TEXAS A&I UNIVERSITY

Kingsville, Texas

TECHNICAL REPORT #3

INVESTIGATION OF THE SPECHT DENSITY ESTIMATOR

DEPARTMENT OF MATHEMATICS

TEXAS A&I UNIVERSITY

RESEARCH GRANT #NGR 44-073-003

TECHNICAL REPORT #3

INVESTIGATION OF THE SPECHT DENSITY ESTIMATOR

F. M. Speed

L. M. Rydl

DEPARTMENT OF MATHEMATICS

July, 1971

TABLE OF CONTENTS

DENSITY ESTIMATION

The purpose of this project is to determine the feasibility of using the Specht density estimator function on the IBM 360/44. Factors such as storage, speed, amount of calculations, size of the smoothing parameter and sample size have an effect on our results. We want to (1) investigate the reliability of the Specht estimator for normal and uniform distributions and (2) show the effects of the smoothing parameter and sample size.

The Specht function which is supposed to estimate a density distribution of an m dimensional random variable, is as follows:

$$f_i(x) = \frac{1}{(2\pi s)^{m/2}} \cdot \frac{1}{K_i} \sum_{j=1}^{K_i} \mathrm{Exp}\left[\frac{-(x - A_{ij})^T (x - A_{ij})}{2s}\right]$$

where A_{ij} $i = 1,\ldots r$ (pattern classes)

$j = 1,\ldots k$, are the m-dimensional samples from the i^{th} class

s is the data-dependent smoothing parameter set by the user.

X is a given vector

T denotes the transpose of a vector.

The Specht function, supposedly has the necessary mathematical properties of a probability density estimator. It is non-negative for any value of x and it integrates to unity over the entire space. As $s \longrightarrow 0$ and the number of training samples increases without limit, the Specht function becomes identical to the true probability density.

In the Sept. 1968 Laboratory for Agricultural Remote Sensing Research Bulletin (LARS) No. 844, two disadvantages are (1) the storage of the entire training sample and (2) the amount of computation required for each classification. The LARS suggests a version of the Specht estimator to eliminate these problems as follows:

$$f_i(x) = \frac{c}{K_i} \sum_{j=1}^{K_i} \exp\left[\frac{-A_j^T A_j}{2s}\right] \exp \frac{A_j^T x}{s}$$

where $c = (2\pi s)^{-m/2} \exp\left[\frac{x^T x}{2s}\right]$, (see Appendix A for derivation). After the first run, this function was corrected by replacing A_j with A_{ij} and by inserting the necessary minus sign in the value of c so that

$$c = 2\pi s^{-m/2} \exp\left[\frac{-x^T x}{2s}\right]$$

To examine the feasibility of using the function suggested be Specht and simplified by LARS, a Fortran IV program was written and run on the IBM 360/44. This program utilized the IBM Subroutines Randu and Gauss to generate random variables from uniform and normal distributions respectively. However, later on Randu and Gauss were replaced with better generators, GEN and RNORM. PLOTK was used to graph the true and estimated functions to show validity of the estimator. A description of the subroutines is given in Appendix B.

To validate the estimator,

$$f_i(x) = \frac{c}{K_i} \sum_{j=1}^{K_i} \exp\left[\frac{-A_j^T A_j}{2s}\right] \exp\left[\frac{A_j^T x}{s}\right]$$

several runs were made varying s, the smoothing parameter and n, the number of training samples. When s was small, .08, and N large, 500, the absolute value of the argument in the second exponential function became larger than one the machine could handle, i.e. 174.673.

At first this restricted our choices for s and N but after investigating the cause we were able to circumvent the problem by using the original Specht estimator. A check was put in the program to set the value of the exponential function equal to zero when the absolute value of the argument, $\left[\frac{A_j^T x}{s}\right]$, was greater than 174.673.

The reason we were getting the argument too large was dependent on both s and N. The smaller s became, the smaller a random variable had to be in the training sample to cause the program interupt. As N became larger, the probability of generating such a number increased.

Consider the following case:

S = .07, A(j) = 3.183, N = 500, x = -4.

$$\frac{A_j^T x}{s} = \frac{(3.183)\ (-4)}{.07} < -174.673$$

this value is not negative at all points therefore $\exp\left[\frac{A_j^T x}{s}\right]$ can't be set to zero. However, if we consider the original estimator we see the argument of the exponential function is negative and this allows us to set the value of the exponential function to zero when its absolute value is greater that 174.673.

Samples of sizes 25,50,100,200,500,1000,2000, and 7000 of normally distributed numbers were generated using GEN and RNORM. The true normal density function was also evaluated at each point x. Then these two values were graphed and the maximum absolute value of the difference between them was calculated.

This procedure was followed for different values of S. We let S take on the values of 1, .5, 125, .15, .1, .09, .08, .07, .06, .03, and .01.

The accuracy of the Specht estimator can be seen in the graphs in Appendix C. These results indicate that our best choice for S was .01 when N equalled 7000. For small sample sizes the value of s has very little effect. In fact some runs showed $S = 1$, $N = 25$ was better than $S = 1$, $N = 50$, 100, or 200.

So as stated, the Specht estimator approaches the true distribution as $S \to 0$ and $N \to \infty$.

Another program was written to use the Specht estimator to generate a uniform distribution. As our graphs reveal, the Specht estimator did not yeild a uniform distribution.

We let S = .09, .1, .15, .125, and .5 and N = 100, 200, 500, 1000, and 7000. The results of these runs were very unsatisfactory. Other runs were made yielding similar results. Graphs 15, 16, 17 show what happens for N = 7000 and S = .09, .1, 15.

CONCLUSION

This preliminary study shows that the Specht estimator is highly dependent upon the choice of N and S. As N gets large, it is still necessary to choose S with case. A too small choice of S will produce very bad results. While the Specht estimator does a fair job of estimating the normal distribution, it is poor for estimating the uniform density. A more detailed study is being done to compare this estimator with other estimators found in the literature.

APPENDIX A

The probability densities are estimated by functions of the form:

$$f_i(x) = \frac{1}{(2\pi s)^{m/2}} \cdot \frac{1}{K_i} \sum_{j=1}^{K_i} \text{EXP}\left[-\frac{(x - A_{ij})^T(x - A_{ij})}{2s}\right] \quad (1)$$

$$\begin{aligned} \text{EXP}[(x - y)^T(x - y)] &= \text{EXP}[(x^T - y^T)(x - y)] \\ &= \text{EXP}[x^Tx - x^Ty - y^Tx + y^Ty] \\ &= \text{EXP}[x^Tx - 2y^Tx + y^Ty] \\ &= \text{EXP}[x^Tx]\ \text{EXP}[-2y^Tx]\ \text{EXP}[y^Ty] \end{aligned}$$

Now $\text{EXP}\left[\frac{-(x - A_{ij})^T(x - A_{ij})}{2s}\right] = \text{EXP}\left[\frac{-x^Tx}{2s}\right] \text{EXP}\left[\frac{2A_{ij}^Tx}{2s}\right] \text{EXP}\left[\frac{-A_{ij}^TA_{ij}}{2s}\right]$

$$= \text{EXP}\left[\frac{-x^Tx}{2s}\right] \text{EXP}\left[\frac{A_{ij}^Tx}{s}\right] \text{EXP}\left[\frac{-A_{ij}^TA_{ij}}{2s}\right]$$

Now Eq(1) can be written:

$$\begin{aligned} f_i(x) &= \frac{1}{(2\pi s)^{m/2}} \cdot \frac{1}{K_i} \sum_{j=1}^{K_i} \text{EXP}\left[\frac{-x^Tx}{2s}\right] \text{EXP}\left[\frac{A_{ij}^Tx}{s}\right] \text{EXP}\left[\frac{-A_{ij}^TA_{ij}}{2s}\right] \\ &= \frac{1}{(2\pi s)^{m/2}} \cdot \frac{1}{K_i} \text{EXP}\left[\frac{-x^Tx}{2s}\right] \sum_{j=1}^{K_i} \text{EXP}\left[\frac{A_{ij}^Tx}{s}\right] \text{EXP}\left[\frac{-A_{ij}^TA_{ij}}{2s}\right] \end{aligned}$$

