WILLIAM MARSH RICE UNIVERSITY
6100 South Main Street
Houston, Texas 77001

ANNUAL REPORT

to

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

on

ROBUST PARAMETER ESTIMATORS FOR COMMUNICATION DATA

Presented by:

P. Papantoni-Kazakos
Electrical Engineering
[redacted]

Grant Number: AFOSR 77-3156
Starting Date: 15 October 1976
Date: December 3, 1977

P. Papantoni-Kazakos
Principal Investigator
(713) 527-8101, Ext. 3579

Summary of Completed Work

The object of this grant is the analysis and design of decision procedures that have stable, good performance in statistically ill-defined environments. Such procedures indicate the way to design powerful receivers for systems whose statistical behavior can not be described precisely (due to incomplete availability of data about the system behavior).

In the framework of this idea the following progress has been already made:

1. Different distance measures have been studied for use as performance criteria for robust estimates. Careful evaluation and comparison of these distances was done and their similarities, advantages and disadvantages were carefully stated. It was observed that some of these distances are more naturally related to the estimation problem and that in cases in which they are equivalent, the designer may use the one that is computationally or structurally more convenient. The use of the Vasershtein distance was proposed and used as stability measure for estimates in statistically contaminated environments. This distance is naturally related to the commonly used performance measures in parameter estimation. Through the use of the Vasershtein distance, qualities of powerful robust (with uniformly good performance inside a family of data statistics) estimates were found, when the observable data are dependent. For dependent data also robust estimates that perform well in the presence of small number of discrete data were studied. Such analysis is valuable in cases that the engineer must make his decision in real time.

Some of the work mentioned in this paragraph is included in this report, while some of it is still in progress.

2. A thorough study of the work already accomplished (by the author as well as other investigators) on nonparametric statistical procedures in the presence of small number of discrete data was done and included in a book on the use of nonparametric procedures in Communication Systems.

3. A feature selection problem was studied, when several distance measures are used as discrimination criteria. This helped for a better understanding of the qualities of the distances. It was found that the feature extraction algorithm is sometimes independent of the criterion. This allows the maintenance of a single feature construction mechanism that works equally well for several systems with different specifications. This feature selection algorithm is then robust.

4. A sequencial procedure for clustered data was proposed and analyzed. This procedure applies to several stages of statistical information about the system and it varies from the known procedures in the fact that data collection costs are included and the data clusters considered are finite in number. The results are therefore nonasynpotic and they apply to any problem in which the data are collected sequentially in clusters and there is a preassigned maximum number of such clusters available. The results have been tested numerically for some systems with given specifications.

5. Hampel's general qualitative definition of robustness of sequences of estimators on memoryless observation processes was generalized to stationary processes. Structural properties of the estimates were found in this case and based on these properties the design of robust estimates that operate on dependent data is now in progress.

6. The constructive analysis of robustness completed by the author is being used now in the performance analysis of communication Networks at Bell Laboratories.

7. The discrimination of Gaussian processes has been studied and efficient computationally methods have been found. This method leads also to efficient discrimination of contaminated Gaussian processes.

In twelve months, one Ph.D. thesis and one book have been partially supported by this grant, three papers have been submitted to journals, four conference presentations have been made, two University and three Bell Laboratories reports have been produced. Finally, two seminars at Bell Telephone Laboratories have been presented.

In what follows, a list of publications supported by this grant, and some of the work accomplished that is not included in the semiannual report dated May 6, 1976, are presented.

Activities Supported by AFOSR

Ph.D. Thesis Completed:

1. R.Y.S. Li, "Methods for Data Reduction," May 1977.

Books Published:

1. P. Papantoni-Kazakos and D. Kazakos, editors and contributors, "Nonparametric Methods in Communications. Selected Topics." Marcel Dekker Inc., New York 1977.

Papers Submitted to Journals:

1. P. Papantoni-Kazakos, "Some Distance Measures and Their use in a Feature Selection Problem."

2. P. Papantoni-Kazakos and R. M. Gray, "Robustness of Estimators on Stationary Observation."

3. D. Kazakos and P. Papantoni-Kazakos, "Asymptotic Discrimination of Gaussian Processes.

Conference Presentations:

1. P. Papantoni-Kazakos, "Some distance measures and their use in feature selection," Eleventh Annual Conference on Information Sciences and Systems, The Johns Hopkins University.

2. P. Papantoni-Kazakos, D. Kazakos and R. Li, "A Kalman Filtering Formulation for the Linear Reduction of Gaus-Markov Data," Eleventh Annual Conference on Information Sciences and Systems, The Johns Hopkins University.

3. D. Kazakos and P. Papantoni-Kazakos, "Robust Rate Distortion," International Symposium on Information Theory, 1977.

4. P. Papantoni-Kazakos, "Some Problems in Communication Networks," Fifteenth Annual Allerton Conference on Cricuit and System Theory, 1977.

University Reports:

1. P. Papantoni-Kazakos, "Some Distance Measures and Their use in Feature Selection," Rice University E.E. Technical Report #7611, November 1976.

2. P. Papantoni-Kazakos, "Some New Performance Criteria in Robust Statistics - Small Sample Robustness," Technical Report #7701, January 1977.

Bell Laboratories Technical Memoranda:

1. P. Papantoni-Kazakos, "Some Distance Measures and Their Use in Feature Selection," TM-77-3452-5, July 12, 1977.

2. P. Papantoni-Kazakos, "Some Performance Criteria Incorporating Data Dependence in Robust Estimation," TM-77-3452-4, July 12, 1977.

3. P. Papantoni-Kazakos and R. M. Gray, "Robustness of Estimation on Stationary Observations," TM-77-3452-7, September 20, 1977.

Seminars Presented:

1. P. Papantoni-Kazakos, "The Vasershtein Distance in the Constructive Analysis of Robust Estimates," Bell Telephone Laboratories, Holmdel, New Jersey, April 1977.

2. P. Papantoni-Kazakos, "Robust Estimators on Stationary Observations," Bell Telephone Laboratories, Holmdel, New Jersey, October 1977.

Comments on the Accomplished Work From Scientists in the Field

The constructive analysis of robustness with the use of a Vasershtein stability criterion has been considered as more naturally incorporation the proper performance criteria in parameter estimation by people at Stanford University and Bell Telephone Laboratories, that I talked to. Also, the extension of the analysis to data evolving from general stationary process (rather than just process with independent data), has been considered important for the understanding of robust estimates in the presence of dependent data structures.

The study and evaluation of different distance measures and their applications to the feature selection problem has been considered valuable by attendies of the 1977 Johns Hopkins Conference. The different distance measures are used as different discriminant measures, each representing a different class of problems. Their uniform evaluation and comparison that has not been done before and the analysis of their value to the feature extraction problem has been considered a nice contribution.

The sequencial decision scheme included in the thesis enclosed here, and more particular its version for two nonparmetric distinct classes has been considered very valuable by scientists in pattern recognition. Its use allows data savings as well as good performance for discrimination between two statistically ill-defined data classes.

Workshops Attended:

1. 1977 Communications Workshop, Tuscon, Arizona, April 1977.

SOME PROBLEMS IN COMMUNICATION NETWORKS

by

P. Papantoni-Kazakos
Bell Laboratories

ABSTRACT

A general discussion is presented on some of the open problems in communication networks. Routing structures and causes for unsuccessful communication through the network are emphasized. Some open problems involving sophisticated parametric as well as robust statistical algorithms are stated.

Work done at Rice University and supported by the Air Force Grant AFOSR 77-3156

I. Description of the Network

To understand some of the problems involved in reliably communicating messages within the network, some basic network operations must be described.

The smallest element (that is of any interest to the network analyst) in a communication network is a center. A center consists of several units that communicate directly with each other. Different centers communicate through a number of routes, where each time the route one particular message is carried on is chosen hierarchically. Each route consists of a number of links that are, in general, connected to each other through tandems (switching offices). Finally, each link consists of a number of single message carriers that are called trunks, while the tandems connect several centers. A message originating at center A (figure 1) and with destination another center B follows a routing hierarchy described as follows:

At first tries the direct route that consists of a single link connecting the two centers (dotted line in figure 1). If all the trunks in this link are functioning properly but are busy, the message tries the next route in the hierarchy (route through A, T_1, B in figure 1). If this route is also well functioning but busy, the message tries the next route in hierarchy and so on, until it reaches the final route available to it (route A T_2T_3B in figure 1). If this last route is busy or malfunctioning, the message fails to go through and a <u>communication failure to B</u> is recorded at A.

The rejection of the message by a particular route due to full occupancy (at the moment) of all trunks involved is called blocking. Under healthy network conditions blocking probabilities can be assigned to each route that correspond to a particular center pair (A,B) and are functions of the A to B communication load, the number of routes connecting $A_e B$, and the number of trunks in each such route.

Suppose now that an "average load" time period is considered and the communication from center A to center B is studied. If in some of the routes between A and B a link is malfunctioning (due to some faulty trunk), and if center A is unaware of the malfunction, messages from A to B will keep trying this link with probability specified by the initial routing structure and the "average load". As a result to that some messages will be killed by the malfunctioning link and communication failures from A to B will be recorded. Therefore, in the presence of faulty links which center A is unaware of, communication failures will be caused that are not just due to overload and are not happening just at the highest in hierarchy route.

The routing structure described above is based on a trade off between economy and communication efficiency. The direct links (dotted line in figure 1) carry usually the highest portion of the message load, while the higher in hierarchy alternate routes are used during traffic picks and they have capacity high enough to secure good communication when such picks are occurring and low enough so that they do not remain idle most of the time.

The performance of the network, as viewed by the users, is measured through its ability to successfully respond to communication attempts. It efficiency as viewed by an outside observer is a combination of two factors: effectiveness in responding to communication demands, and average degree of occupancy.

2. Some Open Problems

We are concentrating here on the performance evaluation of the network. The following major question arises in this case:

Is it possible to evaluate the network performance at a particular time, if yes what kind of data are required and how can such an evaluation be effective without utilizing an excessive amount of information? Also, how can malfunctions be localized or even predicted with the use of economically attractive methods?

In two Bell Labs technical memoranda that have not been cleared for publication yet, the author analyzes the use of limited center-to-center successful and unsuccessful communication completions for locating the faulty links that cause the failures, and for continuously monitoring the quality in communication throughout the network. The algorithms used are economical not only because they only utilize a limited amount of information, but also because they are one step memory and computationally efficient. However, assumptions as to the routing structures have been made that in some cases need relaxation. Specifically, "average load" time periods are observed and the routing probabilities are then considered unchanged. But,

even during such time periods the load fluctuations may be momentarily substantial, in which case "robust" algorithms that are mostly insensitive to such load variations must be developed.

Furthermore, the effect of malfunctioning network links to the relationship between communication messages through them must be studied further. Specifically, such links may cause "partial message killing" as well as interference between messages. These effects result in additional reduction of the communication quality within the network.

References

1. P. Papantoni-Kazakos, "Use of link Call Completions in Quality Control" Bell Laboratories technical memorandum in typing process.

2. P. Papantoni-Kazakos, "The potential of end-to-end call completion measurements in Network trouble localization and quality control," Bell Laboratories technical memorandum in process.

3. J.S. Kaufman, "Faculty-Trunk Detection Algorithms Using EADAS-ICUR Traffic Data," Bell System Technical Journal, Vol. 56, No. 6, pp. 919-976.

4. D. Kazakos, "Recursive Estimation of Prior Probabilities Using a Mixture," IEEE Trans. Inform. Theory, vol. IT-23(2), pp. 203-211.

5. P. Papantoni-Kazakos, "Some Performance Criteria Incorporating Data Dependence in Robust Estimation," Bell Laboratories Technical Memorandum TM-77-3452-4, July 12, 1977.

6. E.S. Page, "Continuous inspection schemes," Biometrika 41, pp. 100-115, 1954.

7. H. Chernoff, Sequential Analysis and Optimal Design, Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

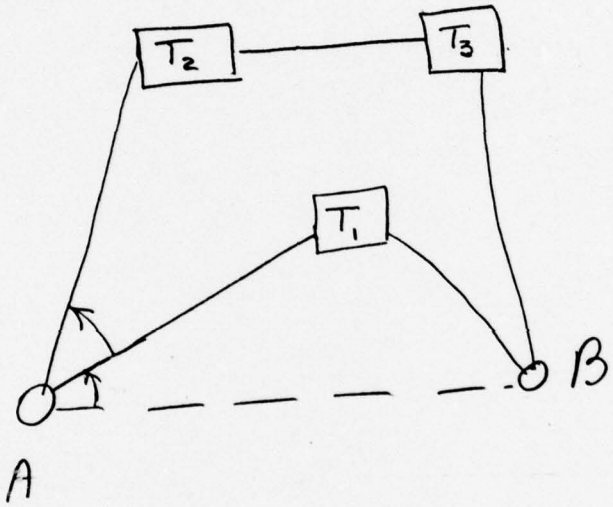

FIGURE 1

Asymptotic Discrimination of Gaussian Processes

Dimitri Kazakos* and Titsa Papantoni-Kazakos**

State University of New York at Buffalo

and

Bell Laboratories

ABSTRACT

We present a theory on the asymptotic approximation of block Toeplitz matrices by block circulant ones. The method is then applied to the calculation of the asymptotic Bhattacharyya distance B_n and divergence J_n between two stationary vector Gaussian processes, in terms of the two spectral density matrices, $F_1(\lambda)$, $F_2(\lambda)$. Specifically,

$$B = \lim_{n \to \infty} n^{-1} B_n = (2\pi)^{-1} \int_0^{2\pi} \log\{|2^{-1}F_1(\lambda) + 2^{-1}F_2(\lambda)| \cdot |F_1(\lambda)|^{-1/2} |F_2(\lambda)|^{-1/2}\} d\lambda$$

$$J = \lim_{n \to \infty} n^{-1} J_n = (2\pi)^{-1} \int_0^{2\pi} \text{trace}\,[F_1(\lambda)F_2^{-1}(\lambda) + F_2(\lambda)F_1^{-1}(\lambda) - 2I] d\lambda$$

The above expressions are useful because of existing upper and lower bounds to the Bayes error of misclassification. Furthermore, they can be considered as distance measures in their own right. The availability of efficient spectral estimation techniques renders them most useful.

* Research supported by NSF Grant ENG 76 20295.

