NASA TECHNICAL MEMORANDUM NASA TM-75565

NONPARAMETRIC ALGORITHMS FOR THE SEARCH OF SIGNALS

T.S. Myshenkova
Academy of Sciences USSR, Institute of Space Research

Translation of "Neparametricheskiye Algoritmy Poiska Signalov," Academy of Sciences USSR, Institute of Space Research, Moscow, Report PR-277, 1976, pp. 1-18.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D.C. 20546 OCTOBER 1978

TABLE OF CONTENTS

NONPARAMETRIC ALGORITHMS FOR THE SEARCH OF SIGNALS

T.S. Myshenkova
Academy of Sciences USSR, Institute of Space Research

Introduction

The purpose of this paper is to develop and justify (insofar as possible) nonparametric algorithms for the search of signals and their isolation. During their mathematical formulation, these problems require certain a priori data about the signal and noise. A priori data narrows down the search (provided it does not reflect the properties of all signals) and allows to isolate only a part of signals whose characteristics satisfy restrictions corresponding to this data. /3*

The search proper assumes that the time, shape, duration, power etc. of the signal which appears in the receiving device is unknown.

Thus, on one hand, by specifying certain noise and signal characteristics with the aim of isolating the latter (using parametric methods based, for example, on threshold characteristics constructed on the basis of a priori data), we are not solving the search problem in general--we are solving the search problem for a completely defined class of signals. On the other hand, by reducing the volume of a priori data to a certain level, we can no longer apply the well-developed parametric methods, which compared with nonparametric methods,

---

*Numbers in the margin indicate pagination in the foreign text.

among other things, use up less machine time and computer memory. /4

I. Test Based on Spearman's Rank Correlation Coefficient

Let us consider a discrete time--dependent stochastic process, i.e.

$$F = F(t_i) \equiv F_i\,, \quad i = 1, 2, \ldots, ; \tag{1}$$

$$t_{i+1} - t_i = \Delta t = const. \tag{2}$$

Let us introduce random variables characterizing the form of the process F in a sample of size n

$$\Delta_{ij} = \begin{cases} 1, & F_i \le F_j \\ 0, & F_i > F_j \end{cases} \tag{3}$$

$$i, j = 1, 2, \ldots, n;$$

$$\delta_i = \sum_{j=1}^{n} \Delta_{ij}\,. \tag{4}$$

It is easily seen that δ_i equals the number of readings F from a sample of size n which are not smaller than a given reading F_i from the same sample.

Suppose that $\inf_{i=\overline{1,n}} \{\delta_i\}$ is attained at $i = i_m$ ($m = 1, \ldots, m_0 < n$), i.e. in the given sample, the readings F_{i_m} are largest in magnitude.

The following quantity /5

$$\gamma_i^m = |i - i_m| \tag{5}$$

characterizes the distance between Fi and Fi_m.

Thus, two random variables δi and γi are associated with each random variable Fi. Clearly if the readings Fi, i=1, 2, ..., n are mutually independent random variables, the same statement can also be made about the random variables δi and γi (each separately).

It is assumed that one characteristic of a signal from a point source is the presence of periodicity in readings arriving in the detector. The question with which period (or periods) the intensity of the arriving signal varies is a problem in spectral analysis of a detected signal. Our problem is a search of signals.

From the presence of periodicities during variation in the intensity of radiation from sources which has been ascertained in many studies, we infer that the signal has a "non-random" deterministic form. This fact will be used in constructing a search algorithm.

The search algorithm is constructed on the basis of the so-called nonparametric independence test. This test is convenient in that unlike most decision problems, it does not require a knowledge of the type of distribution of the studied process.

This test is based on the Spearman rank correlation coefficient.

A rule which staisfies this test is as follows: Suppose /6
that only the following is known about a random sample of n pairs $(X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n)$. X_k occupies the position A_k in order of decreasing magnitude, and Y_k occupies the position B_k in order of decreasing magnitude (k-1, 2, ..., n). If the random variables X and Y are independent, the statistic

$$R = 1 - \frac{6}{n(n^2-1)} \sum_{k=1}^{n} (A_k - B_k)^2 \qquad (6)$$

has an asymptotically normal distribution with mean 0 and variance $1/n-1$ as $n \to \infty$.

