest value of

$$BIC = \log p(x^n|\hat{\theta}) - \frac{d}{2}\log n. \tag{1.2}$$

The difference between (1.1) and (1.2) is in the size of the penalty term. The BIC penalizes models with more parameters more than the AIC does. Thus, generally, the AIC will give models with more parameters.

What do these MSP's mean? Akaike (1977) said his criterion was motivated by entropy considerations. Nevertheless, the AIC is essentially equivalent to Mallows' C_p familiar from regression, see Shibata (1981), as well as to cross-validation and generalized-cross-validation, see Li (1987). It is worth recalling that the AIC is inconsistent for model selection, see Woodroofe (1982) and Hannan (1980), but that Shibata (1981) established a sense in which the AIC seems to be asymptotically optimal for choosing the number of terms to include in a linear model when the dimension of the model is permitted to increase. See Hannan and Quinn (1979) for a dependent case with a different modification. Moreover, Haughton (1988) seems to agree with Geisser and Eddy (1979) that the inconsistency may not affect the use of the AIC for prediction purposes. Indeed, there is some evidence that AIC is optimal in certain predictive contexts.

The BIC arises from seeking the mode of a posterior density. Suppose we have a prior Π on a discrete class of models indexed by i. If each model is equipped with a prior density w for its parameter then one can form the posterior density $\Pi(i|x^n)$. The mode of this density is a natural choice for a model. However, it is easier, and asymptotically equivalent, to maximize

$$\log m(x^n) - \frac{d}{2} \log n \tag{1.3}$$

where $m(x^n) = \int w(\theta) p(x^n|\theta) d\theta$, and d is the dimension of θ. In turn, (1.3) leads to (1.2) by a Laplace expansion argument, which also reveals the penalty term $(d/2)\log n$, see Haughton (1988). Using a Bayes factor argument, Schwarz (1978) establishes the optimality of the BIC for exponential families.

Thus, there is a sort of predictive optimality which one might heuristically associate to the AIC and a sort of hypothesis testing optimality one might heuristically associate to the BIC. Furthermore, one might expect that the AIC will perform better than the BIC when the true model has many parameters and that the BIC will perform better than the AIC when the true model has few parameters. The consequence is that the AIC and the BIC have different classes of parametric families associated to them on which they may be expected to perform well. Nevertheless, the question remains: If we know little about the class in which the true model lies should we use AIC or BIC? Furthermore, what if we are unable even to determine whether the intuition behind the AIC or the BIC is relevant to the problem under investigation?

For instance, one might believe that some kind of coding optimality is relevant, especially if one's goal is to estimate a density and wants high sensitivity to tail behavior. In this minimum description length (MDL) context, Barron and Cover (1990), and Rissanen (1996) the authors minimize a data driven analogue of coding redundancy to choose a model. In this formulation one must explicitly specify the class of functions over which the optimization will be done. The size of the penalty term and the asymptotic form of the risk is determined largely by the class. The MDL approach, and its variants, goes back to Barron (1985), and Rissanen (1978). See also Wallace and Freeman (1987). In the fully parametric setting it has the same $(d/2)\log n$ penalty term as the BIC, as well as analogous asymptotic properties, see Barron and Cover (1990). One can argue that, when they are both defined, the MDL refines and extends the BIC by providing an interpretation for the prior and for the objective function in terms of code length. The MDL may perform slightly better than the BIC in some coding contexts because it uses an optimal constant term to track the coding redundancy better. However, the coding argument justifying the MDL is at present unrelated to the optimality of the BIC due to Schwarz (1978) and the entropy motivation of the AIC leads to a different penalty term from the BIC.

Given that these approaches (AIC, BIC, MDL,...) are only some of the possible approaches to model selection that one might take, some authors have sought to contrast MSP's by establishing general properties of various collections of MSP's. Often these have been based on the nature of the penalty term.

The AIC and BIC are members of a general class of MSP's studied by Bethel and Shumway (1988) who established consistency for a large class of penalty terms. Consider

$$\log p(x^n|\hat{\theta}) - df_m(n), \tag{1.4}$$

where $f_m(n)$ is a function of the sample size n, for each model class m we entertain. We assume that MSP's with sufficiently different f_m's are optimal in sufficiently different senses that the parametric families they are unlikely to choose the same parametric family, when they are allowed to, at least for small and moderate sample sizes. For instance, when f_m is of the form $o(n)$ and unbounded the results in Bethel and Shumway (1988) give consistency.

More recently, Yang and Barron (1998) provided general results for MSP's of the form

$$-\sum \log p(x_i|\hat{\theta}^{(k)}) + \lambda_k d_k + \nu C_k. \tag{1.5}$$

The first term in (1.5) is minus the maximized log-likelihood. The middle term is the product of d_k the dimension of the parameter in the k^{th} model and a constant λ_k which is interpretable in some cases as a dimensionality constant of the model related to the metric entropy. The third term is a description complexity penalty, and corresponds to a Bayesian prior. One of the main results in Yang and Barron (1998) gives conditions under which the expected squared Hellinger distance is bounded by a real factor times an index of resolvability. This index is similar to the minimization of the expected value of (1.5) over a class of parametric families. Yang and Barron (1998) also note that the middle term in (1.5) can be related to the bias correction interpretation of the AIC, and to the BIC, the difference being the size of the penalty on the number of parameters.

There are many other MSP's that have been examined. A partial list includes: informational complexity, see Bozdogan et al. (1997), informational minimaxity, see Barron and Xie (1996), minimally informative likelihoods, see Yuan and Clarke (1999). However, our point here is not to investigate the optimality of a specific MSP, nor to study classes of MSP. Our point is to present a general method which can combine these techniques in a prequential context. Our method presumes knowledge of the approximate optimality properties of an MSP and assumes that MSP's have already been grouped into class which are represented by a canonical member. Our method does not assume substantial knowledge of the data generating mechanism.

Essentially, we resolve the question of which MSP to use by letting the data decide – or more precisely, introduce a statistic to decide which of a set of canonical representatives for classes should be used. The key problem we address here is the optimality of this approach so as to establish that it is better to choose an MSP in a data driven way, and given such a choice decide how long, in sequential settings, to continue using it, bailing out when its predictions are too far wrong. We implicitly answer the question of what classes of models to use by using classes defined from the MSP.

In brief, we assume we have a sequence of outcomes and given the first n of them we try to predict the $n+1$ outcome. The prediction technique we develop here associates a class of models to each of a collection of canonical MSP's and then and uses a statistic to choose one of the MSP's. Then we use the MSP chosen to choose a model, estimating the parameters in it and using that model to predict the next outcome. Upon receipt of the next outcome, one may update the parameter estimates (if the prediction was good), or reuse the MSP to choose a new model (if the prediction was not good) or rechoose the MSP, thereby repeating the whole procedure (if the sequence of predictions has been bad enough for long enough). The adequacy of prediction is measured by recent error (the most recent squared difference between the predicted value and the actual value) and by a cumulative error (the sum of squared differences between predicted values and actual values from the first use of the current MSP to the present). A main part of the specification of the procedure will be identifying thresholds for rechoosing the model and rechoosing the MSP. Clearly, this approach can be refined by physical modeling. Here, however, we have relied on interpretations of the MSP's.