Now let $c = \frac{1}{(2\pi s)^{m/2}} \text{EXP}\left[\frac{-x^Tx}{2s}\right]$

And then Eq(1) can be written:

$$f_i(x) = \frac{c}{K_i} \sum_{j=1}^{K_i} \text{EXP}\left[\frac{A_{ij}^Tx}{s}\right] \text{EXP}\left[\frac{-A_{ij}^TA_{ij}}{2s}\right]$$

APPENDIX B

This appendix contains a discription of the subroutines used in this report as well as a listing of each program. These programs include:

1) GEN
2) RNORM
3) PLOTK

SUBROUTINE GEN(IL,IX,IY,U,YFL)

where

IL = $2^9 + 1$

IX - initial value

IY - replaces IX for the next random number

U = 3

YFL - Random number from uniform distribution

```
      SUBROUTINE GEN(IL,IX,IY,U,YFL)
      IY = IL*IX + U
      IF(IY)5,6,6
    5 IY = IY + 2147483647 + 1
    6 YFL = IY
      YFL = YFL*.4656613E - 9
      RETURN
      END
```

SUBROUTINE RNONM(AM,S,IX,RN1,RN2)

where

AM - is the mean of the distribution

S - is the standard deviation

IX - is used to initialize GEN

RN1
RN2 - is the pair of normally distributed random numbers

```
SUBROUTINE RNORM(AM,S,IX,RN1,RN2)

Call GEN(IL,IX,IY,U,U1)

IX = IY

Call GEN(IL,IX,IY,U,U2)

PI2 = (3.14159)*2

Z = SQRT(-2.*S*ALOG(U1))

RN1 = Z*COS(PI2*U2) + AM

RN2 = Z*SIN(PI2*U2) + AM

RETURN

END
```

SUBROUTINE PLOTK(T,SS,N,NP,NA,NV,NAME)

PURPOSE - Plot the data points in an array.

T - Is an array in the main program which holds the type of distribution to be printed in the heading.

SS - This is the user supplied smoothing parameter to be printed in the heading.

N - This is the user supplied number of training points to be printed in the heading.

NP - The number of points to be plotted, less than or equal to 100.

NA - The length of the array.

NV - Number of columns in array (Number of variables to be graphed, 1st. one is the abscissa, 2nd. one plotted as 1)

NAME - Is an array that holds the data points to be plotted.

```
      SUBROUTINE PLOTK(T,SS,N,NP,NA,NV,NAME)
      INTEGER P
      REAL NAME (100)
      DIMENSION DF(20)
      DIMENSION SCALE(6),LINE(103),INJ(9),T(2)
      WRITE(3,12) T,N,SS
12    FORMAT(1H1,T68,'GRAPH'//T50,'PLOTS OF TRUE AND ESTIMATED DENSITY F
     1UNCTIONS'/T51,'1 DENOTES TRUE',2A4,', 2 DENOTES ESTIMATED'/T60,N
     2= ',I6,T71,'S =',F5.2)
      KBLNK=1077952576
      KSTAR=1547714624
      INJ(1)=-247447488
      INJ(2)=-230670272
      INJ(3)=-213893056
      INJ(4)=-197115840
      INJ(5)=-180338624
      INJ(6)=-163561408
      INJ(7)=-146784192
      INJ(8)=-130006976
      INJ(9)=-113229760
      KPLUS=1312833600
      KMNUS=1614823488
C     DETERMINE SCALE FACTORS
      K=NA+1
      K1=NA+NP
      YMAX=NAME(K)
      YMIN=NAME(K)
      KV=NV-1
      DO 1000 I=1,KV
      DO 25 J=K,K1
      ARRAY=NAME(J)
      IF(YMAX-ARRAY)10,15,15
10    YMAX=ARRAY
15    IF(YMIN-ARRAY)25,25,20
20    YMIN=ARRAY
25    CONTINUE
      K=((I+1)*NA)+1
      K1=(I+1)*NA+NP
      IF(YMIN .GT. 0.) YMIN = 0.
1000  CONTINUE
      H1=YMAX-YMIN
      H12=H1*.2
      H1=100./H1
      SCALE(1)=YMIN
      DO 30 I=2,6
30    SCALE(I)=SCALE(I-1)+H12
C     PRINT SCALES AND BORDERS
      WRITE(8,1001)SCALE
1001  FORMAT(1H0,12X,F8.2,5(12X,F8.2))
      DO 35 I=1,103
35    LINE(I)=KMNUS
      DO 40 I=2,102,10
40    LINE(I)=KPLUS
      WRITE (8,1002)LINE
1002  FORMAT(9X,'ABS',5X,103A1)
C     PLOT MAIN BODY OF GRAPH
      DO 100 I=1,NP
      DO 102 L=1,103
102   LINE(L)=KBLNK
```

```
      LINE(1)=KMNUS
      LINE(103)=KMNUS
      X=NAME(I)
      K4=NV-1
      DO 90 J=1,K4
      P=(NA*J)+I
      NORMV=(NAME(P)-YMIN)*H1+2.5
      IF(NORMV) 200,200,80
200   NORMV = 2
80    LINE(NORMV)=INJ(J)
90    CONTINUE
      WRITE(8,1003)X,LINE
1003  FORMAT(3X,F10.2,4X,103A1)
100   CONTINUE
C     PRINT BOTTOM LINE AND SCALE
      DO 105 I=3,101
105   LINE (I)=KMNUS
      DO 110 I=2,102,10
110   LINE(I)=KPLUS
      WRITE(8,1002) LINE
      WRITE (8,1004) SCALE
1004  FORMAT(12X,F8.2,5(12XF8.2))
      RETURN
      END
```

APPENDIX C

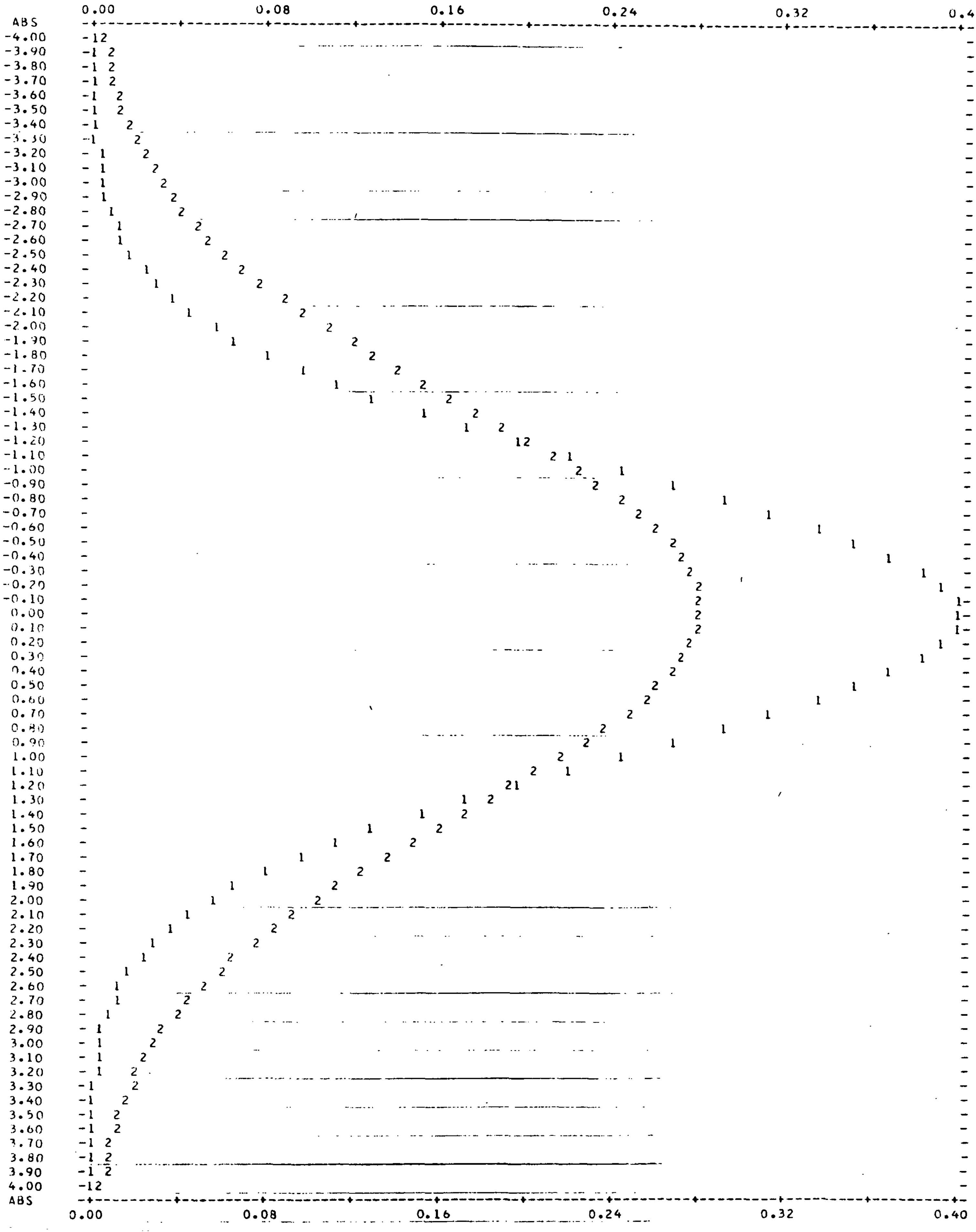
In graphs 1-6, we have held the number of sample points at 7000 and varied the smoothing parameter, s, from 1. down to .01. It is evident that as s becomes smaller, our estimated values (graph with 2's) very closely fit the true curve (graphed wtih 1's).

