NEAREST NEIGHBOR RULE CLASSIFICATION
OF TIME SERIES IN EXPLORATORY POPULATION
SCREENING PROBLEMS

BY

WILL GERSCH

TECHNICAL REPORT NO. 384
DECEMBER 9, 1986

PREPARED UNDER CONTRACTS
N00014–86–K–0156 and N00014–83–K–0238
(NR–042–267)
FOR THE OFFICE OF NAVAL RESEARCH

Reproduction in Whole or in Part is Permitted
for any purpose of the United States Government

Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
NEAREST NEIGHBOR RULE CLASSIFICATION
OF TIME SERIES IN EXPLORATORY POPULATION
SCREENING PROBLEMS

BY

WILL GERSCH

TECHNICAL REPORT NO. 384
DECEMBER 9, 1986

Prepared Under Contracts
N00014-86-K-0156 and N00014-83-K-0238
(NR-042-267)
For the Office of Naval Research
Herbert Solomon, Project Director

Reproduction in Whole or in Part is Permitted
for any purpose of the United States Government

Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
NEAREST NEIGHBOR RULE CLASSIFICATION OF TIME SERIES  
IN EXPLORATORY POPULATION SCREENING PROBLEMS  
By  
Will Gersch  

1. **Introduction, Background and Motivation.**  

A general time series classification problem is considered in the setting of an exploratory population screening problem. In that setting, a modest number of categorically labeled time series from different individuals is assumed to be available. Typically, the data is characterized by broad inter-subject time series variability within each category. The scientific question of interest is to ascertain whether or not there is sufficient information, in that particular variety of time series observations, to reliably distinguish between individuals in the alternative population categories.  

We consider a nearest neighbor (NN) time series classification rule. A measure of dissimilarity between the new to-be-classified time series and each member of the set of labeled sample time series is computed. The new time series is classified with the label of the particular labeled sample time series which is least dissimilar. The measure of dissimilarity between the new time series and each of the labeled sample time series is computed as an estimate of the Kullback Leibler number between the time series as if the time series were Gaussian distributed.  

This dissimilarity measure is demonstrated to have sufficient metric properties for the statistical properties of NN rule classification to apply. Thus, under reasonable assumptions about the distribution of the categorical population time series, the probability of misclassification can be reliably assessed with only a relatively small number of labeled sample time series by employing a leave out one at-a-time cross validation of the labeled sample time series data set.  

In Section 2, ANALYSIS, the nearest neighbor Kullback Leibler type dissimilarity measure classification rule of time series is described. The metric properties of the dissimilarity measure are developed. The discussion is specialized to stationary time series. Time and frequency domain formulas for the Kullback Leibler measure between stationary multivariate Gaussian time series are developed. The time domain formulas yield a practical realization of the dissimilarity measure computations. An example of the classification of the level of anesthesia of humans in surgery by the analysis of stationary time series electroencephalograms (EEGs) by our NN-KL number dissimilarity rule is worked in Section 3. Additional considerations including the question of the arbitrariness of EEG time series normalization and a data thinning procedure for alleviating the large data storage and large computing burden, known to be associated with NN rule classification,
also appear in Section 3. A summary and discussion including comparison of the KL-NN time series classification rule and maximum likelihood Akaike AIC and Shore-Johnson entropy criterion classification rules, additional comments on data normalizations and another KL-NN time series classification rule applications appear in Section 4.

2. Analysis.

2.1 The Nearest Neighbor-Kullback Leibler Type Dissimilarity Measure Classification Rule.

Consider the labeled time series

\[
\begin{pmatrix}
    z^{(1)} \\
    \theta^{(1)}
\end{pmatrix}, \ldots, \begin{pmatrix}
    z^{(N)} \\
    \theta^{(N)}
\end{pmatrix}
\]

\[
x^{(m)} = (x^{(m)}(1), \ldots, x^{(m)}(T)), m = 1, \ldots, N
\]

\[
\theta^{(m)} \in \{1, \ldots, N\}
\]

and a new to-be-classified time series \(z^{(0)} = (z^{(0)}(1), \ldots, z^{(0)}(T))\). \(x^{(m)}\) and \(z^{(0)}\) denote \(T\)-duration \(d\)-variate time series and \(\theta^{(m)}\) denotes the label or category of the \(m\)th time series for \(m = 1, \ldots, N\).

Let \(\delta(z^{(0)}, x^{(m)})\), \(m = 1, \ldots, N\) be a measure of dissimilarity between the new to-be-classified time series \(z^{(0)}\) and the labeled time series \(x^{(m)}\), for \(m = 1, \ldots, N\). The NN classification rule is

\[
\text{If } \delta(z^{(0)}, x^{(m)}) \leq \delta(z^{(0)}, x^{(m')}), m = 1, \ldots, N
\]

Then \(\theta^{(0)} = \theta^{(m')}\).

