OPTIMUM DESIGN OF SERIAL MEASUREMENT TREES

ROBERT AHLBRANDT, HARI K. IYER AND DOMINIC F. VECCHIA

Marion Merrell Dow Inc., Colorado State University and the National Institute of Standards and Technology

ABSTRACT

Suppose some or all of \( p \) predictor variables \( X_1, X_2, \ldots, X_p \) may be measured and their values used to predict a response \( Y \). When measurement costs are considered, it may be advantageous to measure predictors serially, deciding adaptively which variable to measure next with an option to stop taking measurements and predict \( Y \) at any point during the measurement process. In this article we derive optimal prediction strategies of a serial type under a cost structure that combines measurement costs with the usual prediction–error loss function. It is assumed that the joint distribution of \( Y, X_1, X_2, \ldots, X_p \) is known. The results provide a theoretical basis upon which estimation of optimum serial measurement plans may be addressed.

Key words and phrases. Bayes rule, CART, loss function, prediction tree.

1. INTRODUCTION

An important objective in designing a prediction tree for regression or classification is the construction of an optimal tree. Usually, tree performance is evaluated in terms of an appropriate loss function, but actually constructing a tree (of a given size) which has minimum expected loss is a problem of considerable computational complexity. Practical procedures for building prediction trees have been proposed by many authors. Almost universally, the procedures are stepwise optimal, operating at each step by splitting each terminal node of a partially grown tree maximize the performance of the new tree. Optimality is almost always expressed exclusively with respect to ordinary prediction–error loss functions.

In this article we consider the design of measurement strategies for constructing prediction
trees which explicitly account for costs incurred by measuring a set of predictors. We derive an optimum measurement strategy in a certain class of serial measurement plans. Here, optimality is with respect to a cost function which depends on both prediction errors and the cost of measurement. An optimum rule prescribes an adaptive serial order for measuring the predictors where, at each measurement opportunity, a decision is made about which (if any) predictor should be measured next. Throughout the present article, it is assumed that the joint distribution of the response and predictor variables is known. Our results provide a theoretical basis upon which estimation of optimum serial measurement plans may be addressed.

Consider a set of $p$ predictor variables $\{X_1, X_2, \ldots, X_p\}$ and a response variable $Y$. The class of serial measurement plans that we consider permits both possible extremes of measurement effort. In exceptional cases, the best strategy may be to use a constant predictor, so none of the predictors need be measured. At the other extreme, some cost structures may require measurement of all of the predictors. If $Y$ is continuous, for example, the best prediction rule under a quadratic loss function (and given measurement costs) may be the usual regression function $E[Y \mid X_1, X_2, \ldots, X_p]$. (In other words, the serial measurement framework includes, in a sense, the usual regression problem.) The more interesting cases, however, are the intermediate possibilities, where a decision may be made to halt the measurement sequence and predict $Y$ based on values of the particular subset of predictors observed thus far. The optimal rule generally requires both the order and number of predictors actually measured to change from one prediction to another.

The measurement strategies that we consider will be termed simple serial measurement plans (SSMP). For each SSMP, there is a Bayes predictor of $Y$ that may be expressed as a prediction tree. The optimum design problem is to construct a SSMP which supports a prediction tree having minimum Bayes risk. A Bayes predictor associated with an optimum SSMP will be called a best serial predictor (BSP) of $Y$. The tree representation of the BSP of $Y$ consists of paths which follow the optimum SSMP for each individual prediction.
A general description of the class of SSMP's, the objective function to be minimized, and the derivation of an optimum plan is presented in section 3. A generalization of SSMP's to a larger class, termed general serial measurement plans, is outlined briefly in the concluding remarks in section 4. We begin in the following section with a discussion of the special case of three predictor variables, which serves to illustrate the essential structure of the serial measurement problem.

Tree-structured prediction rules provide appealing alternatives to conventional linear modelling techniques. There is an extensive literature on both theory and applications of prediction trees, dating back to Morgan and Sonquist's (1963) development of stepwise optimal trees in least-squares regression. A more general treatment of both classification and regression trees is discussed in Breiman et al. (1984). Chou (1991) generalizes procedures presented in Breiman et al., and lists references on the implementation of prediction trees in a wide variety of applications.

