and upon applying the vector calculus identity

$$\nabla \cdot (\psi \bar{A}) = \bar{A} \cdot \nabla\psi + \psi \nabla \cdot \bar{A}$$

to the first term, (13) follows.

Corollary: The divergence of $\hat{s}(x)$ equals its variance at x; i.e.,

$$\nabla \cdot \hat{s}(x) = V(x). \tag{14}$$

This corollary relates spatial variations in $\hat{s}(x)$ to conditional expectations at x; and, in addition, it enables us to state that the CME is completely specified by its variance function $V(x)$. This statement is a consequence of Helmholtz's theorem, which states that a vector is completely specified by its divergence and curl. (Recall that as $\hat{s}(x)$ is conservative, $\nabla \times \hat{s}(x) = 0$.)

The primary value of Property 3 and its corollary is that they relate both the likelihood ratio and CME to a conditional variance function of the signal. Hence, by viewing $L(x)$ as a potential function and $\hat{s}(x)$ as a conservative vector field, the mathematics of potential theory can be applied to solving problems in signal detection theory.

III. Conclusions

This correspondence has noted a fundamental property relating optimum detection and CME for random signals in white Gaussian noise for discrete-time processes, and has discussed the role of the estimation–correlation operation in forming an optimum decision statistic.

By viewing the log-likelihood ratio as a potential function, the CME of the signal was shown to constitute a conservative vector field. This concept was used to show the intimate connection between spatial variation (divergence) of the CME and the conditional signal variance. These results suggest that the mathematics of potential theory might play an important role in furthering the theory of signal detectability.

C. P. Hatsell[3]
Dep. Elec. Eng.
USAF Inst. Technol.
Wright-Patterson AFB, Ohio
L. W. Nolte
Dep. Elec. Eng.
Duke Univ.
Durham, N.C. 27706

[3] Formerly with the Department of Electrical Engineering, Duke University, Durham, N.C.

Note on Optimal Selection of Independent Binary-Valued Features for Pattern Recognition

***Abstract*—Given a set of conditionally independent binary-valued features, a counter example is given to a possible claim that the best subset of features must contain the best single feature.**

Recently, Elashoff *et al.*[1] showed that for optimal selection of a subset of independent binary-valued features, the features generally may not be evaluated independently. Specifically, an example is given[1] in which, given three independent variables x_1, x_2, and x_3 such that $\varepsilon(x_1) < \varepsilon(x_2) < \varepsilon(x_3)$, where $\varepsilon(x_i)$ is the error probability when the ith variable alone is used, the first and third variables are jointly better than the first and second variables. In other words, $\varepsilon(x_1,x_3) < \varepsilon(x_1,x_2)$, where $\varepsilon(x_i,x_j)$ is the probability of error when the ith and jth variables are used together. In this note, the results of Elashoff *et al.* are carried one step further, and it is shown that the best pair of variables need not contain the best single variable.

Let there be two equiprobable pattern classes C_1 and C_2, and let $\alpha_i = P(x_i = 1 \mid C_1)$ and $\beta_i = P(x_i = 1 \mid C_2)$, $i = 1,2,3$, where $P(x_i = 1 \mid C_j)$ is the conditional probability that the ith variable takes on the value ONE conditioned on the jth pattern class. As in Elashoff *et al.*, let the following assumptions be made:

1) $\varepsilon(x_1) < \varepsilon(x_2) < \varepsilon(x_3)$;
2) $\alpha_i < \beta_i$, $i = 1,2,3$;
3) $\beta_1 - \alpha_1 > \beta_2 - \alpha_2 > \beta_3 - \alpha_3$.