In our case, the random variable δ_k constructed above represents the position number of the random variable F_k in order of decreasing magnitude, and the random variable γ_k, which characterizes the neutral position of random variables F, represents the position number of the random variable F_k ordered by its distance from Fi_m.

This criterion allows to accept or reject the independence hypothesis at a particular significance level. Specifically, for any $n>1$, the independence hypothesis is rejected at significance level $\leq \alpha$ if

$$R > \frac{u_{1-\alpha}}{\sqrt{n-1}} \quad \text{(one-sided test)}$$

$$|R| > \frac{u_{1-\alpha/2}}{\sqrt{n-1}} \quad \text{(two-sided test)} \qquad (7)$$

where $u_{1-\alpha}$ is the $(1-\alpha)$-th quantile for the random variable u which is normally distributed with mean 0 and variance $1/n-1$.

In this approach the only restriction imposed on the noise process is the statistical independence of the random variables δ and γ (or a very low known correlation allowing to specify α nonarbitrarily). Clearly this restriction is equivalent to the requirement of mutually independent readings. Because of its simplicity, such a model is used sufficiently /7
widely. The real question is: is it realistic?

2. Algorithm Using One-Sided Sign Test

Let us consider another approach to the search problem.

In the absence of a sufficient (or even necessary statistic), no a priori statement can be made about signal characteristics required for the search. Such statistics as sample means and the standard deviation will not be used to test one hypothesis or another about the distribution, since:

a) the search proper assumes that the distribution of the signal is unknown;

b) the statistical characteristics of the noise do not remain constant--the process is nonstationary;

c) although the distribution of the random variable F_i (concretely the number of particles which arrived in the detector during time $t_{i+1}-t_i=\Delta t$) approaches asymptotically the normal distribution (Central Limit Theorem) when Δt is increased (see [2]), which justifies one hypothesis, signals whose duration is less than Δt are no longer detected.

One signal characteristic which will be included in the search is its duration, which among other things, depends on the conditions in which the experiment is carried out.

Thus let

$$\overline{F}_i = \frac{1}{n} \sum_{\kappa=0}^{n} F_{i+\kappa} , \tag{8}$$

$$d_{im} = F_{i+m} - \overline{F}_i , \; m = 0, 1, 2, \ldots, \mathcal{T}_0 , \tag{9}$$

(suppose T_{st} denotes the time during which the stochastic process F can be regarded as stationary. Then a natural bound on $\mathcal{T}_0$ is

$$\mathcal{T}_0 \leq \left[\frac{T_{st}}{\Delta t} \right] \tag{10}$$

/8

where [X] is the integer part of the number X).

Next,

$$\gamma_{im} = \begin{cases} 1, & d_{im} \geq 0 \\ 0, & d_{im} < 0 \end{cases} \tag{11}$$

$$\beta_{i\kappa}^{N} = \left(\sum_{j=\kappa}^{\kappa+N} \gamma_{ij} \right), \tag{12}$$

$$\kappa = i, i+1, \ldots, \mathcal{T}_0 - N.$$

N is the assumed duration of the signal and n is the size of the sample over which the sample means $\overline{F}_i$ must satisfy

$$N < n < \mathcal{T}_0 . \tag{13}$$

It can be easily verified that the assumptions made (about noise) in regard to:

a) the symmetry of the random variable F_1,

b) the independence of the random variables (readings) F_1,

c) the fact that

$$\mathcal{P}\{\alpha_{im}=0\}=0$$

imply

$$\mathcal{P}\{\alpha_{im}>0\}=\mathcal{P}\{\alpha_{im}<0\}=\frac{1}{2}, \tag{14}$$

$$\mathcal{P}\{\beta_{ik}^{N}=P/\text{noise}\}=C_N^P\left(\frac{1}{2}\right)^N, \tag{15}$$

$$\mathcal{P}\{\beta_{ik}^{N}\geq P/\text{noise}\}=\sum_{t=P+1}^{N}C_N^t\left(\frac{1}{2}\right)^N=\mathcal{P}(P). \tag{16}$$

Items a), b), c) constitute the tested H_0 hypothesis. /9

Let P_α be the smallest value of P for which

$$\mathcal{P}(P)\leq\alpha \tag{17}$$

Then the one-sided test (so-called sign test) requires rejection of the H_0 hypothesis at significance level $\leq\alpha$ if

$$\beta_{ik}^{N}>P_\alpha \tag{18}$$

Thus, the algorithm based on this test presupposes that:

1) the noise process satisfies a), b), c).