A heuristic version of this technique has been computationally implemented in de Luna and Skouras (1999). Crediting Dawid (1992 p. 117) for the technique, de Luna and Skouras (1999) uses the relative cumulative predictive loss to choose between the AIC and BIC and establishes its consistency. The three computed examples they develop, and the simulation study they perform, suggest the method is better than using either the AIC or BIC alone. In fact, the technique used in de Luna and Skouras (1999) was first described in Clarke (1997), and here we build on the extensive computational work of de Luna and Skouras (1999) to clarify the sense in which combining MSP's does better in general.

The setting of online prediction is essential for our approach because accurate prediction is the main way that the adequacy of a model must be reflected – regardless of the goals of an analysis. Indeed, good prediction is a defining feature of empirical science. Moreover, good prediction is a test of any subsidiary aim: If the goal of an analysis is to estimate a parameter then any good estimate of a parameter should give good predictions. If the goal of the analysis is model identification then the best model should give the best predictions. If the goal of the analysis is hypothesis testing, then any rejected model should give worse predictions than any accepted model. Thus the criterion of good prediction is central to the statistical enterprise.

Here, the prediction technique we develop is in the spirit of the predictive sequential – 'prequential' – approach of A. P. Dawid and co-authors. This alternative approach to prediction abandons the goal of selecting the true model and seeks only as small a predictive error as possible. In practice, this often leads to consistency, indicating how strong the criterion of good prediction is. The prequential approach has been developed in a series of papers by A. P. Dawid and co-authors, see for instance Dawid (1992, 1984), Seillier-Moiseiwitsch and Dawid (1993) amongst many others. Most recently Skouras and Dawid (1998) study the efficiency of point prediction systems. A key point made by Dawid in these works is that the performance of a method must be assessed independently of the method to avoid conflict of interest: One can anticipate that a method will be biased in its own favor thereby making it harder to find other potentially better methods.

Making the prediction online simulates the scientific approach of refining models so they become ever more accurate. One tries one MSP and then gets to try another if the first one doesn't work well. However, one must distinguish between the adequacy of the MSP and the adequacy of a model it chooses. If the MSP is good but the model chosen does poorly then again one still must refine the choice in the light of more data. Permitting occasional jumps from MSP to MSP may speed this process by permitting the use of a new MSP and a new model at the same timestep.

The main benefit of our method and the prequential setting is its generality. It is intended for problems where we have little pre-experimental information, but can rely on getting ever more data. We do not restrict the models or MSP's available for our use: All we must do is specify them. Often this is reasonable: In some time series contexts, for instance, we don't know what assumptions are valid and when they are valid, outside of particular special cases. In addition to the generality, the method here reduces to special cases as expected. We comment that one can imagine using different statistics to choose an MSP and thereby wanting to develop an MSP-selection principle. Such hierarchies probably provide diminishing returns.

A specific benefit from the generality is that the assumptions are so weak that Bayesian and frequentist methods can be combined. Indeed, as a parallel to wave-particle duality in physics we can go back and forth between them as the data indicate. That is, the prequential approach permits one to use the Bayesian model for stochastic variation at some times and the frequentist model for stochastic variations at others. Here, we present our methods in a Bayesian context but this is not essential. In particular, in a linear models context, one can compare the predictive performance of random effects models (an example of a hierarchical Bayes model) with a class of fixed effects models (based on the frequentist paradigm). The hypothesis test of Dawid (1986) to decide whether to use a random effects model or a fixed effects model would then be a suitable way to choose an MSP. In this case, we treat the Bayes and frequentist models for statistical variation empirically, based on how they perform in the real world. One can anticipate that in many cases the two approaches will give equivalent predictions although conceptually distinct. (The classes of models associated to the Bayes and frequentist approaches would be interesting to identify, but only resolve the divide for specific loss function. Moreover, mixed models would be difficult to interpret.)

In the next section we give the details of our strategy, along with our heuristic justifications. In Section 3, we give theoretical results: We give conditions under which our method of combining different MSP's provides better predictions than any of the individual MSP's from which it is formed. Section 4 discusses the potential benefits from omitting some data points. Finally, in a concluding section we identify some of the remaining gaps and questions to address the broader issues of modeling and prediction.

2. GENERAL DESCRIPTION OF THE TECHNIQUE

The technique we present here was first described heuristically in Clarke (1997). Later, de Luna and Skouras (1999) computationally implemented a special, heuristic case of the technique in a time series context. We begin by defining the technique from Clarke (1997) rigorously. This rigor will permit establishment of an optimality result in Section 3.

2.1 *Formulation of the Method*

Formally, suppose we have k techniques for model selection denoted $MSP_i(y^n)$ for $i = 1, ...k$, where y^n is the data stream $y_1, ..., y_n$. Here, an MSP is a rule by which one associates to y^n a parametric family to be used to model the data sequence. The parametric family can then give a prediction for the next outcome Y_{n+1}. We will assume that each parametric family is equipped with a unique prior; the family may or may not include explanatory variables. For now, we assume there are no explanatory variables but we release this assumption in Section 3. We denote the collection of prior likelihood pairs we are willing to consider by $\tilde{F}$ with elements $\tilde{f}_i$ of the form $w(\theta)q(y|\theta)$. (We use the Bayesian framework for the convenience of working with $m(y^{n+1}|y^n)$ rather than $q(y_{n+1}|\hat{\theta}(y^n))$. The predictive densities have also been identified by Aitchison 1975 as optimal under relative entropy which locally often behaves like squared error loss.)

Ideally, we want to choose $\tilde{F}$ to be the collection of all smooth images of finite dimensional real hyperplanes in the collection of all probability distributions on a measurable space (a set equipped with a sigma field), with a density. Since we cannot deal with uncountably many parametric families well, we extract from $\tilde{F}$ a finite list of models $F = \{f_1, ..., f_n\}$ from which we will choose.

We suppose that the members of F are representative in the sense that no member of $\tilde{F}$ is too far away from some member of F. For instance, in the case of IID data and no covariates, one can imagine transmitting messages from $\tilde{F}$ and seeking a family of representatives F that achieves the rate distortion function lower bound, an optimality criterion from data compression, see Berger (1971). Intuitively, small F's give high compression and distortion but low complexity whereas larger F's will give less compression and distortion but higher complexity. The parametrization in the true model will help determine what representatives to choose.

If there are k MSP's, $MSP_1, ..., MSP_k$ then we partition $\tilde{F}$ into k subsets $\tilde{F}_1, ..., \tilde{F}_k$. This induces a partition $F_1, ...F_k$ of F. The partition of $\tilde{F}$ into $\tilde{F}_i$'s by the MSP's is defined by choosing squared error loss and setting

$$\begin{aligned}\tilde{F}_i = \tilde{F}_{i,n} = \{w(\theta)q(y|\theta) | &E_m(E_i(Y_{n+1}|Y^n) - E_{wq}(Y_{n+1}|Y^n))^2 \\ &\leq \min_j E_m(E_i(Y_{n+1}|Y^n) - E_{wq}(Y_{n+1}|Y^n))^2\} \qquad (2.1)\end{aligned}$$

where E_m denotes expectation with respect to

$$m(y^n) = \int w(\theta)q(y^n|\theta)d\theta \tag{2.2}$$

for appropriate n, and

$$E_i(Y_{n+1}|Y^n = y^n) = E_{MSP_i(y^n)}(Y_{n+1}|Y^n = y^n) = \int y_{n+1}p(y_{n+1}|\theta)w(\theta|y^n)d\theta dy_{n+1}, \tag{2.3}$$

in which $p(y_j|\theta)$ is the parametric family chosen by MSP_i upon receipt of $Y^n = y^n$. Here, $w(\theta|y^n)$ is the posterior for θ given y^n using the prior $w(\theta)$ and the parametric family $p(y_j|\theta)$ chosen by MSP_i upon receipt of $Y^n = y^n$. Finally,

$$E_{wq}(Y_{n+1}|Y^n) = \int y_{n+1}q(y_{n+1}|\theta)w_t(\theta|Y^n)d\theta, \tag{2.4}$$

in which $q(y_i|\theta)$ is the true parametric family used in the true posterior w_t as well as in the likelihood for y_{n+1}. Note that the data y^n are used to choose the MSP and to give a prediction from the model chosen by the MSP.