That is, the label of the new series \(z^{(0)}\) is the same as the label of its nearest dissimilarity measure time series.

The specification of a dissimilarity measure between time series is critical. We employ an estimate of the Kullback Leibler number between time series that is computed as if the time series were Gaussian distributed. Let \(X_0\) and \(X_m\) be two \(d\)-vector random variables with the probability density functions \(f_0\) and \(f_m\) respectively. Then, the \(I\)-divergence between \(f_0\) and \(f_m\) is, (Kullback, 1959):

\[
(2.1.1) \quad I(f_0, f_m) = \int_{-\infty}^{\infty} f_0(x) \log \frac{f_0(x)}{f_m(x)} dx.
\]

In particular, let \(X_0 \sim N(\mu_0, \Sigma_0)\) and \(X_m \sim N(\mu_m, \Sigma_m)\). That is let \(X_0\) and \(X_m\) each be normally distributed with \(d\)-component mean vectors \(\mu_0, \mu_m\) and \(d \times d\) covariance matrices \(\Sigma_0, \Sigma_m\) respectively. In that case, (Kullback, 1959),

\[
(2.1.2) \quad 2 I(f_0, f_m) = \log \frac{\Sigma_m}{\Sigma_0} + tr\Sigma_m^{-1}\Sigma_0 + tr\Sigma_m^{-1}(\mu_0 - \mu_m)(\mu_0 - \mu_m)' - d.
\]
In (2.1.2) and subsequently, the notation $|A|, \text{tr}(A), A^{-1}A'$ denotes respectively the determinant, trace, inverse and transpose of the matrix $A$.

Consider the $d$-variate $T$-duration labeled sample time series $x^{(m)}, m = 1, \ldots, N$ and the new time series $x^{(0)}$. Let $\hat{\mu}_j = (\hat{\mu}_j(1), \ldots, \hat{\mu}_j(T))'$; $\hat{\Sigma}_j, j = 0$ or $m$ be the sample mean and sample covariance matrices respectively of $x^{(j)}$ or estimates thereof computed from an ensemble of realizations or from a single realization of the $x^{(j)}$ process. Then, let

\begin{equation}
2\delta(x^{(0)}, x^{(m)}) = \log \frac{\left| \hat{\Sigma}_m \right|}{\left| \hat{\Sigma}_0 \right|} + \text{tr} \hat{\Sigma}_m^{-1}\hat{\Sigma}_0 + \text{tr} \hat{\Sigma}_m^{-1}(\hat{\mu}_0 - \hat{\mu}_m)(\hat{\mu}_0 - \hat{\mu}_m)' - dT.
\end{equation}

(2.1.3)

denote a measure of the dissimilarity computed between the sample time series $x^{(0)}$ and $x^{(m)}$. That is, the dissimilarity measure $\delta(x^{(0)}, x^{(m)})$ in (2.1.3) is computed from the sample time series to mimic (2.1.2), as if the time series were Gaussian distributed.

**Comment.** The $I$-divergence of Kullback Leibler information number (also the information for discrimination, information gain or entropy of $f_0$ relative to $f_m$) has a basic role in the information theoretic approach to statistics, and in statistical physics as the maximization of entropy (Kullback, 1959; Good, 1963; Jaynes, 1957; Akaike, 1977). The $I$-divergence does not satisfy the triangle inequality and is not a metric. Certain analogies do exist between the properties of probability density functions and Euclidean geometry, wherein $I$-divergence plays the role of squared Euclidean distance, (Csiszar, 1975).

### 2.2 The Metric Properties of $\delta(x^{(0)}, x^{(m)})$.

A desirable property of the dissimilarity measure in (2.1.3) is that it have sufficient metric properties so that the statistical properties of nearest neighbor rule classification apply. (See Cover and Hart, 1966.) The necessary properties are that i) $\delta(x^{(0)}, x^{(0)}) = 0$, ii) $\delta(x^{(0)}, x^{(m)}) > 0$ for any $x^{(m)} \neq x^{(0)}$, and iii) the minimum value of the dissimilarity measure $\delta(x^{(0)}, x^{(m)}) \to 0$, as $N$ the number of labeled samples increases indefinitely. Property i) is immediate from (2.1.3), Property ii) is a general property of the KL number, (Kullback, 1959), Property iii): Let the sample $Td \times Td$ covariance matrix $\hat{\Sigma}$ be distributed in accordance with distribution $F$. $F$ is assumed to be absolutely continuous. Let $\hat{\Sigma}_0, \hat{\Sigma}_1, \hat{\Sigma}_2, \ldots$ be I.I.D. random variables from that distribution. Then the space $R^{Td\times Td}$, on which the sample covariances are defined, is a separable metric space and the minimum Euclidean distance between the sample covariances goes to zero as $N$, the number of labeled samples from $F$, goes to infinity. That is with $m'$ the index of $\hat{\Sigma}_m$ that is the NN of $\hat{\Sigma}_0, \| \hat{\Sigma}_0 - \hat{\Sigma}_m \| \to 0$ with increasing $N$. Since $\delta(\hat{\Sigma}_0, \hat{\Sigma}_m)$ is a continuous function of $\hat{\Sigma}_m$, $\delta(\hat{\Sigma}_0, \hat{\Sigma}_m) \to 0$, (Royden, 1968).
Comment: Under reasonable regularity conditions, the finite labeled sample cross-validation leave out one at-a-time cross-validation of the labeled sample data set has $O(1/N)$ properties for the estimation of the probability of misclassification, (Rogers and Wagner, 1978). That property permits the implicit conjecture in the exploratory population screening problem investigation, that there is sufficient evidence in the measurement data to achieve statistically satisfactory discrimination, to be tested with only a moderate number of labeled sample time series.

