Sankhyā: The Indian Journal of Statistics
2001, Volume 63, Series A, Pt. 2, pp. 229–249

COMBINING MODEL SELECTION PROCEDURES FOR ONLINE PREDICTION

By B. CLARKE
University of British Columbia, Vancouver, Canada

SUMMARY. 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 is associated with a collection of models for which it is best suited. This means one can use the data 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, re-use 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. **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 one. Many people advocate a specific MSP for general use. However, other people, with equally good reasons, advocate a different MSP. They can't all be right. The discrepancy can be cleared up by recognizing that one MSP may be better for a given class of models than another MSP is. Also, one MSP may be better for one purpose, say prediction, than another MSP is for another purpose, say parameter estimation. This means the physical meaning and statistical interpretation of MSP's cannot be ignored. So, 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, you should keep your options open: You should search over various MSP's, and the classes of models associated to them and evaluate performance by a criterion which is good, independently of the purpose of the modeling.

Here, we justify a technique for how to choose an MSP for use in an online prediction setting. Our technique permits different MSP's at different times depending on how well they perform. The main strength of our technique is that it makes

Paper received November 2000; revised March 2001.

AMS (1991) *subject classification.* Primary 62M20; secondary 62L99.

Key words and phrases. Prequential statistics, model selection, model mis-specification, model meta-selection, online prediction

very weak assumptions on the data generating mechanism, satisfies a form of what has been termed the 'prequential' principle see Dawid (1984), and asymptotically performs no worse than using the 'best' predictor.

First, 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. 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.

Here, the prediction technique we develop is in the spirit of the predictive sequential – 'prequential' – approach of A. P. Dawid and co-authors. This 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. 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 others. More recently Skouras and Dawid (1998) study the efficiency of point prediction system and Wong (2000) and Wong and Clarke (2000) propose a prediction technique that outperforms Bayes in some small sample contexts.

The prequential approach has two main principles. One is that the statistical problem is a sequential game in which any model is to be evaluated by the quality of the forecasts it produces for the next outcome using the specific data set at hand. This focuses attention on predictions and ensures inferences will be primarily empirical rather than based on model assumptions. The second, also called the 'prequential' principle, see Dawid (1984) is that a forecast should be assessed by a method which compares forecasts to realized outcomes in a way that is independent of the model used to make the forecast. That is, one wants to avoid using the model to evaluate the performance of the model. Otherwise put, there should be a common performance criterion used to judge all models.

In brief, we assume we have sequential data and, given the first n of them, we predict the $n+1$ outcome. The prediction technique we develop here associates a class of models to each of a collection of MSP's and uses a statistic to choose one of them. 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), 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 predictions have been bad enough for long enough). The adequacy of prediction is measured by a recent error and by a cumulative error. A main part of the specification of the procedure will be identifying thresholds for rechoosing the model and rechoosing the MSP. (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 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.

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 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 adaptive 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.

To make this concrete consider the Akaike Information Criterion, AIC, and the Bayesian Information Criterion, BIC. The AIC, Akaike (1977), chooses the member of a class of parametric families having the largest value of

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

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

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

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 (1.1) was motivated by entropy considerations. Nevertheless, the AIC is equivalent to Mallows' C_p, see Shibata (1981), as well as to cross-validation and generalized-cross-validation, see Li (1987). Recall that the AIC is inconsistent for model selection, see Woodroofe (1982) and Hannan (1980), but that Shibata (1981) established that the AIC is 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. Moreover, Haughton (1988) agrees with Geisser and Eddy (1979) that the inconsistency may not affect the use of the AIC for prediction. Indeed, there is evidence that AIC is optimal in certain predictive contexts, see Shao (1997) and Li (1987).

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

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

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, see Haughton (1988). Using a Bayes factor argument, Schwarz (1978) establishes the optimality of the BIC for exponential families when the dimension remains bounded.