** Research supported by Air Force Grant AFOSR 77-3156.

I. INTRODUCTION

It is well known that the Bayes decision rule is the optimal one in deciding between two statistical hypotheses with known prior probabilities and conditional probability density functions. One of the most common statistical models for data is the Gaussian random process. In assessing the performance of the statistical classifier using the Bayes decision rule, one is faced with the difficult task of evaluating its performance through the available expression for the probability of misclassification, Pe:

$$Pe = \int_{R^{kxn}} \min[\pi_1 f_1(x^n), \pi_2 f_2(x^n)]dx^n \qquad (1)$$

where

$$x^n = [x_1 \ldots x_n] \quad , \quad x_i \in R^k,$$

$f_1(x^n)$, $f_2(x^n)$ are the conditional p.d.f. and π_1, π_2 are the prior probabilities of hypotheses H_1, H_2. Clearly, numerical integration techniques have to be used. Due to the high dimension of the integration region, numerical techniques are costly and they do not provide understanding of the influence of several parameters of interest in f_1, f_2 to Pe. For example, if one wishes to reduce the data by some feature selection techniques, the expression (1) cannot by useful in choosing the optimal transformation of x^n. Also, (1) does not provide any feeling as to the incremental reduction of Pe as n grows.

The following pair of bounds to Pe is known: [1] - [4]

$$2^{-1}\pi_1\pi_2 \exp\{-2B_n\} < Pe < (\pi_1\pi_2)^{1/2} \exp\{-B_n\} \qquad (2)$$

$$8^{-1} \exp\{-2^{-1} J_n\} \leq Pe \tag{3}$$

where:

$$B_n = -\log \int [f_1(x^n) f_2(x^n)]^{1/2} dx^n \tag{4}$$

$$J_n = \int [f_1(x^n) - f_2(x^n)] \log[f_1(x^n) f_2^{-1}(x^n)] dx^n \tag{5}$$

A lower bound tighter than (3) has been developed in [5]. It has been shown in [6] that no upper bound to Pe in terms of J_n exists. In the present paper, we will develop asymptotic expressions for

$$B = \lim_{n \to \infty} n^{-1} B_n \quad , \quad J = \lim_{n \to \infty} n^{-1} J_n \tag{6}$$

in terms of the spectral density matrices $F_1(\lambda)$, $F_2(\lambda)$. The motivation lies in the fact that the spectral densities are among the first characteristics of a process to be measured, and very efficient spectral estimation techniques are available in the statistical literature. [7] - [10].

The technique to be used is, we think, interesting by itself, and useful in other applications. It is based on the asymptotic approximation of block Toeplitz matrices by block circulant ones. A similiar technique was used in [11] - [13] for evaluation of rate-distortion functions.

The distance measures derived have several potential applications, which are discussed briefly in the section that follows. Those are:

(a) Feature selection in high-dimensional observation spaces in pattern recognition.

(b) Clustering algorithms for the same situations as in (a).

(c) Reduction of remote sensing data.

(d) Speech processing.

(e) Biomedical EEG signal analysis.

(f) Tone Detection in Telephone Networks.

II. ASYMPTOTIC APPROXIMATIONS

A block Toeplitz knxkn matrix $\overline{R_n}$ has the form:

$$\overline{R_n} = \begin{bmatrix} R_o & R_{-1} & \cdot & \cdot & \cdot & R_{-n+1} \\ R_1 & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot & \cdot \\ \cdot & & & \cdot & & R_{-1} \\ R_{n-1} & \cdot & \cdot & \cdot & R_1 & R_o \end{bmatrix} \tag{7}$$

where R_i are kxk matrices. A block circulant knxkn matrix $\overline{C_n}$ has the form:

$$\overline{C_n} = \begin{bmatrix} C_o & C_1 & C_2 & \cdot & \cdot & \cdot & C_{n-1} \\ C_{n-1} & C_o & C_1 & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & C_2 \\ \cdot & & & & \cdot & \cdot & C_1 \\ C_1 & \cdot & \cdot & \cdot & \cdot & C_{n-1} & C_o \end{bmatrix} \tag{8}$$

where C_i are kxk matrices. Consider the knxkn matrix

$$V_n = \frac{1}{\sqrt{n}} \begin{bmatrix} I_k & I_k & I_k & \cdots & I_k \\ I_k & wI_k & w^2 I_k & \cdots & w^{n-1} I_k \\ I_k & w^2 I_k & w^4 I_k & \cdots & w^{2(n-1)} I_k \\ \vdots & \vdots & \vdots & & \vdots \\ I_k & w^m I_k & w^{2m} I_k & \cdots & w^{(n-1)m} I_k \\ \vdots & \vdots & \vdots & & \vdots \\ I_k & w^{n-1} I_k & w^{2(n-1)} I_k & & w^{(n-1)^2} I_k \end{bmatrix} \tag{9}$$

where $w = \exp(i2\pi n^{-1})$.

It can be easily shown that V_n is a unitary matrix, i.e.:

$$V_n^{-1} = V_n^t \quad , \quad |V_n| = 1 \tag{10}$$

Consider now the matrix

$$\tilde{C}_n = V_n^{-1} \overline{C_n} V_n = V_n^t \overline{C_n} V_n \tag{11}$$

Performing the multiplication (11), we observe that $\tilde{C}_n$ is block diagonal:

$$\tilde{C}_n = \begin{bmatrix} C(0) & & & & \\ & C(2\pi n^{-1}) & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & C(2\pi n^{-1}(n-1)) \end{bmatrix} \tag{12}$$

where, C(u) is a kxk matrix function defined as:

$$C(u) = \sum_{m=0}^{n-1} C_m \exp(-imu) \tag{13}$$

Consider, now the problem of finding the eigenvalues of $\bar{C}_n$. They are solutions of the equation in x:

$$0 = |\bar{C}_n - xI_{kn}| = |V_n^{-1}(C_n - xI_{kn})V_n^t| = |\tilde{C}_n - xI_{kn}| =$$

$$= \prod_{m=0}^{n-1} |C(2\pi mn^{-1}) - xI_k| \tag{14}$$

Thus, the kn eigenvalues of C_n are identical to the union of n sets of eigenvalues of the matrices $\{C(2\pi mn^{-1}) , m = 0, \ldots , n-1\}$. Let us now define the weak norm of an sxs matrix $A = \{a_{ij}\}$ as:

$$\lfloor A \rfloor = [s^{-1} \sum_{i=1}^{s} \sum_{j=1}^{s} |a_{ij}|^2]^{1/2} = [s^{-1} \sum_{i=1}^{s} |q_i|^2]^{1/2} \tag{15}$$

where $(q_1 \ldots q_s)$ are the eigenvalues of A. Also, we define the strong norm $\| A \|$ as:

$$\| A \| = \max_i |q_i| \tag{16}$$

If a_{ij} are kxk matrices and A is ksxks, we still have:

$$\lfloor A \rfloor^2 = s^{-1} \sum_{i=1}^{s} \sum_{j=1}^{s} \lfloor a_{ij} \rfloor^2 \tag{17}$$

Let $\{A_n\}$, $\{B_n\}$ be two sequences of Hermitian sxs matrices. We say that they exhibit "mutual approximation", denoted by $A_n \sim B_n$, if:

a) $\| A_n \|$, $\| B_n \|$, $\lfloor A_n \rfloor$, $\lfloor B_n \rfloor$ are all bounded from above by a finite number M independent of n.

b) $$\lim_{n \to \infty} \lfloor A_n - B_n \rfloor = 0 \tag{18}$$

Let $\{a_k^{(n)}\}_{k=1}^n$, $\{b_k^{(n)}\}_{k=1}^n$ be the eigenvalues of A_n, B_n correspondingly. We say that the sets $\{a_k^{(n)}\}$, $\{b_k^{(n)}\}$ are "asymptotically equally distributed" in the interval [-M,M] if

$$|a_k^{(n)}| \leq M \quad , \quad |b_k^{(n)}| \leq M \ , \forall \ k,n$$

and for any continuous function $f(\cdot)$ on [-M,M] we have

$$\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} [f(a_k^{(n)}) - f(b_k^{(n)})] = 0 \tag{19}$$

The following theorem of Grenander and Szego [14] will be used.

Theorem 1: Let $\{A_n\}$; $\{B_n\}$ be two sequences of Hermitian matrices with eigenvalues $\{a_k^{(n)}\}$, $\{b_k^{(n)}\}$. If $A_n \sim B_n$, and either $\lim_{n\to\infty} |A_n|$ or $\lim_{n\to\infty} |B_n|$ exists, then $\{a_k^{(n)}\}$, $\{b_k^{(n)}\}$ are asymptotically equally distributed.

We have found until now that the kn eigenvalues of a block circulant matrix of the type (8) are grouped according to (14). However, we are interested in asymptotic expressions of eigenvalues of covariance matrices of the type (7) with $R_{-k} = R_k$. We will therefore approximate the block symmetric Toeplitz matrix

$$\overline{R_n} = \begin{bmatrix} R_o & R_1 & \cdot & \cdot & R_{n-1} \\ R_1 & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & R_1 \\ R_{n-1} & \cdot & \cdot & R_1 & R_o \end{bmatrix} \tag{20}$$

by the block circulant one.

There is no unique approximation by a block circulant. A convenient one will be chosen next. It is a matrix generalization of the circulant approximation used by Grenander and Szego [14]. Let $F(\lambda)$ be the spectral density of the stationary random process in question. We assume that

$$\sup_{\lambda}|F(\lambda)| \leq M < +\infty$$
$$0 < m \leq \inf_{\lambda}|F(\lambda)| \tag{21}$$

Let

$$R_k^1 = \begin{cases} (1-|k|/p)\ R_k & \text{for } |k| < p \leq n \\ 0 & \text{Otherwise} \end{cases} \tag{22}$$

$$\bar{R}_n^1 = \{R^1_{|i-j|}\} \tag{23}$$

and

$$F_p(\lambda) = \sum_{k=-p}^{p} (1-|k|/p)\ R_k e^{-ik\lambda} = \sum_{k=-p}^{p} R_k^1\ e^{-ik\lambda} \tag{24}$$

Consider the block-circulant matrix

$$\bar{L}_n = V_n L_n V_n^{-1} \tag{25}$$

where L_n is a block diagonal, with diagonal blocks:

$$\{L_n\}_{mm} = F_p(2\pi m n^{-1}) \tag{26}$$

and where V_n is given by (9). It is easily shown that:

$$\{\bar{L}_n\}_{ms} = n^{-1} \sum_{j=1}^{n} e^{2\pi i j(m-s)n^{-1}}\ F_p(2\pi j n^{-1}) \tag{27}$$

We need to calculate the differences

$$\left|\bar{L}_n - \bar{R}_n^1\right|^2 \quad , \quad \left|\bar{R}_n^1 - \bar{R}_n\right|^2$$

We have:

$$\left|\bar{L}_n - \bar{R}_n^1\right|^2 = 2n^{-1} \sum_{m-1}^{n} m\left|R_m^1\right|^2 \leq 2pn^{-1} \sum_{m-1}^{\infty} \left|R_m\right|^2 \tag{28}$$

$$\left|\bar{R}_n^1 - \bar{R}_n\right|^2 \leq 2n^{-1} \sum_{m=1}^{p} m^2 p^{-2}(p-m)\left|R_m\right|^2 + 2n^{-1} \sum_{m=p+1}^{\infty} (n-m)\left|R_m\right|^2 \tag{29}$$

Due to (21) we have

$$\sum_{m=0}^{\infty} \left|R_m\right|^2 < +\infty$$

and thus for a given $\varepsilon > 0$ we can pick a p so large that $\sum_{m=p+1}^{\infty} \left|R_m\right|^2 < \varepsilon$. By choosing first p and then n sufficiently large, we can make the distance $\left|\bar{L}_n - \bar{R}_n\right|$ sufficiently small, i.e.:

$$\left|\bar{L}_n - \bar{R}_n\right| \leq k_1 n^{-1/2} + \left[k_2 n^{-1} + 2\varepsilon\right]^{1/2} \tag{30}$$

$\bar{L}_n$ is Hermitian and bounded, and its nk eigenvalues are the union of the eigenvalues of the n matrices $\{F_p(2\pi m n^{-1}) \ , \ m = 1, 2, \ldots, n\}$.

Let $h_q(p,u)$ be the q^{th} largest eigenvalue of $F_p(u)$. It can be easily shown that $h_q(p,u)$ is a continuous function of u. According to Theorem 1, for any continuously differentiable function g on [m,M], we have:

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} g[h_q(p, 2\pi m n^{-1})] = (2\pi)^{-1} \int_0^{2\pi} g[h_q(p,u)]du \tag{31}$$

Also,

$$\left| \int_0^{2\pi} \{g[h_q(p,u)] - g[h_q(u)]\}du \right| \leq A \cdot \int_0^{2\pi} \left|h_q(p,u) - h_q(u)\right| du \tag{32}$$

where $h_q(u)$ is the qth eigenvalue of F(u) and A is a bound on the first derivative of g. Thus, as $p \to \infty$, the right side of (32) goes to 0.

In conclusion, if $p,n \to \infty$ in the manner prescribed in the development of (30), we will have:

$$\lim_{n,p\to\infty} n^{-1} \sum_{m=1}^{n} g[h_q(p, 2\pi m n^{-1})] = (2\pi)^{-1} \int_0^{2\pi} g[h_q(u)]du \qquad (33)$$

(33) was developed by an asymptotic approximation of the block circulant matrix $\overline{R}_n$ by the block circulant matrix $\overline{L}_n$. Simpler block circulant approximations to $\overline{R}_n$ may be developed, along the lines of [15], [16],[17] but this would require more restrictive conditions on $F_1(\lambda)$, $F_2(\lambda)$ than (21).

In the following section we will apply equation (33) to the calculation of the asymptotic expressions given in the abstract.