2.3 Kullback-Leibler Number Formulas for Multivariate Stationary Gaussian Time Series.

The formula for the dissimilarity between the $T$-duration $d$-variable sample stationary time series $x^{(0)}$ and $x^{(m)}$ in (2.1.3) indicated operations on $Td \times Td$ matrices. For the $d$ and $T$ that appear in some data analysis situations, that computation is forbidding. Time domain and frequency domain formulas for the KL numbers between stationary Gaussian time series are expressed in parametric model forms. In particular, KL number type dissimilarity measures can be computed from autoregressive (AR) models of the time series in operations on $d \times d$ matrices.

2.3.1 Time Series Representations. Let $\{x^{(0)}(t)\}$ and $\{x^{(m)}(t)\}$ denote $d$-variate zero-mean stationary ergodic Gaussian time series with corresponding probability density functions $f^{(0)}, f^{(m)}$ and $d \times d$ matrix covariance functions $\Gamma^{(0)}(k), \Gamma^{(m)}(k)$ and power spectral density matrices $S_0(f)$ and $S_m(f)$ respectively. Identify the time series $x^{(i)}(t)$ parametrically in terms of the Wold (moving average) and the AR representations, (Whittle, 1963).

\[
x^{(i)}(t) = h^{(i)}(t) * \varepsilon^{(i)}(t), \quad i = 0, 1, 2, \ldots, N
\]

\[
A^{(i)}(t) * x^{(i)}(t) = \varepsilon^{(i)}(t); \\
E\{\varepsilon^{(i)}(t)\} = 0; E\{\varepsilon^{(i)}(t + k)\varepsilon^{(i)}(t)\} = V_i \delta_{k,0}.
\]

In (2.3.1), the symbol $*$ denotes the convolution operations, $E$ is the expectation operator and $h^{(i)}(t)$ and $A^{(i)}(t)$ are respectively the $d \times d$ impulse response matrix and AR matrix coefficients. Denote the action of the AR operator defined by $A^{(m)}(t)$ on $x^{(0)}(t)$ by

\[
A^{(m)}(t) * x^{(0)}(t) = e^{(0,m)}(t).
\]

In (2.3.2), $e^{(0,m)}(t)$ has an interpretation as a zero-mean "residual" time series in the conventional sense of a regression analysis. Its zero-lag covariance matrix is

\[
E\{e^{(0,m)}(t)e^{(0,m)}(t)\} = V_0^m.
\]
Employing the notation of (2.3.2) in (2.3.1)

\[ A^{(m)}(t) \ast h^{(0)}(t) \ast e^{(0)}(t) = e^{(0,m)}(t) \]

\[ h^{(0,m)}(t) \ast e^{(0)}(t) = e^{(0,m)}(t). \]

(2.3.4)

In (2.3.4), \( h^{(0,m)}(t) \) designates the impulse response of the cascade of filters \( A^{(m)}(t) \) and \( h^{(0)}(t) \). Using elementary linear operations,

\[ h^{(0,m)}(t) = h^{(0)}(t) + \sum_{i=1}^{\infty} A^{(m)}(i) h^{(0)}(t-1), t = 0, 1 \ldots . \]

(2.3.5)

2.3.2 Kullback-Leibler Number Formulas. Time and frequency domain formulas for the Kullback-Leibler number between those Gaussian time series are

(2.3.6a) \( 2I(f^{(0)}, f^{(m)}) = \log \frac{|V_m|}{|V_0|} + \text{tr} \left[ \sum_{i=0}^{\infty} h^{(0,m)}(t) V_0 h^{(0,m)}(t) V_m^{-1} \right] - d \)

(2.3.6b) \[ = \log \frac{|V_m|}{|V_0|} + \text{tr} \left[ \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A^{(m)}(i) \Gamma^{(0)}(j-1) A^{(m)}(j) V_m^{-1} \right] - d \]

(2.3.6c) \[ = \log \frac{|V_m|}{|V_0|} + \text{tr} \left[ \int_{-1/2}^{1/2} S_0(f) S_m(f)^{-1} \right] - d. \]

Proof. Equation (2.1.2) can be expressed as

(2.3.7) \[ 2I(f^{(0)}, f^{(m)}) = \log \frac{|V_m|}{|V_0|} + \text{tr}[V_m^0 V_m^{-1}] - d. \]

The first two parametric time-domain formulas for the Kullback-Leibler numbers between Gaussian time series in (2.3.6) may be obtained from (2.3.7) by replacing \( V_m^0 \) by its definition, (2.3.3), and then substituting \( e^{(0,m)} \) by its representations in (2.3.4) and (2.3.2) and taking the indicated expectations.