2. STRUCTURE OF THE SERIAL PROCEDURE

Breiman et al. (1984, Chapters 9, 10) develop the theoretical framework for classification and regression tree (CART) methodology when the joint cumulative distribution function (cdf) $F$ of $(Y, X_1, X_2, \ldots, X_p)$ is assumed known. They consider a (finite) partition $\tilde{T}$ of the space $\mathcal{X}$ of possible measurement vectors, and study Bayes predictors that are constant on each member $t \in \tilde{T}$. In their case, the Bayes risk depends on the particular partition $\tilde{T}$, but does not depend on the actual tree leading to that partition. When one accounts explicitly for measurement costs, however, the risks of Bayes predictors corresponding to SSMP's will generally depend on their associated trees. The following example illustrates this point.
EXAMPLE 1. Let $S = [-1,1] \times [-1,1]$ and suppose $(X_1, X_2)$ is distributed on $S$ with density function

$$f(x_1, x_2) = \begin{cases} 
1/20 & \text{if } (x_1, x_2) \in S_1 = (0,1] \times (0,1] \\
1/4 & \text{if } (x_1, x_2) \in S_2 = [-1,0] \times (0,1] \\
1/2 & \text{if } (x_1, x_2) \in S_3 = [-1,0] \times [-1,0] \\
1/5 & \text{if } (x_1, x_2) \in S_4 = (0,1] \times [-1,0]
\end{cases}$$

and suppose that the random variable $Y$ is such that

$$Y \mid (X_1, X_2) \in S_1 = 1 \text{ a.s.}$$

$$Y \mid (X_1, X_2) \in S_1^c = 2 \text{ a.s.}$$

Consider the partition $\tilde{T} = \{S_1, S_1^c\}$ of the domain $S$. The Bayes rule corresponding to $\tilde{T}$ is

$$a(x_1, x_2) = \begin{cases} 
1 & \text{if } (x_1, x_2) \in S_1 \\
2 & \text{if } (x_1, x_2) \in S_1^c
\end{cases}$$

with Bayes risk equal to zero. Next consider the two prediction trees shown in Figure 1, and let $c_1$ and $c_2$ be the respective costs of measuring $X_1$ and $X_2$. The expected cost associated with the upper tree in Figure 1 ($X_1$ measured first) is $c_1 + c_2/4$, while that of the lower tree ($X_2$ measured first) is $3c_1/10 + c_2$. So, while both trees give rise to the same partition $\tilde{T}$, the plan which measures $X_1$ first is a better measurement design than the other whenever $14c_1 < 15c_2$.

Figure 1 here

The serial measurement design problem is motivated by the preceding example. If the predictor variables may be measured in any order, and if the cost of measurement varies among the predictors, then an optimum measurement strategy is one which minimizes the total expected cost. That is, optimality is defined with respect to a cost function which is the sum of an ordinary loss function and the cost of measurement. In the remainder of this section, the particular case of three predictors will be used to illustrate the basic structure of SSMP’s.
Let \( Y, X_1, X_2, X_3 \) be random variables with joint cdf \( F \), and for a given predictor \( \hat{Y} \), let \( L(Y, \hat{Y}) \) denote a general loss function. Furthermore, suppose there exist real valued functions \( \mu_0, \mu_i(X_i) \) for \( i \in \{1, 2, 3\} \), \( \mu_{ij}(X_i, X_j) \) for \( i \neq j \) and \( i, j \in \{1, 2, 3\} \), and \( \mu_{ijk}(X_i, X_j, X_k) \) for \( ijk \) a permutation of \( \{1, 2, 3\} \) which minimize

\[
E[L(Y, \mu_0)]
\]

\[
E[L(Y, \mu_i(X_i)) \mid X_i = x_i] \forall x_i \in D_i
\]

\[
E[L(Y, \mu_{ij}(X_i, X_j)) \mid X_i = x_i, X_j = x_j] \forall (x_i, x_j) \in D_i \times D_j
\]

\[
E[L(Y, \mu_{ijk}(X_i, X_j, X_k)) \mid X_i = x_i, X_j = x_j, X_k = x_k] \forall (x_i, x_j, x_k) \in D_i \times D_j \times D_k
\]

for permissible choices of \( i, j, k \) and given domains \( D_1, D_2, D_3 \). We have deliberately defined equivalent expressions for the usual set of eight fixed-subset predictors of \( Y \). So we may write, for example, \( \mu_{21}(b, a) \) rather than \( \mu_{12}(a, b) \). Allowing permuted forms of the predictors is a notational convenience in the present discussion, but also serves to emphasize the order-dependent structure of the SSMP's discussed in the next section. (To be consistent with later notation we should write, for instance, \( \mu_{(i,j)}(X_i, X_j) \). However, there should be no confusion due to our notation for the special case.)