III. ASYMPTOTIC EXPRESSIONS

We are now considering two stationary, k-dimensional vector Gaussian processes. Let $x^n = (x_1 \ldots x_n)$ be a sequence of n vector, zero mean observations, and let the corresponding covariance matrices be:

$$\bar{R}_{nj} = E[x^n (x^n)^t | H_j] \quad , \quad j = 1, 2$$

$$\bar{R}_{nj} = \begin{bmatrix} R_{oj} & R_{1j} & \cdot & \cdot & R_{n-1,j} \\ R_{1j} & R_{oj} & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ R_{n-1,j} & & \cdot & \cdot & R_{oj} \end{bmatrix} \tag{34}$$

The Bhattacharyya distance B_n is then expressed as:

$$2B_n = \log|2^{-1}\bar{R}_{n1} + 2^{-1}\bar{R}_{n2}| - 2^{-1}\log|\bar{R}_{n1}| - 2^{-1}|\bar{R}_{n2}| \tag{35}$$

and the divergence:

$$2J_n = \text{trace}\,[\bar{R}_{n1}\bar{R}_{n2}^{-1} + \bar{R}_{n1}^{-1}\bar{R}_{n2} - 2I] \tag{36}$$

For the calculation of B, we observe that:

$$n^{-1}\log|\bar{R}_{n1}| = n^{-1} \sum_{i=1}^{nk} \log d_{in}$$

where $\{d_{in}\}$ are the kn eigenvalues of $\bar{R}_{n1}$. According to the theory, they are asymptotically equally distributed to the eigenvalues of a block circulant approximation $\bar{L}_n$. Thus,

$$\bar{R}_{n1}\bar{R}_{n2}^{-1} \sim \bar{L}_{n1}\bar{L}_{n2}^{-1} = V_n L_{n1} V_n^{-1} V_n L_{n2}^{-1} V_n^{-1} = V_n L_{n1} L_{n2}^{-1} V_n^{-1} \tag{40}$$

where V_n is the unitary matrix (9), and L_{n1}, L_{n2} are block diagonal matrices corresponding to the two p-modifications of the spectral densities as specified by (23), (24). Thus, the mth diagonal block of $L_{n1}L_{n2}^{-1}$ is:

$$\{L_{n1}L_{n2}^{-1}\}_{mm} = F_{p1}(2\pi mn^{-1})\, F_{p2}^{-1}(2\pi mn^{-1}) \tag{41}$$

where

$$F_{ps}(\lambda) = \sum_{k=-p}^{p} (1-|k|/p)\, R_{ks} e^{-ik\lambda} \quad , \quad s = 1,2 \tag{42}$$

The kn eigenvalues of $\bar{R}_{n1}\bar{R}_{n2}^{-1}$ are thus asymptotically equally distributed to the union of eigenvalues of the n kxk matrices

$$\{F_{p1}(2\pi mn^{-1})F_{p2}^{-1}(2\pi mn^{-1}) \quad , \quad m = 1, \ldots, n\}$$

Let $h_q(p,2\pi mn^{-1})$ be the qth ordered eigenvalue of the matrix $F_{p1}(2\pi mn^{-1})F_{p2}^{-1}(2\pi mn^{-1})$. Using (33) with $g(x) = x$, we find:

$$\lim_{n,p\to\infty} n^{-1} \sum_{m=1}^{n} h_q(p,2\pi mn^{-1}) = (2\pi)^{-1} \int_0^{2\pi} h_q(u)du \tag{43}$$

where, $h_q(u)$ is the qth ordered eigenvalue of $F_1(u)F_2^{-1}(u)$. Summing over q, we have:

$$\lim_{n\to\infty} n^{-1} \text{trace}\, \bar{R}_{1n}\bar{R}_{2n}^{-1} = (2\pi)^{-1} \int_0^{2\pi} \text{trace}\, F_1(\lambda)F_2^{-1}(\lambda)d\lambda \tag{44}$$

Collecting terms, we find:

$$2J = (2\pi)^{-1} \int_0^{2\pi} \text{trace}[F_1(\lambda)F_2^{-1}(\lambda) + F_2(\lambda)F_1^{-1}(\lambda) - 2I]d\lambda \tag{45}$$

It is interesting that (45) can stand on its own as a distance measure. (45) has found applicability is speech processing, for measuring the

distance between two sounds in a subjectively meaningful way [19]. In [19], the scalar case k=1 was utilized.

Other applications of the two new distance measures are envisioned in EEG signal analysis [20]-[22]. It is also plausible that a distance measure of the type (43) or (38) may be a good clustering criterion in the space of spectral densities. If one wishes to "cluster" EEG's for the purpose of identifying "disease clusters", (38) and (43) may be useful measures due to the availability of spectral density estimates of EEG's [20]. Furthermore, the association of (38), (43) with the probability of misclassification, is an intuitively appealing factor.

It can be easily shown that $J(F_1,F_2)$ is convex in F_2 for fixed F_1, while B is neither convex nor concave. This is an advantage in favor of J. On the other hand, B provides better bounding expressions to the probability of misclassification than J does.

The geometry of J_n in the space of probability measures has been analyzed in [23], [24], and several convenient geometrical properties were established.

A criticism against B_n is that it is not a true distance measure because it does not satisfy the triangular condition. However, we shall show that there is a one-to-one correspondence of B_n to a proper distance measure.

Let

$$H_n = H_n(f_1(x^n),f_2(x^n)) = \{ \int [f_1^{1/2}(x^n) - f_2^{1/2}(x^n)]^2 dx^n \}^{1/2} \qquad (46)$$

H_n is the Hellinger distance [1] between $f_1(x^n)$, $f_2(x^n)$, and obviously satisfies the triangle condition. Furthermore,

$$2^{-1}H_n^2 = 1 - \int [f_1(x^n)f_2(x^n)]^{1/2}dx^n = 1 - \exp(-B_n) \tag{47}$$

Drawing further from the results of [14], we can show that:

$$\left| n^{-1}B_n - B \right| \leq k_3 n^{-1/2} \tag{48}$$

Thus,

$$\exp(-B_n) = b^n d_n^{-1}$$

where $b = \exp(-B)$, $d_n^{-1} = \exp(c_o n^{-1/2})$. Substituting in (45), we have:

$$b = d_n[1-2^{-1}H_n^2]^{1/n} \tag{49}$$

As $n \to \infty$, $d_n \to 1$, thus

$$b \cong [1-2^{-1}H_n^2]^{1/n} \tag{50}$$

Equations (47), (48) give a one-to-one correspondence between b and H_n. The function relating them is dependent on n, but has a simple structure. In Figure 1 the functional relationship is drawn.

Experts in the areas of feature selection and clustering techniques in Statistical Pattern Recognition have observed [25] - [29] that both methods are computationally expensive and exhibit strange behavior in high-dimensional measurement spaces. It is envisioned that the following two-step techniques may alleviate the above problems.

(I) Clustering

(Ia) Estimation of the spectral density $F_i(\lambda)$ from the ith observation record

$$\{x_i(k),\ k \geq 0\} \quad , \quad i = 1, \ldots, M$$

(Ib) Clustering the M records in the space of the measured $F_i(\lambda)$ and using as distance measures the numbers B or J.

(II) Feature Selection

(IIa) Same as (Ia). For M=2.

(IIb) Find the linear kxs, s < k transformation A that reduces the amount of data and maximizes J or B.

The point of view in proposing tha above methods is as follows. In practice, any time series $\{x_i(k),\ 1 \leq k \leq T\}$ observed for a long time interval T, exhibits statistical dependence only for pairs of samples that are neighbors in time. In clustering T-dimensional vectors by the usual techniques, a large T will impose substantial computing resource requirements. Furthermore, the convergence of clustering procedures may be in doubt, and the statistical independence between distant samples is not utilized. It is believed that the idea of fitting models to the data and then clustering the models according to J or B in a low-dimensional space may alleviate the stated difficulties.

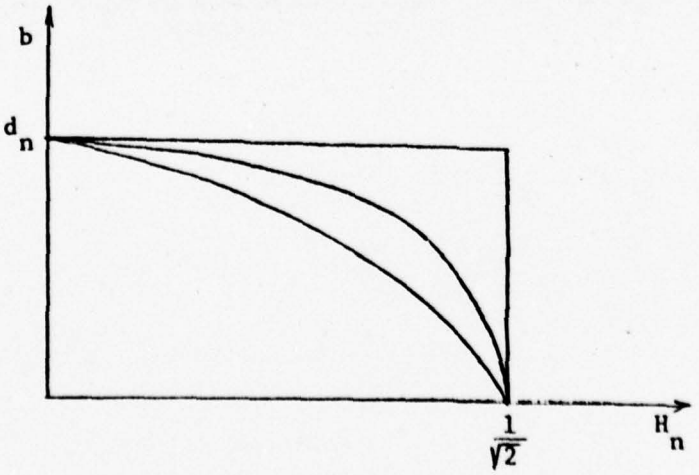

Figure 1

REFERENCES

[1] T. Kadota and L. A. Shepp, "On the Best Finite Set of Linear Observables for Discriminating Two Gaussian Signals", IEEE Trans. on Infor. Theory, Vol. IT-13, pp 278-285, April 1967.

[2] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection", IEEE Trans. on Comm. Tech, Vol. COM-15, pp 52-60, February, 1967.

[3] K. Matusita, "On the notion of Affinity of Several Distributions and Some of its Applications", Ann. Inst. Stat. Math., Vol. 19, pp 181-192, 1967.

[4] D. G. Lainiotis and S. K. Park, "Probability of Error Bounds", IEEE Trans. on Syst., Man, Cybern., Vol. SMC-7, pp 175-178, April 1971.

[5] G. T. Toussaint, "On Some Measures of Information and their Application to Pattern Recognition", in Proc. Conf. Measures of Information and their Application, Indian Inst. of Technology, Bombay, August 1974.

[6] J. W. Van Ness, "Dimensionality and Classification Performance with Independent Coordinates", IEEE Trans. on Syst., Man, Cybern., Vol. SMC-7, no. 7, pp 560-564, July 1977.

[7] E. Parzen, "On Empirical Multiple Time Series Analysis", Proc. Fifth Berkeley Symp. Math Statist. Prob. 1 (L. LeCam and J. Neyman, Eds.) Univ. of California Press, Berkeley, Calif., pp 305-340, 1967.

[8] E. Parzen, "Multiple Time Series Modeling", in Multivariate Analysis, II, (P. R. Krishnaiah, Ed.), Academic Press, pp 389-409, 1969.

[9] E. Parzen, "Some Recent Advances in Time Series Modeling", IEEE Trans, on Automatic Control, Vol. AC-19, no. 6, pp 723-729, December 1974.

[10] E. Parzen, "The Role of Spectral Analysis in Time-Series Analysis", Rev. Inst. Internat. Statist., Vol. 35, pp 125-141.

[11] W. Toms,"Information Rates of Dynamic Systems", Ph.D. Thesis at Cornell Unviersity, 1971.

[12] W. Toms and T. Berger, "Information Rates of Stchastically Driven Dynamic Systems", IEEE Trans on Infor. Theory, Vol. IT-17, no. 1, pp 113-114, January 1971.

[13] R. M. Gray, "Information Rates of Autogressive Processes", IEEE Trans. on Infor. Theory, Vol. IT-16, pp 412-421, July 1970.

[14] U. Grenander amd G. Szego. Toeplitz Forms And Their Applications. Berkeley California: Univ. of California Press, 1958.

[15] R. M. Gray, "On the Asymptotic Eigenvalue Distribution of Toeplitz Matrices", IEEE Trans. on Infor. Theory, Vol. IT-18, no. 6, pp 725-730, November 1972.

[16] R. M. Gray, "Toeplitz and Circulant Matrices: II", Technical Report no. 6504-1, Information Systems Lab., Stanford University, April 1977.

[17] J. Pearl, "On Coding and Filtering Stationary Signals by Discrete Fourier Transforms", IEEE Trans. on Infor. Theory, Vol. IT-19, no. 2, pp 229-232, March 1973.

[18] V. V. Fedorov, Theory of Optimal Experiments. Academic Press, 1972.

[19] A. H. Gray and J. D. Markel, "Distance Measures for Speech Processing", IEEE Trans. on Acoust. Speech, Sig. Proc., Vol. ASSP-24, no. 5, pp 380-391, October 1976.

[20] L. G. Pinson and D. G. Childers, "Frequency Wavenumber Spectrum Anaysis of EEG Multielectrode Array Data", IEEE Trans. on Biomedical Eng., Vol. BME-21, pp 192-206, May 1974.

[21] W. Gersh and J. Yonemoto, "Automatic Classification of EEGS: A Parametric Model. New Features for Classification Approach", Proceedings of the 1977 Joint Automatic Control Conference.

[22] W. Gersh and J. Yonemoto, "Parametric Time Series Models for Multivariate EEG Analysis", Computers and Biomedical Research, Vol. 10, pp 113-125, 1977.

[23] I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems", Ann. Prob., Vol. 3, no. 1, pp 146-158, February 1975.

[24] R. Blahut, "The Geometry of Probability Space", IEEE International Symposium of Information Theory, Ithaca, New York, and unpublished manuscript, October 1977.

[25] J. Tou and R. P. Heydorn, "Some Approaches to Optimum Feature Extraction", in Computer and Information Sciences, Vol. 2, J. T. Tou Ed., Academic Press, New York, 1967.

[26] T. L. Henderson and D. G. Lainiotis, "Application of State-Variable Techniques to Optimal Feature Extraction-Multichannel Analog Data", IEEE Trans. on Infor. Theory, Vol. IT-16, no. 4, pp 396-406, July 1970.

[27] K. Fukunaga, Personal Communication during the 1977 IEEE International Symposium on Information Theory.

[28] W. L. G. Koontz, P. M. Narendra, and K. Fukunaga, "A Graph Theoretic Approach to Nonparametric Cluster Analysis", IEEE Trans. on Computer, Vol. C-25, no. 9, pp 936-944, September 1976.

[29] P. Papantoni-Kazakos, R. Li, and D. Kazakos, "Linear Dimensionality Reduction of Stationary Vector Gaussian Processes", Fifteenth Annual Allerton Conference on Communications, Control, and Computing, September 1977.

Bell Laboratories

Cover Sheet for Technical Memorandum

The information contained herein is for the use of employees of Bell Laboratories and is not for publication (see GEI 13.9-3)

Title- Robustness of Estimators on Stationary Observation‡

Date- September 20, 1977

TM- 77-3452-7

Other Keywords- Generalized Ornstein distance

Author(s)	Location and Room	Extension
P. Papantoni-Kazakos*	HO 2B-402	5347
Robert M. Gray†		

Charging Case- 49008-61

Filing Case- 49008

ABSTRACT

Hampel's general qualitative definition of robustness of sequences of estimators on memoryless observation processes is generalized to stationary ergodic processes by substituting the generalized Ornstein (or $\bar{\rho}$) distance for the marginal Prohorov distance as the measure of "closeness" of observations. More general sequences of estimators are allowed. The approach yields results analagous to those of Hampel for the more general case considered, often provides strict generalizations of Hampel's results, and in some cases yields simpler proofs.

---

*P. Papantoni-Kazakos completed this work at Rice University, Houston, Texas.

†R. M. Gray is with Stanford University.

‡This research was supported by the Air Force Office of Scientific Research under AFOSRA Grant 77-3156 and AFOSR Contract F44 620-73-C-0065.