2) any significant deveiation from noise (last formula (17)) is the signal.

3. Use of Simplified Statistics

We designed and debugged programs written in Fortran IV which are utilized to test item 1) for various noise processes (using data recorded on telemetry tapes).

In the given algorithm, statistic (12) requires a large number of time consuming calculations. From this standpoint, the algorithm can be substantially simplified if one of the following random variables is introduced instead of α (see [9]) and γ (see [11]):
either

$$\Delta_{ij} = \begin{cases} 1, & F_i \le F_j \\ 0, & F_i > F_j, \end{cases} \tag{19}$$

where

$$i, j: \ |i - j| \le \mathcal{T}_0$$

or

$$\theta_i = \begin{cases} 1, & F_i \le F_{i+1} \\ 0, & F_i > F_{i+1}, \end{cases}$$

$$i = 1, 2, \ldots, . \tag{20}$$

/10

The corresponding statistics are

$$\beta_{ik}^{N} = \sum_{j=i+k}^{i+k+N} \Delta_{ij}, \tag{21}$$

$$k: \ i + k + N \le \mathcal{T}_0,$$

or

$$\beta_k^N = \sum_{i=k}^{k+N} \theta_i,$$

$$k: \ k + N \le \mathcal{T}_0, \tag{22}$$

where N as before, is assumed to be the duration of the signal, k is a current value of the subscript satisfying the inequality

$$k + N \le \mathcal{T}_0.$$

It is clear that the test formulated above for testing the H_0 hypothesis can be applied to each of these statistics.

A program written in Fortran IV was developed and debugged for calculating statistics (21), (22).

4. Results of Computerized Processing of Various Experimental Data

Page 12 gives the program for calculating the statistic $R\sqrt{n-1}$ (see [6]). The results of the calculations are illustrated in Fig. 1. The array R corresponds to this statistic in the program. The calculations were carried out for the case n=30. The noise process and the process containing the signal were read into the program in the form of samples, each consisting of 50 readings. Altogether 20 values were obtained from the array R (R(k), $k=\overline{1,20}$) for each case, i.e. 20 samples, each consisting of 30 readings formed from 50 readings that were read in.

The given significance level α can be specified as follows: /11

$$\alpha \ : \ U_{1-\alpha} = \max_{1\le k\le 20} |R_{\text{noise}}(k)|).$$

From Fig. 1 it is evident that in the case of the signal, R in absolute value is greater than the R value for noise nearly everywhere.

Page 13 gives program No. 2, which calculates the statistics β^N_{1k} (corresponding to array MI) and P(p) (corresponding to array FR). The statistic β^N_{1k} was calculated for N=20.

The arrays that were read in (noise and signal) had the same dimension as in program No. 1, i.e. 50. Fig. 2 illustrates (besides the obtained curves) the operation of the algorithm: α is given (horizontal line) and Pα is obtained (see 17). It is evident that if the significance level α is specified in this manner, nearly the entire signal lies in the region $\beta^N_{1k} > P\alpha$, which signifies the presence of the latter.

Conclusion

The fourth section presents the results of calculations of the statistics $R\sqrt{n-1}$ (corresponding to array R in program No. 1), β^N_{1k} (corresponding to array M1 in Program No. 2) and of the value of P given by expression (16) (corresponding to array FR in program No. 2) for noise processes and processes containing the signal (data from telemetry tapes--"Filin experiment"). In [7] this signal was isolated in a different way using the correlation method.

The preliminary operations involving centering and summing centered readings over four channels used up more machine time than processing of the same array in this study. For a given significance level more accurate data on utilized machine time can be used to compare the efficiencies of the discussed algorithms with each other and with the algorithm in reference [7].

In conclusion the author expresses his deep gratitude to the pioneer of this study, V.I. Berezin and also to Ye.N. Kiseleva for her great contribution made in writing and debugging programs, processing data and formating results.