The next set of graphs, 7-10, shows what happens when we fixed s at .05 and varied n from 100 to 1000. Graphs 11-14, exhibit the same information for s = .07.

The next set of graphs 15-17, shows the results of using the Specht estimator to estimate the uniform density. Hence N = 7000, and s = .09, .10, .50.

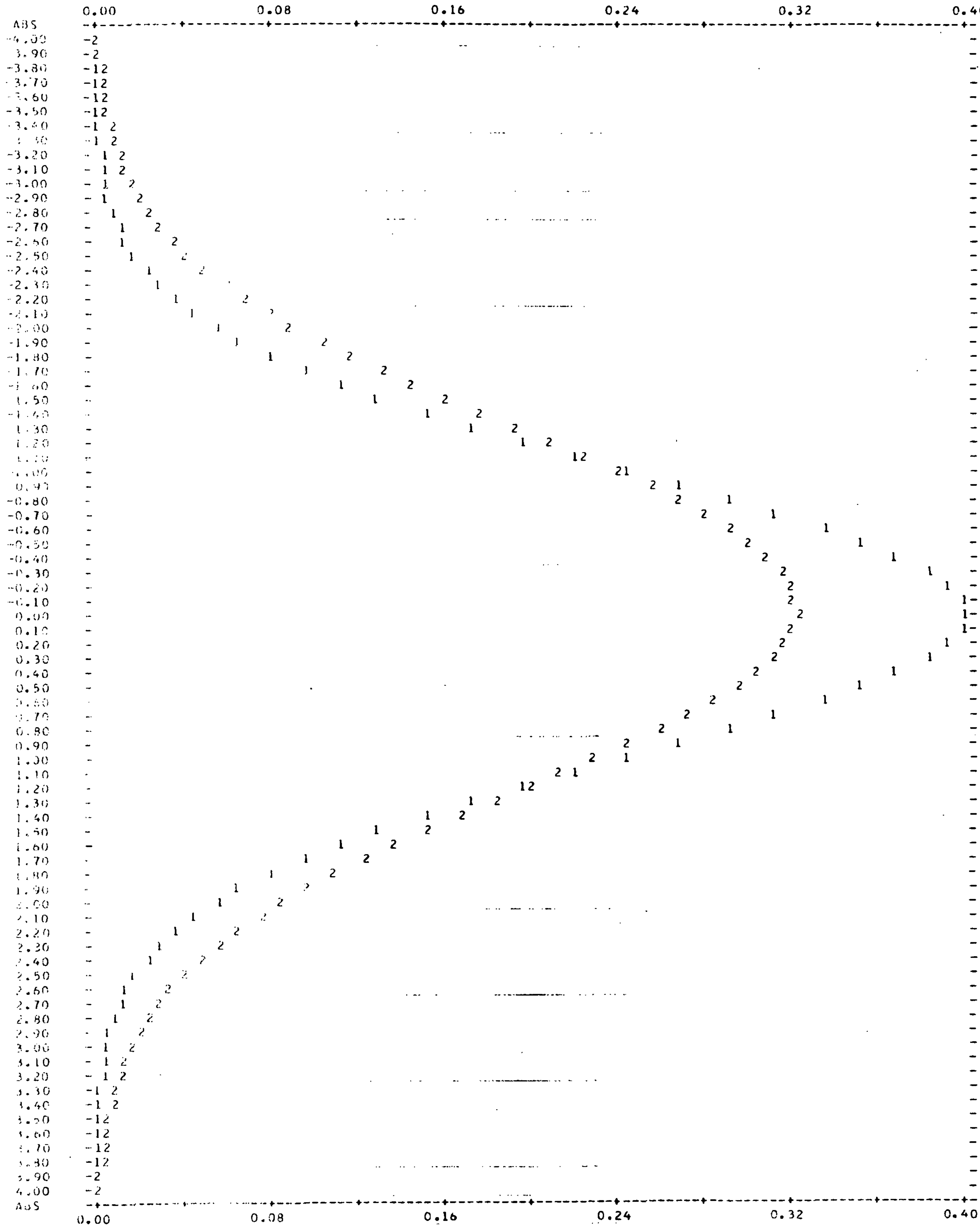
GRAPH 1

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 1.00



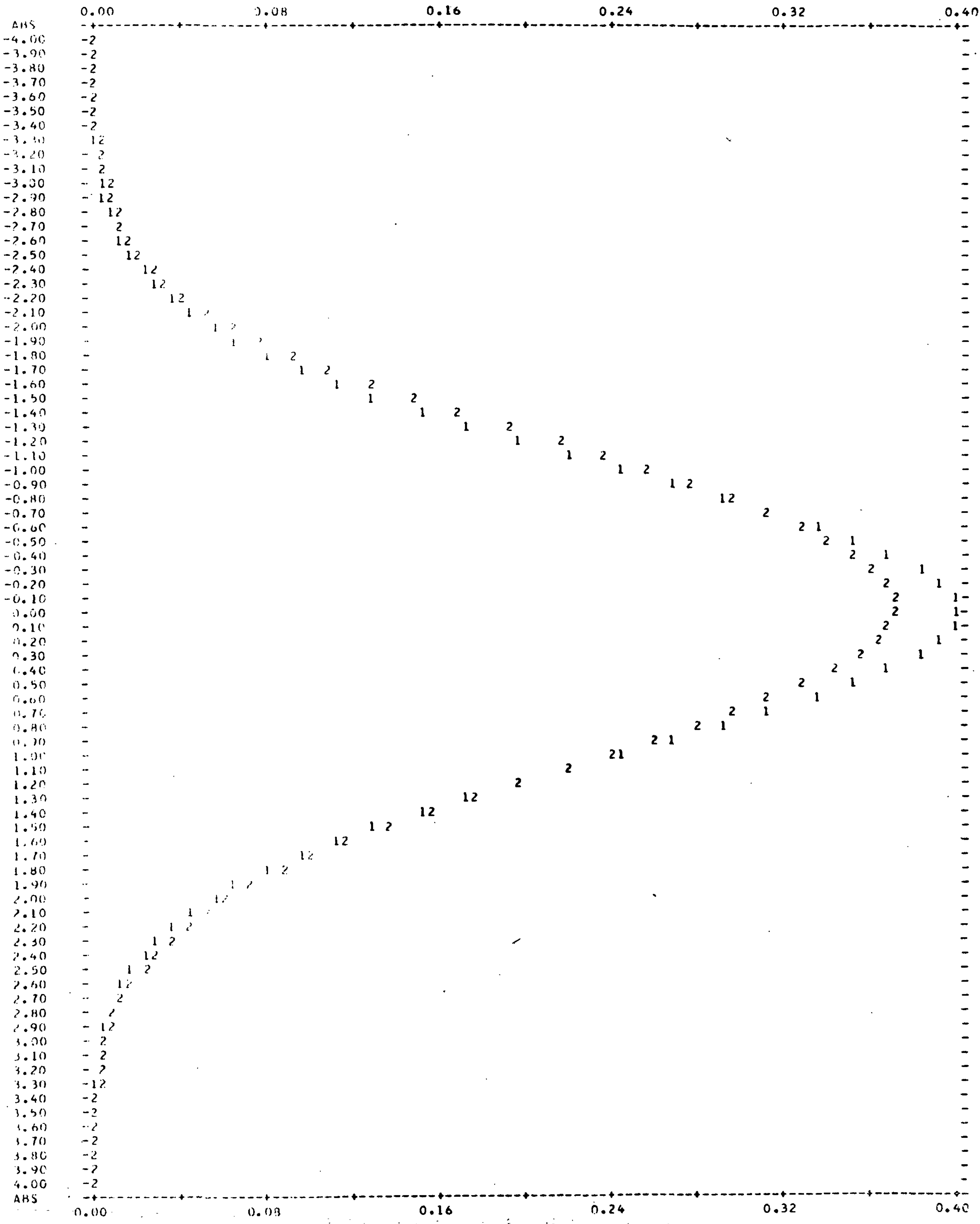
GRAPH 2

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.50



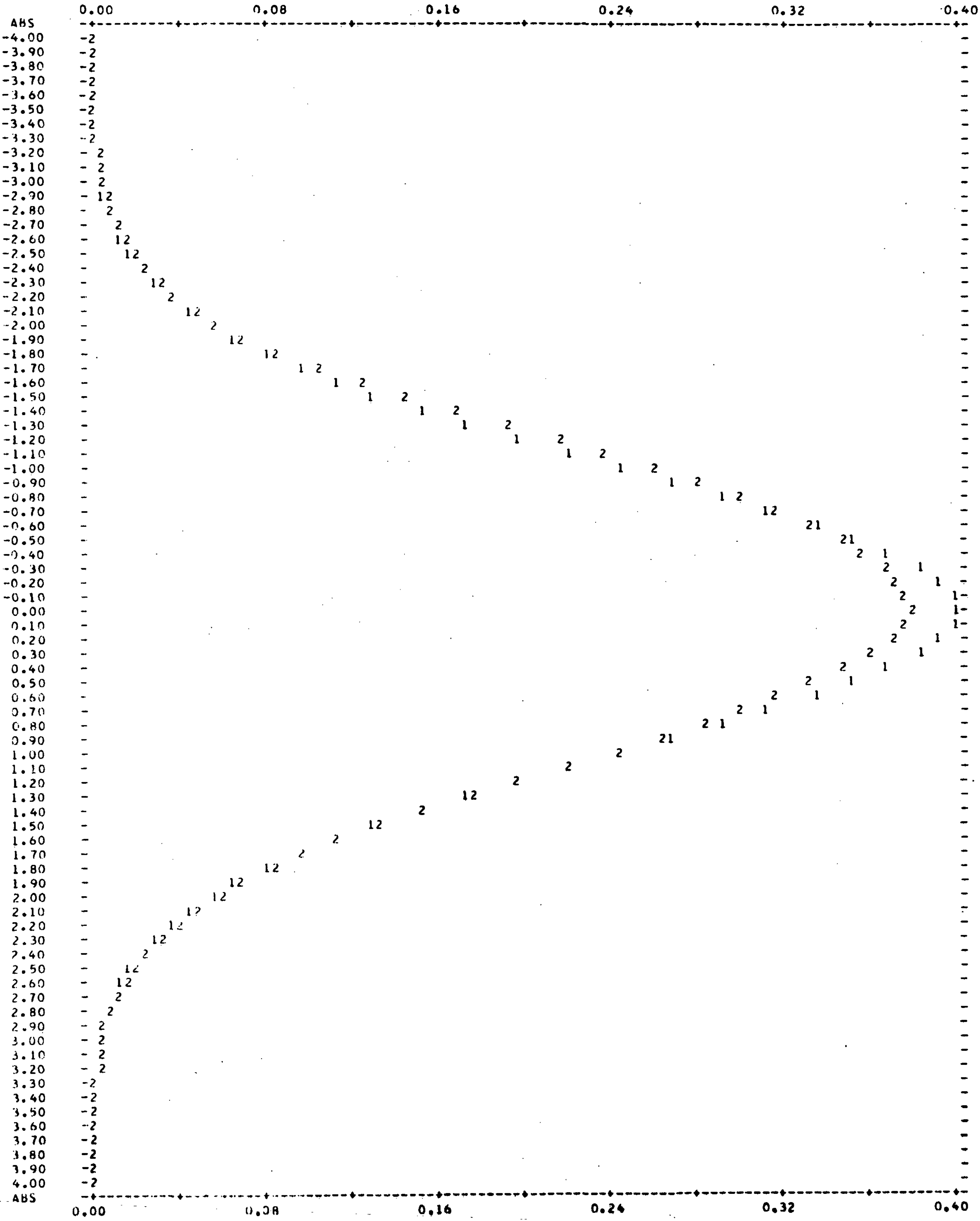