Suppose that \( c_1, c_2, \) and \( c_3 \) denote the respective costs associated with measuring the predictors \( X_1, X_2, \) and \( X_3 \). Suppose further that the loss function \( L(\cdot, \cdot) \) is given in the same units as the measurement costs. Then the total expected cost due to prediction error and measurement costs for respective subsets of the predictors of \( Y \) are given by

\[
\begin{align*}
C_0 & = E[E[L(Y, \mu_0)] \\
C_i & = E[E[L(Y, \mu_i(X_i)) \mid X_i]] + c_i \\
C_{ij} & = E[E[L(Y, \mu_{ij}(X_i, X_j)) \mid X_i, X_j]] + c_i + c_j \\
C_{ijk} & = E[E[L(Y, \mu_{ijk}(X_i, X_j, X_k)) \mid X_i, X_j, X_k]] + c_i + c_j + c_k
\end{align*}
\]

for permissible values of the indices \( i, j \) and \( k \). So, the subset of predictors which minimizes the total expected cost provides a well-defined solution to the problem of choosing a best fixed-subset prediction equation when we explicitly account for known measurement costs.
Example 2. Suppose \( X_1, X_2, X_3 \) are independently distributed such that \( P(X_i = -1) = P(X_i = 1) = 1/2, \ i = 1, 2, 3 \) and the conditional distribution of \( Y \) given \( X_1, X_2, X_3 \) is \( N(-X_1 + 2X_2 + 2X_3 - 2X_1X_2 + X_1X_3, 2 + X_1) \). Consider the squared-error loss function: 
\[ L(Y, \hat{Y}) = (Y - \hat{Y})^2, \]
and suppose \( c_1 = 3, c_2 = 0, \) and \( c_3 = 2 \) are the respective costs of measuring the three predictors. Then the total expected costs (1)-(4) for the eight fixed subset predictors of \( Y \) are: \( C_0 = 16, C_1 = 18, C_2 = 12, C_3 = 14, C_{12} = 10, C_{13} = 15, C_{23} = 10, \) and \( C_{123} = 7. \) Hence, the best fixed-subset predictor for the given cost structure is based on the subset \( \{X_1, X_2, X_3\} \).

Further analysis of Example 2 (presented later) is based on the serial measurement strategy: If the predictor variables may be measured in any order, then we can construct a predictor of \( Y \) with total expected cost that is no greater than (and often strictly less than) the total expected cost of the best fixed-subset predictor of \( Y \). The class of SSMP’s containing an optimum serial predictor is illustrated below in both graphical and algebraic forms.

For permissible values of \( i, j, k \), let \( X_{ij} = (X_i, X_j) \) and \( X_{ijk} = (X_i, X_j, X_k) \). We will write \( \mu_i \) for \( \mu_i(X_i) \), \( \mu_{ij} \) for \( \mu_{ij}(X_{ij}) \), and \( \mu_{ijk} \) for \( \mu_{ijk}(X_{ijk}) \) in the following discussion when there is no potential for confusion. Figure 2 shows, in skeleton form, the possible simple serial measurement paths and resulting predictors of \( Y \) for each of four possible SSMP’s for the present special case. In the figure, a SSMP corresponding to the strategy where \( X_i \) is measured first is indicated by the entry point \( (i) \), while the entry \( (0) \) denotes the SSMP based on no measurements.

Figure 2 here
The class of predictors of \( Y \) associated with the sequential measurement paths in Figure 2 will be denoted by \( \mathcal{Y} \). Let \( N_3 = \{1, 2, 3\} \) and for fixed \( i, j \in N_3 \), let \( k \) be the single element of \( N_3 - \{i, j\} \). Then a typical predictor \( \hat{Y} \in \mathcal{Y} \) is given by

\[
(5) \quad \hat{Y} = \mu_0 I_{\{0\}}(d) + \sum_{i=1}^3 \left( \mu_i I_{D_{i,0}}(X_i) + \sum_{j \in N_3 - \{i\}} \left[ \mu_{ij} I_{D_{ij,0}}(X_{ij}) + \mu_{ijk} I_{D_{ij,k}}(X_{ij}) \right] I_{D_{ij}}(X_i) \right) I_{\{i\}}(d)
\]