Thus, there is a sort of predictive optimality which one might associate to the AIC and a sort of hypothesis testing optimality one might 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. Thus, the AIC and BIC are expected to perform well on different classes of parametric families. However, if we cannot choose the right class of models and cannot tell if the intuition behind the AIC or BIC is relevant, which – if either – should we use?

Recently, an interesting answer to this question has been supplied, showing that the situation for the AIC and BIC is not as simple as the foregoing discussion assumes. Indeed, Mukhopadhyay (2000) reveals that the current use of (1.1) and (1.2) is misleading. The BIC arises when zero-one loss is used but needs to be adjusted to accommodate increasing dimensions. This can be done as in Mukhopadhyay (2000, Sec. 2.2.2, 2.2.3). It is seen there that correcting the asymptotic approximation leads to a generalized BIC. Moreover, an empirical Bayes model selection rule is equivalent to the AIC (Mukhopadhyay 2000, Sec. 3.3, 3.4) and this can be outpeformed by an unconstrained empirical Bayes rule (in which the empirical Bayes estimates replace least squares estimates), regardless of whether the true model is in the model space, see Mukhopadhyay (2000, 3.5, 3.6). Thus, one can argue that, in squared error prediction loss, a correct version of the BIC may be optimal, in principle obviating conventional use and justification of the AIC.

For contrast, suppose you want to estimate a density and are concerned about tail behavior. Knowing that coding criteria involve logarithms of density ratios that are sensitive to tail behavior you might hope that some kind of information criterion is relevant. So, consider the minimum description length (MDL) context. Barron and Cover (1990), and Rissanen (1996) minimize a data driven analogue of coding redundancy to choose a model. One still must optimize over a specific class of functions, and the size of the penalty term and the risk is determined 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). 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. 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.

These MSP's (AIC, BIC, MDL,...) are only a few of the MSP's authors have proposed. There are many others. A partial list includes: informational complexity, see Bozdogan et al. (1997), informational minimaxity, see Barron and Xie (2000), minimally informative likelihoods, see Yuan and Clarke (1999). However, our point is to combine MSP's in a prequential context. This means we want to use knowledge of the optimality properties of MSP's in place of assumptions about the data generating mechanism. This is one reason some authors have sought to establish general properties of collections of MSP's, usually based on the penalty term.

The AIC and BIC are members of a 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. When f_m is $o(n)$, and unbounded, Bethel and Shumway (1988) give consistency. Here, we will assume that MSP's with sufficiently different f_m's are optimal in sufficiently different senses that they are unlikely to choose the same parametric family.

A partial characterization of MSP's begun by Li (1987) was continued by Shao (1997). He defined a generalized information criterion GIC_{λ_n} as the sum of a squared error term and a complexity penalty, with λ_n representing the relative weighting of the two terms. Shao (1997, Sec. 4) identifies three classes of MSP's. The first sets $\lambda = 2$ and contains Mallows' C_p, the AIC, delete-1 cross-validation and generalized cross-validation; these may be appropriate when there is no fixed dimension correct model. The second has $\lambda \to \infty$ and contains the delete-d form of cross validation for $d/n \to \infty$; it may be appropriate when there is a fixed dimension correct model. The third has any fixed $\lambda > 2$ and contains the delete-d cross-validation with $d/n \to \tau$ for some $\tau \in (0,1)$. It represents a trade off between the first two classes. We suggest that this might be a good triple of classes to use with the technique we present below.

In a different context, 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 as a dimensionality constant. The third term is a complexity penalty, like 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 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 (1.5) can be related to the bias correction interpretation of the AIC, and to the BIC.

We comment that the inclusiveness of the assumptions here permits us to combine Bayesian and frequentist methods as in the AIC, BIC case. Indeed, we can go back and forth between them as the data indicate. Thus, although we present our methods in a Bayesian context 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 compare the Bayes and frequentist models by how they perform in a predictive context in the real world, an evaluation criterion that is independent of the modeling assumptions. In many cases the two approaches will give equivalent predictions although remain conceptually distinct.