GRAPH 3

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.10



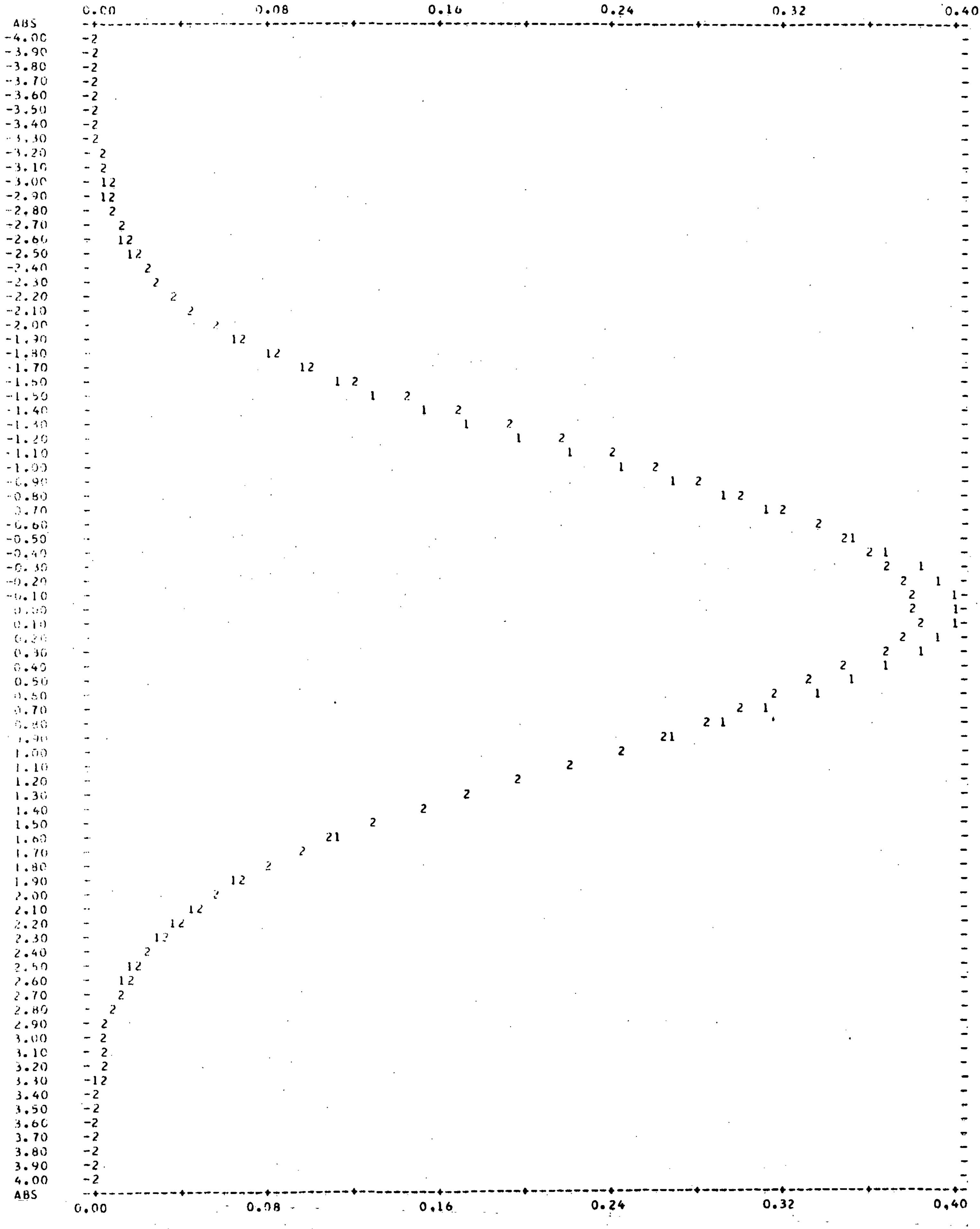
GRAPH 4

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.05



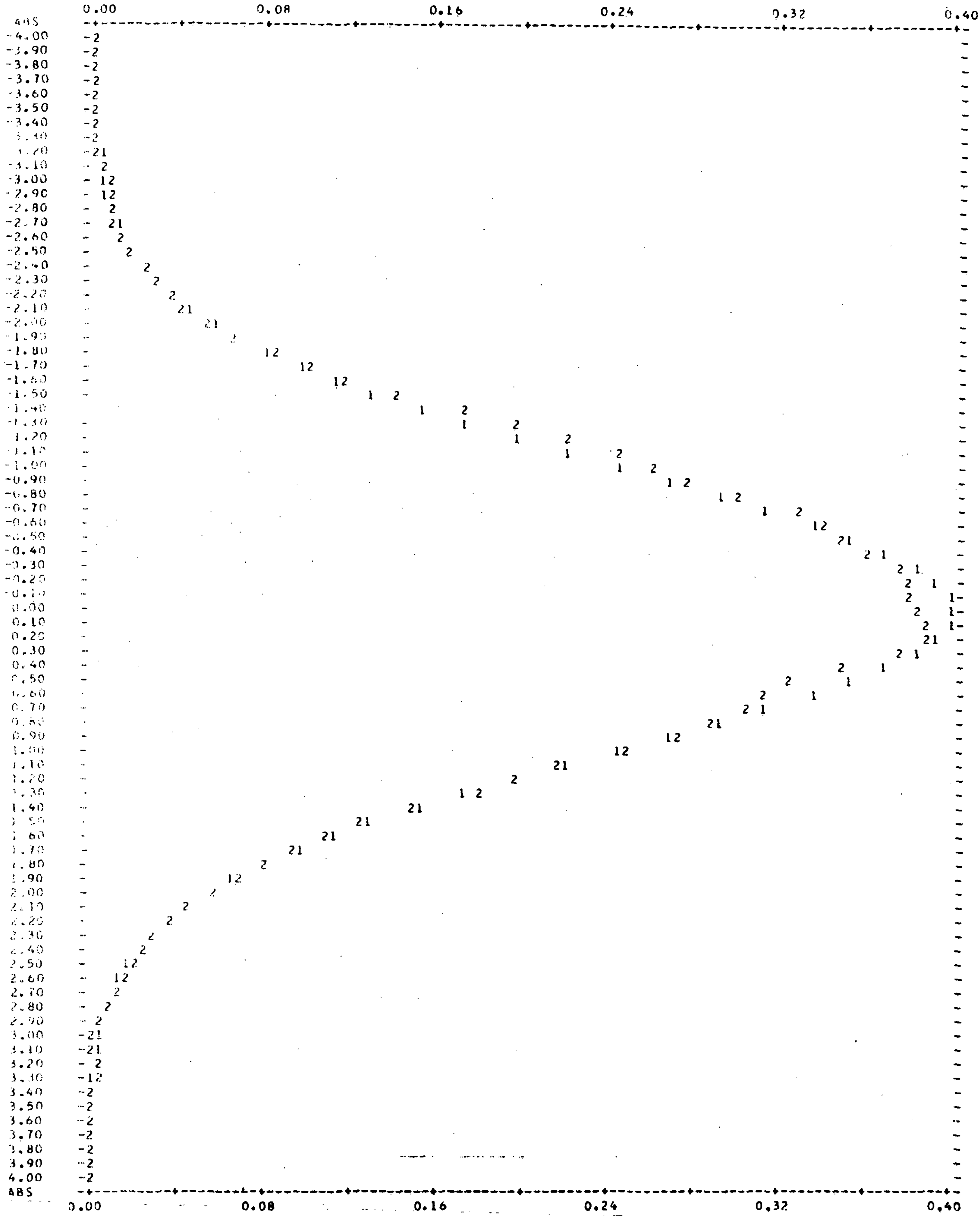
GRAPH 5

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.03