It is seen that $\tilde{F}_i$ is the set of models whose posterior means are best matched, under squared error loss, by the posterior means from models chosen by MSP_i. This is reasonable because the posterior mean of the true model is the optimal predictor of Y_{n+1} using Y^n under squared error loss. We have used squared error loss for its mathematical convenience in illustrating the proposed technique.

Note that now we have associated to each MSP a collection of parametric families for which it performs better than the other MSP's under consideration. The parametric families associated to an MSP in this way will be called its catchment area, the members of which it (the MSP) is most suited to choosing when one of them is true.

For instance, with independent data, the BIC satisfies an optimality property for exponential families whereas the AIC is not even consistent. However, the AIC may be more appropriate in prediction contexts where one wants to permit models with more parameters. Thus, we have reason to believe that some MSP's are better at choosing different types of models, when they are true, than other MSP's are. In the AIC versus BIC case this may be related to the different penalties on the number of parameters: The AIC will probably 'find' a model with many parameters faster than the BIC will. The BIC will probably 'find' a model with few parameters faster than the AIC.

Admittedly, the parametrization of a model can nevertheless influence model selection with any MSP and this will be largely arbitrary. Nevertheless, the complexity of the model, in terms of number of parameters, may be regarded as an aspect of physical modeling based on the number of influences on the response. As a generality, if we choose a model with fewer parameters one would expect it to approximate the true model poorly. If we choose a model with many parameters, its complexity will degrade its predictive performance.

In expressions (2.1) and (2.4) we have used the notion of a true parametric family. We take this to mean that the data generating mechanism is in one of a class of similar data generating mechanisms which, for physical modeling reasons, can be represented by different true parameter values in the conventional sense. Implicitly, we regard identifying a parametric family to represent the class data generating mechanisms as the central problem in model selection. The subsequent estimation of the parameters is dissociable from model selection in this sense. In this way we avoid nonidentifiability.

Now we have a collection of MSP's and an optimal catchment area for each. We also have predictors whose arguments are the data, and the MSP. It remains to make the choice of MSP into a function of the data. So, let $T = T(Y^n)$ take values from 1 to k to identify one of the k MSP's for each $Y^n = y^n$. This T is intended to choose the MSP which should be most effective at choosing a model for y^n. We use T to improve the prediction of Y_{n+1} by using Y^n to select the best MSP first, using that MSP to get a prediction. Thus, in parallel to a data sequence Y_1,...,Y_n we have a prediction sequence $\hat{Y}_1$,...,$\hat{Y}_n$ in which $\hat{Y}_{n+1}$ predicts Y_{n+1}. We set

$$\hat{Y}_{n+1} = \hat{Y}_{n+1,T(y^n)} = E_{MSP_{T(y^n)}}(Y_{n+1}|Y^n = y^n). \tag{2.5}$$

2.2 *Evaluating How the Method Performs*

Aside from specifying T, Section 2.1 provides a perfectly well defined procedure for generating a prediction sequence. To obey the prequential principle we next want to develop a way to evaluate its performance independently of its construction. Depending on what the evaluation reveals there will be several options open to us.

Our evaluation rests on two indices of predictive performance. First we will define a current error, $CURE$, and a current threshold CUT that we hope is greater than $CURE$. Then we will define a cumulative sum of squared errors for an MSP, CSE, and a conditional variance for the cumulative sum of squared errors, $CVCSE$. We will want the CSE to be less than a mean plus a function of the the $CVCSE$.

When $wq = w(\theta)q(y|\theta)$ is true, we assess how well $\hat{Y}_{n+1,i}$ (where $MSP(y^n) = i$) has predicted Y_{n+1} by evaluating the conditional expectation of the current squared error

$$CURE = (Y_{n+1} - E_{MSP_{T(y^n)}}(Y_{n+1}|y^n))^2. \tag{2.6}$$

holding y^n and $w(\theta)q(y|\theta)$ fixed. This gives

$$E_{(Y_{n+1}|y^n),wq}(Y_{n+1} - E_{MSP_{T(y^n)}}(Y_{n+1}|y^n))^2$$

$$= \int (y_{n+1} - E_{MSP_{T(y^n)}}(Y_{n+1}|y^n))^2 m(y_{n+1}|y^n) dy_{n+1}, \tag{2.7}$$

in which

$$m(y_{n+1}|y^n) = \int q(y_{n+1}|\theta) \frac{w(\theta)q(y^n|\theta)}{\int w(\theta')q(y^n|\theta')d\theta'} d\theta. \tag{2.8}$$

We also want the conditional variance of the current squared error (2.6). This is

$$Var_{(Y_{n+1}|y^n),wq}((Y_{n+1} - E_{MSP_{T(y^n)}}(Y_{n+1}|y^n))^2). \tag{2.9}$$

In practice, however, wq is not known so we cannot take expectations with respect to it.

Note that (2.7), (2.8) and (2.9) depend on the true but unknown model wq. It is tempting to replace wq by the model chosen by the MSP. However, this would violate the prequential principle. Arguably, this violation is minor, but for the sake of intellectual rigor we get around this problem by replacing the conditional density (2.8) based on wq by the mean of the k conditional densities for $(y_{n+1}|y^n)$ obtained from the k prior likelihood pairs chosen by the k MSP's. This choice is independent of T, gives the benefits of averaging, and partially addresses model uncertainty since the models chosen by the different MSP's come from disjoint sets. This is indicated by changing the subscript from wq to avg. (A more involved averaging would be over all elements of F.) Thus we have

$$E_{(Y_{n+1}|y^n),avg}(Y_{n+1} - E_{MSP_{T(y^n)}}(Y_{n+1}|y^n))^2$$

$$= \frac{1}{k}\sum_{l=1}^{k}\int_{\Theta_l}\int_X (y_{n+1} - E_{MSP_{T(y^n)}}(Y_{n+1}|y^n))^2 p_l(y_{n+1}|\theta_l) w_l(\theta_l|y^n) dy_{n+1} d\theta_l \tag{2.10}$$

in which $p_l(y_j|\theta_l)$ is chosen by MSP_l. We define the variance similarly and denote it

$$Var_{(Y_{n+1}|y^n),avg}((Y_{n+1} - E_{MSP_{T(y^n)}}(Y_{n+1}|y^n))^2) \tag{2.11}$$

Note that once the k MSP's and sets F_i have been chosen, and a data sequence y^n given, the expressions (2.10) and (2.11) are fully defined.