REFERENCES

1. Weisskopf, M.C., Kahn, S.M., "Short-term Time variability of Cygnus X-1." Columbia Astrophysics Laboratory, Columbia University, New York, March 1975.

2. Garsky, H., "Observations of Galactic X-ray Sources," American Science and Engineering. Cambridge, Massachusetts.

3. Borovkov, A.A., Kurs teorii veroyatnosti [Course in Probability Theory], Moscow, Nauka Press, 1972.

4. Hajek, J. and Sidak, Z., Teoriya rangovykh kriteriyev [Theory of Rank Tests], Moscow, Nauka Press, 1971.

5. Germain, K., Programmirovaniye na IBM/360 [Programming the IBM/360], Moscow, Mir Press, 1978.

6. Korn, G. and Korn, T., Spravochnik po matematike dlya nauchnykh rabotnikov i inzhenerov [Handbook of Mathematics for Scientists and Engineers], Moscow, Nauka Press, 1978.

7. Berezin, V.I., Myshenkova, T.S. et at., Vydeleniye i detektirovaniye rentgenovskikh istochnikov [Isolation and Detection of X-ray Sources], Proceedings of the International Conference on Telemetry, USA, Washington, 1975.

PROGRAM NO. 1

```
      DIMENSION ML(50,50),MEL(50)
      DIMENSION P(50)
      READ(1,10) P
   10 FORMAT(24F3.2)
      MC=50
      KC=30
      MCP=MC-KC
      K1=KC+1
      K2=K1**3-K1
      N1=KC-
      P1=SQRT(FLOAT(N
      DO   I=1,MC
      DO 4 J=1,MC
      PR=P(I)-P(J)
      IF(PR) 3,3,2
    3 ML(I,J)=
      GO TO 4
    2 ML(I,J)=
    4 CONTINUE
      CONTINUE
      DO 5 K=  MC
      M K
      KL=K+
      DO 6 I=K,KL
      L=0
      DO     =K,KL
    7 L=L+ML(I,J)
    6 MEL(I
      M1=MEL(K)
      IM=K
      DO 14  =K,KL
      M2=M1-MEL(I
      IF(M2) 8,15 1
    8 M1=MEL(I)
      IM=I
      GO TO  4
   15 M1=M
      IM=I
   14 CONTINUE
      J2=0
      DO 9 J=K,KL
      J1=ABS(FLOAT(J-K))
      M3=ABS(FLOAT(MEL(J)-J1))
      M3=M3**2
    9 J2=J2+M3
      R=6*J2
      R=R/K2
      R=1-R
      R=R*P
      WRITE(3,11) K,J2,R
   11 FORMAT(I3,I5,F8  )
    5 CONTINUE
      END
```

PROGRAM NO. 2


```
      DIMENSION M(49)
      DIMENSION M1(30)
      DIMENSION FR(30)
      DIMENSION P(50)
      READ(1,10) P
   10 FORMAT(24F3.2)
      MC=50
      MCR=MC-1
      N=20
      MCP=MC-N
      DO 1 I=1,MCR
      PR=P(I)-P(I+1)
      IF(PR) 2,2,3
    2 M(I)=1
      GO TO 1
    3 M(I)=0
    1 CONTINUE
      DO 4 I=1,MCP
      I1=I+N-1
      L=0
      DO 5 K=I,I1
    5 L=L+M(K)
    4 M1(I)=L
      WRITE(3,10) M1
   10 FORMAT(5H M1=,16I5)
      F1=1.
      DO 6 K=1,N
      F=K
    6 F1=F1*F
      FN=F1
      DO 7 I=1,MCP
      N1=M1(I)+1
      F8=0.
      DO 8 J=N1,N
      F2=1.
      DO 9 K=1,J
      F=K
    9 F2=F2*F
      F4=F2
      J1=N-J
      F3=1.
      DO 13 K=1,J1
      F=K
   13 F3=F3*F
      F5=F3
      F6=F4*F5
      F7=FN/F6
      F8=F8+F7
    8 CONTINUE
      DO 15 K=1,N
   15 F8=F8/2.
    7 FR(I)=F8
      WRITE(3,11) FR
   11 FORMAT(5H FR=,5F20.14)
      END
```