In the next section we give the details of our strategy, along with 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. We also show that our method reduces to the standard method under the usual assumption. 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 what we call the adaptive predictor by rigorously specifying the technique from Clarke (1997). 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 a parametric family equipped with a unique prior to y^n. The parametric family and prior generate a prediction for the next outcome Y_{n+1}. For now, we assume the family has 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 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 densities on a measurable space with respect to a fixed dominating measure. Since we cannot deal with uncountably many parametric families, we extract from $\tilde{F}$ a finite list of models $F = \{f_1, ..., f_\ell\}$ from which we will choose. Note that ℓ is not dependent on n. 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. Intuitively, small F's give higher approximation error but low complexity whereas larger F's will give smaller approximation error but higher complexity.

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

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

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 $p(y_j|\theta)$. 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}.

It is seen that $\tilde{F}_i$ is the set of models with predictive means that are best matched, under squared error loss, by the predictive means from models chosen by MSP_i. This is reasonable because the predictive mean of the true model is the optimal predictor of Y_{n+1} using Y^n. The point of the $\tilde{F}_i$'s is to associate to each MSP a collection of parametric families for which it performs better than the other MSP's. The $\tilde{F}_i$'s will be called catchment areas.

For instance, with independent data, the BIC satisfies an optimality property for exponential families when the dimension is bounded 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. 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.

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 various parameter values. In general, if we approximate a true parametric family by a parametric family with fewer parameters we expect the discrepancy to lead to bias. If we approximate a true parametric family by another with many parameters, the complexity will degrade predictive performance. Thus, we regard identifying a parametric family to represent the class of data generating mechanisms as the central problem in model selection. Subsequent estimation of the parameters is dissociable from model selection and avoids nonidentifiability.

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. Intuitively, it is enough for T to identify the optimal catchment area. 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

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

and refer to it as the adaptive predictor.

2.2 *Evaluating How Well the Adaptive Predictor Performs.* Aside from choosing T, Section 2.1 provides a well defined procedure for generating a prediction sequence. To obey the prequential principle we evaluate its performance independently of its construction. Our evaluation rests on two indices of predictive performance. First we will define a current error, $CURE$, and a current threshold CUT. 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

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

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

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

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. 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. Thus we have

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

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

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

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

Since we are using $\hat{Y}_{n+1} = \hat{Y}_{n+1}(y^n)$ we compare the current error

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

to the current threshold

$$CUT = (2.10) + 3L_1\sqrt{(2.11)} \quad (2.13)$$

in which L_1 is a scalar 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 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.

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 this in Section 4.

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

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

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

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

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

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

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

$$CSE > SCCT = \sup_{wq \in F_i} (CECSE(wq) + L_2 \sqrt{CVCSE(wq)}) \qquad (2.17)$$

where F_i is the catchment area of MSP_i, SCCT is the supremal cumulative conditional threshold, $SCCT = SCCT(i, n)$, and $L_2 = L_2(n)$ is a slowly increasing function of n. If one catchment area has only independent models and another catchment area has dependent models then the rate of decrease of the standard error for elements of the two classes may differ. In such cases, one would permit $L_2 = L_2(n, i)$ where $i = 1, ..., k$ indicates the MSP, or equivalently its catchment area.

Use of L_2 compensates for the assumption, implicit in the sum in (2.16), that the prediction errors are independent. In fact, they are not independent so (2.16) alone will typically give an unjustifiably small standard error. In (2.17) the rate of decrease in the standard error is $L_2(n)/\sqrt{n}$, so a slow increase in L_2 can be used to inflate the standard error to a realistic size in dependent cases. Since similar considerations apply to L_1 in (2.13), we consider L_2 only.