Since we are using $\hat{Y}_{n+1} = \hat{Y}_{n+1}(y^n)$ to predict Y_{n+1}, we calculate $(y_{n+1} - \hat{y}_{n+1})^2$ and compare it to thresholds of the form 'mean plus or minus a factor times the standard error'. In particular, we compare the current error

$$CURE = (y_{n+1} - \hat{y}_{n+1})^2 \tag{2.12}$$

to the current threshold

$$CUT = (2.10) + 3K\sqrt{(2.11)} \tag{2.13}$$

in which K is a factor to be chosen later. Now, we want $CURE \leq CUT$ for good prediction. The reverse event $CURE > CUT$ means that the model we have used gave a prediction too far from the actual data point y_{n+1}. When this occurs, we may excuse it as a random fluctuation or we may want to take remedial action. For instance, we might want the option of using a different model for our next prediction. We can rechoose the model using the same MSP or rechoose the MSP and use it to rechoose the model. It is possible that we end up with the same MSP choosing the same old model, however, we have required that choice to compete against the alternatives.

To distinguish these two cases – rechoose the model with the same MSP and rechoose both the MSP and the model – we note that before using a different MSP we want to be sure that our current model class is really inadequate. Thus, we find the cumulative error that has arisen from the use of the MSP. Note that since expectations have so far been over Y_{n+1} with respect to the models chosen by k MSP's we have neglected somewhat the effect of T, even though we used T to choose the MSP from which to get a model for predictions. This gap can be partially addressed by the choice of terms included in the cumulative error sum. Obvious possibilities are 1) One can use the cumulative errors of only the most recent uses of the MSP chosen by T, 2) One can use the cumulative errors of all uses of the MSP, or 3) One can use the cumulative sum of all prediction errors that one would have made had one used that MSP all the time. The form of the cumulative error one uses will depend on the assumptions one makes: For IID data it makes sense to use to use 3). For stationary dependent data or independent but non-stationary data we would suggest 2) and for truly inchoate data sequences 1) might be the best choice. We return to the issue of data retention/deletion later.

The cumulative sum of errors for an MSP that we consider is

$$CSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,T})^2, \tag{2.14}$$

in which it is understood that the sum is over some well specified collection of uses of the MSP, actual or hypothetical. Generically, we have denoted the number of such uses by n. For instance, we might have predicted $y_1, ..., y_n$ by use of MSP_i chosen because $T(y^1) = ... = T(y^n) = i$. If $T(y^{n+1}) \neq i$ then we might change the MSP and possibly wish to reset n to 1. How often we evaluate T – at each timestep as in this instance or only for selected timesteps – will have implications for how often we permit ourselves to change the MSP.

We compare the CSE for an MSP to a threshold analogous to (2.13). Thus we require a mean and variance for (2.14). We define

$$CECSE(wq) = \frac{1}{n}\sum_{i=1}^{n} E_{wq}((Y_i - E_{MSP_{T(y^{i-1})}}(Y_i|y^{i-1}))^2|T(y^{i-1}) = t) \tag{2.15}$$

to be the conditional expectation of the CSE. We have written (2.15) as if T had chosen the same MSP for n timesteps in a row and we have deleted the data predating the last change of MSP. As with the form of the expression (2.14), one can imagine several non-equivalent ways to form the sum in (2.15). Similarly, the conditional variance of the cumulative sum of errors is

$$CVCSE(wq) = \frac{1}{n}\sum_{i=1}^{n} Var_{wq}((Y_i - E_{MSP_{T(y^{i-1})}}(Y_i|y^{i-1}))^2|T(y^{i-1}) = t) \tag{2.16}$$

Note that in (2.15) and (2.16) the conditioning changes from term to term to reflect the accumulation of data; neither (2.15) nor (2.16) is the actual expectation or variance of the CSE over the whole sample space. In addition, the sum in (2.15) presumes the predictions are independent; this is typically false. We can, and do, correct for this by the inclusion in (2.17) of a sample size dependent constant $K = K(n)$ which increases slowly with n. This is reasonable because if one assumes independence when in fact the data are dependent one usually gets an unjustifiably small standard error. We regard the constant K as a correction for this. An alternative is to model the sequential dependence structure carefully. However, the motivation for the technique in this paper is, in part, to address cases where such modeling is not possible. We hope (2.14) and (2.15) are practical surrogates for the actual quantities.

As with $CURE$, wq in (2.16) is unknown. Rather than replacing wq with an average of models, we want to examine the variation in the error due to the MSP directly. Since all we want is a threshold, we take a supremum. Thus, we say the MSP $T(y^n) = i$ as inadequate when

$$CSE > SCCT = \sup_{wq \in F_i} (CECSE(wq) + K\sqrt{CVCSE(wq)}) \tag{2.17}$$

where F_i is the catchment area of MSP_i and SCCT is the supremal cumulative conditional threshold, $SCCT = SCCT(i, n)$.

We argue that (2.13) and (2.17) satisfy a weak form of the prequential principle. These performance assessments are not entirely independent of the procedure generating the predictions. However, they are dependent only on the aggregate properties of the catchment areas. (The K in 2.17 depends only on classes also.)

Now, if we begin at timestep 0 and choose MSP_i to predict y_1 at timestep 1, and then continue using MSP_i – whether out of modeling arguments or because $T(y^2) = ...T(y^n - 1) = i$, then at time step n there are 4 possible actions we might take to predict y_{n+1}. They can be recorded as follows:

1. We might get

$$CURE \leq CUT, CSE \leq SCCT,$$

indicating good prediction in the present and a history of good prediction. In this case, we use the current data point to update the parameter estimates of the model currently in use. We continue to use this model to generate a prediction for time $n + 1$.

2. We might get

$$CURE \geq CUT, CSE \leq SCCT,$$

indicating a bad prediction in the present but a history of good prediction. This leads us to hope that the problem is with the lowest element of the prediction, the choice of model. The problem may be more serious in the sense that we have a higher level problem, namely a bad MSP, but having a good history suggests that the MSP is still adequate. In this case, we re-use the MSP to rechoose the model. Then we estimate the parameters in the new model, using all data up to the present and get a prediction from it for the next time step.

3. We might get

$$CURE \geq CUT, CSE \geq SCCT$$

indicating a bad prediction in the present, and a history of bad enough predictions that the cumulative error is inflated. Together, these bad predictions suggest the higher level problem that the catchment area of the MSP is just not capturing the the phenomenon. In this case, we rechoose the MSP and then use the new MSP to choose a new model. We use this newly chosen model from the newly chosen MSP to get a prediction for the next time step. We implicitly assume that we rechoose the MSP using the run of data from the last change of MSP to the present.

4. The final possibility is that we get

$$CURE \leq CUT, CSE \geq SCCT.$$

This indicates the unusual case that we got a good prediction from a bad MSP. In practice we choose the thresholds so that this will be mathematically impossible, or its probability will be is very small. We return to this point in Section 3.

We comment that setting $SCCT = 0$ puts us automatically in cases 3 or 4; this corresponds to rechoosing the MSP at each timestep which is in effect what de Luna and Skouras (1999) did. At the cost of more frequent and larger evaluations this eliminates the use of (2.14), (2.15), (2.16). By contrast, setting $CUT = \infty$ puts us automatically in cases 1 or 4. Since 4 is heuristically ruled out, we are left with case 1: We never rechoose the MSP. This provides a sense in which the present procedure generalizes existing methods to a sequential procedure.

3. THEORETICAL RESULTS

As we have seen it is often possible to associate to an MSP a physical interpretation which motivates its use and a catchment area on which its use is both natural and optimal. Consequently, MSP's that have similar interpretations, or have nearly the same sets of models associated to them may perform similarly. Some of these equivalences (AIC - C_p for instance) have already been noted.