The frequency domain formula for the Kullback-Leibler number between stationary time series with Gaussian distributions (2.3.6) is obtained by operator manipulations and Parseval’s theorem. For example, consider only the expression \( \text{tr}[V_m^0 V_m^{-1}] \) in (2.3.7). From (2.3.3) and (2.3.4)

\[ E\{\text{tr}[e^{(0,m)}(t)e^{(0,m)}(t)' V_m^{-1}]\} \]

(2.3.8) \[ = \text{tr}[h^{(0)}(t) \ast V_0 \ast h^{(0)}(t)' \ast A^{(m)}(t) \ast V_m^{-1} \ast A^{(m)}(t)'] \]

\[ = \text{tr} \int_{-1/2}^{1/2} S_0(f) S_m^{-1}(f) df. \]
2.3.3 Computations of $\delta(x^{(0)}, x^{(m)})$ Between Stationary Time Series.

The formula for the dissimilarity between the $T$-duration $d$ variable sample stationary time series $x^{(0)}$ and $x^{(m)}$ in (2.1.3) indicates operations on $Td \times Td$ matrices. The Kullback-Leibler number formula (2.3.6b) expressed in terms of an AR model representation involves operations on $d \times d$ matrices. Therefore, as an alternative to (2.1.3) consider mimicking the second time domain formula in (2.3.6b) by

$$
2\delta(x^{(0)}, x^{(m)}) = \log \frac{|\hat{V}_m|}{|\hat{V}_0|} + \text{tr}
\left( \sum_{i=0}^{p_m} \sum_{j=1}^{p_m} \hat{A}^{(m)}(i) \hat{\Gamma}^{(0)}(j-1) \hat{A}^{(m)'}(j) \hat{V}_m^{-1} \right) - d.
$$

Equation (2.3.9) only involves operation on $d \times d$ matrices. The finite duration multivariate time series $x^{(0)}$ and $x^{(m)}$ are modeled by finite order AR models. In (2.3.9), the hatted quantities are estimates of the corresponding theoretical quantities in (2.3.6b) and $p_m$ is the order of the AR modeled time series $x^{(m)}$. Here, we indicate a Whittle recursive AR model computation-Akaike AIC criterion order selection procedure for fitting the multivariate AR models to data, (Whittle, 1963, Akaike, 1976, Gersch and Yonemoto, 1977). Let $x^{(0)}(t); t = 1, \ldots, T$ denote the observed $d$-variate time series. For $j = 0$ or $m$ let

$$
\hat{x}^{(j)} = \frac{1}{T} \sum_{t=1}^{T} x^{(j)}(t),
$$

$$
\hat{\Gamma}^{(j)}(k) = \frac{1}{T} \sum_{t=1}^{T-k} [x^{(j)}(t + k) - \hat{x}^{(j)}][x^{(j)}(t) - \hat{x}^{(j)}]',
$$

$k = 0, 1, \ldots$

denote the sample mean and sample covariance of the $j$-th time series for $j = 0$ or $m$. Then, the Whittle-Akaike procedure AR model of order $p_j$ fitted to $x^{(j)}$, the $j$-th time series, satisfies

$$
\sum_{i=0}^{p_j} \hat{A}^{(j)}(i)x^{(j)}(t - I) = \epsilon^{(j)}(t); A^{(j)}(0) = I_{d \times d}
$$

$$
E[\epsilon^{(j)}(t)] = 0, E[\epsilon^{(j)}(t + k)\epsilon^{(j)}(t)'] = \hat{V}_j \delta_{k,0}
$$

$$
\sum_{i=0}^{p_m} \hat{A}^{(j)}(i)\hat{\Gamma}^{(j)}(k - i) = \hat{V}_j \delta_{n,0} \text{ for } k = 0, 1, \ldots, p_m.
$$

Comments (1). The AR models of the labeled time series can be interpreted as templates of those time series. Correspondingly, NN rule classification of time series has a template matching interpretation. The new series is compared against the templates of the labeled sample time series. The most similar template is the one for which the dissimilarity measure is smallest.