where \( d \in \{0, 1, 2, 3\} \) and, for example, choosing \( d = i(i > 0) \) designates \( X_i \in D_i \) as the first predictor to be measured. Associated with this choice is a partition of \( D_i \) comprising the subsets \( D_{i,0} \), \( D_{ij} \), and \( D_{ij,k} \) (for \( i, j, k \) distinct), which represent the regions in the domain of \( X_i \) leading, respectively, to immediate prediction of \( Y \) by \( \mu_i \), to measurement of \( X_j \), or to measurement of \( X_k \). If the value of \( X_i \) leads, for instance, to measurement of \( X_j \), the subsets \( D_{ij,0} \) and \( D_{ij,k} \) appearing in (5) partition \( D_i \times D_j \) into regions designating prediction of \( Y \) by \( \mu_{ij} \) or measurement of \( X_k \) and prediction by \( \mu_{ijk} \), respectively. Furthermore, since \( \mu_{ij} = \mu_{ji} \), we also require that \( (X_i, X_j) \in D_{ij,k} \) if and only if \( (X_j, X_i) \in D_{ji,k} \) for distinct \( i, j \in N_3 \) and \( k \in N_3 - \{i, j\} \cup \{0\} \).

Clearly, any fixed-subset predictor of \( Y \) is contained in the class \( \mathcal{Y} \) by appropriate choice of \( d \) and the various partitioning subsets. For example, if the subset \( \{X_1, X_2\} \) should always be used to predict \( Y \), then \( D \) is set equal to 1 or 2. Assuming that \( d = 1 \), then we choose \( D_{1,2} = D_1, D_{1,0} = D_{1,3} = \emptyset, D_{12,0} = D_1 \times D_2, \) and \( D_{12,3} = \emptyset \), which leads to the predictor \( \hat{Y} = \mu_{12}(X_1, X_2) \). The remaining fixed-subset predictors are obtained in similar fashion.

To illustrate the cost savings of a serial measurement strategy, we return to the situation of Example 2.

**Example 3.** Consider the joint distribution, loss function, and measurement costs in Example 2, and recall that the best fixed-subset of predictors was \( \{X_1, X_2, X_3\} \) with an associated total expected cost of \( C_{123} = 7 \). Let \( \hat{Y} \) be the SSMP represented by the equation

\[
\hat{Y} = \begin{bmatrix} 0 \cdot I_{\{0\}}(X_2, X_3) + \left( 2 \cdot I_{\{1\}}(X_1) + 6 \cdot I_{\{1\}}(X_1) \right) I_{\{1\}}(X_2) \\ 9 \cdot I_{\{1\}}(X_2, X_1) + \left( 0 \cdot I_{\{1\}}(X_3) + 6 \cdot I_{\{1\}}(X_3) \right) I_{\{1\}}(X_2) \end{bmatrix} I_{\{1\}}(X_2)
\]
The total expected cost associated with this SSMP is $C = 6$. Using the results presented in the next section, it can be shown that the domains above correspond to an optimum SSMP.

Serial prediction trees like the one illustrated in the example can be useful in many practical situations. For example, a physician may want to predict a patient's probability of having a heart attack. First, the patient's blood pressure—a relatively inexpensive measurement—is observed. If the blood pressure is above a certain level, the physician may order more expensive blood chemistry tests. Finally, an electrocardiogram may be ordered, depending on the results of earlier tests. In this way the need for more information may be balanced with the cost of obtaining that information to minimize the total expected loss due to prediction error and measurement costs. Notice in this example that the procedure illustrated above explicitly accounts for the possibility that an electrocardiogram may be the preferred second test, depending on the patient's observed blood pressure.

3. SERIAL MEASUREMENT OF SEVERAL PREDICTORS

Suppose $p$ predictor variables \( \{X_i \in D_i, i = 1, 2, \ldots, p\} \) with associated measurement costs \( c_1, c_2, \ldots, c_p \) may be observed in any desired order to predict an unknown response \( Y \). For a given predictor \( \hat{Y} \), let \( L(Y, \hat{Y}) \) denote a general loss function, specified in the same units as the measurement costs.

With \( p \) predictors, there are \( \sum_{i=0}^{p} \frac{p^i}{(p-i)!} \) possible paths in a SSMP for measuring the predictors. For instance, the four prediction trees in Figure 2 (where \( p = 3 \)) contain 16 potential measurement paths. The following definitions provide the framework for constructing a best serial predictor of \( Y \).