Here, we abstract from this setting to consider a set of MSP's, each with a set of models on which to use it, the sets assumed disjoint. Our main result is that in a mean squared error sense, the procedure in Section 2 is asymptotically no worse than using any fixed MSP. That is, changing from MSP to MSP by use of a sensible T can only decrease the asymptotic mean squared error. This means that an adaptive strategy outperforms any fixed strategy which is one of its ingredients.

3.1 *Optimality of the Method over Individual MSP's*

For ease of exposition, suppose $k = 2$, that is, we have MSP_1 and MSP_2, with catchment areas $F_{1,n}$ and $F_{2,n}$, respectively, defined as in (2.1) by the loss function, so that $T(Y^n) = 1$ or 2. The case $k \geq 3$ is similar. Our result, informally, is that if T can be used to identify the right catchment area asymptotically then using T to choose an MSP as in Section 2 gives a smaller asymptotic expected squared error than the constant use of either of the MSP's from which T chooses. More formally, we have the following.

Definition: The function $T(Y^n)$ choosing one of k MSP's is consistent for the sequence of catchment areas $F_{i,n}$ of the MSP's if and only if for any i and any sequence wq_n in $F_{i,n}$, the indicator function $\chi_{T(Y^n)=i}$ converges to 1 in wq_n probability.

That is, the consistency of T means that T chooses the right MSP, or set F_i, regardless of which element in F_i is true. Note that we have dropped the subscript on the catchment area: We do this for brevity and to reflect the implicit assumption that when implementing this technique the catchment areas for $n+1$ must be chosen so that they are broadly compatible with the catchment areas at time n.

Theorem 3.1: Let T be any consistent choice for MSP's for the catchment areas of the MSP's defined by squared error loss and suppose we recalculate T at each timestep using all accumulated data to choose an MSP. Suppose all the prior-likelihood pairs in the sets F_1 and F_2 have uniformly bounded second moments, that is, there is an $M > 0$ so that for all densities wq and all times i, $E_{wq} Y_i^2 < M$. Then, for any $wq \in F$,

$$\liminf_{n \to \infty} [E_{Y^{n+1}} (Y_{n+1} - E_{MSP_i}(Y_{n+1}|Y^n))^2$$

$$-E_{Y^{n+1}} (Y_{n+1} - E_{MSP_{T(Y^n)}}(Y_{n+1}|Y^n))^2] \geq 0, \tag{3.1}$$

in which the expectation is taken with respect to the mixture distribution of Y^{n+1}, i.e., w.r.t. $\int w(\theta) q(y^{n+1}|\theta) d\theta$.

Remark 1: There are many consistent choices for T. One, used in de Luna and Skouras (1999), is the relative cumulative predictive loss. Indeed, de Luna and Skouras (1999) establishes consistency for the catchment areas they use, using all past predictions. Alternatively, one can define T to be the choice of catchment area closest to the empirical distribution function. One can also use statistics from hypothesis tests to choose a catchment area provided that the probability of type one and type two errors goes to zero. These techniques generalize to 3 or more MSP's.

Remark 2: The assumption that the argument of T is the entire data string up to the time of prediction can be relaxed. It is enough that T be consistent. Consequently, it will usually be enough to take the argument of T to be all those outcomes y_i for which in the past MSP_i was actually used.

Proof: Let

$$D(k, T, wq) = (E_{wq}(Y_{n+1}|Y^n) - E_{MSP_k}(Y_{n+1}|Y^n))^2$$

$$-(E_{wq}(Y_{n+1}|Y^n) - E_{MSP_T}(Y_n + 1|Y^n))^2, \tag{3.2}$$

and let Δ denote the difference within the liminf of (3.1). That is, set

$$\Delta = E_{Y^{n+1}}(Y_{n+1} - E_{MSP_k}(Y_{n+1}|Y^n))^2 - E_{Y^{n+1}}(Y_{n+1} - E_{MSP_T}(Y_{n+1}|Y^n))^2. \tag{3.3}$$

Then, by adding and subtracting $E_{wq}(Y_{n+1}|Y^n)$, it is seen that

$$\Delta = E_{Y^n} D(k, T, wq). \tag{3.4}$$

(The two squared terms cancel each other, and both rectangular terms are zero.)

For consistent T we have, under assumption 1, that

$$E_{Y^n} \left(D(k, T, wq) - D(k, i, wq) \right) \to 0, \tag{3.5}$$

when $wq \in F_i$. So, adding and subtracting $E_{Y^n} D(k, i, wq)$ in (3.4) means it is enough to examine the asymptotics of $E_{Y^n} D(k, i, wq)$. This, however, is nearly trivial: To see that (3.1) holds it is enough to note that for fixed i, MSP_k satisfies

$$E_{Y^n} D(i, i, wq) \leq E_{Y^n} D(k, i, wq),$$

because (2.1) guarantees MSP_i is the best MSP to use when an element of F_i is true. $\square$

Thus, we have that using a consistent T improves the squared error performance of predictors. This is partially because we are enlarging the collection of models from which we can choose by associating them with MSP's, but also because we are only using any given MSP for the models where it beats out the other MSP's we are willing to consider.

In Section 2 we indicated that $K(n)$ in (2.17) should be chosen so that the fourth possibility (that we could get $CURE \leq CUT$ while $CSE \geq SCCT$) is ruled out. Our next result shows one sense in which this is possible. We write $SCCT(i)$ to emphasize the dependence on MSP_i.

Theorem 3.2. Suppose a Central Limit Theorem holds for the sequence of cumulative sums CSE in (2.14) with rate $\phi(n)/\sqrt{n}$, where $\phi(n)$ is increasing, when $wq \in F_i$. Let $\psi(n)$ be a non-decreasing sequence. Then, for $K(n) = \psi(n)\phi(n)$ we have that for $wq \in F_i$

$$P_{wq}(CSE \geq SCCT(i)) \to 0.$$

Remark 1: It is seen that Theorem 3.1 does not depend on the detailed structure of the procedure T we have defined in Section 2; Theorem 3.1 only requires consistency. By contrast, Theorem 3.2 uses the detailed structure of the procedure as can be seen in the proof. Indeed, Theorem 3.2 gives that for $wq \in F_i$, $P_{wq}(CURE \leq CUT, CSE \geq SCCT(i)) \leq P_{wq}(CSE \geq SCCT(i))$, so that the present result really only uses the procedure of Section 2 for $SCCT$. The initial level of the structure based on CU is analogously based on the CLT, and is used primarily to identify the optimal conditional expectation. Thus, the three levels of parameter estimation, model choice by an MSP, and choice of MSP are dissociable.

Remark 2: The choice of $\phi(n)$ depends on the rate in the CLT for the catchment area. If the data is independent and one models with a catchment area of independent models then $\phi(n) = 1$ is a sensible choice and $\psi(n)$ can be constant (or increasing very slowly) to give a probability from a normal percentile (or zero). If the data are dependent then $\phi(n)$ will typically be increasing and the choice of ψ may be more problematic. Clearly, different catchment areas may have different ϕ's but one may still use the same ψ.