(2) Time series classification is extensively practiced in non-exploratory scoring problem situations in speech processing and in EEG classification, (References in Gray et al., 1980 and Gevins,
In speech classification, a frequent practice is to model each of the labeled sample and new time series by pre-assigned order AR models and to classify using a NN-Euclidean distance measure between the AR model coefficients. Those classification procedures are feature analysis-discriminant analysis rules. The classification of EEG's using spectrum analysis domain features, such as energy in particular frequency bands and spectral coherence between channels in those frequency bands are also feature analysis-discriminant analysis procedures. The poignant remark "...the problem for which that solution is optimum, is not known." (Cover, 1973) is applicable here. The point is that in exploratory population screening problems, care must be exercised in computing dissimilarity between time series. Otherwise the near optimum probability of misclassification property of true NN classification rules on finite labeled sample data will not be preserved. (Rogers and Wagner, 1978).

(3) A frequency domain formula for the Kullback Leibler number between Gaussian distributed time series was probably first due to Pinsker, (1964). Subsequently frequency domain formulas were developed by Shumway and Unger, (1974), and Hawkes and Moore, (1976) and B.D.O. Anderson et al., (1978). The first time domain formula for the Kullback Leibler number between scalar time series is probably due to Itakura and Saito, (1968). Akaike, (1976) shows a different state space model development for a version of the second time domain formula in (2.3.6).

3. An Anesthesia Level Classification By EEG Analysis

Population Screening Problem Example.

An exploratory EEG time series data-population screening classification problem is treated by the nearest neighbor rule approach. The category or state of an individual is to be classified by comparison of his or her EEG with EEGs taken from other individuals. The automatic classification of anesthesia levels L1 and L3, respectively the anesthesia levels insufficient for and sufficient for deep surgery by machine computations on the EEG alone is considered. Extension of the nearest neighbor rule approach to distinguish between more than two categories or anesthesia levels does not involve any new concepts.

The anesthesia level EEG data originated in an experiment at Vancouver General Hospital, 280 epochs of visually screened halothane-nitrous oxide anesthesia level labeled EEGs were collected from twenty different individuals in surgery. The non-EEG criteria determined anesthesia levels were classified by a single anesthesiologist to eliminate the problem of inter-EEG-rater variability. Details of the experimental surgical anesthesia situation and a review of the status of automatic classification of anesthesia levels using EEG data appear elsewhere (McEwen, 1975 a, b and Gersch
et al., 1979). The data consisted of 64 second recordings of four channel EEG epoch data, (F4-C4, F3-C3, C4-02, and C3-01 in the 10-20 EEG system), analogue-FM recorded through a 0.54 to 30 Hz. bandpass filter and subsequently digitally transcribed at the rate of 128 samples/second. An examination of the available data suggested that we confine our attention to a two category classification problem to classify the anesthesia levels L1 and L3 respectively, the anesthesia levels that are insufficient and just sufficient for deep surgery. The data selected for analysis was the 73 EEG epochs comprised of all the 35-L1 EEG epochs available and 38-L3 EEG data epochs (in sets of 2-3 epochs per individual) from a total of 18 different individuals. The analysis was performed on the first twenty second intervals of each EEG data epoch at a reduced data rate of 128/3 samples per second on $d = 4$ EEG data channel and $d = 2$ EEG data channel (C4-02 and C3-01) data. This constitutes the labeled sample data base.

The implicit conjecture in the EEG population screening problem is that there is sufficient information in the EEG alone to achieve clinically acceptable levels of discrimination between categorical EEG states. The credibility of this conjecture is strained by evidence of the broad inter-subject categorical EEG variability. Figure 1, 2-channel twenty second anesthesia level L1 and L3 EEG epochs from five different subjects suggest that the EEG of an individual does differ in the L1 and L3 anesthesia level states and also illustrate broad inter-subject EEG variability. The L1 EEGs appear to be relatively homogeneous “fast” EEGs whereas the L3 EEGs include fast, slow, regular and irregular EEGs. No obvious visual properties of the EEGs distinguish the L1 and L3 EEGs from each other.

To achieve a baseline appraisal of the achievable discriminability between the L1 and L3 anesthesia level EEG sample populations, the EEG epochs of a single individual at a time were deleted from the 18 individual-73 epoch labeled sample EEG data. Each of the deleted-individual’s EEG epochs was classified against the remaining 17 individual labeled EEG sample population using the KI-NN and KL-kNN rules. (kNN rules classify $x$ by assigning it the label most frequently represented among the $k$ nearest samples, Duda and Hart, (1973)). The results obtained are shown in Table 1.

The entries in the table indicate the number of classification errors and the percentage of correct classification for the best $d = 2$ EEG channel and $d = 4$ EEG channel KL-kNN classification performance. The best classification results for the $d = 2$ and $d = 4$ EEG data channels was 85% and 89% overall correct classification respectively.