For simplicity of notation, let $l_i = (\beta_i - \alpha_i)$, $h_i = \frac{1}{2}(1 - \alpha_i - \beta_i)$, and $D_{ij} = |h_i| - |h_j|$. It is shown by Elashoff *et al.* that for two conditionally independent variables x_i and x_j, the minimum error probability is given by

$$\varepsilon(x_i,x_j) = \tfrac{1}{2}[\varepsilon(x_i) + \varepsilon(x_j) - l_i\,|h_j| - l_j\,|h_i|], \tag{1}$$

where $\varepsilon(x_k) = \frac{1}{2}[1 - (\beta_k - \alpha_k)]$ for $k = i,j$. From (1) and conditions 1), 2), and 3) above, it can easily be shown that for $\varepsilon(x_1,x_2) < \varepsilon(x_2,x_3)$, a sufficient condition is given by

$$|h_1| > |h_3| \tag{2}$$

and a necessary and sufficient condition is given by

$$D_{31} < \frac{1}{2}\left(\frac{l_1 - l_3}{l_2}\right)(1 + 2\,|h_2|). \tag{3}$$

Consider as an example, three features x_1, x_2, and x_3 chosen so as to violate (3) such that $\varepsilon(x_1) < \varepsilon(x_2) < \varepsilon(x_3)$. Such examples are not difficult to find. The probabilities of the three features conditioned on the two pattern classes are given by

$$\alpha_1 = 0.10 \qquad \alpha_2 = 0.05 \qquad \alpha_3 = 0.01$$

$$\beta_1 = 0.90 \qquad \beta_2 = 0.80 \qquad \beta_3 = 0.71.$$

Substituting these figures into (1) yields the following results:

$$\varepsilon(x_1) = 10 \text{ percent} \qquad \varepsilon(x_1,x_2) = 8.25 \text{ percent}$$

$$\varepsilon(x_2) = 12.5 \text{ percent} \qquad \varepsilon(x_1,x_3) = 6.9 \text{ percent}$$

$$\varepsilon(x_3) = 15 \text{ percent} \qquad \varepsilon(x_2,x_3) = 5.875 \text{ percent}.$$

From these results it is observed that, although all pairs of features are better than the best single feature, the pair consisting of the two worst single features is much better than the pair consisting of the two best single features. Furthermore, the best pair does not contain the best single feature; in fact, the best pair is made up of the worst single features.

Godfried T. Toussaint
Dep. Elec. Eng.
Univ. British Columbia
Vancouver, B.C., Canada

Manuscript received August 7, 1970; revised January 27, 1971. This research was supported in part by the National Research Council of Canada under Grant NRC A-3308 and by the Defence Research Board of Canada under Grant DRB 2801-30.

[1] J. D. Elashoff, R. M. Elashoff, and G. E. Goldman, "On the choice of variables in classification problems with dichotomous variables," *Biometrika*, vol. 54, 1967, pp. 668–670.

Comments on "A Modified Figure of Merit for Feature Selection in Pattern Recognition"

In a recent correspondence [1] a modification of the conventional mutual-information effectiveness criterion for feature selection in pattern recognition was described. However, there seems to be some confusion between selecting a subset of features and selecting features individually. This apparent confusion may confuse the reader further

Manuscript received August 15, 1970; revised January 27, 1971.

about the independence assumptions and about some of the conclusions concerning the merit measure.

The mutual-information criterion for a subset of m features, using the notation of [1], is given by

$$I(C \mid \bar{X}) = \sum_{i=1}^{n} \sum_{j=1}^{q^m} P(\bar{X} = j \mid C = i)P(C = i) \cdot \log_2 \frac{P(\bar{X} = j \mid C = i)}{P(\bar{X} = j)}, \quad (1)$$

where $\bar{X}$ has m components and each component can take on q values. On the other hand, the mutual-information criterion for any one feature X^k, where $k = 1,2,\cdots,m$, is given by

$$I(C \mid X^k) = \sum_{i=1}^{n} \sum_{j=1}^{q} P(X^k = j \mid C = i)P(C = i) \cdot \log_2 \frac{P(X^k = j \mid C = i)}{P(X^k = j)}. \quad (2)$$

Equation (2) is equivalent to the expression for the mutual-information criterion given by (12) in [1], and the merit measure is based on this equation.