Proof: For $wq \in F_i$, and $SCCT(i)$ as in (2.17), we can remove the supremum to get

$$P_{wq}(CSE > \sup_{wq}(CECSE(wq) + K(n)\sqrt{CVCSE(wq)}))$$

$$\leq P_{wq}(CSE > (CECSE(wq) + \frac{\psi(n)\phi(n)}{\sqrt{n}}\sqrt{nCVCSE(wq)})). \qquad (3.6)$$

By the CLT, $\phi(n)/\sqrt{n}\sqrt{nCVCSE(wq)}$ converges to a constant. By the LLN, the $CECSE$ converges to a constant also. Thus, since $CECSE$ is positive, the slow increase in $\psi(n)$ forces (3.6) to go to zero. $\square$

We observe that

$$T(Y^n) = \chi_{CSE_n \leq SCCT_n} T(Y^{n-1}) + \chi_{CSE_n > SCCT_n} T(Y^n) \tag{3.7}$$

and that $CSE > SCCT$ means we rechoose the MSP, whereas the complement $CSE \leq SCCT$ means $T(Y^n) = T(Y^{n+1})$ That is, (3.7) is a sort of recurrence relation for an MSP, showing when we use the same MSP repeatedly.

3.2 *Inclusion of Finitely Many Explanatory Variables*

In our new setting, suppose we have up to l explanatory variables. That is, we believe Y_i is distributed as $f(x_{1,i}, ..., x_{l,i}, \theta)$ and it is understood that the first l entries in the parameter θ are coefficients of $X_i = (X_{1,i} ... X_{l,i})$. Our task will be to predict Y_{n+1} using $y_1, ..., y_n$ and $X_1, ..., X_{n+1}$. Note that we use X_{n+1} as well as $X^n = (X_1, ..., X_n)$ and Y^n to predict Y_{n+1}. We do not rule out the case that one of the X_i's is a lagged version of Y_n.

Now, the optimal predictor under squared error loss is

$$\hat{Y}_{n+1} = E_{MSP_{T(Y^n,X^n)}}(Y_{n+1}|Y^n = y^n, X^{n+1})$$

$$= \int y_{n+1} p(y_{n+1}|\theta, X_{n+1}) w(\theta|y^n, X^n)) d\theta dy_{n+1} \tag{3.8}$$

which is a parallel to (2.3). In (3.8), the parametric family $p(y_j|\theta, X_j)$ is chosen by $MSP_{T(Y^n,X^n)}$ upon receipt of $Y^n = y^n$ and X^n. Likewise, $w(\theta|y^n, X^n)$ is the posterior for θ given y^n and X^n using the prior $w(\theta)$ and the parametric family $p(y_j|\theta, X_j)$ chosen by $MSP_{T(Y^n,X^n)}$ upon receipt of $Y^n = y^n$ and X^n.

Here we are assuming that a nonstochastic countably infinite sequence of design points X_1..., X_n,... at which measurements will be made, has been fixed before the data $Y_1, ...Y_n$ are collected. This is unrealistic in that design points can be chosen adaptively, however, we ignore this rather than approximating it by putting a distribution on the X_i's.

Now, analogous to (2.6), and (2.7) to assess $\hat{Y}_{n+1}$ we examine the conditional variance holding y^n, X^n and $w(\theta)q(y|\theta, X)$ fixed. This is

$$E_{(Y_{n+1}|y^n,X^{n+1}),wq}(Y_{n+1} - \hat{Y}_{n+1})^2$$

$$= \int (y_{n+1} - E_{MSP_{T(y^n,X^n)}}(Y_{n+1}|y^n, X^{n+1}))^2 m(y_{n+1}|y^n, X^{n+1}) dy_{n+1}, \tag{3.9}$$

where

$$m(y_{n+1}|y^n, X^{n+1}) = \int q(y_{n+1}|\theta, X_{n+1}) \frac{w(\theta)q(y^n|\theta, X^n)}{\int w(\theta')q(y^n|\theta', X^n)d\theta'} d\theta. \qquad (3.10))$$

Replacing wq in (3.9) by an average as in (2.10) gives a form for the conditional mean. The analogue to (2.9) is

$$Var_{(Y_{n+1}|y^n,X^{n+1}),T}((Y_{n+1}-E_{MSP_{T(y^n,X^n)}}(Y_{n+1}|y^n,X^{n+1}))^2), \tag{3.11}$$

in which again wq can be replace by the average.

The forms of CURE, CUT, and CSE are otherwise unchanged, however, $SCCT$ is now based on the conditional expectation of the cumulative sum of errors

$$CECSE(wq)=\frac{1}{n}\sum_{i=1}^{n}E_{wq}((Y_i-E_{MSP_{T(y^{i-1},X^{i-1})}}(Y_i|y^{i-1},X^{i-1}))^2|T(y^{i-1},X^{i-1})=t) \tag{3.12}$$

and the conditional variance of the cumulative sum of errors

$$CVCSE(wq)=\frac{1}{n}\sum_{i=1}^{n}Var_{wq}((Y_i-E_{MSP_{T(y^{i-1},X^{i-1})}}(Y_i|y^{i-1},X^{i}))^2|T(y^{i-1},X^{i-1})=t). \tag{3.13}$$

At each time n we have the same four possible actions as before. We show that again choosing an MSP adaptively by the use of a consistent T performs better than any one of the MSP's and the fourth possibility has asymptotic probability zero of being chosen.

Theorem 3.3: Optimality: Assume the hypotheses of Theorem 3.1, and that the ranges of the explanatory variables are compact.

Then, for any $wq \in F$,

$$\liminf_{n\to\infty}[E_{Y^{n+1}}(Y_{n+1}-E_{MSP_i}(Y_{n+1}|Y^n,X^n))^2$$

$$-E_{Y^{n+1}}(Y_{n+1}-E_{MSP_{T(Y^n,X^n)}}(Y_{n+1}|Y^n,X^n))^2]\geq 0, \tag{3.14}$$

in which the expectation is taken with respect to the mixture distribution of Y^{n+1}, i.e., w.r.t. $\int w(\theta)q(y^{n+1}|\theta,X^{n+1})d\theta$.

Compatibility: Assume the hypotheses of Theorem 3.2 and that the ranges of the explanatory variables are compact. Then, for any $wq \in F_i$,

$$P_{wq}(CSE \geq SCCT(i)) \to 0.$$

Proof: The proofs of Theorems 3.1 and 3.2 transfer to this new setting with the changes in definition described from (3.8) to (3.13). □

We comment that there is nothing sacred about squared error loss. It is seen that one can replace squared error loss in (2.1) by any other loss function to get a different partition $\{F_1,...,F_k\}$. Likewise, the optimal predictor changes from the conditional mean. Then, one would use the new loss function to assess the difference between the predictor and the next outcome, analogously to (2.6). Continuing in this fashion it is seen that the technique carries over to a large class of loss functions. We suggest that the results here also extend.

4. DATA RETENTION AND MODEL MIS-SPECIFICATION

It has generally been axiomatic in statistics that one wants to retain as much data as possible. In fact, this intuition is not entirely correct. For instance, the amount of data one should retain has been observed to be dependent on the goals of the analysis. A concrete example is given by Fearn (1992) in a calibration setting. Consider the regression of Y on X: One gets a smaller MSE for predictions if one spreads the X's over the entire domain rather than letting them pile up in the centre. Thus, if we have a lot of data and it tends to accumulate centrally, we may get a better MSE performance for predictions if we throw out some of the data. A lot of data in this context means that even after we have thrown out a significant fraction of the data there is so much left that we can estimate the regression coefficients with good precision. In effect, throwing out some of the naturally occurring data moves our results closer to what we would have got with a designed experiment that spread the X's uniformly over the interval or, ideally, put all of them at the endpoints. On the other hand, if one really wanted to estimate the parameters one would of course retain all the data possible.