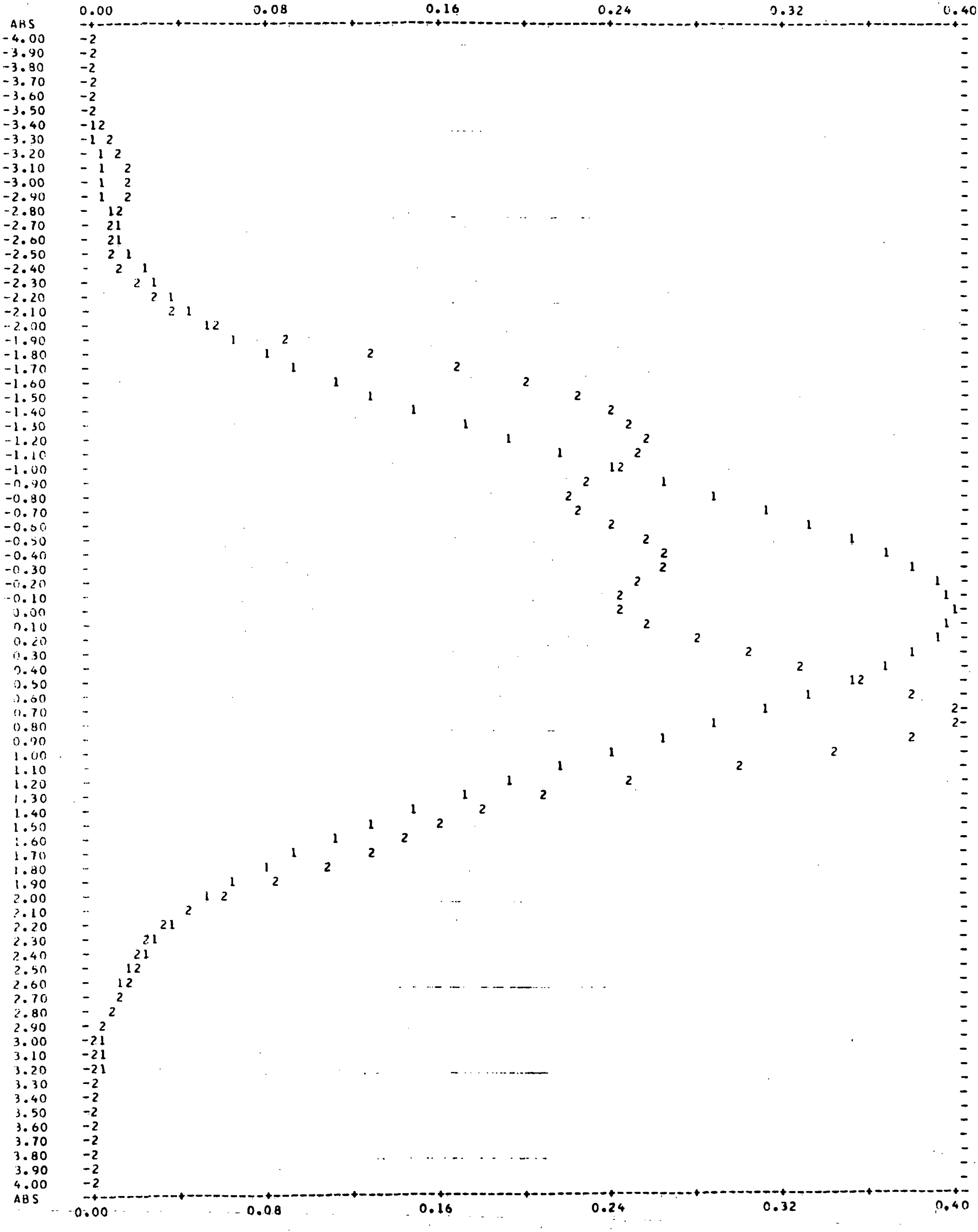
GRAPH 6

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.01



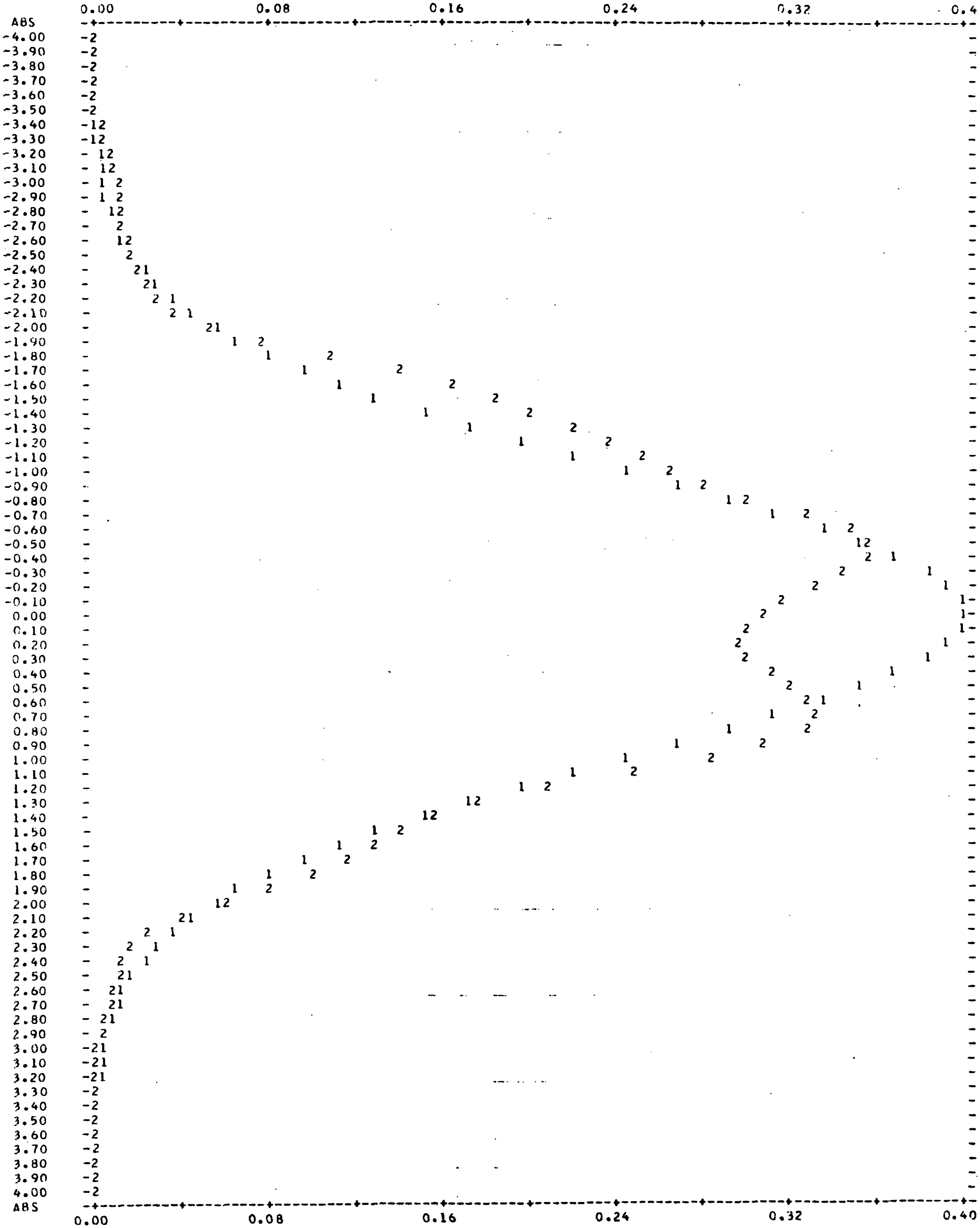
GRAPH 7

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 100 S = 0.05



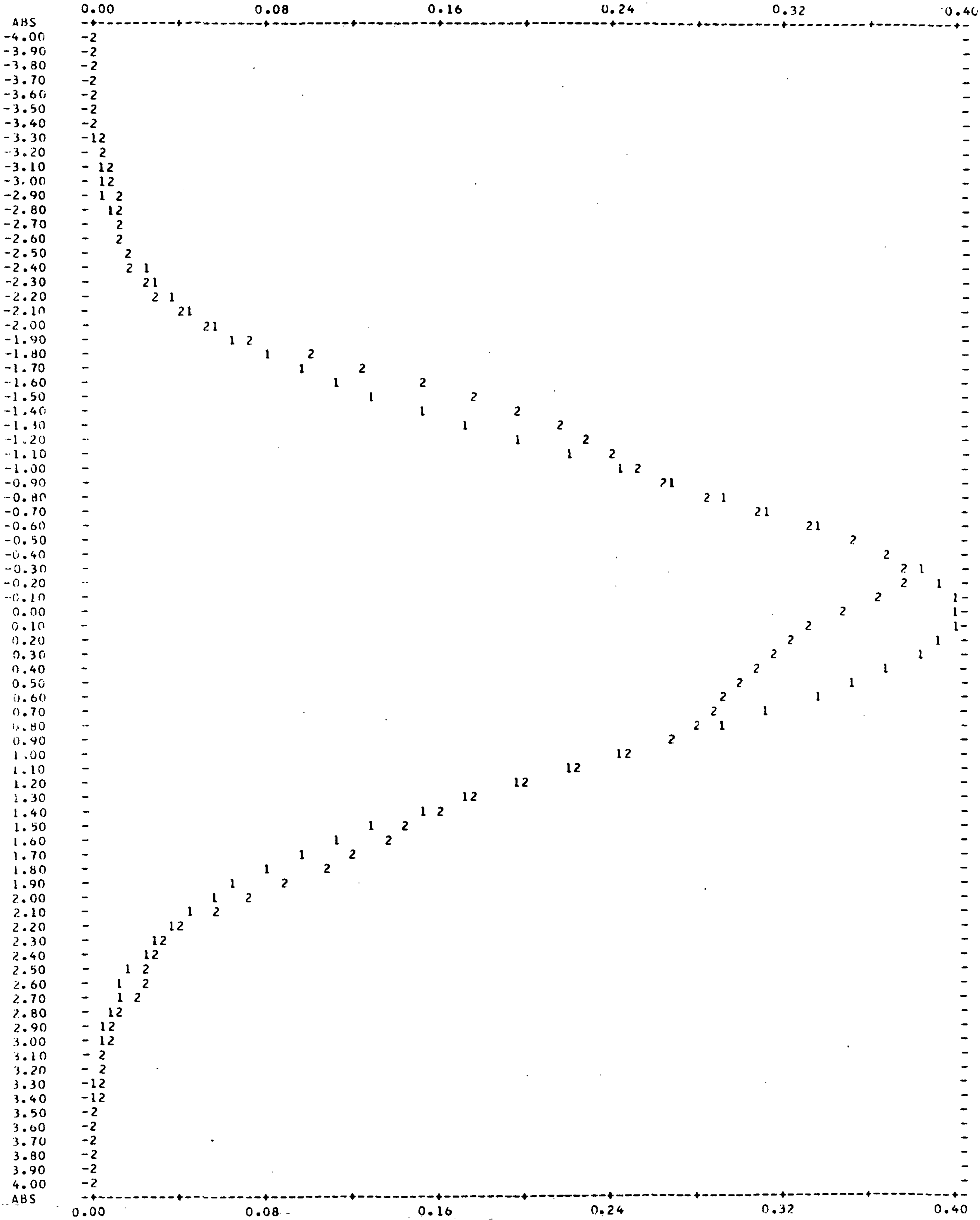