In [1], the authors state that the independence assumption is implied in the following relationships:

$$P(\bar{X}) = \prod_{k=1}^{m} P(X^k), \quad (3)$$

$$P(\bar{X} \mid C) = \prod_{k=1}^{m} P(X^k \mid C). \quad (4)$$

It should be noted that in pattern recognition, the assumption of independent features usually implies only (4). In fact, the assumption of (3) results in an entirely different situation from that resulting when (4) is assumed. Furthermore, Barabash [2] showed that the conditions (3) and (4) cannot, in general, be fulfilled simultaneously.

The mutual information of (1) is defined as follows:

$$I(C \mid \bar{X}) = H(C) - H(C \mid \bar{X}). \quad (5)$$

Since the measure is symmetric in its two arguments [3], (5) may be written as

$$I(C \mid \bar{X}) = H(\bar{X}) - H(\bar{X} \mid C). \quad (6)$$

It is clear from (6) that in order to obtain a maximum amount of information, features should be selected so as to maximize $H(\bar{X})$ and minimize $H(\bar{X} \mid C)$. It is well known that

$$H(\bar{X}) \leq \sum_{k=1}^{m} H(X^k)$$

with equality holding when (3) is satisfied, and

$$H(\bar{X} \mid C) \leq \sum_{k=1}^{m} H(X^k \mid C)$$

with equality holding when (4) is satisfied. Therefore, in order to maximize (6) or (1), features should be selected so as to satisfy (3) and violate (4). However, in practice it is difficult to find situations in which distributions satisfy (3) and violate (4). The much more reasonable and common method of approach is to disregard (3) altogether and to assume (4). Such distributions are more easily found in practice. Furthermore, when

$$P(X^k = i \mid C = j) \neq P(X^k = i \mid C = l)$$

for $j, l = 1,\cdots,n, j \neq l$ $i = 1,\cdots,q$, and $k = 1,\cdots,m$, then assumption (3) implies that (4) is no longer valid, and vice versa, in which case both (3) and (4) cannot be assumed simultaneously. Since this result is considered in detail in [2], a proof will not be presented here.

In the section on statistical relationships, the authors state, "In all phases of this correspondence the independence assumption is upheld and only one feature is considered at a time." This statement is confusing because it seems to imply that features can be considered one at a time because the independence assumption, implied by (3) and (4), is upheld, which is incorrect, as will be shown below.

When a feature subset of m features is selected with a criterion such as the entropy given by

$$H(\bar{X}) = -\sum_{j=1}^{q^m} P(\bar{X} = j) \log_2 P(\bar{X} = j), \quad (7)$$

(3) is usually assumed to reduce the computation. When (3) is assumed, then

$$H(\bar{X}) = \sum_{k=1}^{m} H(X^k), \quad (8)$$

and the features can be evaluated one at a time using $H(X^k)$. Similarly, when a feature subset of m features is selected with a criterion such as the average entropy given by

$$H(\bar{X} \mid C) = -\sum_{i=1}^{n} P(C = i) \sum_{j=1}^{q^m} P(\bar{X} = j \mid C = i) \cdot \log_2 P(\bar{X} = j \mid C = i), \quad (9)$$

(4) is usually assumed for the same reasons. When (4) is assumed, then

$$H(\bar{X} \mid C) = \sum_{k=1}^{m} H(X^k \mid C), \quad (10)$$

and the features can be evaluated one at a time using $H(X^k \mid C)$. In both of the above examples, the entropy criteria are additive under the corresponding independence assumptions. It would be desirable, when using the mutual-information criterion of (1) to select a subset of m features, to be able to select the features individually using (2) as was done in [1]. This could be done if the following condition were upheld:

$$I(C \mid \bar{X}) = \sum_{k=1}^{m} I(C \mid X^k). \quad (11)$$

Unfortunately, the mutual-information criterion of (1) cannot be made additive under the independence assumptions for the following reasons. If, on the one hand, either (3) only, or (4) only, is assumed, then it can easily be shown that (1) does not decompose, i.e., (11) does not hold, although (1) is simplified by either of the two assumptions. If, on the other hand, both (3) and (4) are assumed, then it can easily be shown that (1) does decompose and (11) is valid, but (3) and (4) are mutually contradictory and cannot be assumed simultaneously. Therefore, if one is going to select a *subset* of features using the mutual-information criterion, one should use (1), and no independence assumptions can correctly lead to using (2). On the other hand, if one wants to decide arbitrarily to select *individual features*, to make up the desired subset, by using the mutual-information criterion, then one should use (2), in which case the independence assumptions (3) and (4) are irrelevant.