Looking closer at Fearn (1992) one sees that the gain in MSE performance is chiefly for predictions made relatively far (more than 2 σ) from the mean. It can be shown, see Fearn (1992), that in this case (concern with MSE far from the mean) regressing Y on X after discarding data gives predictions that are better than what one would get from regressing X on Y and inverting. In other cases, one would insist on regressing Y on X for physical reasons – it just may not make sense to regress the other way. The issue here is that the extra data in the centre, while representative of the population for which one wants to predict, is 'misleading' because it is *too* representative: The improvement in central performance (which is good in the most commonly occurring setting) is at the expense of how well we handle atypical incoming data points.

Thus we have a situation in which throwing out data in response to model mis-specification (or the concern about it) gives better predictive performance. This suggests a general principle: One wants to retain all the data *only* when model-misspecification is negligible.

Why does this make sense? Consider the case where one model, say P, is true and we unwisely use a parametric family P_θ which does not contain P. It is well known that $P_{\hat{\theta}}$ will converge to the member of the parametric family, P_{θ^*} closest to P in relative entropy distance. However, the issue is to compare the distance between P and P_{θ^*} and the distance between P_{θ^*} and $P_{\hat{\theta}}$. If it is possible to throw out data to make $P_{\hat{\theta}}$ closer to P rather than closer to P_{θ^*} it would be helpful. Of course, P is unknown, but the principle remains because of Fearn (1992) – without knowing the true model, we still know how to allocate the X's optimally.

By contrast, de Luna and Skouras (1999) is a time series setting in which model mis-specification is assumed not to exist and they rechoose the MSP at each time step using all the accumulated data. In effect, they set both SCCT's to zero. This is sensible if we are certain we have identified a class of distributions which contains the true distribution, and all the data is representative of the same identifiable member of the family, and we can uncover it rapidly as data accumulate. (Rapidly means we don't get trapped in a local optimum.) This is the same as the setting of Theorem 3.1. In effect, we uncover the true model so fast that the estimated model gives useful predictions, i.e., prediction is a function of good model selection.

The procedure developed in Section 2 is intended to be useful even when model mis-specification is a problem, and it is suboptimal to retain all data when one changes from one MSP to another, or reuses the same MSP for many time steps. This assumes all models are wrong but may be useful, so one cannot assume away model mis-specification by sufficiently detailed modeling (which may not be feasible).

Consider an example which is at the opposite extreme of de Luna and Skouras (1999) and Theorem 3.1: Suppose the data stream one is trying to predict is a sequence of strings of finite length. Suppose the length of the strings is variable but as a practical matter it is impossible (or worthless) to model the lengths. Also, assume that the data from different strings are unrelated, with unrelated distributions, possibly in different catchment areas. Within a string, the data follows a distribution known to be in a relatively small class. Now, it would be natural to suspect that a change in MSP is associated with the arrival of a new string. Although, the reverse need not be true. (Not every new string necessarily forces a change in the MSP: Two successive strings may have distributions in the same catchment area and therefore both are best found by the same MSP.)

In this example, we have independence and nonidenticality from string to string so it makes sense to throw out all data preceding the most recent change in MSP. Consequently, earlier data cannot help make predictions. Our technique can accommodate this: Once the MSP has been rejected, one can use only the most recent data that led to the rejection of the MSP (the data where $CURE$ exceeded CUT so often that CSE ended up exceeding $SCCT$) to rechoose an MSP by recalculating T. Then one uses this MSP until it's rejection is forced by too many errors of too great a magnitude.

Consider a variant on the sequence-of-strings example. Suppose, that within a string the data are dependent and the degree of dependence assumed by the models in the catchment areas mistakenly underrepresents it. That is, the data is more dependent than you think. Then if one has k data points it will be heuristically equivalent to k' data points where $k' < k$ in terms of a model which has less dependence. (This can be formalized in some cases, see Clarke 1996.) Otherwise put, you will believe you know more than you do as a consequence of mis-specifying the dependence structure. In this case, to get standard errors for prediction that are closer to the ones one would get from using the true model were it known one would have to throw out some data – and this is within a string! In short, this is a case in which using all the data in a wrong model does worse than using less data in the same wrong model. By contrast, in the rare case that the data is more independent than you think then you will have been conservative. This is suboptimal, but will not lead you astray so badly.

Now, a key question in a predictive context is how much data to retain, as well as what to do with it. In the present procedure, there are three places that data are used: Choosing an MSP by use of T, making a prediction (use the MSP to choose a model, then estimate the parameters), and evaluating the thresholds to assess performance. There are several settings in which different strategies for the use of data may be optimal, and in many cases it will be unclear which to use because the optimal strategies are dependent on the unknown model class. All that we can say is that in model mis-specification often it will be good to throw out data, but how to do this is unclear.

As a generality, when using T to choose an MSP, we suggest it is better to retain more recent data, or functions of data that are most tied to recent data e.g., the last few residuals in a regression setting. Also, as a generality, it may be better to use thresholds and form predictions from all previous uses of the MSP chosen, i.e., if a different MSP had been used at a timestep in the past do not use the data point from that timestep when calculating the CSE and $SCCT$ or for making the next prediction. That is, it is probably best to use the same data for both predictions and thresholds. Note that in this proposal, one does not in general rechoose the MSP at each timestep as in de Luna and Skouras (1999). Instead, one uses a chosen MSP until its use leads to a CSE that is larger than its $SCCT$, rechoosing the MSP only when it failed. This is intermediate between using full data retention to rechoose at every timestep and throwing out all data from past MSP's as in the first sequence of strings example. This is intended to approximate the most typically optimal procedure and be not too bad in other cases.

With Fearn (1992) in mind, we suggest refining the procedure for making predictions and setting thresholds by not retaining all data from the previous uses of the MSP. That is, above a threshold on sample size to ensure the precision of estimates, we should throw out central data between changes of MSP particularly as it recedes into the past, especially in a regression context. This is similar to throwing out outliers, except that the central values are the outliers relative to detecting when to change the method of prediction. This makes sense because we want our predictions to be good when the X's are far from their mean value: Those are the cases mostly likely to make us want to change models or MSP's. Indeed, one expects to find lack of fit as one extrapolates outside the domain of the explanatory variables. The error structure – not changing MSP's until CSE exceeds $SCCT$ – means that before changing models we are sure our current model and MSP really are inadequate.

An additional problem with rechoosing the MSP at every timestep is that the variation introduced by the use of T is unexamined: Possibly we end up overfitting the data because we have allowed so many parametric families. This is the same argument as is used for not performing too many hypothesis tests, or estimating too many parameters from a fixed sample.

Thus, we are arguing that there are circumstances in which it is better not to use all the data all the time and in such cases our algorithm here (which uses the same MSP until an error criterion is met, but records which MSP was used at each timestep) is better than rechoosing the MSP at every timestep, the other extreme. The improvement will be in computational complexity also, but our argument is based on better prediction in settings where model mis-specification is unavoidable.

5. DISCUSSION

This paper has three main points. The first is to present a general form of a technique for online prediction that combines the use of several model selection principles (MSP's), and partial retention of data in response to model mis-specification. The second point was to establish that the the use of several MSP's gives better squared error predictive performance than restriction to a single model MSP. The third point was to argue that model mis-specification may imply the desirability of not using all the data all the time. Indeed, we argue that throwing out data points may be useful in predictive context.