The objectives of this exploratory population screening anesthesia level classification by EEG analysis study appear to have been met. With only a moderate sized label sample data base, the
TABLE 1: DELETE ONE-SUBJECT-AT-A-TIME, KL-NN RULES RESULTS

\begin{tabular}{|c|c|c|c|}
\hline
 & \textbf{KL - 3NN; }d = 2 & & \textbf{KLNN; }d = 4 \\
\hline
\textbf{Errors, %} & \textbf{Correct} & & \\
\textbf{Labeled} & \textbf{L1} & \textbf{L3} & \\
\textbf{EEGS} & & & \\
\hline
\text{L1} & 1 & - & \\
35 epochs & 97\% & - & \\
\hline
\text{L3} & - & 10 & \\
38 epochs & - & 74\% & \\
\hline
\end{tabular}

results obtained suggest that the population screening anesthesia level classification by EEG analysis scenario has substantial possibilities for clinical applications. The classification performance obtained using the NN-KL number rule yielded smaller probability of misclassification than obtained earlier using the same data by spectral domain-feature analysis discrimination (McEwen, 1975a). These results are also among the best yet obtained in time series classification, Gevins, (1980).

\textbf{Comments (1).} Classification of time series is subject to arbitrary conventions and normalizations. The alternative normalizations of the EEG that are possible in KL-NN rule classification are explicit in the dissimilarity measure formula (2.3.9). The alternative normalizations influence the relative dominance of the first and second terms in that equation. Three rational EEG normalizations are 1) Fix the EEG amplitude level to correspond to some convenient EEG machine level of amplification. 2) Normalize each EEG to have the same mean square power. For example, the zero-lag covariance matrices \( \Gamma^{(m)}(0) \) can be made to be identical for each of the labeled sample EEGs and the new EEG using linear transformations. In that case, discrimination is based upon the differences in the structure of the sample correlation matrices of the EEG time series and the residual variance or average power that is not explained by the AR models of the EEGs. 3) Normalize the EEG to have the same “generator” or average source power. That is, by linear transformation make \( |V_m| \) identical for each of the labeled sample EEGs and the new EEG. The first term in (2.3.9) is thus reduced to zero and discrimination between EEGs is based upon the differences in the normalized sample covariance matrices of the EEGs.

Some simple alternative EEG normalization computational experiments were conducted. The results are not included here. We observed that KL-NN classification performance was not particularly vulnerable to the alternative EEG normalizations. The computations reported in this paper
are based upon the first normalizing alternative.

(2) Implementation of the KL-NN classification rule requires that each of the labeled sample EEGs time series be stored and that a KL number dissimilarity measure be computed between the new EEG and each labeled sample EEG. In some circumstances such data storage and/or computational requirements may be so excessive as to render the KL-NN rules technically or economically unfeasible. These burdens might be relieved by a data thinning procedure in which groups or clusters of EEG time series in each labeled state could be replaced by a prototypic element EEG time series. Classification would be achieved by KL-NN rules based upon the dissimilarity between the new EEG and each of the prototypic EEG members.

An agglomerative data thinning procedure was designed to have particular statistical properties and also achieve data storage and computational burden reductions. The properties of this procedure are: 1) The procedure converges on all finite data sets. 2) Each of the labeled sample EEG time series is required to be correctly classified by KL-NN classification against the EEG prototypes. This procedure achieved a simultaneous data thinning cluster analysis of each labeled sample categorical EEG class and retained the classification performance of the entire labeled sample data base.

4. Summary and Discussion.

A nearest neighbor time series classification rule was introduced. The dissimilarity between a new to-be-classified time series and each of the members of the labeled sample time series is computed as an estimate of the Kullback Leibler number as if the time series were Gaussian distributed. This NN-KL time series classification rule was demonstrated to have sufficient metric properties for the statistical properties of NN rule classification to hold. An economical procedure for computing KL numbers between stationary Gaussian time series, via AR modeling of the time series and an example of the application of the NN-KL rule to the classification of EEG time series in a population screening problem were shown. The probability of misclassification achieved using the NN-KL rule was smaller than that achieved using spectral features discrimination rules on the same data.

In the following we discuss interpretations of the NN-KL number classification rules as maximum likelihood and entropy criterion rules. Additional aspects of data normalizations and another, non-population screening NN-KL number rule application are also discussed.
4.1 The NN-KL Classification Rule as an ML Rule, and Akaike AIC Rule and as a Shore-Johnson Rule.

The NN-KL number time series classification rule has an interpretation as a classification rule based upon distributional assumptions. If the prior class probabilities and the to-be-classified and labeled time series conditional data distributions were known, the exploratory population screening problem objective of minimizing the probability of misclassification, could be achieved by a Bayes classification rule, (Duda and Hart, (1973), for example). For simplicity of presentation, and because it was reasonable in the EEG time series classification problem that motivated our work, we assumed equal class priors. In that case, the maximum likelihood (ML) rule is optimal if only \( x^{(0)} = x^{(m)} \) or \( f_0 = f_m \) for some \( m, \) in \( m = 1, \ldots, N. \) That is, the NN-KL number rule is an ML rule only if the distribution of the time series to-be-classified is in the class of the labeled sample time series distributions. As suggested by the illustration of the EEG time series in Figure 1, in the exploratory population screening problem, it is extremely unlikely that \( x^{(0)} = x^{(m)} \) for any \( x^{(m)} \) in the labeled time series.