GRAPH 8

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 200 S = 0.05



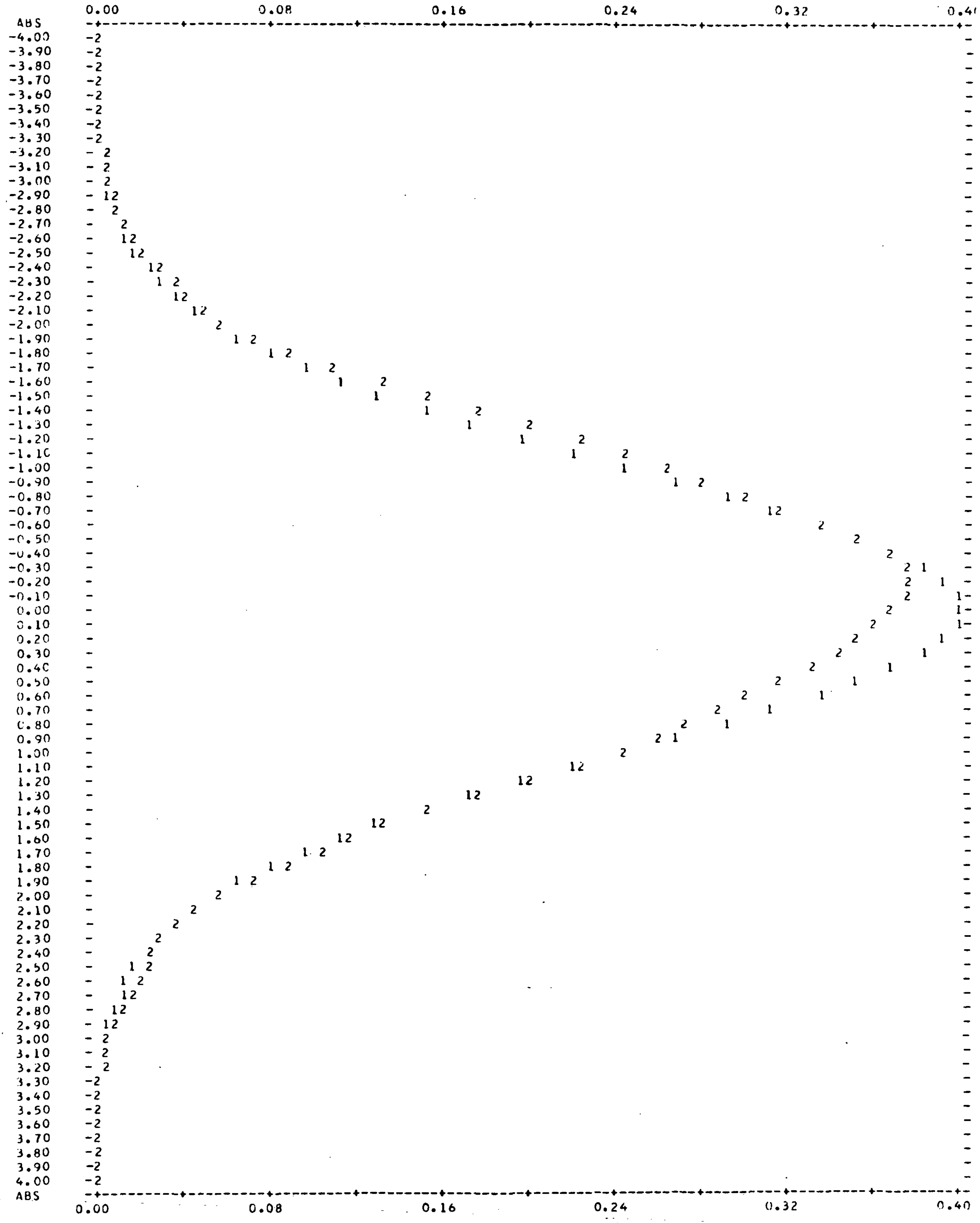
GRAPH 9

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 500 S = 0.05



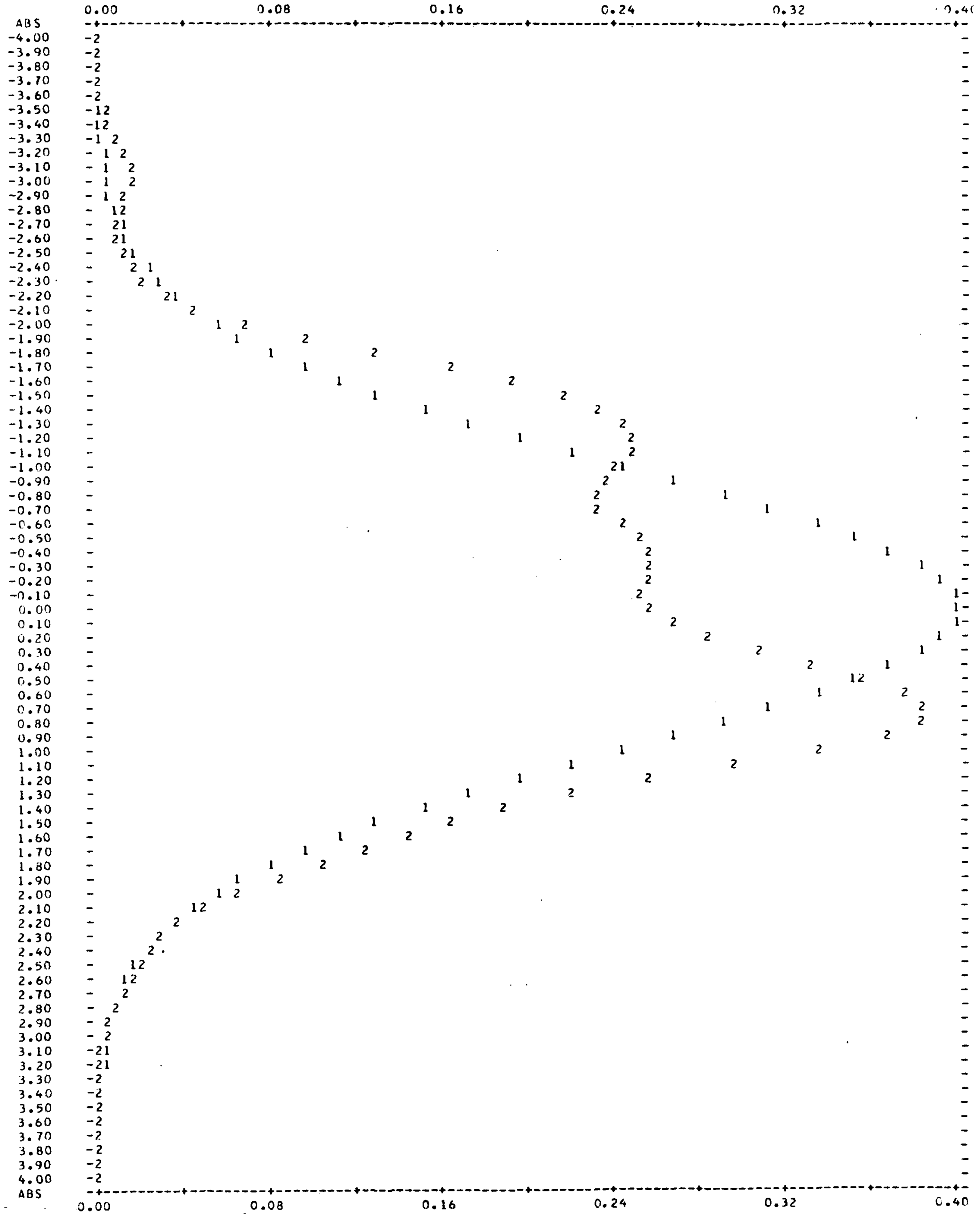
GRAPH 10