Pages Text 32 Other 2 Total 34

No. Figures - No. Tables - No. Refs. 15

BELL TELEPHONE LABORATORIES, INC.

TM-77-3452-7

DISTRIBUTION
(REFER GEI 13.9-3)

COMPLETE MEMORANDUM TO

CORRESPONDENCE FILES

OFFICIAL FILE COPY
PLUS ONE COPY FOR
EACH ADDITIONAL FILING
CASE REFERENCED

DATE FILE COPY
(FORM E-1328)

10 REFERENCE COPIES

ADLEMAN,RICHARD
AHERN,W C
AMSTER,JAMES M
ANTONIAK,CHARLES E
•BUCHNER,MORGAN M,JR
COZINE,CLAIRE D
DESCLOUX,A
•DORROS,IRWIN
DUFF,CAROLINE M
DUNCAN,D P
ESTBERG,DONALD G
FORYS,L J
FREDERICKS,A A
FUHRMANN,STEVE WAYNE
HATCH,R W
•HAYWARD,W S,JR
HEALY,J D
HOLTZMAN,JACK M
HORING,S
KAZAKOS,PANAYOTA
LA MACCHIA,J T
LAUE,RICHARD V
LOPIPARO,PETER
MAHOOD,G K
MARLOW,NORMAN A
MC CUMBER,D E
MERCER,ROBERT A
MESSERLI,E J
<REHERT,ALLEN F
REISNER,G A
RUPPEL,A F
SILBIGER,HERMAN R
SULLIVAN,J L
SULLIVAN,MARY A
SYKES,JACK S
TESTAVERDE,JANET
•WEBER,JOSEPH H
•WILSON,M P
WINTHROP,JOEL A
•ZYDNEY,HERBERT M
40 NAMES

COVER SHEET ONLY TO

CORRESPONDENCE FILES

4 COPIES PLUS ONE
COPY FOR EACH FILING
CASE

AARON,M R
ABED,M I
ABRAHAM,S A
ACAMPORA,A S
AHO,ALFRED V
ALTERMAN,MICHAEL E
AMRON,IRVING
ANDERSON,E J
ANDERSON,L G
ARNOLD,GEORGE W
ARREDONDO,G A
ARTHURS,E
ATAL,BISHNU S
BACH,MAURICE J
BALL,MARSHALL
BARNETT,W T
BASSETT,ROBERT M
BECKER,RICHARD A
BENES,V E
BENNING,R D
BERGLAND,G D
BHARUCHA,BEHRAM H
BOBILIN,R T
BODNER,H A
BOGERT,BRUCE P
BOYCE,W M
BOZA,L B
BRANDENBURG,L H
BRELSFORD,WILLIAM M
BROWN,A B,JR
BROWN,W STANLEY
BRUSH,GARY G
BUCZNY,F A,JR
BULFER,A F
BURKE,P J
BUSHNELL,W J
BYER,TREVOR C
CALDWELL,W NEAL
CAREY,J H
CARLIN,JAMES W
CARROLL,J DOUGLAS
CASTELLANO,MARY ANN
CAVINESS,JOHN D
CHADDHA,R L
CHAMBERS,J M
CHANG,J K
CHAO,MIN-TE
CHICK,ARTHUR J
CHIEN,TA-MU
CHILDS,CAROLYN

COVER SHEET ONLY TO

CHRISTENSEN,KENT R
CHU,PAUL H N
CLAYTON,D P
CLEVELAND,WILLIAM S
COCHRANE,J I
COCHRAN,JAMES A
COHEN,HARVEY
<COLE,LOUIS M
COLTON,JOHN R
COOPER,C A
CYGANSKI,DAVID
DALEY,WILLIAM E
DAVIDSON,CHARLES LEWIS
DAVID,ALEXANDER J
DAVIS,D R
DE FAZIO,M J
DE LESSIO,N K
DEUTSCH,DAVID N
DEVLIN,SUSAN J
DI GIACOMO,J G
DOMBROWSKI,F J
DORFMAN,PHILIP J
DUNN,DOUGLAS M
DUTTWEILER,D L
ECKBERG,A E
ECKLER,A ROSS
EDDY,T W
EISENBERG,MARTIN
EVANS,MELVIN J
EVERHART,J R
FAULHABER,GERALD R
FIKE,ROBERT L
FLEISCHER,H I
FONTENOT,MICHAEL L
<FORD,G A
FOSCHINI,G J
FOWLKES,EDWARD B
FRANKS,RICHARD L
FRANK,AMALIE J
FRANK,RUDOLPH J
FREEMAN,K G
FRETWELL,L J,JR
FRIDAY,DENNIS S
FRIEDMAN,KENNETH A
FROEHLICH,F E
FUCHS,EDWARD
GABBE,JOHN D
GALE,WILLIAM A
GAREY,MICHAEL R
GAY,FRANCIS A
GIBB,KENNETH R
GIBSON,ALLEN E
GIFFELS,CARL A
GILBERT,E N
GITLIN,R D
GNANADESIKAN,R
GODFREY,A BLANTON

COVER SHEET ONLY TO

GOLABEK,RUTH T
GOLDMAN,BARRY M
GOLDMAN,JOEL
GOLDSTEIN,A JAY
GOPINATH,B
GOTZ,BEN
GRACZYK,J F
GRAEF,R P
GRAHAM,R L
GRAU,T G
GRAYSON,C F,JR
GREENBAUM,HOWARD J
GROFF,R H
GROSS,ALAN M
GRUCHOWSKI,M P
GUANCIAL,EDWARD,JR
GUNTHER,F L
GUREY,STEPHEN M
GUST,V J
GUTHERY,S B
GUTSCHERA,K D
HAISCH,H F,JR
HALFIN,S
HALL,H M,JR
HALL,M A
<HALL,MILTON S,JR
HALL,WILLIAM G
HANDLER,GARY J
HANKINS,RICHARD W
HAREL,DANY
HASKELL,BARRY G
HAUPT,W PAUL
HAUSE,A DICKSON
HAWKINS,RICHARD B
HAYDEN,DONALD F,JR
HEFFES,H
HEJNY,GEORGE J
HEYMAN,DANIEL P
HIGHLAND,P A,JR
HINDELITER,R G
HIRSCH,D
HOADLEY,BRUCE
HOOKE,JOHN A
HORENKAMP,JOHN J
HOU,TIEN-FANG
<HO,TIEN-LIN
HWANG,FRANK KWANGMING
JACKSON,ROBERT A
JACOBS,H S
JACOBS,IRA
JAGERMAN,D
JAIN,ARIDAMAN KUMAR
JAKUBEK,R J
JENSEN,BLISS D
JOHNSON,KEITH W
JULESZ,BELA
KADOTA,T T

COVER SHEET ONLY TO

KAISER,J F
KALBO,A L
KAMMIRE,C A
KANTNER,HOMER E
KAPLAN,DAVID L
KAPPEL,JOSEPH G
KARR,M A
KATCHINOFF,ELLEN
KATTKE,G W
<KATZ,M J
KAUFMAN,J S
KAYEL,R G
KEARNS,TIMOTHY J,JR
KEILIN,JOSEPH E
KEINATH,R C
KERNIGHAN,BRIAN W
KERSTEN,PAUL R
KESSLER,JAMES E
KETTENRING,JON R
KLANCER,H W
KLEINER,BEAT
KNOLL,RONALD L
KOBRAN,ALAN S
KORBE,WILLIAM P,II
KOSMAN,ROBERT A
KRAMER,S A
KRANZMANN,R F
KRISHNAN,K R
KRUSKAL,JOSEPH B
KYLIN,JOHN C
LAMBERT,CONSTANCE A
LAMBERT,PETER F
LAMB,LARRY C
LANDWEHR,JAMES M
LANZEROTTI,L J
LENAHAN,TERRENCE A
LEVINSON,STEPHEN E
LEWINE,R N
LEWINSKI,D A
LIEBESMAN,B S
LIEN,MONTE D
LIMB,JOHN O
LIM,Y S
LIN,SING H
LIOU,MING L
LIU,K S
LUCKY,R W
LUDERER,GOTTFRIED W R
LUDWIG,JAMES J
LUDWIG,R L
LUM,M P
LUSS,HANAN
LUTZ,KENNETH J
LYCKLAMA A NYEHOLT,H
MAC DOUGALL,LAURENCE L
MACOCK,DEBORAH Y
MAGEE,FRANCIS R,JR
...
391 TOTAL

• NAMED BY AUTHOR > CITED AS REFERENCE < REQUESTED BY READER (NAMES WITHOUT PREFIX WERE SELECTED USING THE AUTHOR'S SUBJECT OR ORGANIZATIONAL SPECIFICATION AS GIVEN BELOW)

MERCURY SPECIFICATION.....

COMPLETE MEMO TO:
345-SUP 3452

COVER SHEET TO:
MAPR* = MATHEMATICS/PROBABILITY/SURVEY PAPERS ONLY
MAPRIT = MATHEMATICS/PROBABILITY/INFORMATION THEORY
STPR* = STATISTICS/PROBABILITY/SURVEY PAPERS ONLY
STPRIT = STATISTICS/PROBABILITY/INFORMATION THEORY
STTH* = STATISTICS/THEORY/SURVEY PAPERS ONLY
STTHCM = STATISTICS/THEORY/COMMUNICATION THEORY

HO CORRESPONDENCE FILES
HO 5C101

TM-77-3452-7
TOTAL PAGES 34

TO GET A COMPLETE COPY:

1. BE SURE YOUR CORRECT ADDRESS IS GIVEN ON THE OTHER SIDE.
2. FOLD THIS SHEET IN HALF WITH THIS SIDE OUT AND STAPLE.
3. CIRCLE THE ADDRESS AT RIGHT. USE NO ENVELOPE.
4. INDICATE WHETHER MICROFICHE OR PAPER IS DESIRED.

PLEASE SEND A COMPLETE

( ) MICROFICHE COPY ( ) PAPER COPY

TO THE ADDRESS SHOWN ON THE OTHER SIDE.

Bell Laboratories

subject: Robustness of Estimators on Stationary Observation‡ - Case 49008

date: September 20, 1977

from: P. Papantoni-Kazakos*
R. M. Gray†
TM77-3452-7

MEMORANDUM FOR FILE

1. Introduction

In his classic paper, Hampel (1971) introduced a definition of robustness in parameter estimation that accurately reflected the intuitive notion that a sequence of estimates of a parameter was robust for an observation process μ if another process ν that was "close" to μ yielded a "close" distribution on the parameter estimates. Hampel considered memoryless or independent, identically distributed (i.i.d.) observation processes and measured their "closeness" by the Prohorov distance on the marginal probability measures. As he considered i.i.d. processes, his underlying parameter depended implicitly only on these unknown marginals. Hampel then proved that weak* continuous functionals on the space of probability distributions defined robust sequences of estimators under his assumptions. He also showed his results could be adapted via an alternative notion of robustness to weakly dependent observations, in particular, observations that were close to memoryless in a Prohorov sense.

A critical part of his derivation was the fact that if two i.i.d. processes μ and ν are close in a marginal Prohorov sense, then one could construct a pair process p having μ and ν as coordinate processes and such that under p the sample distributions of two coordinate n-tuples x^n produced by ν and y^n produced by ν were close in a Prohorov sense with high probability.

*P. Papantoni-Kazakos completed this work at Rice University, Houston, Texas.
†R. M. Gray is with Stanford University.
‡This research was supported by the Air Force Office of Scientific Research under AFOSRA Grant 77-3156 and AFOSR Contract F44 620-73-C-0065.

During the past few years, a generalization of Ornstein's $\bar{d}$ distance of ergodic theory (called the $\bar{\rho}$, "rho-bar," or generalized Ornstein distance) has been shown to provide a similar control for sample distributions for general stationary and ergodic processes and, largely as a result, has found numerous applications in information theory (see, e.g., Gray, Neuhoff and Shields (1975), Gray, Neuhoff, and Onura (1975), Gray, Neuhoff and Ornstein (1975)). In this paper we show that using the $\bar{\rho}$ distance as a measure of closeness of the observation processes, there is a natural qualitative definition of robustness for all stationary ergodic processes, that a weakened version of Hampel's weak-continuous estimator sequence implies robustness and that all of Hampel's results have analogs in this more general case. Our formalism does not quite contain Hampel's in the case of i.i.d. processes and parameters depending only on the marginal probabilities, but is a strict generalization in some cases such as when the metric on the observation alphabet is bounded or when the class of probability measures considered is constrained to have a finite second moment (see Lemma 2.1).

We also note that we need not confine estimates to take values in R^k as Hampel does, but instead we only require that the parameter alphabet be a complete, separable metric (Polish) space. Hence function valued parameter spaces are allowed.

As a side result, some easy generalizations of the convergence of sample distributions [Parthasarathy (1967)] for stationary and ergodic processes are developed.

2. Preliminaries

Let $(\Omega, \mathcal{B}_\Omega)$ be a measurable space such that Ω is a complete, separable metric space (or Polish space) with metric ρ and $\mathcal{B}_\Omega$ is the Borel σ-field generated by the open sets under ρ. Since Ω is separable, there is a countable collection of sets $\mathcal{G}_\Omega = \{G_i;\ i = 1,2,\ldots\}$ such that $\mathcal{B}_\Omega = \sigma(\mathcal{G}_\Omega)$, that is, $\mathcal{B}_\Omega$ is the σ-field generated by $\mathcal{G}_\Omega$. Let Ω^n be the space of n-tuples with coordinates in Ω and Ω^∞ the space of sequences $\omega = (\ldots,\omega_{-1},\omega_0,\omega_1,\ldots)$, $\omega_i \in \Omega$ all i. Let $\mathcal{B}_\Omega^n$ be the σ-field of subsets generated by all rectangles of the form $\prod_{i=0}^{n-1} B_i$, $B_i \in \mathcal{B}_\Omega$ (since Ω is Polish $\mathcal{B}_\Omega^n = \sigma(\mathcal{G}_\Omega^n)$, the σ-field generated by rectangles with $B_i \in \mathcal{G}_\Omega$). Let $\mathcal{B}_\Omega^\infty$ be the σ-field generated by all rectangles of the form $B = \{\omega:\omega_i \in B_i,\ n \le i \le m\}$, $B_i \in \mathcal{B}_\Omega$. Let μ be a probability measure on the measurable space $(\Omega^\infty, \mathcal{B}_\Omega^\infty)$ yielding a probability space $(\Omega^\infty,\mathcal{B}_\Omega^\infty,\mu)$. The sequence of coordinate functions $X_n:\Omega^\infty \to \Omega$ defined by $X_n(\omega) = \omega_n$, $n = \ldots,-1,0,1,\ldots$ on $(\Omega^\infty,\mathcal{B}_\Omega^\infty,\mu)$ forms a random process and is denoted either by $[\Omega, \mu, X]$ to emphasize alphabet Ω, measure μ, and name X, or simply by μ to emphasize measure, or by $\{X_n\}$ to emphasize name.