In general, it is unclear how to choose L_2. One technique follows from extending the approach of Zidek and Wang (2000). For the first n' data points form the empirical distribution $\hat{F}_{n'}$. Generate many new independent sequences of n' data points from $\hat{F}_{n'}$. For each sequence form the expression in (2.15). A histogram of these values gives the sampling distribution of (2.15) as a random variable. One can then take the variance of the sampling distribution. Doing this for $n' = 2, ..., n$ gives a sequence of pairs $(\log Var_{n'}, \log n')$. One can fit a simple linear regression model to these $n-1$ pairs so the coefficient of $\log n'$ can be transformed to an exponent of n.

The validity of the linear regression can be tested by seeing if a first (or higher) order autocorrelation structure, or a moving average structure gives different results. The dependence of the data going into the initial $\hat{F}_{n'}$ may be tested, when n' is large enough, by seeing if leaving out one or two data point makes only a small difference compared to the independent case. If the difference is small this confirms dependence and one would leave out a small number of well chosen data points so that the information in the remaining data could be used to form an $\hat{F}_{n''}$ where n'' represents the number of data points which, when treated as independent, have information equivalent in magnitude to the information in the n' original dependent data points. One is led to do this because the adequacy of the empirical distribution as an estimator for the true distribution relies on independence.

It is seen that (2.13) and (2.17) satisfy a weak form of the prequential principle in that CUT and $SCCT$ are partially independent of the procedure generating the predictions. They are not entirely independent of the procedure, however, because they depend on the aggregate properties of the catchment areas. This is a weakness that may be difficult to overcome because of the generality of the model spaces that might be considered.

2.3 *A Tentative Algorithm.* 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 ways 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 use the updated 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. There might be 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. In this case, we rechoose the MSP and then use the new MSP to choose a new model. We use this newly chosen model to get a prediction for the next time step.

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 in cases 3 or 4; this corresponds to rechoosing the MSP at each timestep, as in de Luna and Skouras (1999). 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.

3. Theoretical Results

For ease of exposition, suppose $k = 2$, so 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.

3.1 *Optimality of the method over individual MSP's.* The effectiveness of the adaptive method depends on T. A good T will give a useful MSP reliably. One criterion for this is the following.

Definition: The function $T(Y^n)$ is consistent for the catchment areas $F_{i,n}$ 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.

The consistency of T means that T chooses the right MSP, or set F_i, regardless of which element in F_i is true. We have dropped the subscript on the catchment area for brevity and to indicate the catchment areas for $n+1$ must be chosen to be compatible with the catchment areas at time n.

Theorem 3.1. *Let T be any consistent choice for MSP's suppose we recalculate T at each timestep using all accumulated data. If all the elements of F_1 and F_2 have uniformly bounded second moments, i.e., that is, there is an $M > 0$ so that for all densities wq and all times i, $E_{wq}Y_i^2 < M$, then we have that for any $wq \in F$,*

$$\liminf_{n\to\infty}[E_{Y^{n+1}}(Y_{n+1} - E_{MSP_i}(Y_{n+1}|Y^n))^2 - E_{Y^{n+1}}(Y_{n+1} - E_{MSP_{T(Y^n)}}(Y_{n+1}|Y^n))^2] \geq 0, \qquad (3.1)$$

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. For instance, de Luna and Skouras (1999) choose T to be the index of the MSP minimizing 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 give the 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.

Remark 2. The assumption that the argument of T is the entire data string up to the time of prediction can be relaxed. For consistency of T it will usually be enough to use those outcomes y_i for which in the past MSP_i was actually used.

Proof. Let

$$\begin{aligned} D(k,T,wq) = &(E_{wq}(Y_{n+1}|Y^n) - E_{MSP_k}(Y_{n+1}|Y^n))^2 \\ &-(E_{wq}(Y_{n+1}|Y^n) - E_{MSP_T}(Y_n+1|Y^n))^2, \end{aligned} \tag{3.2}$$

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 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 is easy: To see that (3.1) holds note that for fixed i, MSP_k satisfies

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

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

Thus, 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, but also because we are only using an MSP where it beats out the other MSP's.