**Definition 1.** Define \( A_0 = \emptyset \). Let \( N_p = \{1, 2, \ldots, p\} \) and for all \( k \in N_p \) define

\[
A_k = \{(\alpha_1, \alpha_2, \ldots, \alpha_k) \mid \alpha_i \text{ are distinct elements of } N_p\}
\]

Also, for each \( A_k = (\alpha_1, \alpha_2, \ldots, \alpha_k) \in A_k \), let \( \bar{A}_k = \{\alpha_1, \alpha_2, \ldots, \alpha_k\} \). The relative complement of \( \bar{A}_k \) in \( N_p \) is denoted by \( \bar{A}_k \).
DEFINITION 2. For each \( k \in N_p, A_k \in A_k \), and \( \alpha_{k+1} \in \tilde{A}_k^c \cup \{0\} \), let \( D_{A_k \cdot \alpha_{k+1}} \) be subsets of \( D_{\alpha_1} \times D_{\alpha_2} \times \cdots \times D_{\alpha_k} \) satisfying

\[
\begin{align*}
(a) & \quad D_{A_k \cdot \alpha_{k+1}} \cap D_{A_k \cdot \alpha'_{k+1}} = \emptyset \text{ for all } \alpha_{k+1}, \alpha'_{k+1} \in \tilde{A}_k^c \cup \{0\} \text{ and } \alpha_{k+1} \neq \alpha'_{k+1} \\
(b) & \quad \bigcup_{\alpha_{k+1} \in \tilde{A}_k^c \cup \{0\}} D_{A_k \cdot \alpha_{k+1}} = D_{\alpha_1} \times D_{\alpha_2} \times \cdots \times D_{\alpha_k}
\end{align*}
\]

Then \( \mathcal{G} \) will denote the class of all collections of domains \( \mathcal{D} \) defined by

\[
(7) \quad \mathcal{D} = \{ D_{A_k \cdot \alpha_{k+1}} \mid k = 1, 2, \ldots, p; \ A_k \in A_k; \ \alpha_{k+1} \in \tilde{A}_k^c \cup \{0\} \}
\]

which satisfy conditions (a) and (b).

DEFINITION 3. Corresponding to a given vector \( A_k = (\alpha_1, \alpha_2, \ldots, \alpha_k) \in A_k \), let \( X_{A_k} = (X_{\alpha_1}, X_{\alpha_2}, \ldots, X_{\alpha_k}) \) denote the associated vector of predictors and let \( x_{A_k} \) be an observed value of \( X_{A_k} \). Then the random variable \( \mu_{A_k}(X_{A_k}) \), abbreviated sometimes as \( \mu_{A_k} \), is a function of \( X_{A_k} \) which minimizes

\[
(8) \quad E[L(Y, \mu_{A_k}(X_{A_k})) \mid X_{A_k} = x_{A_k}]
\]

for every \( x_{A_k} \in D_{\alpha_1} \times D_{\alpha_2} \times \cdots \times D_{\alpha_k} \). Also, \( \mu_0 \) is defined to be a constant which minimizes \( E[L(Y, \mu_0)] \).

Suppose that \( A_k = (\alpha_1, \alpha_2, \ldots, \alpha_k) \) and \( A'_k = (\alpha'_1, \alpha'_2, \ldots, \alpha'_k) \) are distinct elements of \( A_k \) in (6). Then if \( \tilde{A}_k = \tilde{A}'_k \) notice that the predictors in (8) satisfy \( \mu_{A_k}(X_{A_k}) = \mu_{A'_k}(X_{A'_k}) \). The following definition presents the general class \( \mathcal{Y} \) of serial predictors of \( Y \) using \( p \) available predictor variables.

DEFINITION 4. For each value of \( d \in \{0, 1, 2, \ldots, p\} \) and collection of regions \( \mathcal{D} \in \mathcal{G} \) in (7), we inductively define a serial predictor of \( Y \) using predictor variables \( X_1, X_2, \ldots, X_p \) as follows:

1. If \( d = 0 \) then define \( \hat{Y} = \mu_0 \). 
2. If \( d = i \) then observe the value of the predictor \( X_i \).

3. Suppose that \( k (\leq p) \) predictors \( X_{c_1}, X_{c_2}, \ldots, X_{c_k} \) have been observed in the listed order, and let \( A_k = (\alpha_1, \alpha_2, \ldots, \alpha_k) \). If the observed value of \( X_{A_k} \) lies in \( D_{A_k,0} \) then define \( \hat{Y} = \mu_{A_k}(X_{A_k}) \); otherwise proceed to step 4.