Let $T:\Omega^\infty \to \Omega^\infty$ denote the shift transformation defined by $X_n(T\omega) = X_{n+1}(\omega)$. The process μ is stationary if $\mu(TF) = \mu(F)$ for all $F \in \mathcal{B}_\Omega^\infty$. The process is ergodic if $TF = F$ implies $\mu(F) = 0$ or 1.

Denote $(\omega_0,\ldots,\omega_{n-1})$ by ω^n and define $X^n:\Omega^\infty \to \Omega^n$ by $X^n(\omega) = (X_0(\omega), X_1(\omega),\ldots,X_{n-1}(\omega)) = \omega^n$. Let μ^n denote the restriction of μ to $(\Omega^n,\mathcal{B}_\Omega^n)$, that is, if $F \in \mathcal{B}_\Omega^n$, then $\mu^n(F) = \mu(X^n)^{-1}(F) = \mu(\omega:\omega^n \in F)$.

Let $\mathcal{m}_s$ denote the class of all stationary processes with alphabet Ω and let $\mathcal{m}_e$ denote the class of all stationary and ergodic processes with alphabet Ω. To avoid confusion we will often use different names with different measures, e.g., typical members of $\mathcal{m}_e$ are $[\Omega, \mu, X]$ and $[\Omega, \nu, Y]$.

A process $[\Omega, \mu, X]$ is said to be i.i.d. if for every rectangle $B = \times_{i=0}^{n-1} B_i$, $B_i \in \mathcal{B}_\Omega$, we have $\mu^n(B) = \prod_{i=0}^{n-1} \mu^1(B_i)$. Let $\mathcal{m}_m$ denote the collection of all i.i.d. or memoryless processes and note that $\mathcal{m}_m \subset \mathcal{m}_e \subset \mathcal{m}_s$.

Given two processes $\mu, \nu \in \mathcal{m}_s$ the generalized Ornstein distance or $\bar{\rho}$ distance between μ and ν can be defined as follows: for $x^n, y^n \in \Omega^n$ set

$$\rho_n(x^n, y^n) = n^{-1} \sum_{i=0}^{n-1} \rho(x_i, y_i)$$

and define $\mathcal{P}(\mu^n, \nu^n)$ as the set of all measures p^n on $(\Omega^n \times \Omega^n, \mathcal{B}_\Omega^n \times \mathcal{B}_\Omega^n)$ having μ^n and ν^n as coordinates, that is, $p^n(\Omega^n \times F) = \nu^n(F)$, $p^n(F \times \Omega^n) = \mu^n(F)$, all $F \in \mathcal{B}_\Omega^n$. Define the n^{th} order distance

$$\bar{\rho}_n(\mu^n, \nu^n) = \inf_{p \in \mathcal{P}(\mu^n, \nu^n)} E_p \rho_n \tag{2.1}$$

and the $\bar{\rho}$ distance by

$$\bar{\rho}(\mu, \nu) = \sup_n \bar{\rho}_n(\mu^n, \nu^n) \quad . \tag{2.2}$$

If with a slight abuse of notion we also let X^n and Y^n denote coordinate functions on $\Omega^n \times \Omega^n$ so that if $z = (x^n, y^n) \in \Omega^n \times \Omega^n$, then $X^n(z) = x^n$, $Y^n(y) = y^n$, then (2.1) also can be written

$$\bar{\rho}_n(\mu^n, \nu^n) = \inf_{p \in \mathcal{P}(\mu^n, \nu^n)} E_p \rho_n(X^n, Y^n)$$

Thus $\bar{\rho}_n(\mu^n,\nu^n)$ measures the smallest possible expected "distortion" between X^n and Y^n over all stochastic links preserving the probabilistic description of each. We note $\bar{\rho}_n$ is the Vasershtein-distance between the random vectors X^n and Y^n described by μ^n and ν^n [Vasershtein, (1969)]. The following are some useful properties of $\bar{\rho}$ for later use.

Properties of $\bar{\rho}$ [Gray, et. al., (1975)]:

(i) $\lim_{n\to\infty} \bar{\rho}_n(\mu^n,\nu^n)$ exists and equals $\sup_n \bar{\rho}_n(\mu^n,\nu^n)$.

(ii) If μ and ν are i.i.d., then $\bar{\rho}(\mu,\nu) = \bar{\rho}_1(\mu',\nu')$.

(iii) $\bar{\rho}(\mu,\nu) \le \bar{\rho}(\mu,\eta) + \bar{\rho}(\eta,\nu)$ (triangle inequality).

(iv) The distance can also be defined as follows: Let $\mathcal{P}_s(\mu,\nu)$ be the collection of all stationary pair processes with coordinate processes μ and ν, that is, all measures p on $(\Omega^\infty\times\Omega^\infty, \mathcal{B}_\Omega^\infty\times\mathcal{B}_\Omega^\infty)$ such that $p(\Omega^\infty\times F) = \nu(F)$, $p(F\times\Omega^\infty) = \mu(F)$, all $F\in\mathcal{B}_\Omega^\infty\times\mathcal{B}_\Omega^\infty$ (where we use T to denote the shift on $\Omega^\infty\times\Omega^\infty$ as well as on Ω^∞). In a similar fashion let $\mathcal{P}_e(\mu,\nu)$ denote the class of all stationary and ergodic pair processes with μ and ν as coordinates. Define the coordinate functions $(X_n,Y_n):\Omega^\infty\times\Omega^\infty\to\Omega\times\Omega$ by $(X_n,Y_n)(x,y) = (X_n(x),Y_n(y)) = (x_n,y_n)$. We have that

$$\bar{\rho}(\mu,\nu) = \inf_{p\in\mathcal{P}_s(\mu,\nu)} E_p\rho(X_0,Y_0) \qquad (2.3a)$$

and if $\mu, \nu \in \mathcal{M}_e$,

$$\bar{\rho}(\mu,\nu) = \inf_{p \in \mathcal{P}_e(\mu,\nu)} E_p \rho(X_0, Y_0) \tag{2.3b}$$

We note that (2.3b) follows from (2.3a) via the ergodic decomposition of stationary processes [see Oxtoby (1952) or Rohlin (1949)].

Another important property of $\bar{\rho}$ is that it is the closest that generic (typical, regular) sequences of μ and ν (those sequences whose sample averages converge to expectations of enough functions to determine the measure) can be made to each other in a limiting ρ_n sense [Gray, et. al., (1975)]. In the next section we develop a result for sample distributions similar to that of Hampel and Parthasarathy since the existing $\bar{\rho}$ result is not directly useful here because it involves a different type of sample average. The basic idea is that $\bar{\rho}$ closeness of two processes will imply that with high probability the process will produce close sample distributions.

Hampel used the Prohorov metric between μ^1 and ν^1 to measure the distance between i.i.d. processes μ and ν. We can define a Prohorov distance between processes using a generalization of Moser, et. al., (1975) and this distance can be easily related to $\bar{\rho}$ by using the Strassen-Dudley form for the Prohorov distance [Strassen (1965), Dudley (1968)]: Define the n^{th} order Prohorov distance

$$\Pi_n(\mu^n,\nu^n) = \inf_{p \in \mathcal{P}(\mu^n,\nu^n)} \inf\{\gamma : p(x^n,y^n : \rho_n(x^n,y^n) > \gamma) \leq \gamma\}, \tag{2.4}$$

which is the Prohorov metric between μ^n and ν^n with respect to the metric ρ_n (which generates the product topology), and

$$\Pi(\mu,\nu) = \sup_n \Pi_n(\mu^n,\nu^n) \quad . \tag{2.5}$$

It is known [Strassen (1965), Dudley (1968)] that a p_n achieving the infimum exists. We have immediately using Chebychev's inequality [as in Dobrushin (1970)] that if p^n achieves $\bar{\rho}_n$ (i.e., $E\, p_n\rho_n = \bar{\rho}_n$, in the Appendix it is shown that the infimum is a minimum for Polish alphabets), then

$$p^n(x^n,y^n:\rho_n(x^n,y^n) > \epsilon) \le E\, p^n\rho_n/\epsilon \le \bar{\rho}_n(\mu^n,\nu^n)/\epsilon$$

and hence choosing $\bar{\rho}_n(\mu^n,\nu^n) = \epsilon^2$ yields

$$p^n(x^n,y^n:\rho_n(x^n,y^n) > \bar{\rho}_n(\mu^n,\nu^n)^{1/2}) \le \bar{\rho}_n(\mu^n,\nu^n)^{1/2}$$

whence

$$\Pi_n(\mu^n,\nu^n)^2 \le \bar{\rho}_n(\mu^n,\nu^n)$$

$$\Pi(\mu,\nu)^2 \le \bar{\rho}(\mu,\nu) \tag{2.6}$$

so that closeness in $\bar{\rho}$ is stronger than closeness in Prohorov. In some cases the two distances generate the same topology, however, as the following easy Lemma shows.

Lemma 2.1

(a) If ρ is bounded, then $\bar{\rho}_n$ and Π_n generate the same topology (and hence so do $\bar{\rho}$ and Π).

(b) Given a class $\mathcal{M}$ of processes μ such that there exists an a^* such that $E\,\mu\rho(X_0,a^*)^2 \le \rho^* < \infty$, then $\bar{\rho}_n$ and Π_n generate the same topology on $\mathcal{M}$.

Proof.

(a) Let ρ_{max} be the largest value of ρ, then if p^n yields Π_n we have

$$Ep_n\rho_n(X^n,Y^n) \le \Pi_n(\mu^n,\nu^n) + p^n(x^n,y^n:\rho_n(x^n,y^n) > \Pi_n(\mu^n,\nu^n))\rho_{max} = \Pi_n(\mu^n,\nu^n)(1+\rho_{max})$$

and hence small Π_n implies small $\bar{\rho}_n$ which with (2.6) proves (a).

(b) We have similar to before that

$$Ep_n\rho_n(X^n,Y^n) \le \Pi_n(\mu^n,\nu^n) + \int dp^n(x^n,y^n)\rho_n(x^n,y^n)$$

$$x^n,y^n:\rho_n(x^n,y^n)>\Pi_n(\mu^n,\nu^n)$$

Let $G = \{x^n,y^n:\rho_n(x^n,y^n) > \Pi_n(\mu^n,\nu^n)\}$ and let 1_G be the indicator function for G. Since ρ_n is a metric, we have from the triangle inequality and the Couchy-Schwartz inequality

$$E\ p_n\rho_n(X^n,Y^n) \le \Pi_n(\mu^n,\nu^n)+E\ p_n\rho_n(X^n,a^{*n})1_G+E\ p_n\rho_n(Y^n,a^{*n})1_G$$

$$\le \Pi_n(\mu^n,\nu^n)+(E\ p_n\rho_n(X^n,a^{*n})^2)^{1/2}(E\ p_n1_G^2)^{1/2}$$

$$+ (E\ p_n\rho_n(Y^n,a^{*n})^2)^{1/2}(E\ p_n1_G^2)^{1/2}$$

$$\le \Pi_n(\mu^n,\nu^n) + 2\rho^{*1/2}p_n(G),$$

$$= \Pi_n(\mu^n,\nu^n)(1+2\rho^{*1/2})$$

completing the proof as before.

We use $\bar{\rho}$ and not Π as a distance measure on observation processes for several reasons, primarily because $\bar{\rho}$ has several properties useful for robustness (and other) studies that Π does not. In particular, (1) the $\sup_n \Pi_n$ need not be achieved in the limit $n \to \infty$ as is $\bar{\rho}$. As a result there is no process definition for $\Pi = \Pi(\mu,\nu)$ analogous to (2.3). This means there need <u>not</u> exist a single stationary p such that $p(x,y:\rho_n(x^n,y^n) > \Pi) \leq \Pi$ for all n. The p yielding $\bar{\rho}$, however, guarantees that $\bar{\rho}(\mu,\nu) = E_p\, \rho(X_0,Y_0) = E_p\, \rho_n(X^n,Y^n)$ and hence via Chebychev's inequality it is true that $p(x,y:\rho_n(x^n,y^n) > \bar{\rho}^{-1/2}) \leq \bar{\rho}^{-1/2}$ for all n. This uniform bound for all n is crucial to prove robustness. (2) If μ and ν are i.i.d., then $\bar{\rho} = \bar{\rho}_1$ and hence marginal closeness of $\bar{\rho}_1$ in such a case guarantees process closeness of $\bar{\rho}$. The analog is not true for Prohorov, that is, it is not true for μ,ν i.i.d. that $\Pi(\mu,\nu) \approx \Pi_1(\mu^1,\nu^1)$. It need not even be true that given $\epsilon > 0$ there exists a δ such that $\Pi_1(\mu^1,\nu^1) < \delta$ implies $\Pi(\mu,\nu) < \epsilon$. Thus marginal closeness of Prohorov does <u>not</u> ensure process closeness for i.i.d. processes. As a result, using $\Pi(\mu,\nu)$ as a closeness notion would <u>not</u> be a strict generalization of Hampel's definition of robustness for i.i.d. processes. (3) The $\bar{\rho}$ distance between processes can often be explicitly evaluated or bounded (as in the Gaussian case) making it useful in applications. No general bounds to Π (except in terms of $\bar{\rho}$ via (2.6)) exist. (4) It is $\bar{\rho}$ and not Π that allows a simple demonstration that close processes likely produce sample functions with close sample distributions (as in the next section). Hampel's Porohorov approach worked in the i.i.d. case because he was able to produce an i.i.d. pair process p with the correct coordinates by

simply taking the i.i.d. process with the marginal yielding $\Pi_1(\mu^1,\nu^1)$. If μ and ν were not i.i.d., p constructed in this way would not have μ and ν as coordinates. The $\bar{\rho}$ avoids this problem since it has an equivalent definition in terms of processes.