Theorem 3.2. *Assume the conditions of Theorem 3.1. In addition, suppose that both MSP's are consistent and let* $q = q_\theta$ *denote the densities in the true parametric family, equipped with prior* w*. Then, in squared error distance with respect to* wq*,* *we have that the adaptive predictor based on T is asymptotically equivalent to the asymptotically best predictor. That is:*

$$E_{MSP(T)}(Y_{n+1}|Y^n) - E_{wq}(Y_{n+1}|Y^n) \overset{L^2}{\to} 0,$$

as $n \to \infty$.

Proof. For simplicity, we use inequalities such as $(a_1, ..., a_k)^2 \leq K \sum a_i^2$ without further comment, letting K vary from occurrence to occurrence. When wq is in the catchment area F_i, we can add and subtract $E_{MSP(i)}(Y_{n+1}|Y^n)$ to get

$$\begin{aligned} &E_{wq}(E_{MSP(T)}(Y_{n+1}|Y^n) - E_{wq}(Y_{n+1}|Y^n))^2 \\ &\quad \leq KE_{wq}(E_{MSP(T)}(Y_{n+1}|Y^n) - E_{MSP(i)}(Y_{n+1}|Y^n))^2 \\ &\quad\quad +KE_{wq}(E_{MSP(i)}(Y_{n+1}|Y^n) - E_{wq}(Y_{n+1}|Y^n))^2. \end{aligned} \tag{3.6}$$

So, it is enough to show both terms in (3.6) go to zero. The first term can be decomposed into sets on which T assumes a specified value to give

$$E_{wq}\left(\sum_j \chi_{T=j} E_{MSP(j)}(Y_{n+1}|Y^n) - E_{MSP(i)}(Y_{n+1}|Y^n)\right)^2. \tag{3.7}$$

The contribution of the term with $j = i$ is zero. Dropping it and using the above inequality, we get

$$K \sum_{j \neq i} E_{wq}\chi_{T=j}(E_{MSP(j)}(Y_{n+1}|Y^n) - E_{MSP(i)}(Y_{n+1}|Y^n))^2 \tag{3.8}$$

as an upper bound on (3.7). In turn, (3.8) is bounded by

$$K \sum_{j \neq i} E_{wq}\chi_{T=j} E_{MSP(j)}(Y_{n+1}|Y^n)^2 + K \sum_{j \neq i} E_{wq}\chi_{T=j} E_{MSP(i)}(Y_{n+1}|Y^n)^2. \tag{3.9}$$

All the terms in (3.9) go to zero: Observe that by dropping the indicator function, the argument of the first expectation is bounded by $E_{MSP(j)}(Y_{n+1}|Y^n)^2$. By the second moment assumption it is seen that $E_{wq}E_{MSP(j)}(Y_{n+1}|Y^n)^2$ is bounded. Consequently, by the dominated convergence theorem, each of the terms in (3.9) goes to zero because, under wq, $\chi_{T=j}$ converges to zero in probability, for $j \neq i$. The second term in (3.6) is similar, but the decomposition is over sets on which $MSP(i)$ gives the members of F_i, and the consistency of $MSP(i)$ permits the domination. □

Thus, when T is consistent, and the MSP's are consistent, the adaptive predictor ultimately uses the same MSP all the time and gets the asymptotically optimal predictor all the time. In this context, the benefit is in small sample behavior. More generally, one does not have consistency – nevermind a stable law to uncover – but the present technique is still applicable.

In Section 2 we indicated that $L_2(n)$ in (2.17) should be chosen so that the fourth possibility is ruled out. Our next result shows this is possible. We write $SCCT(i)$ to emphasize the dependence on MSP_i.