4. If the observed value of \( X_{A_k} \) lies in \( D_{A_k,\alpha_{k+1}} \), where \( \alpha_{k+1} \in \bar{A}_k \), then observe the value of \( X_{\alpha_{k+1}} \) and return to step 3.

The resulting predictor is denoted by \( \hat{Y}_d(\mathcal{D}) \), and the class of predictors \( \mathcal{Y} \) is defined to be the set of all predictors \( \hat{Y}_d(\mathcal{D}) \) with \( d \in \{0, 1, 2, \ldots, p\} \) and \( \mathcal{D} \in \mathcal{G} \).

The operator \( \bowtie \) is used below to denote concatenation of two vectors. For example, if \( U = (1, 2, 4) \) and \( V = (3, 5) \) then \( U \bowtie V = (1, 2, 4, 3, 5) \). In the following definitions empty products are defined to be one, while empty sums are defined to be zero.

**Definition 5.** For each \( k \in N_{p-1}, A_k \in A_k, \alpha_{k+1} \in \bar{A}_k, \) and \( m \in N_p - N_k \), let

\[
B(A_k, \alpha_{k+1}, m) = \{ B \mid B = A_k \bowtie \alpha_{k+1} \bowtie (\alpha_{k+2}, \ldots, \alpha_m) \text{ is an } m\text{-vector} \}
\]

where \( \alpha_{k+2}, \alpha_{k+3}, \ldots, \alpha_m \) are distinct elements of

\[
[\bar{A}_k \cup \{\alpha_{k+1}\}]^c
\]

This set defines the possible measurement sequences of length \( m \) given the first \( k+1 \) variables to be measured.

**Definition 6.** For a given set of regions \( \mathcal{D} \in \mathcal{G} \) in (7) and for each \( k \in \{0\} \cup N_{p-1} \), let \( A_k \in A_k \) and \( \alpha_{k+1} \in \bar{A}_k \) be fixed. Define

\[
\mathcal{L}(\alpha_{k+1} : X_{A_k}, A_k, \mathcal{D}) = \sum_{i=1}^{k} c_{\alpha_i} + \sum_{m=k+1}^{p} \left[ \sum_{B \in B(A_k, \alpha_{k+1}, m)} \left( L(Y, \mu_B) + \sum_{i \in B - \bar{A}_k} c_i \right) \right]
\]

\[
\cdot \left( \prod_{r=k+1}^{m-1} I(X_{A_k \bowtie (\alpha_{k+1}, \ldots, \alpha_r)} \in D_{A_k \bowtie (\alpha_{k+1}, \ldots, \alpha_r)}, \alpha_{r+1}) \right) \cdot I(X_B \in D_{B,0})
\]

10
Then the total expected cost associated with a predictor in Definition 4 when the first \( k \) predictors measured are \( X_{\alpha_1}, X_{\alpha_2}, \ldots, X_{\alpha_k} \) with values \( x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_k} \), and the \((k + 1)\)st predictor to be measured is \( X_{\alpha_{k+1}} \) is given by

\[
C(\alpha_{k+1} : x_{A_k}, A_k, D) = E[L(\alpha_{k+1} : X_{A_k}, A_k, D) | X_{A_k} = x_{A_k}]
\]

where the expectation is with respect to the distribution of \( X_{(\alpha_{k+1}, \ldots, \alpha_p)} \) given \( X_{A_k} = x_{A_k} \) and \( \{\alpha_{k+1}, \ldots, \alpha_p\} = N_p - \hat{A}_k \).

In the following definition we choose a value \( d^* \in \{0, 1, \ldots, p\} \) and a collection of regions \( D^* \in \mathcal{G} \) from (7) which will be shown to yield a best serial predictor of \( Y \).

**DEFINITION 7.** The collection \( D^* \) of regions is defined inductively as follows:

1. \( D^*_{A_p, 0} = D_{\alpha_1} \times D_{\alpha_2} \times \cdots \times D_{\alpha_p} \). Let \( D^*_p = \{D^*_{A_p, 0}\} \).

2. For \( n \in N_{p-1} \), suppose \( D^*_{A_k, \alpha_{k+1}} \) and \( D^*_{A_k, 0} \) have been defined for \( k = n + 1, n + 2, \ldots, p \).