As a final observation, one could also define a Prohorov distance on processes via

$$\rho_\infty(x,y) = \sum_{i=-\infty}^{\infty} 2^{-|i|}\, \rho(x_i,y_i)/(1+\rho(x_i,y_i))$$

$$\Pi'(\mu,\nu) = \inf_{p \in \mathcal{P}_s(\mu,\nu)} \inf\{r: p(x,y:\rho_\infty(x,y) > r) \le r\}$$

This distance generates the weak topology on $\mathcal{P}_s$, but it is of limited use because if "favors" times near zero in determining the metric ρ_∞. It is the limiting behavior of $n^{-1}\sum_{i=0}^{n-1} \rho(x_i,y_i)$ and not ρ_∞ that is important in most applications (such as robustness and problems in information theory). In particular, small $\bar{\rho}$ will be seen to force $n^{-1}\sum_{i=0}^{n-1} \rho(x_i,y_i)$ to be small for all n with high probability, Π' is not "strong" enough to imply this.

Even though we have argued that $\bar{\rho}$ is the appropriate distance measure on processes, the Prohorov metric is quite adequate as a measure of distance of random variables, and hence for many intermediate steps we will use the weaker Prohorov distance to follow Hampel's basic approach where possible.

3. Sample Distributions

Hampel (1975) following Parthasarathy (1967) considers only marginal sample distributions of the following kind: Given an n-tuple $x^n \in \Omega^n$, define the measure $\mu^1_{x^n}$ on $(\Omega,\mathcal{B}_\Omega)$ by assigning probability n^{-1} to each x_i, $i = 0,1,\ldots,n-1$ (if, say, k of the x_i are identical, this point gets probability k/n). This assignment gives a measure $\mu^1_{x^n}$ on $(\Omega,\mathcal{B}_\Omega)$, via

$$\mu^1_{x^n}(F) = \sum_{i:x_i \in F} n^{-1}$$

Parthasarathy (1967) proves that for an i.i.d. process μ,

$$\Pi_1(\mu^1_{x^n},\mu^1) \underset{n\to\infty}{\to} 0 \quad , \quad \mu\text{-a.e.} \tag{3.1}$$

We shall wish to consider more general processes and parameters depending on the whole process and not just the marginal μ^1. Hence we wish to estimate more than just the marginal μ^1 from x^n. Given an n-tuple $x^n \in \Omega^n$ form an estimate of the entire underlying process as follows: Form the periodic string $\bar{x} = (\ldots,x^n,x^n,x^n,\ldots)$, that is, $\bar{x}_k = x_{k \bmod n}$ Define the measure μ_{x^n} on $(\Omega^\infty,\mathcal{B}^\infty_\Omega)$ by placing probability n^{-1} on each string $T^i\bar{x}$, $i = 0,1,\ldots,n-1$ (grouping together identical strings as before), that is,

$$\mu_{x^n}(F) = \sum_{i:T^i\bar{x} \in F} n^{-1}, \quad \text{all } F \in \mathcal{B}^\infty_\Omega \tag{3.2}$$

The process is periodic as defined by Parthasarathy (1961) since $\mu_{x^n}(F \cap T^nF) = \mu_{x^n}(F)$, all $F \in \mathcal{B}^\infty_\Omega$. It is also easily seen to be

stationary from (3.2). Furthermore, if $TG = G$ and hence $T^{-1}G = G$, then if $T^{i}\bar{x} \in G$ for any i, then $T^{j}\bar{x} \in G$ for all j and hence $\mu_{x^n}(G) = 0$ or 1 and the process is ergodic. The process μ_{x^n} has restrictions $\mu_{x^n}^{k}$ which assign measure n^{-1} to each k-tuple obtained by viewing k adjacent symbols within x^n or an "overlap" k-tuple constructed by $(x_i,\ldots,x_{n-1},x_0,\ldots,x_{k+i-n})$, $i = n-k+1,\ldots,n-1$. In particular, $\mu_{x^n}^{1}$ is the same as the Parthasarathy marginal sample distribution. Note that only if $k \leq n$ are the sample distributions "trustworthy," but it is in fact the sample distributions $\mu_{x^n}^{k}$, $n \geq k$, that will be most important. This raises an alternate (and more common) approach of given x^n, define the restrictions (and not a process) $\mu_{x^n}^{k}$ by assigning $(n-k)^{-1}$ to each of the $(n-k)$ k-tuple within x^n. We do not take this approach since (1) it is useful to have a process implying all the restrictions; (2) it is convenient to have n^{-1} be the probability of the atoms for all k and the resulting proofs are simpler; and (3) properties of periodic processes make it easy to demonstrate that a certain seemingly reasonable conjecture is in fact false. The two approaches obviously yield identical results for fixed k and large n since the overlap effects die out as $n \to \infty$.

The main result of this section is the following generalization of (3.1).

Lemma 3.1

If $[\Omega,\mu,X]$ is stationary and ergodic, then for any fixed k

$$\lim_{n\to\infty} \Pi_k(\mu_{x_n^k},\mu^k) = 0 \quad , \qquad \mu\text{-a.e.} \tag{3.3}$$

If, in addition, there exists a reference letter a^* such that

$$E_\mu \rho(X_0,a^*) \le \rho^* < \infty \quad , \tag{3.4}$$

then for any fixed k

$$\lim_{n\to\infty} \bar{\rho}_k(\mu_{x_n^k},\mu^k) = 0 \quad , \qquad \mu\text{-a.e.} \tag{3.5}$$

Proof.

For any $G \in \mathcal{B}_\Omega^n$ the Birkhoff ergodic theorem states that with μ-probability one

$$\mu_{x_n^k}(G) = n^{-1}\left\{\sum_{i=0}^{n-1-1} 1_G(x_i^k) + \sum_{i=n-k+1}^{n-1} 1_G(x_i,\ldots,x_{n-1},x_0,\ldots,x_{i+k-n})\right\} \underset{n\to\infty}{\to} \mu^k(G)$$

Hence since $\mathcal{G}_\Omega^n$ is countable, there is a set $A \in \Omega^\infty$ such that

$$\mu(A) = 1$$

and if $x \in A$,

$$\mu_{x_n^k}(G) \underset{n\to\infty}{\to} \mu^k(G) \quad , \qquad \text{all } G \in \mathcal{G}_A^n \ .$$

Since $\mathcal{G}_A^n$ generates $\mathcal{B}_\Omega^n$, we have from Billingsley (1968) that for $x \in A$, $\mu_{x_n^k} \to \mu^k$ weakly and hence (3.3) holds.

If in addition (3.4) holds, let

$$B = \{u^k,y^k : \rho_k(u^k,y^k) > \Pi_k(\mu_{x_n^k},\mu^k)\}$$

let $p \in \mathcal{P}(\mu_{x^n}^k, \mu^k)$ yield $\Pi_k(\mu_x^k, \mu^k)$, that is, $p(B) \le \Pi_k(\mu_{x^n}^k, \mu^k)$, let U^k denote the coordinate random vector of p corresponding to $\mu_{x^n}^k$ and V^k that corresponding to μ^k. We then have

$$\bar{\rho}_k(\mu_{x^n}^k, \mu^k) \le E_p \rho_k(U^k, V^k) \le \Pi_k(\mu_{x^n}^k, \mu^k) + E_p \rho_k(U^k, V^k) 1_B(U^k, V^k)$$

$$\le \Pi_k(\mu_{x^n}^k, \mu^k) + k^{-1} \sum_{i=0}^{k-1} E_p \rho(U_i, a^*) 1_B(U^k, V^k)$$

$$- k^{-1} \sum_{i=0}^{k-1} E_p \rho(V_i, a^*) 1_B(U^k, V^k)$$

and hence from the Cauchy-Schwartz inequality and the stationarity of μ_{x^n} and μ

$$\bar{\rho}_k(\mu_{x^n}^k, \mu^k) \le \Pi_k(\mu_{x^n}^k, \mu^k)(1 + k^{-1} \sum_{i=0}^{k-1} (E_p \rho(V_i, a^*)^2)^{1/2}$$

$$+ k^{-1} \sum_{i=0}^{k-1} (E_p \rho(V_i, a^*)^2)^{1/2})$$

$$= \Pi_k(\mu_{x^n}^k, \mu^k)(1 + (E_p \rho(U_0, a^*)^2)^{1/2} + (E_p \rho(V_0, a^*)^2)^{1/2})$$

$$\le \Pi_k(\mu_{x^n}^k, \mu^k)(1 + (n^{-1} \sum_{i=0}^{n-1} \rho(x_i, a^*)^2)^{1/2} + \rho^{*1/2})$$

As $n \to \infty$, the sum goes to $E_\mu \rho(v_0, a^*)^2 \le \rho^*$ and hence with μ probability one

$$\lim_{n \to \infty} \bar{\rho}_k(\mu_{x^n}^k, \mu^k) \le \lim_{n \to \infty} \Pi_k(\mu_{x^n}^k, \mu^k)(1 + 2\rho^{*1/2}) = 0 ,$$

completing the proof.

One might hope that a stronger result would hold to the effect that $\bar{\rho}(\mu_{x^n}, \mu) \to 0$ or $\Pi(\mu_{x^n}, \mu) \to 0$ μ-a.e. That $\bar{\rho}(\mu_{x^n}, \mu) \to 0$ is impossible, however, even for general finite alphabet processes since in that case with ρ being the Hamming (discrete) metric convergence in $\bar{\rho}$ (in this case called $\bar{d}$ and being Ornstein's distance) implies convergence in entropy [Shields (1975)], yet periodic processes have entropy zero and hence cannot converge in $\bar{\rho}$ to a process with nonzero entropy. Furthermore, in this case we have seen that $\bar{\rho}$ and Π are equivalent metrics and hence it is not possible for $\Pi(\mu_{x^n}, \mu) \to 0$ μ-a.e. for nontrivial processes. Roughly speaking, sample distributions can describe the k^{th} order restrictions of a process to arbitrary accuracy as $n \to \infty$ and any fixed k, but they cannot approximate the k^{th} order restrictions for <u>all</u> k simultaneously, thereby forcing $\sup_k \Pi_k$ to zero. This observation leads to some of the definitions generalizing those of Hampel to stationary ergodic processes.

4. Sequences of Estimators

A sequence of estimators $\{S_n\}$ is a sequence of measurable mappings $S_n: \Omega^n \to \Lambda$, $n = 1,2,\ldots$ where the parameter space Λ is a Polish space with metric d and $\mathcal{B}_\Lambda$ is the Borel σ-field of subsets of Λ. Unlike Hampel, we do not consider S_n to depend on its argument x^n only through $\mu^1_{x^n}$, that is, $S_n(x^n)$ is not assumed to be invariant under permutations of x^n. In addition, Λ need not be $\mathcal{R}^k$ with the Euclidean metric as in Hampel, allowing more general function spaces. In some cases there will exist a "true" value $S(\mu)$ of the parameter of the process μ being estimated by the sequence $\{S_n\}$. Analogous to a special case considered by Hampel, if $S: \mathcal{m}_e \to \Lambda$ is the mapping giving the "true" parameter, one candidate for the sequence of estimators is $S_n(x^n) = S(\mu_{x^n})$, the parameter associated with the periodic process obtained from the sample n-tuple. Examples are the sample mean $(S_n(x^n) = n^{-1} \sum_{i=0}^{n-1} x_i)$ and sample correlation $(S_n(x^n) = n^{-1} \sum_{i=0}^{n-1} x_{i \bmod n} x_{(i+\tau) \bmod n})$ which are simply the mean and correlation of the process μ_{x^n}. Certain results analogous to those of Hampel will be proved for this special case.

Definition

(i) A parameter $S: \mathcal{m}_S \to \Lambda$ is said to be weakly continuous at μ with respect to the $\bar{\rho}$ distance if given $\epsilon > 0$ there exists a $\delta > 0$ such that $\bar{\rho}(\mu,\nu) < \delta$ implies $d(S(\mu),S(\nu)) < \epsilon$.

(ii) A parameter $S: \mathcal{m}_S \to \Lambda$ is said to be strongly continuous with respect to the $\bar{\rho}$ distance if given $\epsilon > 0$ there exists a positive integer k and a $\delta > 0$ such that if $\bar{\rho}_k(\mu^k,\nu^k) < \delta$, then

$d(S(\mu),S(\nu)) < \epsilon$.

(iii) A parameter $S:\mathfrak{m}_S \to \Lambda$ is said to be, simply, strongly continuous (or strongly continuous with respect to the Prohorov distance) if given $\epsilon > 0$ there exists a positive integer k and a $\delta > 0$ such that if $\Pi_k(\mu^k,\nu^k) < \delta$, then $d(S(\mu),S(\nu)) < \epsilon$.

It follows from the properties of the distance that strong continuity $\Rightarrow$ strong continuity with respect to the $\bar{\rho}$ distance $\Rightarrow$ weak continuity with respect to the $\bar{\rho}$ distance.

The strong notions of continuity are required when considering sample distributions as there the conditions of Π_k or $\bar{\rho}_k$ being small can be met, while the condition of small $\bar{\rho}$ in general cannot.

If under μ a sequence of estimators $\{S_n\}$ converges in probability (under μ) to a value $S_\infty(\mu)$, that is, if for all $\epsilon > 0$

$$\lim_{n \to \infty} \mu(x:d(S_n(x^n),S_\infty(\mu)) > \epsilon) = 0 , \qquad (4.1)$$

then we say $\{S_n\}$ is consistent for $S_\infty(\mu)$ under μ. As pointed out by Hampel, $S_\infty(\mu)$ need not be the same as the "true" parameter value $S(\mu)$, but in such a case $S_\infty(\mu)$ might be a better definition of the "true" parameter given the S_n.

A sequence of estimators $\{S_n\}$ on a process μ induces a family of probability measures $\mu^n S_n^{-1}$ on $(\Lambda,\mathcal{B}_\Lambda)$ defined by

$$\mu^n S_n^{-1}(F) = \mu^n(S_n^{-1}(F)) , \quad \text{all } F \in \mathcal{B}_\Lambda \qquad (4.2)$$

Lemma 4.1

If $S:\mathcal{M}_S \to \Lambda$ is

(i) a strongly continuous parameter at $\mu \in \mathcal{M}_e$, or

(ii) a strongly continuous parameter at $\mu \in \mathcal{M}_e$ with respect to the $\bar{\rho}$ distance and there exists a reference letter in the sense of (3.4),

then the sequence of estimators $\{S_n\}$ given by $S_n(x^n) = S(\mu_{x^n})$ is consistent for S at μ.

Proof.

(i) Given $\epsilon > 0$, chose k, δ such that $\Pi_k(\mu^k,\nu^k) < \delta$ implies $d(S(\mu),S(\nu)) < \epsilon$. From Lemma 3.1, there is an n_0 sufficiently large to ensure that if $n \geq n_0$, then $\mu(x:\Pi_k(\mu_{x^n}^k,\mu^k) > \delta) \leq \epsilon$, $n \geq n_0$, and hence $\mu(x:d(S_n(x^n),S(\mu)) > \epsilon) \leq \mu(x:\Pi_k(\mu_{x^n}^k,\mu^k) > \delta) \leq \epsilon$, $n \geq n_0$, completing the proof.