Theorem 3.3. *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 nondecreasing, when $wq \in F_i$. Let $\psi(n)$ be a non-decreasing sequence. Then, as long as $L_2(n) = \psi(n)\phi(n)$ in increasing we have that for $wq \in F_i$*

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

Remark. The choice of $\phi(n)$ depends on the rate in the CLT for the catchment area. Suppose the data is independent and has a catchment area of independent models. Choosing $\phi(n) = 1$ means $\psi(n)$ must be increasing because, if $\psi(n)$ is constant one gets a normal percentile in the limit below. If one chooses $\phi(n) = \sqrt{\log n}$, however, then $\psi(n)$ can be constant. If the data are dependent then $\phi(n)$ will be increasing and the choice of ψ may be more problematic.

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

$$P_{wq}(CSE > \sup_{wq}(CECSE(wq) + L_2(n)\sqrt{CVCSE(wq)}))$$

$$\leq P_{wq}(CSE > (CECSE(wq) + \frac{\psi(n)\phi(n)}{\sqrt{n}}\sqrt{nCVCSE(wq)})). \qquad (3.10)$$

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 (3.10) to go to zero. □

Remark. Theorem 3.3 uses the detailed structure of the procedure in Section 2, whereas Theorem 3.1 only requires consistency. We see that Theorem 3.3 rules out the fourth possibility because it gives that for every $wq \in F_i$, $P_{wq}(CURE \leq CUT, CSE \geq SCCT(i)) \leq P_{wq}(CSE \geq SCCT(i))$, which goes to zero. The choice of CUT is also based on a CLT so the three levels of parameter estimation, model choice by an MSP, and choice of MSP are dissociable.

3.2 *Inclusion of finitely many explanatory variables.* Now, suppose we have l explanatory variables so Y_i is distributed as $f(x_{1,i},...,x_{l,i},\theta)$ plus an error term 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 is to predict Y_{n+1} using $y_1,...,y_n$ and $X_1,...,X_{n+1}$. The optimal predictor under squared error loss is

$$\begin{aligned} \hat{Y}_{n+1} &= E_{MSP_{T(Y^n,X^n)}}(Y_{n+1}|Y^n = y^n, X^{n+1}) \\ &= \int y_{n+1}p(y_{n+1}|\theta, X_{n+1})w(\theta|y^n, X^n))d\theta dy_{n+1} \qquad (3.11) \end{aligned}$$

which is a parallel to (2.3). In (3.11), 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 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

$$\begin{aligned} &E_{(Y_{n+1}|y^n,X^{n+1}),wq}(Y_{n+1}-\hat{Y}_{n+1})^2 \\ &\quad = \int (y_{n+1} - E_{MSP_{T(y^n,X^n)}}(Y_{n+1}|y^n, X^{n+1}))^2 m(y_{n+1}|y^n, X^{n+1}) dy_{n+1}, \end{aligned} \tag{3.12}$$

where

$$m(y_{n+1}|y^n, X^{n+1}) = \int q(y_{n+1}|\theta, X_{n+1}) \frac{w(\theta)q(y^n|\theta, X^n)}{\int w(\theta')q(y^n|\theta', X^n)d\theta'} d\theta. \tag{3.13}$$

Replacing wq in (3.12) by an average as in (2.10) gives a form for the conditional mean. The analogue to (2.9) is

$$Var_{(Y_{n+1}|y^n,X^{n+1}),T}((Y_{n+1} - E_{MSP_{T(y^n,X^n)}}(Y_{n+1}|y^n, X^{n+1}))^2), \tag{3.14}$$

in which again wq can be replaced 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

$$CECSE(wq) = \frac{1}{n}\sum_{i=1}^{n} E_{wq}((Y_i - E_{MSP_{T(y^{i-1},X^{i-1})}}(Y_i|y^{i-1}, X^{i-1}))^2 | T(y^{i-1}, X^{i-1}) = t) \tag{3.15}$$

and the conditional variance of the cumulative sum of errors

$$CVCSE(wq) = \frac{1}{n}\sum_{i=1}^{n} Var_{wq}((Y_i - E_{MSP_{T(y^{i-1},X^{i-1})}}(Y_i|y^{i-1}, X^{i}))^2 | T(y^{i-1}, X^{i-1}) = t). \tag{3.16}$$

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, reduces to the usual predictor, and the fourth possibility has asymptotic probability zero of being chosen.