There are several issues that impinge on this that deserve comment. First, the present method should interact well with Bayesian model averaging. One can, for instance choose a neighborhood around the model chosen by our procedure, and average over it to get predictions. Alternatively, one can choose a neighborhood around each of the models chosen by each MSP, average within each of those neighborhoods and then average over the local averages. The benefits of Bayesian model averaging are probably dissociable from the benefits of the present method.

A natural way to define neighborhoods in the predictive context is by using Shannon's Mutual Information to topologize the collection of all models. One would hope that models that are close in information are also close physically, that is they do not represent strikingly different physical assumptions. Then, averaging over a collection of neighborhoods centered at models chosen by distinct MSP's would represent greater averaging than merely averaging over the neighborhood around a model chosen by the best MSP. In short, we would have a collections of Occam's windows around the models chosen by our MSP's and any two models being averaged would be either in the same neighborhood (hence physically compatible) or in two disjoint neighborhoods centred on physically incompatible models. We suggest that local averaging for physical compatibility will be enough in a prequential context.

Second, there are technical issues that require further work. Is there a clear example showing that the catchment area of one MSP (say BIC with models having few parameters) is meaningfully different from the catchment area of another MSP (say AIC with models having many parameters)? Is it reasonable to do as we have done in terms of regarding $F_{n,i}$ as being F_i? This assumption permitted us to imagine using the natural extensions of representatives of the catchment area for all n. We anticipate there are cases in which $F_{i,n}$ will be stable as n increases or at least can be characterized as a function of n as n increases; this was done implicitly by de Luna and Skouras (1999). Moreover, it is not clear how many MSP's one should use. One answer is as many as have catchment areas that are physically relevant. However, too few or two many will lead to different problems in some cases.

Third, the details of applications in many specific cases beyond de Luna and Skouras (1999) remain to be worked out. How much past data to use, which of the past data to use, and how to use it remain unexplored outside several examples and heuristics. The full generality of the method remains to be tested. We have not formally explored the optimal frequency of reselecting the MSP. We suggest that reselecting the MSP at every time step will be suboptimal in many cases. Moreover, if there is model mis-specification we may want to throw out data. However, we probably want to retain as much as possible of the data accumulated forcing us to change MSP's, and preferentially retain noncentral data that led us to the use of the new MSP in earlier stages when recalculating error thresholds and predicting the next outcome. Clearly, the more often we re-choose our MSP the more the information in our data will be used to choose a catchment area and there will be less information to permit use of the MSP and parameter estimation. Ultimately, this may weaken our predictions by inflating their variance. This is a sense in which the variation in T remains to be examined and indicates that re-choosing the MSP too often may be sub-optimal.

Finally, the central principle might be that the more frequently we wish to use T to rechoose the MSP, the more data we must retain, and so the less model mis-specification we can tolerate. Equivalently, the less frequently we re-use T, the less data we need to retain, and the more model mis-specification we can tolerate while still getting predictions that are no worse.

6. ACKNOWLEDGEMENTS

The author is grateful to Goetz Trenkler at the Department of Statistics at the University of Dortmund for helpful discussions and important background on prediction that motivated this project. This research was partially supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475 at the Statistics Department at the University of Dortmund and by NSERC Operating Grant 5-81506. Paul Gustafson at the Department of Statistics at the University of British Columbia and Ao Yuan at the Massachussetts General Hospital read and commented on an earlier version of this manuscript.

REFERENCES

Aitchison, J. (1975) Goodness of prediction fit. *Biometrika*, **62**, 547-554.

Akaike, H. (1977) On entropy maximization principles. An *Applications of Statistics*, Krishnaiah, P. K. Ed. pp. 27-41, North Holland, Amsterdam.

Barron, A. R. (1985) *Logically Smooth Density Estimation.* Ph.D. thesis, Department of Electrical and Computer Engineering, Stanford University.

Barron, A.R. and Cover, T. (1990) Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, **44**, 1034-1054.

Barron, A. R. and Xie, Q. (1996) Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inform. Theory*, to appear.

Berger, T. (1971) *Rate Distortion Theory.* Prentice Hall, Englewood Cliffs, New Jersey.

Bethel, J. and Shumway, R. (1988) Asymptotic Properties of Information Theoretic Methods of Model Selection. University of California at Davis, Division of Statistics Technical Report #112.

Bozdogan, H., Bearse, P. M. and Schottmann, A. M. (1997). Empirical econometric modeling of food consumption using a new informational complexity approach. *J. Appl. Econ.*, **12**, 563-586.

Clarke, B. (1996) Implications of reference priors for prior information and sample size. *J. Amer. Statist. Assoc.*, **91**, 173-184.

Clarke, B. (1997) Online Forecasting Proposal. *Technical Report 3/1997* at the Statistics Department of the University of Dortmund.

Dawid, A. P. (1992) Prequential data analysis. In: *Current Issues in Statistical Inference: Essays in Honor of D. Basu.* Ghosh, M. and Pathak, P. K. Eds., pp. 113-126, Inst. Math. Statist. Lecture Notes.

Dawid, A. P (1984) Statistical theory: The prequential approach. *J. Roy. Statist. Soc. Ser. B*, **147**, 278-292.

Dawid, A. P (1986) Symmetry analysis of the mixed model. Research Report #53, Department of Statistical Science, University College London.

Fearn, T.(1992). Flat or Natural? A Note on the Choice of Calibration Samples. In: *Near Infra-Red Spectroscopy: Bridging the Gap between Data Analysis and NIR Applications.* Hildrum et al. Eds. Ellis Howard Publishers, New York. pp. 61–67.

Geisser, S. and Eddy, W. (1979) A predictive approach to model selection. *J. Amer. Statist. Assoc.*, **74**, 153-160.

Hannan, E. J. (1980) The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071-1081.

Hannan, E. J. and Quin, B. G. (1979) The determination of the order of an autoregression. *J. R. Statist. Soc., B*, **41**, 190-195.

Haughton, D. (1988) On the choice of a model to fit data in an exponential family. *Ann. Statist.*, **16**, 342-355.

Li, K. C.,(1987). Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set. *Ann. Statist.*, **15**, No. 3, 958-975.

Rissanen, J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465-471.

Rissanen, J. (1996) Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, **47**, 40-47.

Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993) On testing the validity of sequential probability forecasts, *J. Amer. Statist. Soc.*, **88**, 355-359.

Schwartz, G. (1978) Estimating the dimension of a model. *Ann. Statist.* , **6**, 461-464.

Shibata, R., (1981). An optimal selection of regression variables. *Biometrika*, **68**, No. 1, 45-54.

Skouras, C. and Dawid, A. P. (1998) On efficient point prediction systems, *J.R. Statist. Soc. B.*, **60** Part 4, pp. 765-780.

Wallace, C. S. and Freeman, P.R. (1987) Estimation and inference by compact coding, with Discussion. *J. Roy. Statist. Ser. B*, **49**, 240-265.

Woodroofe, M. (1982) On model selection and the arcsine laws. *Ann. Statist.*, **10**, 1182-1194.

Yang, Y. and Barron, A. R. (1998) An asymptotic property of model selection criteria. *I.E.E.E. Trans. Inform. Theory*, **44**, No. 1, 95-116.

Yuan, A. and Clarke, B. (1999) A minimally informative likelihood for decision analysis and robustness in location families. To appear in *Can. J. Statist.*.