Under the assumption that \( x^{(0)} \neq x^{(m)} \) for any \( m, m = 1, \ldots, N \) the NN-KL number rule has an expected log likelihood classification rule interpretation. Similarly, the Akaike AIC-maximum expected entropy procedure (Akaike 1977, 1983), has a classification rule interpretation. The AIC of the \( m \)-th model, given the data \( x^{(0)} \) is:

\[
(4.1) \quad \text{AIC}(m; x^{(0)}) = -2 \log \text{(likelihood model } m) + 2(\# \text{ free parameters}).
\]

The minimum AIC procedure is: i) Compute AIC \((m, x^{(0)})\) for each \( m, m = 1, \ldots N \). ii) Label \( x^{(0)} \) with the label of that labeled time series for which the AIC is smallest. In the application of the the AIC to classification, the AIC is computed with zero free parameters. Then, under the assumption of normality -2 log likelihood of the \( m \)-th labeled time series is

\[
(4.2) \quad -2 \log \text{likelihood}(x^{(m)}; x^{(0)}) = T \log 2\pi + T \log |\hat{\Sigma}_m| + tr(\hat{\Sigma}_m^{-1}\hat{\Sigma}_0).
\]

To within constant terms, (4.2) is identical to the NN-KL dissimilarity measure in (2.1.3), (for stationary time series, \( \mu_0 = \mu_m \)).

The minimum cross-entropy classification rule, Shore and Johnson, (1980), (1982) can readily be seen to be identical to an AIC classification rule. Equation (17) of the 1982 paper, is,

\[
(4.3) \quad H(q^+, \hat{q}_*) = H(q^+, q) + H(q, \hat{q}_*).
\]

12
In (4.3) $H(q^+, q_\ast)$ is the dissimilarity due to approximating the true distribution $q^+$ by an assumed (normal) distribution $q_i$, and $H(q, \hat{q}_\ast)$ is the dissimilarity due to choosing the NN (normal) distribution $\hat{q}_\ast$ that is closest to $q^+$. The minimum cross entropy classification procedure selects that normal distribution model $\hat{q}_\ast$ that is closest to the true distribution $q^+$. This is achieved by the minimum of $H(q^+, \hat{q}_\ast)$ over the $\hat{q}_\ast$.

The Akaike and Shore and Johnson entropy criterion classification rules have the important property that the distribution of the data $z^{(0)}$ need not be in the class of alternative classification distributions as is required with ML classification rules. Thus, the entropy criterion procedures offer a justification for the common but formally incorrect use of ML classification procedure in such situations.

Also, we note that the entropy criterion interpretation of kNNT classification rules suggests an alternative approach to the kNNT rule classification. Each of the labeled time series (the past experience) potentially contains some information about how to classify the new time series $z^{(0)}$. Adding the log likelihoods of each of the labeled time series in each categorical class is an entropy criterion-information theoretic alternative to contemporary kNNT decision rules. The number of kNNT’s that are relevant to the classification of a new time series can differ for each time series. The non-information theoretic coin-toss decision in the case of ties is eliminated in the suggested alternative criterion. Akaike, (1978), indicates the likelihood interpretation of the AIC that is relevant for this approach.

4.2 More on Data Normalizations.

A slight difference in the NN-KL and entropy criterion dissimilarity measure formulas, (2.1.3) and (4.2) and (4.3) respectively, that is potentially of considerable practical importance is in the inclusion of the term $\log |\hat{\Sigma}_0|$ in (2.1.3). Equation (2.1.3) exhibits a precise decomposition of the contribution of particular terms to the total information available for discrimination. In the classification of stationary time series, the terms $\hat{\mu}_0$ and $\hat{\mu}_m$ in (2.1.3) are identically zero and clearly the term $\log(|\hat{\Sigma}_m|/|\hat{\Sigma}_0|)$ is independent of $\hat{\Sigma}_0$ for all $m$, as in (4.2). Notice that by a covariance normalizing linear transformation of each labeled time series, $|\hat{\Sigma}_m|$ can be equated to $|\hat{\Sigma}_0|$. In that case the contribution $\log(|\hat{\Sigma}_m|/|\hat{\Sigma}_0|)$ is identically zero and the classification is based solely on the term $\text{tr}(|\hat{\Sigma}^{-1}_m|\hat{\Sigma}_0)$. In EEG time series, the physiological interpretation of that normalization is that the time series are equated for EEG “generator” power. Another normalization alternative is to equate the time series for mean square power. That is, equate the zero-lag covariance matrices $\Gamma^{(m)}(0)$ for $m = 0, 1, \ldots, N$. The circumstances under which particular normalizations are optimal
for discrimination are a potentially interesting empirical research topic.