Had the authors of [1] discussed selection of features individually rather than the selection of a subset of m features, and had they left out any discussion of independence assumptions, their correspondence would have been clear because then the mutual-information criterion is given by (12) in [1]. The confusion arises because the authors discuss the selection of *subsets* of features, for which the mutual-information criterion is given by (1), and then end up using the merit measure that is based on (2), which follows from (1) only under the mutually contradictory assumptions (3) and (4).

The conclusion that the merit measure does better than the mutual-information criterion may also confuse some readers. The authors state that *subsets of features* were consistently better when selected using the merit measure than when using (2). Although these results are of interest, (2) is not the mutual-information criterion for subsets of features. It would be of interest to find out if *individual features* were consistently better when selected using the merit measure than when using (2). Furthermore, subsets of features selected with the mutual-information criterion of (1) may be consistently better than subsets selected with the merit measure.

Finally, it should be stated that these comments do not negate the validity of the merit measure. Rather, they may help to explain why the merit measures does better than (2) when selecting subsets of

features. However, in order to appreciate the merit measure more fully, experiments would have to be performed using (1) and a comparison of (1) with the merit measure with respect to computational complexity would have to be made.

GODFRIED T. TOUSSAINT
Dep. Elec. Eng.
Univ. British Columbia
Vancouver, B.C., Canada

REFERENCES

[1] J. E. Paul, Jr., A. J. Goetze, and J. B. O'Neal, Jr., "A modified figure of merit for feature selection in pattern recognition," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-16, July 1970, pp. 504–507.
[2] Y. L. Barabash, "On properties of symbol recognition," *Eng. Cybern.* (USSR), Sept./Oct. 1965, pp. 71–77.
[3] H. F. Ryan, "The information content measure as a performance criterion for feature selection," in *IEEE Proc. 7th Symp. Adaptive Processes,* Los Angeles, Calif., Dec. 16–18, 1968.

Unsupervised Estimation of Signals With Intersymbol Interference

***Abstract*—A decision-directed unsupervised estimation algorithm is used in a receiver that processes signals with unknown intersymbol interference. The estimator utilizes the statistical structure of this dependent sample problem to reduce the computational complexity. Asymptotic and experimental dynamic probability of error results are presented.**

I. INTRODUCTION

In seeking to maximize the flow of information through a digital data communications system with a band-limited channel, it is necessary to operate at rates where energy from one transmitted signal baud is smeared into following bauds. This intersymbol interference between L adjacent bauds can greatly reduce the performance of the communications system unless compensated for at the receiver. Approaches that rely on linear methods such as tapped delay lines or shift registers have been in use for some time [1]–[5]; however, approaches that seek to minimize error without the linearity constraint [6]–[9] can offer a significant improvement in performance. Simulation results in [9] give a practical channel example where the optimum nonlinear sequential detector with sufficient "look ahead" at future samples has a considerably lower probability of error than a transversal equalizer.

In many practical applications the effect of the channel on the transmitted signals is unknown and time varying. For a particular channel model, the channel-dependent parameters in the receiver's decision procedure can be estimated either by using isolated pulses as sounding signals (e.g., see [2]) or by using the unknown signals received during actual data transmission. In the latter case, since the received signals are of unknown classification, such an estimation approach is called *unsupervised estimation.*

This correspondence investigates the application of a suboptimum unsupervised nonlinear receiver to intersymbol interference in linear channels that are slowly varying relative to the convergence rate of the estimation algorithm. The emphasis of the approach is on a simply structured receiver that can follow fairly rapid changes in channel parameters such as occur in some fading communications channels. Dynamic performance simulations are presented in Section V showing examples where the convergence rate of the receiver to the unknown channel parameters is almost as fast as the convergence rate of a sample mean. This is in contrast with earlier parameter adjustment algorithms based on gradient techniques [2]–[5], [9], which are much slower.