Theorem 3.4. *I) Optimality: Assume the hypotheses of Theorem 3.1, and that the ranges of the explanatory variables are compact.*

Then, for any $wq \in F$,

$$\begin{aligned} \liminf_{n\to\infty} [&E_{Y^{n+1}}(Y_{n+1} - E_{MSP_i}(Y_{n+1}|Y^n, X^n))^2 \\ &- \; E_{Y^{n+1}}(Y_{n+1} - E_{MSP_{T(Y^n,X^n)}}(Y_{n+1}|Y^n, X^n))^2] \geq 0, \end{aligned} \tag{3.17}$$

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

II) Reduction: Under the assumptions of Theorem 3.2 and part I), the adaptive predictor reduces to the usual predictor. That is,

$$E_{MSP(T)}(Y_{n+1}|Y^n, X^n) - E_{wq}(Y_{n+1}|Y^n, X^n) \xrightarrow{L^2} 0, \tag{3.18}$$

as $n \to \infty$, for any $wq \in F$.

III) Simplification: Assume the hypotheses of Theorem 3.3 and that the ranges of the explanatory variables are compact. Then, for any $wq \in F_i$,

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

so the fourth possibility never happens asymptotically.

Proof. The proofs of Theorems 3.1, 3.2 and 3.3 transfer to this new setting with the changes in definition described from (3.11) to (3.16). □

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).

4. Data Retention and Model Mis-specification

It is axiomatic in statistics that one wants to use as much data as possible. However, this intuition is not entirely correct. For instance, the amount of data one should retain can depend on the goals of the analysis. In a calibration setting, Fearn (1992) considers the regression of Y on X: He notes that if we have a lot of data and it tends to accumulate centrally, we get a better MSE performance for predictions if we throw out some of the central data. A lot of data in this context means that we can estimate the regression coefficients with good precision. Throwing out some central data moves our predictions 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 retain all the data.

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. Fearn (1992) shows that when one is most concerned 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. 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 is at the expense of how well we handle atypical incoming data points.

Thus throwing out data in response to model mis-specification can give 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. They rechoose the MSP at each time step using all the accumulated data. This is sensible if we have 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 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 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.

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 cannot be modeled. 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. 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.

Since we have independence and nonidenticality from string to string it makes sense to throw out all data preceding the most recent change in MSP, because 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 underrepresents it. Then, you will believe you know more than you do as a consequence of the mis-specified dependence structure. To get standard errors for prediction that are closer to the ones one would get from using the true model 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, if the data is more independent than you think you will have been conservative.

Now, the key question in a predictive context is which 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.

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. 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 fails. 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 the cases most likely to make us want to change models or MSP's are those for which the X are far from their mean value.

An additional problem with rechoosing the MSP at every timestep is that the variation introduced by the use of T is unexamined: We may end up overfitting the data. Thus, there are circumstances in which it is better not to use all the data all the time and in such cases our algorithm here is better than rechoosing the MSP at every timestep. 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 MSP's. The second point was to establish that the use of several MSP's gives better squared error predictive performance than restriction to a single MSP. The third point was to argue that predicting in the presence of model mis-specification may be improved by omitting some data.

There are several issues that impinge on this and 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 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. Models that are close in information should represent similar physical assumptions.

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 $F_{i,n}$ 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; too few or too many will lead to different problems.

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. We may want to retain preferentially the recent noncentral data that led us change MSP's. 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. This may weaken our predictions by inflating their variance.

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.

Acknowledgements. The author is grateful to Goetz Trenkler at the Department of Econometrics 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. The author is also grateful to an anonymous referee for helpful comments on the manuscript.