A non-trivial example that illustrates the relevance of the contribution of the individual terms in (2.1.3) to the information for discrimination in the classification of ensembles of trials of non-stationary evoked potential time series is treated in Gersch and Brotherton, (1981). For the particular experimental situation considered, it was shown that there was very much more information for discrimination in the covariance or background EEG than there was in the average evoked potentials, (the second and third terms in 2.1.3 respectively). This example is a counter-example to thirty years practice of a commonly accepted average evoked potential dogma for EEG discrimination.

4.3 NN-KL Rules in Normalized Baseline Classification: Another Application.

Examples of the potentially important practical value of the NN-KL rule as an ML classification rule in the classification production line assembled auto engines appears in Gersch et al., (1983). Each new production line engine is to be classified as normal, as having one of a particular large class of known disorders or as in the category of a not previously seen object. Statistically, every single new engine in each of the categorical classification states is more like every other engine in the same state than it is like any engine in a different categorical state. Such problems may be called "normalized baseline" classification problems. The statistical model description of the measurements from a single trial on a single engine, with intentionally created assembly flaws, becomes the prototype model for all other engines in the same engine flaw category.

This situation is very different than the population screening problem. It is essentially a ML classification problem. ML time series classification problems become one-to-one with the problem of modeling time series. The application of NN-KL number type ML classification rules to a variety of stationary and nonstationary time series that were obtained from auto engine sensors is shown in Gersch et al., (1983). An important property of the NN-KL classification procedures is that they offer an estimate of misclassification probability that can be based on only the set of single prototypic trial samples. The important estimate of the probability of misclassification is not available in the "signature analysis" procedures that are currently used for classification of production line inspection flaws, (Volin, 1979).

Acknowledgements. The work reported here was supported in part by grants from the National Science Foundation. The careful reading of the manuscript by the referees is gratefully appreciated. Permission to use Tables and Figure 1 from Gersch et al, Science, 205, 193-195, has been granted.
References


LEGEND FIGURE 1: Twenty second epochs of bipolar anesthesia level L1 and L3 EEGs (C4-02, C3-01) from 5 individuals. (L1: Anesthesia level insufficient for deep surgery, L3: Anesthesia level sufficient for deep surgery.)

VGH - HA (F 4, L1, S40) 0-20 SECS.

VGH - HA (F 22, L1, S42) 0-20 SECS.

VGH - HA (F 97, L1, S52) 0-20 SECS.

VGH - HA (F 144, L1, S71) 0-20 SECS.

VGH - HA (F134, L1, S73) 0-20 SECS.

VGH - HA (F 2, L3, S40) 0-20 SECS.

VGH - HA (F 26, L3, S42) 0-20 SECS.

VGH - HA (F101, L3, S52) 0-20 SECS.

VGH - HA (F145, L3, S71) 0-20 SECS.

VGH - HA (F170, L3, S73) 0-20 SECS.
**Nearest Neighbor Rule Classification of Time Series in Exploratory Population Screening Problems**

**Author(s):**
W. Gersch

**Performing Organization Name and Address:**
Department of Statistics
Stanford University
Stanford, CA 94305

**Controlling Office Name and Address:**
Office of Naval Research
Statistics & Probability Program Code 1111

**Report Date:**
December 9, 1986

**Number of Pages:**
20

**DISTRIBUTION STATEMENT (of this report):**
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

**KEY WORDS:**
AIC; classification; entropy; Kullback-Leibler information; nearest neighbor rules; population screening; time series.

**ABSTRACT:**
PLEASE SEE FOLLOWING PAGE.
20. ABSTRACT

In exploratory time series classification problems only a modest number of categorically labeled sample time series is available. The scientific conjecture in this situation is that there is sufficient information in the time series to distinguish between the alternative population categories. The principal objective of the analysis is to determine the separability of the alternative categorical classes, i.e., to estimate the minimum achievable probability of misclassification. A nearest neighbor time series classification rule is advocated for the solution of this problem. In our approach, the dissimilarity between a new-to-be-classified time series and each of the members of the labeled sample time series is computed as an estimate of the Kullback Leibler number, as if the time series were Gaussian distributed. Time and frequency domain formulas for the Kullback Leibler number between stationary multivariate Gaussian time series yield a practical parametric model computational realization of our procedure. An application is shown in an exploratory classification of the anesthesia level of humans in surgery by the analysis of stationary multivariate EEG time series. The probability of misclassification error is obtained by a leave out one-at-a-time cross validation classification of the labeled sample time series. The KL-NN classification rule is a maximum entropy criterion classification rule.