The regions \( D^*_{A_n, \alpha_{n+1}} \) are defined for each \( A_n = (\alpha_1, \alpha_2, \ldots, \alpha_n) \) and \( \alpha_{n+1} \in \tilde{A}^c_n \cup \{0\} \) by assigning \( x_{A_n} \in D_{\alpha_1} \times D_{\alpha_2} \times \cdots \times D_{\alpha_n} \) to the set \( D^*_{A_n, \alpha_{n+1}} \) if \( \alpha_{n+1} \) is the smallest index such that

\[
C(\alpha_{n+1} : x_{A_n}, A_n, D^*) \leq C(\alpha'_{n+1} : x_{A_n}, A_n, D^*)
\]

for all \( \alpha'_{n+1} \in \tilde{A}^c_n \cup \{0\}, \alpha_{n+1} \neq \alpha'_{n+1} \). Note that calculation of the inequality only requires expectations (10) that depend on domains \( D^*_{A_k, \alpha_{k+1}} \) and \( D^*_{A_k, 0} \) already defined for every \( k > n \). Let \( D^*_n = D^*_p \cap \{D^*_{A_n, \alpha_{n+1}} \mid A_n \in A_n, \alpha_{n+1} \in \tilde{A}^c_n \cup \{0\}\} \).

Define \( D^* = D^*_1 \) and let \( C^*_i \) (\( i = 0, 1, \ldots, p \)) be the total expected cost using \( D^* \) when \( X_i \) is measured first. Then \( d^* \) is defined to be the smallest index \( i \) for which \( C^*_i = \min \{C^*_0, C^*_1, C^*_2, \ldots, C^*_p\} \).
Theorem Let \( c_1, c_2, \ldots, c_p \) be the measurement costs associated with the available predictors \( X_1, X_2, \ldots, X_p \) of a response \( Y \) and let the general loss function \( L(Y, \hat{Y}) \) of a predictor \( \hat{Y} \) be specified in the same units as the costs. Consider the value \( d^* \) and collection of regions \( D^* \) constructed in Definition 7. Then \( \hat{Y}_d(D^*) \) is a best serial predictor (BSP) of \( Y \) in the class \( \mathcal{Y} \) of predictors in Definition 4. That is, \( \hat{Y}_d(D^*) \) minimizes the total expected cost due to prediction errors and measurement costs in the class of serial predictors.

Proof. We show by induction that the regions constructed in Definition 7 are optimal. For fixed \( n \in N_{p-1} \), suppose the optimal regions \( D^{*}_{A_k \cdot \alpha_{k+1}} \) and \( D^{*}_{A_k \cdot 0} \) have been defined for \( k = n + 1, n + 2, \ldots, p \). Observe that these regions are optimal regardless of the choice of \( D^{*}_{A_k \cdot \alpha_{k+1}} \) for \( k \leq n \), given that one is currently at the node \( A_n \). Consider admissible regions \( D_{A_n \cdot 0} \) and \( D_{A_n \cdot \alpha_{n+1}} \) for all \( \alpha_{n+1} \in \tilde{A}^c_n \). We must show that the expected cost using \( d \) and

\[
D_1 = \left\{ D^{*}_{A_k \cdot \alpha_{k+1}} \mid \alpha_{k+1} \in \tilde{A}^c_k \cup \{0\}, k \leq n \right\} \cup \left\{ D^{*}_{A_k \cdot \alpha_{k+1}} \mid \alpha_{k+1} \in \tilde{A}^c_k \cup \{0\}, k > n \right\}
\]

is greater than or equal to the expected cost associated with using \( d \) and

\[
D_2 = \left\{ D^{*}_{A_k \cdot \alpha_{k+1}} \mid \alpha_{k+1} \in \tilde{A}^c_k \cup \{0\}, k < n \right\} \cup \left\{ D^{*}_{A_k \cdot \alpha_{k+1}} \mid \alpha_{k+1} \in \tilde{A}^c_k \cup \{0\}, k \geq n \right\}.
\]