(ii) As in (i) with Π_k replaced by $\bar{\rho}_k$.

Lastly, let Π_d denote the Prohorov distance between measures on $(\Lambda,\mathcal{B}_\Lambda)$ with respect to the metric d.

5. Robust Sequences

Definition.

Given a collection of processes $\mathcal{M} \subset \mathcal{M}_s$, a sequence of estimators $\{S_n\}$ is robust for $\mathcal{M}$ at a process μ if given $\epsilon > 0$ there is a $\delta > 0$ such that for all n and all processes $\nu \in \mathcal{M}$

$$(A)\quad \bar{\rho}(\mu,\nu) < \delta \Rightarrow \Pi_d(\mu^n S_n^{-1}, \nu^n S_n^{-1}) < \epsilon \quad .$$

The definition is intuitively the same as Hampel's: A robust sequence is one for which close observation processes imply uniformly (over n) close estimate distributions. Hampel defines robustness only at i.i.d. processes and only for $\mathcal{M}_m$, the class of all i.i.d. processes. In the case of $\mathcal{M}_m$, (A) is equivalent to $\bar{\rho}(\mu,\nu) = \bar{\rho}_1(\mu^1,\nu^1)$, the marginal distance being small. Since $\Pi_1(\mu^1,\nu^1)^2 \le \bar{\rho}_1(\mu^1,\nu^1)$, robustness at an i.i.d. process for $\mathcal{M}_m$ in our sense is slightly weaker than Hampel's robustness. If ρ is bounded or we add the constraint to $\mathcal{M}_m$ that there exist a reference letter as in Lemma 2.1, then for $\mathcal{M}_m$ the two notions for robustness at an i.i.d. process are equivalent.

The following auxilliary definitions will prove useful.

Definition.

(i) A sequence of estimators $\{S_n\}$ is asymptotically robust for a collection $\mathcal{M} \subset \mathcal{M}_s$ at μ if given $\epsilon > 0$ there is a $\delta > 0$ and an n_0 such that for all $n \ge n_0$ and processes $\nu \in \mathcal{M}$ (A) holds true.

(ii) A sequence of estimators $\{S_n\}$ is small sample robust for a collection $\mathcal{M} \subset \mathcal{M}_s$ at μ if for any integer n_0 and any

$\epsilon > 0$ there is a $\delta > 0$ such that (A) holds for all $n = 1,2,\ldots,n_0$.

Lemma 5.1

If a sequence $\{S_n\}$ is both asymptotically robust and small sample robust for η at μ, then it is robust for η at μ.

Proof.

Given $\epsilon > 0$, choose δ_1, n_0 such that (A) is satisfied for $n \geq n_0$ and then δ_2 so that (A) is satisfied for $n \leq n_0$ and set $\delta = \min(\delta_1, \delta_2)$.

The following technical definition is an asymptotic weakened version of Hampel's condition (B) and will play a similar role.

Definition.

Condition (B) is said to be asymptotically satisfied for a sequence of estimators $\{S_n\}$ and a process μ if given $\epsilon > 0$, $\eta > 0$ there exist positive integers k and n_0 and a $\delta > 0$ and for all $n \geq n_0$ a set $F_n \in \mathcal{B}_\Omega^n$ such that

$$\mu^n(F_n) > 1 - \eta \tag{5.2}$$

and if $x^n \in F_n$, $y^n \in \Omega^n$, and

$$\Pi_k(\mu_{x^n}^k, \mu_{y^n}^k) < \delta \quad , \tag{5.3}$$

where Π_k is the Prohorov distance with respect to ρ_k as in (2.4), then

$$d(S_n(x^n), S_n(y^n)) < \epsilon \quad . \tag{5.4}$$

If we forced $n_0 = 1$, then the above condition would be identical to Hampel's except for the fact that we allow a general k (which may depend on ϵ and n) while he requires $k = 1$. Hence our condition is weaker (his condition (B) implies ours, but not conversely). The following is analogous to Hampel's Lemma 1.

Lemma 5.2

If $\mu \in m_s$ and $\{S_n\}$ asymptotically satisfy condition (B), then $\{S_n\}$ is asymptotically robust at μ.

Proof.

Choose ϵ as in (A). For (B) use the same ϵ, set $\eta = \epsilon/2$ and let k, δ_E, n_0, F_n be the promised objects for $n \geq n_0$. Choose $\delta = \min(\frac{\epsilon}{3}, \epsilon^2/4)$. The key to the proof is that given x^n and y^n, the measure p' on $(\Omega^n, \mathcal{B}_\Omega^n)^2$ which assigns probability n^{-1} to each pair k-tuple x_i^k, $i = 0,1,\ldots,n-k$, $(x_i,\ldots,x_{n-1},x_0,\ldots,x_{i+k-n})$, $i = n-k+1,\ldots,n-1$, is in $\mathcal{P}(\mu_{x^n}^k, \mu_{y^n}^k)$ and hence

$$\bar{\rho}_k(\mu_{x^n}^k, \mu_{y^n}^k) \leq E_{p'}\rho_k = n^{-1}\sum_{i=0}^{n-k}\rho_k(x_i^k,y_i^k) + n^{-1}\sum_{i=n-k+1}^{h-1}\rho_k(x_i,\ldots;x_{i+k-n};$$

$$y_i,\ldots,y_{i+k-n})$$

$$= n^{-1}\sum_{i=0}^{n-1}\rho(x_i,y_i) = \rho_n(x^n,y^n) \qquad (5.5)$$

for all k (and hence $\bar{\rho}(\mu_{x^n}, \mu_{y^n})$ is small if $\rho_n(x^n,y^n)$ is).

Let p be the stationary process yielding $E_p\rho(X_0,Y_0) = \bar{\rho}(\mu,\nu)$ we have

$$E_p \bar{\rho}_k(\mu^k_{X^n}, \mu^k_{Y^n}) \leq E_p(n^{-1} \sum_{i=0}^{n-1} \rho(X_i, Y_i)$$

$$= \bar{\rho}(\mu, \nu) \leq \delta \tag{5.6}$$

and hence from Chebychev's inequality

$$p(x,y: \bar{\rho}_k(\mu^k_{x^n}, \mu^k_{y^n}) > \delta^{1/2}) < \delta^{1/2}$$

whence

$$p(x,y: \Pi_k(\mu^k_{x^n}, \mu^k_{y^n}) > \delta_B) \leq$$

$$p(x,y: \Pi_k(\mu^k_{x^n}, \mu^k_{y^n}) > \delta^{1/4}) < \delta^{1/2}$$

and

$$p(x,y: x^n \in F_n, \Pi_k(\mu^k_{x^n}, \mu^k_{y^n}) < \delta_B^{1/2}) >$$

$$1-\eta-\delta^{1/4} \geq 1-\epsilon/2-\epsilon/2 = 1-\epsilon$$

which from (B) implies that with probability $1-\epsilon$ $d(S_n(x^n), S_n(y^n)) < \epsilon$ and hence $\Pi_d(\mu^n S_n^{-1}, \nu^n S_n^{-1}) \leq \epsilon$, completing the proof.

The following definition is a weakened version of one of Hampel's corresponding definitions.

Definition.

A sequence of estimators $\{S_n\}$ is continuous at μ if given $\epsilon > 0$, there exist positive integers k, n_0 and a $\delta > 0$ such that if $n,m \geq n_0$, $x^n \in \Omega^n$, $y^m \in \Omega^m$, and

$$\Pi_k(\mu^k_{x^n}, \mu^k) < \delta$$

$$\Pi_k(\mu^k_{y^m}, \mu^k) < \delta \tag{5.7}$$

then

$$d(S_n(x^n), S_m(y^m)) < \epsilon \quad . \tag{5.8}$$

If a single k works for all ϵ, we say $\{S_n\}$ is continuous of order k at μ (or continuous at μ^k).

Hampel's definition of continuity of an estimator sequence is what we call continuity of order 1 (or at μ^1). Hampel essentially restricts his estimator sequence to depend only on the marginal properties of the process. Analogous to our strong continuity of parameters, we allow the estimator sequence to depend on higher order properties, but for a given $\epsilon > 0$ there must be a finite k such that matching sample distributions of order k to the underlying μ^k forces the estimators to match up for long observation sequences.

Analogous to Hampel's special case, if a parameter $S:\pi_\theta \to \Lambda$ is strongly continuous, then the sequence of estimators $\{S_n\}$ defined by $S_n(x^n) = S(\underset{x}{\ }_n)$ is continuous.

The following lemma is a strict generalization of Hampel's Lemma 2 since our continuity notion for $\{S_n\}$ is weaker than his.

Lemma 5.3

If $\{S_n\}$ is continuous at $\mu \in \pi_\theta$, then, under μ, $\{S_n\}$ is consistent for some $S_\infty(\mu)$, that is, for any $\delta > 0$

$$\lim_{n \to \infty} \mu(x: d(S_n(x^n), S_\infty(\mu)) > \delta) = 0 \quad .$$

Proof.

For a sequence $\epsilon_i \downarrow 0$ choose $\delta_i \downarrow 0$ and $n_i \uparrow \infty$ such that the continuity condition is fulfilled for $n,m > n_i$ (for each i). Define for positive integers k,n and $\delta > 0$ the set

$$B_n(k,\delta) = \{x^n : \Pi_k(\mu_{x^n}^k, \mu^k) < \delta\}$$

and note from Lemma 3.1 that for fixed k,δ

$$\lim_{n \to \infty} \mu^n(B_n(k,\delta)) = 1 \tag{5.9}$$

From the continuity condition, if $x^n \in B_n(k_i,\delta_i)$, $y^m \in B_m(k_i,\delta_i)$, $n,m \geq n_i$, then $d(S_n(x^n),S_m(y^m)) < \epsilon$; and hence the set

$$G_i = \bigcup_{n \geq n_i} \bigcup_{x^n \in B_n(k_i,\delta_i)} S_n(x^n) \subset \Lambda \tag{5.10}$$

has diameter $\operatorname{diam}(G_i) \leq 2\epsilon_i$. Defining the set $S_n(B_n(k_i,\delta_i)) = \bigcup_{x^n \in B_n(k_i,\delta_i)} S_n(x^n)$, (5.10) can also be written

$$G_i = \bigcup_{n,m \geq n_i} S_n(B_n(k_i,\delta_i)) \quad .$$

Note also that since all spaces are Polish, measurability of S_n implies $B_n(k,\delta) \in \mathcal{B}_\Lambda$. Define the set

$$A_i' = \bigcap_{j=1}^{i} G_j = \bigcup_{n,m \geq n_i} \bigcap_{j=1}^{i} S_n(B_n(k_j,\delta_j))$$

and let A_i denote the closure of A_i' (A_i will play the role of Hampel's A_i in our case). The A_i are closed and monotone decreasing since $A_i > A_{i+1}$ and $\operatorname{diam} A_i \leq 2\epsilon_i \downarrow 0$. Furthermore, the sets A_i are nonempty as can be seen as follows: For fixed i and $n \geq n_i$, we

have from (5.9) that

$$\mu(x:S_n(x^n) \in A_i) \geq \mu(x:S_n(x^n) \in A_i')$$

$$\geq \mu(x:S_n(x^n) \in \bigcap_{j=1}^{i} S_n(B_n(k_j,\delta_j)))$$

$$\geq \mu(x:x^n \in \bigcap_{j=1}^{i} B_n(k_j,\delta_j)) \underset{n\to\infty}{\to} 1 \qquad (5.11)$$

and hence A_i cannot be empty. Since Λ is complete and the A_i are closed, monotone decreasing, and empty, from the Cantor intersection theorem, there exists a single point, say $S_\infty(\mu)$, such that $A_i \downarrow S_\infty(\mu)$. Coupled with (5.11), this proves the lemma.

Corollary 5.1: Given $\{S_n\}$, μ, $S_\infty(\mu)$ as in Lemma 5.3, given $\epsilon > 0$ there exists a δ, k, n_0 such that if $n \geq n_0$ and

$$\Pi_k(\mu_{x^n}^k, \mu^k) < \delta \quad ,$$

then $d(S_\infty(\mu), S_n(x^n)) < \epsilon$.

Proof.

Using the notation of the previous proof, choose i so large that $\epsilon \geq 2\epsilon_i$ and set $\delta = \delta_i$, $n \geq n_i$, $\Pi_{k_i}(\mu_{x^n}^{k_i}, \mu^{k_i}) < \delta_i$ implies $S_n(x^n) \in G_i = \bigcup_{n \geq n_i} S_n(B_n(k_i,\delta_i)) \supset A_i'$. Since $S_\infty(\mu) \in A_i$ and diam $G_i < 2\epsilon_i$, this implies $d(S_n(x^n), S_\infty(\mu)) < 2\epsilon_i \leq \epsilon$, completing the proof.

The previous corollary simply makes explicit a fact useful for the next result that is obvious in Hampel's case.

The following theorem is the main result of this paper and is the analog to Hampel's theorem for stationary and ergodic processes and the general sequence of estimators here considered. We show that continuity of $\{S_n\}$ implies asymptotically robust and continuity of the S_n considered as point functions implies small sample robust.

Theorem 5.1

Let a sequence of estimators $\{S_n\}$ and a $\mu \in \mathcal{M}_e$ be such that

(i) S_n is continuous as a point function on Ω^n for every n, that is, given n, $x^n \in \Omega^n$, $\epsilon > 0$, there exists a $\delta = \delta(n, x^n, \epsilon)$ such that $\rho_n(x^n, y^n) \leq \delta$ implies $d(S_n(x^n), S_n(y^n)) < \epsilon$.

(ii) $\{S_n\}$ is continuous at μ, μ stationary and ergodic.

Then $\{S_n\}$ is robust for $\mathcal{M}_e$ at μ.

Comments. Condition (i) might appear different from that of Hampel since we use $\rho_n(x^n, y^n) = n^{-1} \sum_{i=0}^{n-1} \rho(x_i, y_i)$ and he uses $\rho_n'(x^n, y^n) = \max_i \rho(x_i, y_i)$. These metrics generate the same (product) topology, however, and hence the notions are equivalent. Recall also that (ii) is weaker than Hampel's corresponding assumption and the observation processes are far more general, but that our conclusion is in general slightly weaker. We also note that for large n our proof parallels Hampel's by proving condition (B). For small n, however, robustness is proved directly from (i) and our proof is simpler than Hampel's.