Manuscript received November 12, 1969; revised February 4, 1971. This work was supported by the Avionics Laboratory, Wright-Patterson AFB, Ohio, under Contract F 33 615-68-C-1577 and was presented at the 1970 International Symposium on Information Theory, Noordwijk, the Netherlands.

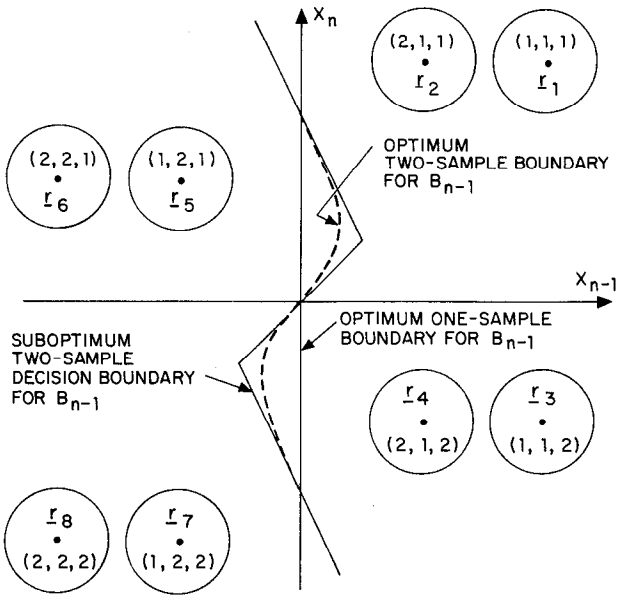

Fig. 1. Decision boundaries and cluster structure for [$v = L = 2$].

II. DIGITAL COMMUNICATIONS SYSTEM MODEL

A block diagram of the M-ary digital-communications-system sampled low-pass equivalent model is given in Fig. 1. One of M symbols $\{b_j\}_{j=0}^{M-1}$. is transmitted during the kth time slot producing B_k. For amplitude modulation the $\{b_j\}$ are real numbers, while for phase-shift keying $\{b_j = \exp(i2\pi j/M)\}_{j=0}^{M-1}$. The symbol sequence is assumed statistically independent. The channel is characterized by a finite-memory linear channel transformation producing R_n, followed by zero-mean additive white Gaussian noise N_n [11]. Hence the received samples X_n can be written

$$X_n = R_n + N_n = \begin{cases} \sum_{k=1}^{n} a_{n-k+1}B_k + N_n, & n < L \\ \sum_{k=n-L+1}^{n} a_{n-k+1}B_k + N_n, & n \geq L, \end{cases} \quad (1)$$

where $\{a\}_{k=1}^{L}$ is the sampled low-pass equivalent channel impulse response.

III. STATISTICAL STRUCTURE OF MULTIBAUD SAMPLE SPACE

For $L > 1$ the sample sequence $\{X_k\}_{k=1}^{n}$ is not statistically independent even though the original symbol sequence and the additive noise were assumed to have this property. Approaches to including the effect of the dependent samples in the decision rule for B_k range from the optimum-compound [6] and sequential-compound [9] decision procedures to suboptimum procedures such as the linear one used by transversal equalizer/thresholders and the nonlinear decision equation that will be used here. Some insight into the statistical structure of this dependent sample problem can be obtained by examining the joint sample space over the last v bauds.

Define the v-dimensional sequence vectors

$$X_n = \begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{n-v+1} \end{pmatrix} \qquad R_n = \begin{pmatrix} R_n \\ R_{n-1} \\ \vdots \\ R_{n-v+1} \end{pmatrix}. \quad (2)$$