Denote the two predictors of \( Y \) associated with \( D_1 \) and \( D_2 \) by \( \hat{Y}_d(D_1) \) and \( \hat{Y}_d(D_2) \), respectively. Fix \( A_n \in A_n \). Since \( d, D_{A_k \cdot 0} \), and \( D_{A_k \cdot \alpha_{k+1}} \) for \( k < n \) are the same for both predictors, to show that the expected loss associated with \( \hat{Y}_d(D_1) \) is at least as large as that associated with \( \hat{Y}_d(D_2) \) it is sufficient to show that

\[
E \left[ \sum_{\alpha_{n+1} \in \tilde{A}^c_n \cup \{0\}} L(\alpha_{n+1} : X_{A_n}, A_n ; D_1) \cdot I(X_{A_n} \in D_{A_n \cdot \alpha_{n+1}}) \right] \\
\geq E \left[ \sum_{\alpha_{n+1} \in \tilde{A}^c_n \cup \{0\}} L(\alpha_{n+1} : X_{A_n}, A_n ; D_2) \cdot I(X_{A_n} \in D^{*}_{A_n \cdot \alpha_{n+1}}) \right]
\]

The lefthand side of this inequality may be written as the conditional expectation

\[
E \left[ E \left( \sum_{\alpha_{n+1} \in \tilde{A}^c_n \cup \{0\}} L(\alpha_{n+1} : X_{A_n}, A_n ; D_1) \cdot I(X_{A_n} \in D_{A_n \cdot \alpha_{n+1}}) \mid X_{A_n} = x_{A_n} \right) \right] \\
= E \left[ \sum_{\alpha_{n+1} \in \tilde{A}^c_n \cup \{0\}} C(\alpha_{n+1} : x_{A_n}, A_n ; D_1) \cdot I(X_{A_n} \in D_{A_n \cdot \alpha_{n+1}}) \right]
\]

12
where we have used Definition 6. Similarly, the righthand side of the inequality is

\[ E \left[ \sum_{\alpha_{n+1} \in \hat{A}_{n} \cup \{0\}} C(\alpha_{n+1} : x_{A_{n}}, A_{n}, D) \cdot I\left( x_{A_{n}} \in D_{A_{n}}^{*} \alpha_{n+1} \right) \right] . \]

For each value \( x_{A_{n}} \) of \( X_{A_{n}} \), if \( x_{A_{n}} \in D_{A_{n}}^{*} \alpha_{n+1} \) and \( x_{A_{n}} \in D_{A_{n}} \alpha_{n+1} \), then

\[ C(\alpha_{n+1} : x_{A_{n}}, A_{n}, D_{1}) = C(\alpha_{n+1} : x_{A_{n}}, A_{n}, D_{2}) \]

But if \( x_{A_{n}} \in D_{A_{n}}^{*} \alpha_{n+1} \) and \( x_{A_{n}} \in D_{A_{n}} \alpha'_{n+1} \) for some \( \alpha'_{n+1} \in \hat{A}_{n} \cup \{0\} \) and \( \alpha'_{n+1} \neq \alpha_{n+1} \), then by the choice of the optimal regions (see condition 2 of Definition 7) we have

\[ C(\alpha_{n+1} : x_{A_{n}}, A_{n}, D_{1}) \geq C(\alpha_{n+1} : x_{A_{n}}, A_{n}, D_{2}) \]

Hence, the expected total cost of \( \hat{Y}_{d}(D_{2}) \) is no greater than the expected total loss of \( \hat{Y}_{d}(D_{1}) \).

Iterating this procedure we conclude that \( D^{*} \) is superior to \( D \) for any other \( D \in G \) in (7).

The optimality of \( d^{*} \) follows from its definition.

Since the class of predictors \( \mathcal{Y} \) contains all fixed subset predictors of \( Y \), the Theorem implies that a BSP based on the regions given in Definition 7 has expected cost no greater than any predictor of \( Y \) based on a fixed subset of the available predictors.

5. CONCLUDING REMARKS

Implementation of the ideas presented here will require estimation of an optimal SSMP based on a training sample. In the case of squared-error loss, one needs a method of estimating the necessary conditional and unconditional moments. For more general loss functions and continuous predictors, density estimation methods would be required (Devroye, 1987). Identification of the collection of regions (Definition 7) needed in the construction of a best serial predictor presents formidable computational problems, even with only a few predictors.

The main assumption concerning the total cost function is that the measurement cost and prediction-error loss are additive. The proof of the theorem is easily modified to accommodate a subadditive cost function, where the cost of measuring a subset of the predictors
simultaneously is no greater than measuring each in succession. The larger class of measurement strategies for subadditive costs could be called general serial measurement plans to distinguish them from the SSMP's discussed in this article.

ACKNOWLEDGMENT

This work is a contribution of the U. S. government and is not subject to copyright in the United States.

REFERENCES


Figure 1. Prediction trees for serial measurement plans in Example 1.
Figure 2. Prediction trees for serial measurement plans with 3 predictors.