Proof.

First choose $\epsilon > 0$, $\eta > 0$ for property (B). From Lemma 5.1 and its corollary there exists $S_\infty(\mu)$, $\delta_0 > 0$, n_1, k such that for $n \geq n_1$

$$\Pi_k(\mu_{x^n}^k, \mu^k) < 2\delta_0 \Rightarrow d(S_\infty(\mu), S_n(x^n)) < \epsilon/2 \quad . \tag{5.12}$$

From Lemma 3.1 there exists an $n_0 \geq n_1$ so large that if $n \geq n_0$,

$$\mu(x:\Pi_k(\mu_{x^n}^k, \mu^k) > \delta_0) < \eta \quad ,$$

For $n \geq n_0$ define $F_n = \{x^n:\Pi_k(\mu_{x^n}^k, \mu^k) < \delta_0\}$ and note that $\mu^n(F^n) \geq 1-\eta$ and if $x^n \in F_n$, $y^n \in \Omega^n$, then if $\Pi_k(\mu_{x^n}^k, \mu_{y^n}^k) < \delta_0$ we have

$$\Pi_k(\mu_{y^n}^k, \mu^k) \leq \Pi_k(\mu_{y^n}^k, \mu_{x^n}^k) + \Pi_k(\mu_{x^n}^k, \mu^k) \leq 2\delta_0$$

and hence from (5.12), $d(S_\infty(\mu), S_n(y^n)) < \epsilon/2$ and therefore

$$d(S_n(x^n), S_n(y^n)) \leq d(S_n(x^n), S_\infty(\mu)) + d(S_\infty(\mu), S_n(y^n)) \leq \epsilon$$

proving condition (B) is asymptotically satisfied and hence by Lemma 5.2 $\{S_n\}$ is asymptotically robust at μ.

We next prove that (i) implies that $\{S_n\}$ is small sample robust at μ, which by Lemma 5.1 will complete the proof. Given $\epsilon > 0$ as before and any n, there exists from Parthasarathy (1967) Thm. 3.2, Ch. 3, a compact set K_n such that

$$\mu^n(K_n) > 1-\epsilon/4, \quad \nu^n(K_n) > 1-\epsilon/4 \quad .$$

Since $S_n:\Omega^n \to \Lambda$, it is uniformly continuous on K_n and hence there is

a δ_n such that for $x^n, y^n \in K_n$, $\rho_n(x^n, y^n) < \delta_n$ implies $d(S_n(x^n), S_n(y^n)) < \epsilon$. Choose δ so small that $\delta \le \min(\delta_i^2, i=1,\ldots,n_0, \epsilon^2/4)$ and let $p \in P_e(\mu,\nu)$ yield $\bar{\rho}(\mu,\nu) = E_p\rho(X_0,Y_0) \le \delta$. We have using the Chebychev inequality that

$$p(x,y: d(S_n(x^n), S_n(y^n)) > \epsilon) \le \mu^n(K_n^c) + \nu^n(K_n^c) + p(x,y: \rho_n(x^n,y^n) > \delta_n)$$

$$\le \epsilon/2 + p(x,y: \rho_n(x^n,y^n) > \delta^{1/2}) \le \epsilon/2 + \delta^{1/2} \le \epsilon$$

and hence

$$\Pi_d(\mu^n S_n^{-1}, \nu^n S_n^{-1}) \le \epsilon ,$$

completing the proof.

The only point in the preceding development where ergodicity was required was in the use of Lemma 3.1 in Lemma 5.3 ensuring sample distributions of the process μ converged to the actual distribution of μ in the sense of (3.3). The resulting consistency of $\{S_n\}$ at μ was then in turn used to prove asymptotic robustness at μ. In particular, if the process μ is ergodic, but we allow the processes ν of Theorem 5.1 to be stationary but not necessarily ergodic, then the entire proof goes through as before giving the following.

Corollary 5.2: Given the conditions of Theorem 5.1, then $\{S_n\}$ is robust for π_s at μ.

That robustness for the class of ergodic processes implies robustness for the class of stationary processes also can be seen from the ergodic decomposition theorem of Rohlin (1949) which states, roughly, that every stationary nonergodic process is a mixture of ergodic processes, that is,

can be viewed as nature first selecting an ergodic process (unknown to the observer) and then sending a sample function from the ergodic process. Thus, if ν is stationary, the observer will actually see some unknown ergodic component, say ν_θ, of ν and hence robustness for ergodic processes will ensure robustness for stationary nonergodic processes.

Corollary 5.3: Let $S:\mathcal{M}_s \to \Lambda$ be such that S is strongly continuous at $\mu \in \mathcal{M}_e$ and $S_n(x^n) = S(\mu_{x^n})$ is a continuous mapping from Ω^n to Λ. Then $\{S_n\}$ is robust for $\mathcal{M}_e$ at μ.

Note that if S is strongly continuous for all μ, then $S_n(x^n) = S(\mu_{x^n})$ is automatically continuous as a point function from (2.6).

Analogous to Hampel's Lemma 3 and corollary we have the following.

Lemma 5.4

If $\{S_n\}$ is robust at $\mu \in \mathcal{M}_e$ and consistent for $S_\infty(\mu)$ at all $\nu \in \mathcal{M}_s$ in a $\bar{\rho}$ neighborhood of μ, then $S_\infty(\mu)$ is weakly continuous at μ.

Proof.

Since $\{S_n\}$ is robust, given $\epsilon > 0$ there is a $\delta > 0$ such that if $\bar{\rho}(\mu,\nu) < \delta$, then $\Pi_d(\mu^n S_n^{-1}, \nu^n S_n^{-1}) < \epsilon$, all n. By consistency, for δ small enough

$$\lim_{n \to \infty} \mu(x: d(S_n(x^n), S_\infty(\mu)) > \epsilon) = 0$$

$$\lim_{n \to \infty} \nu(y: d(S_n(y^n), S_\infty(\nu)) > \epsilon) = 0$$

and hence if α_1 is the measure on $(\Lambda,\mathcal{B}_\Lambda)$ assigning probability one to the point $S_\infty(\mu)$ and α_2 that assigning probability one to $S_\infty(\nu)$,

$$\Pi_d(\mu^n S_n^{-1},\alpha_1) \underset{n\to\infty}{\to} 0$$

$$\Pi_d(\nu^n S_n^{-1},\alpha_2) \underset{n\to\infty}{\to} 0$$

and hence $\Pi_d(\alpha_1,\alpha_2) \le \epsilon$. Since α_1 and α_2 are degenerate, however, $\Pi_d(\alpha_1,\alpha_2) \le \epsilon$ only if $d(S_\infty(\mu),S_\infty(\nu)) \le \epsilon$, proving the lemma.

Corollary 5.4: If $\{S_n\}$ is robust and continuous for all $\mu \in \mathcal{M}_e$, then $S_\infty(\cdot)$ is weakly continuous at all μ.

6. Discussion and Applications

Our approach allows the construction of robust estimators for parameters included in the K^{th} order (K finite, fixed) restriction $(\Omega^K, \mathcal{B}_\Omega^K, \mu^K)$ of an ergodic stationary process $[\Omega, \mu, X]$. Such parameters are the moments of order less than or equal to K.

The M-estimation $S_\infty(\mu)$ of a scalar parameter S included in $(\Omega^K, \mathcal{B}_\Omega^K, \mu^K)$ will be now the solution (if it exists) of the expression [Huber (1964), Huber (1972)]

$$\int_{\mathcal{B}_\Omega^K} \psi(x_1, \ldots, x_K, S_\infty(\mu)) \mu^K(dx_1, \ldots, dx_K) = 0 \qquad (6.1)$$

As in the i.i.d. case, the sequence of estimators $\{S_n\}$ defined by ψ $(S_n : \sum_{i=1}^{n-K} \psi(x_i, \ldots, x_{i+K}, S_n) = 0)$ is robust if the solution is (6.1) is unique and ψ is bounded. In other words, one should look for bounded, "smooth" functions ψ with zero μ^K expectation.

For the robust estimation of a location parameter, in particular, M-estimators, L-estimators or R-estimators, can be used again [Huber (1972)], where the first order restriction $[\Omega^1, \mathcal{B}_\Omega^1, \mu^1]$ of the ergodic stationary process $[\Omega, \mu, X]$ is considered. For the M-estimators, we may use the K^{th} restriction $[\Omega^K, \mathcal{B}_\Omega^K, \mu^K]$ instead and recover the estimate from the expression:

$$\int_{\mathcal{B}_\Omega^K} \psi(x_i - S_\infty(\mu), \ldots, x_K - S_\infty(\mu)) \, \mu^K(dx_1, \ldots, dx_K) = 0 \qquad (6.2)$$

The asymptotic distribution of the estimate $S_\infty(\mu)$ can be found by methods similar to the ones used by Huber (1964).

New estimators determined through new functionals of the data may be considered, where the properties of the functionals may be determined through the conditions in Theorem 5.1.

P. Papantoni-Kazakos

HO-3452-PPK-tmg

Atts.
References
Appendix A

REFERENCES

Billingsley, P. (1968), Convergence of Probability Measures, Wiley, New York.

Dudley, R.M. (1968), "Distances of Probability Measures and Random Variables," Ann. of Math. Stat., 39, pp. 1563-1572.

Gray, R.M., D.L. Neuhoff, and J.K. Omura (1975), "Process Definitions of Distortion-Rate Functions and Source Coding Theorems," IEEE Trans. On Info. They., IT-21, pp. 524-532.

Gray, R.M., D.L. Neuhoff, and P.C. Shields (1975), "A Generalization of Ornstein's $\bar{d}$ Distance with Applications to Information Theory," Ann. Prob., 3, pp. 315-328.

Hampel, F.R. (1971), "A General Qualitative Definition of Robustness," Ann. of Math. Stat., 42, pp. 1887-1896.

Huber, P.J. (1964), "Robust Estimation of a Location Parameter," Ann. Math. Stat., 35, pp. 73-101.

Huber, P.J. (1972), "Robust Statics. A Review," Ann. Math. Stat., 43, pp. 1041-1067.

Moser, J., E. Phillips, and S. Varadhan (1975), Ergodic Theorey: A Seminar, Courant Institute of Math. Sciences, New York.

Oxtoby, . (1952), "Ergodic Sets," Bull. Amer. Math. Soc., 58, pp. 116-136.

Parthasarathy, K.R. (1961), "On the Category of Ergodic Measures," Ill. J. Math., 5, pp. 648-656.

Parthasarathy, K.R. (1968), Probability Measures on Metric Spaces, Academic Press, New York.

Rohlin, V.A. (1949), "Selected Topics from the Metric Theory of Dynamical Systems," Uspehi Mat. Nauk., 4, pp. 57-128; (AMS Translations (2), 49, pp. 171-240.

Shields, P.C. (1975), The Theory of Bernoulli Shifts, Univ. of Chicago Press.

Strassen, V. (1965), "The Existence of Probability Measures with Given Marginals," Ann. Math. Stat., 36, pp. 423-429.

Vashershtein, L.N. (1969), "Markov Processes on Countable Product Space Describing Large Systems of Automata," Problemy Peredachi Informatsii, 5, pp. 64-73.

APPENDIX A: Equations (2.1) and (2.3a) are actually minima. (The proof is due to P.C. Shields.)

Since Ω and hence Ω^∞ are complete, separable metric spaces, any measure μ on $(\Omega^\infty, \mathcal{B}_\Omega^\infty)$ is tight, that is, for any $\epsilon > 0$ there is a compact set F such that $\mu(F) \geq 1-\epsilon$ [Parthassarathy (1967), Thm. 3.2, p. 29]. If one has a family of measures such that given ϵ there is a compact set F such that all members of the family place measure at least $1 - \epsilon$ on F, then the family is compact in the weak topology [Parthasarathy (1967), Thm. 6.7, p. 47]. Given μ, ν choose compact $F \in \mathcal{B}_\Omega^\infty$ such that $\mu(F) \geq 1 - \epsilon/2$, $\nu(F) \geq 1 - \epsilon/2$, then if $p \in \mathcal{P}_s(\mu,\nu)$, $p(F \times F) \geq 1 - \epsilon$ and $F \times F$ is compact. Thus $\mathcal{P}_s(\mu,\nu)$ is compact in the weak topology and a sequence $p_n \in \mathcal{P}_s(\mu,\nu)$ such that

$$E_{\tau_n} \rho(X_0, Y_0) \leq \bar{\rho}(\mu,\nu) + 1/n$$

will have a subsequence -- say ρ_{n_k} -- that converges in the weak topology to a limiting p. The limit $p \in \mathcal{P}_s(\mu,\nu)$ and $E_p \rho(X_0, Y_0) = \bar{\rho}(\mu,\nu)$, completing the proof. The same argument applied to $(\Omega^n, \mathcal{B}_\Omega^n)$ shows that $\bar{\rho}_n$ is also actually a minimum.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR-78-0777	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER 9
4. TITLE (and Subtitle) ROBUST PARAMETER ESTIMATORS FOR COMMUNICATION DATA		5. TYPE OF REPORT & PERIOD COVERED Interim Annual rept. 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) P. Papantoni-Kazakos		8. CONTRACT OR GRANT NUMBER(s) AFOSR-77-3156
9. PERFORMING ORGANIZATION NAME AND ADDRESS Rice University Department of Electrical Engineering Houston, Texas 77001		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Rsch/NM Bolling AFB, DC 20332		12. REPORT DATE Dec 1977 13. NUMBER OF PAGES 41
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

PE 61102F WUAFOSR 2304A5

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The object of this grant is the analysis & design of decision procedures that have stable, good performance in statistically ill-defined environments. Such procedures indicate the way to design powerful receivers for systems whose statistical behavior cannot be described precisely (due to incomplete availability of data about the system behavior). Different distance measures have been studied for use as performance criteria for robust estimates. Careful evaluation and comparison of these distances was done and their

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

403 244

20. Abstract

similarities, advantages and disadvantages were carefully stated. A thorough study of the work already accomplished (by the author as well as other investigators) on nonparametric statistical procedures in the presence of small number of discrete data was done and included in a book on the use of nonparametric procedures in Communication Systems. A feature selection problem was studied, when serveral distance measures are used as discrimination criteria. A sequential procedure for clustered data was proposed and analyzed. Hampel's general qualitative definition of robustness of sequences of estimators on memoryless observation processes was generalized to stationary processes. The constructive analysis of robustness completed by the author is being used now in the performance analysis of communication Networks at Bell Laboratories.