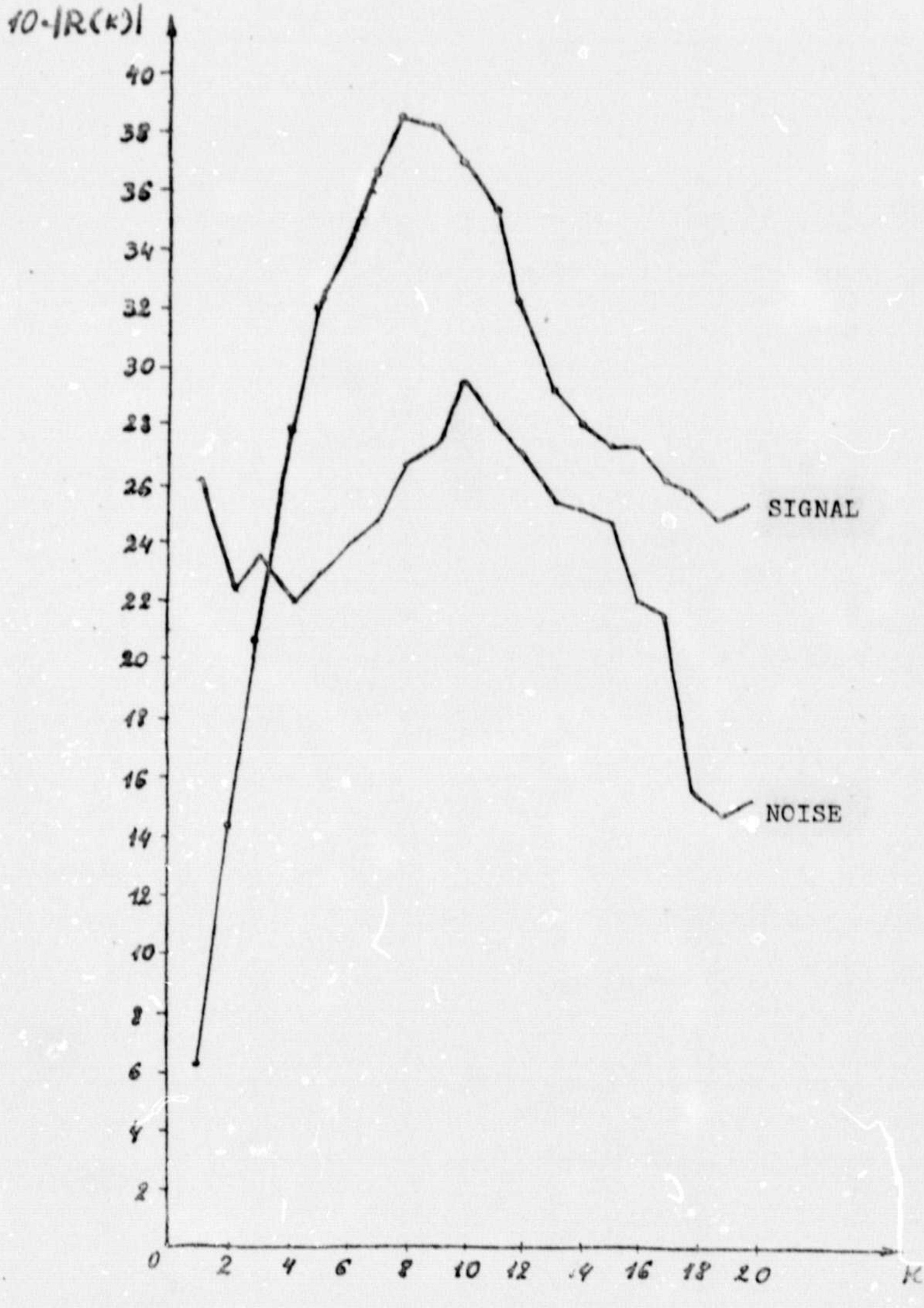
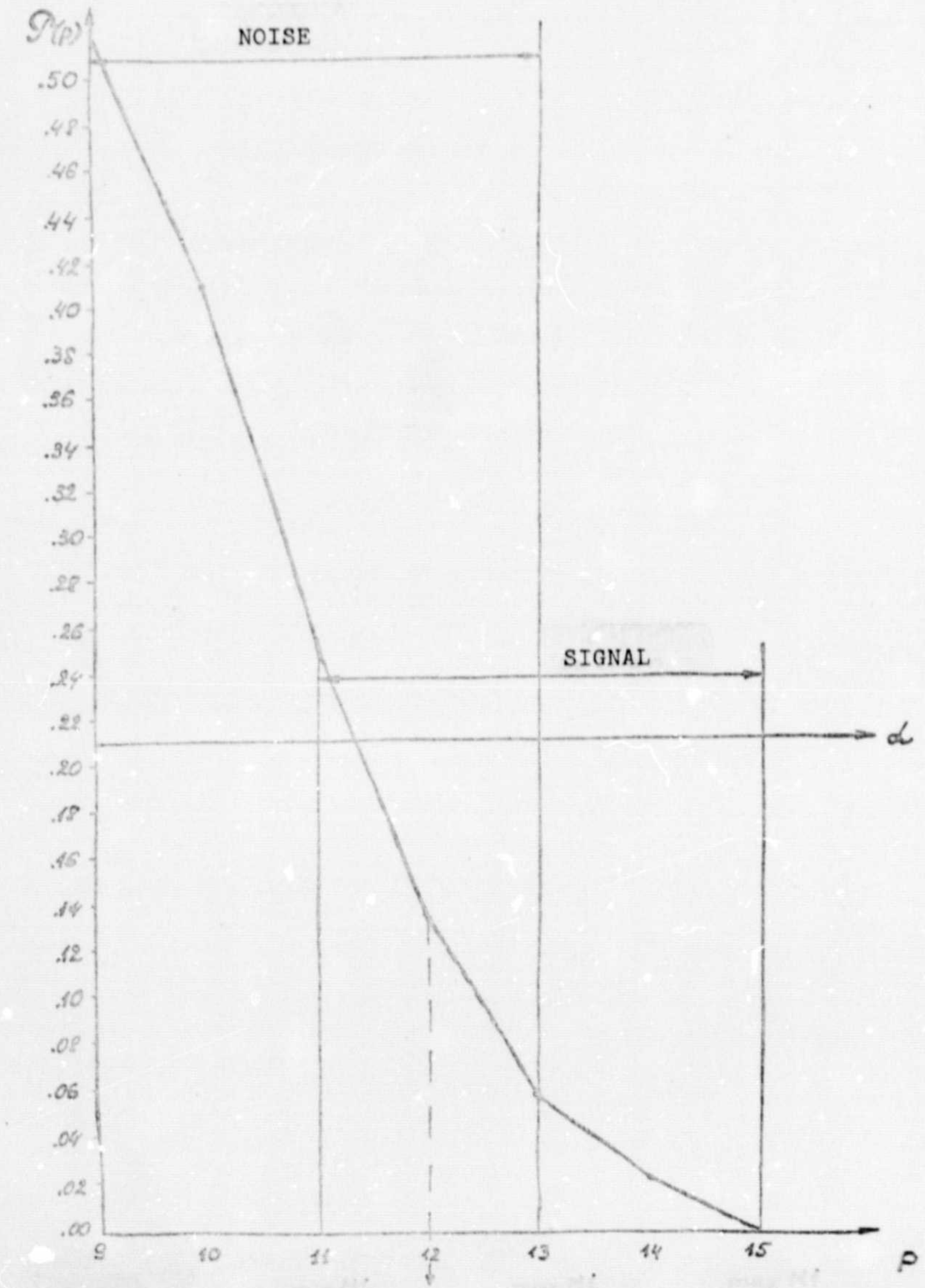

10·|R(k)|
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
2
4
6
8
10
12
14
16
18
20
K
SIGNAL
NOISE

FIG. 1



𝒫(p)
NOISE
SIGNAL
.50
.48
.46
.44
.42
.40
.38
.36
.34
.32
.30
.28
.26
.24
.22
.20
.18
.16
.14
.12
.10
.08
.06
.04
.02
.00
9
10
11
12
13
14
15
p
α
NOISE
SIGNAL
Pα
NOISE
SIGNAL

FIG. 2