References

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

Akaike, H. (1977). On entropy maximization principles. In *Applications of Statistics*, P.R. Krishnaiah, ed., 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 Transactions on Information Theory*, **44**, 1034–1054.

Barron, A.R. and Xie, Q. (2000). Asymptotic minimax regret for data compression, gambling, and prediction, *IEEE Transactions on Information Theory*, **46**, 431–445.

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.* Technical Report **112**, Division of Statistics, University of California at Davis.

Bozdogan, H., Bearse, P.M. and Schottmann, A.M. (1997). Empirical econometric modeling of food consumption using a new informational complexity approach, *Journal of Applied Economics*, **12**, 563–586.

Clarke, B. (1997). *Online Forecasting Proposal.* Technical Report Sonderforschungsbereit **475**, University of Dortmund.

Dawid, A.P. (1984). Statistical theory: the prequential approach, *Journal of Royal Statistical Society Series B*, **147**, 278–292.

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

Dawid, A.P. (1992). Prequential data analysis. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, M. Ghosh and P.K. Pathak, eds., IMS Lecture Notes, Institute of Mathematical Statistics, Hayward, CA, 113–126.

De Luna, X. and Skouras, K. (1999). Model metaselection. Research Report 203, Dept. of Statistical Science, UCL. (Download from http://www.econ.umu.se/ xavier-de.luna/)

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, 61–67.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection, *Journal of the American Statistical Association*, **74**, 153–160.

Hannan, E.J. (1980). The estimation of the order of an ARMA process, *Annals of Statistics*, **8**, 1071–1081.

Hannan, E.J. and Quin, B.G. (1979). The determination of the order of an autoregression, *Journal of the Royal Statistical Society Series B*, **41**, 190–195.

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

Li, K.C. (1987). Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set, *Annals of Statistics*, **15**, 958–975.

Mukhopadhyay, N. (2000). *Bayesian Model Selection for High Dimensional Models with Prediction Error Loss and 0-1 Loss.* Ph.D. Thesis, Department of Statistics, Purdue University.

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

Rissanen, J. (1996). Fisher information and stochastic complexity, *IEEE Transactions on Information Theory*, **47**, 40–47.

Schwartz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

Seillier-Moiseiwitsch, F. and Dawid, A.P. (1993). On testing the validity of sequential probability forecasts, *Journal of the American Statistical Association*, **88**, 355–359.

Shao, J. (1997). An asymptotic theory for linear model selection, *Statistica Sinica*, **7**, 221–261.

Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45–54.

Skouras, C. and Dawid, A.P. (1998). On efficient point prediction systems, *Journal of the Royal Statistical Society Series B*, **60**, 765–780.

Wallace, C.S. and Freeman, P.R. (1987). Estimation and inference by compact coding, with Discussion, *Journal of the Royal Statistical Society Series B*, **49**, 240–265.

Wong, H. (2000). *Transient improvement over Bayes prediction under model uncertainty.* PhD Thesis, Department of Statistics, University of British Columbia.

Wong, H. and Clarke, B. (2000). Small sample improvement over Bayes prediction in the presence of model uncertainty, Submitted.

Woodroofe, M. (1982). On model selection and the arcsine laws, *Annals of Statistics*, **10**, 1182–1194.

Yang, Y. and Barron, A.R. (1998). An asymptotic property of model selection criteria, *IEEE Transactions on Information Theory*, **44**, 95–116.

Yuan, A. and Clarke, B. (1999). A minimally informative likelihood for decision analysis: illustration and robustness, *Canadian Journal of Statistics*, **27**, 649–665.

Zidek, J. and Wang, S. (2000). Comment on: "The estimating function bootstrap" by Kalbfleisch and Hu, *Canadian Journal of Statistics* **28**, 482–485.

B. Clarke
Department of Statistics
University of British Columbia
6356 Agricultural Road, Rm. 333
Vancouver, B.C. Canada V6T 1Z2.
E-mail:bertrand@stat.ubc.ca