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 1000 S = 0.05



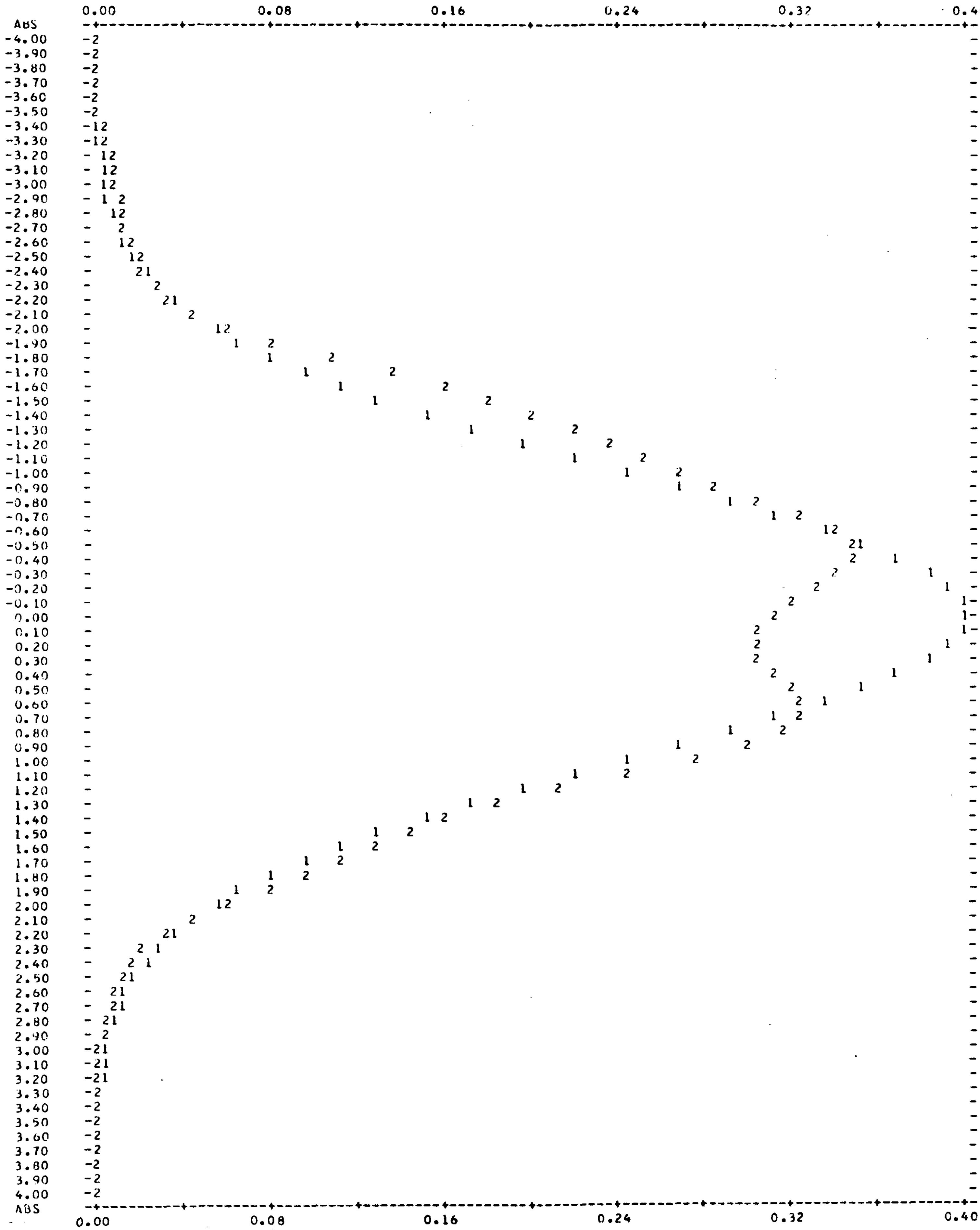
GRAPH 11

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 100 S = 0.07



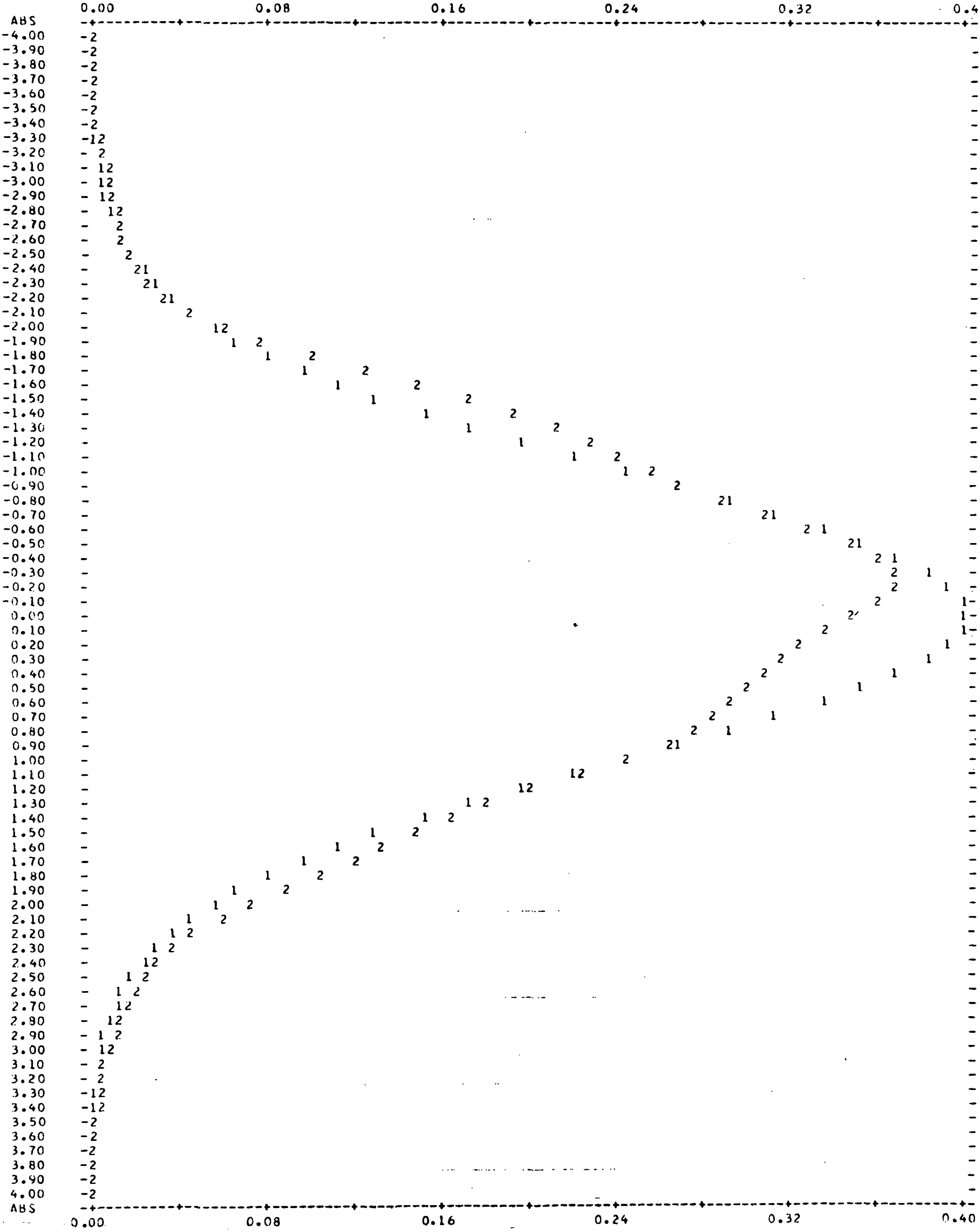
GRAPH 12

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 200 S = 0.07



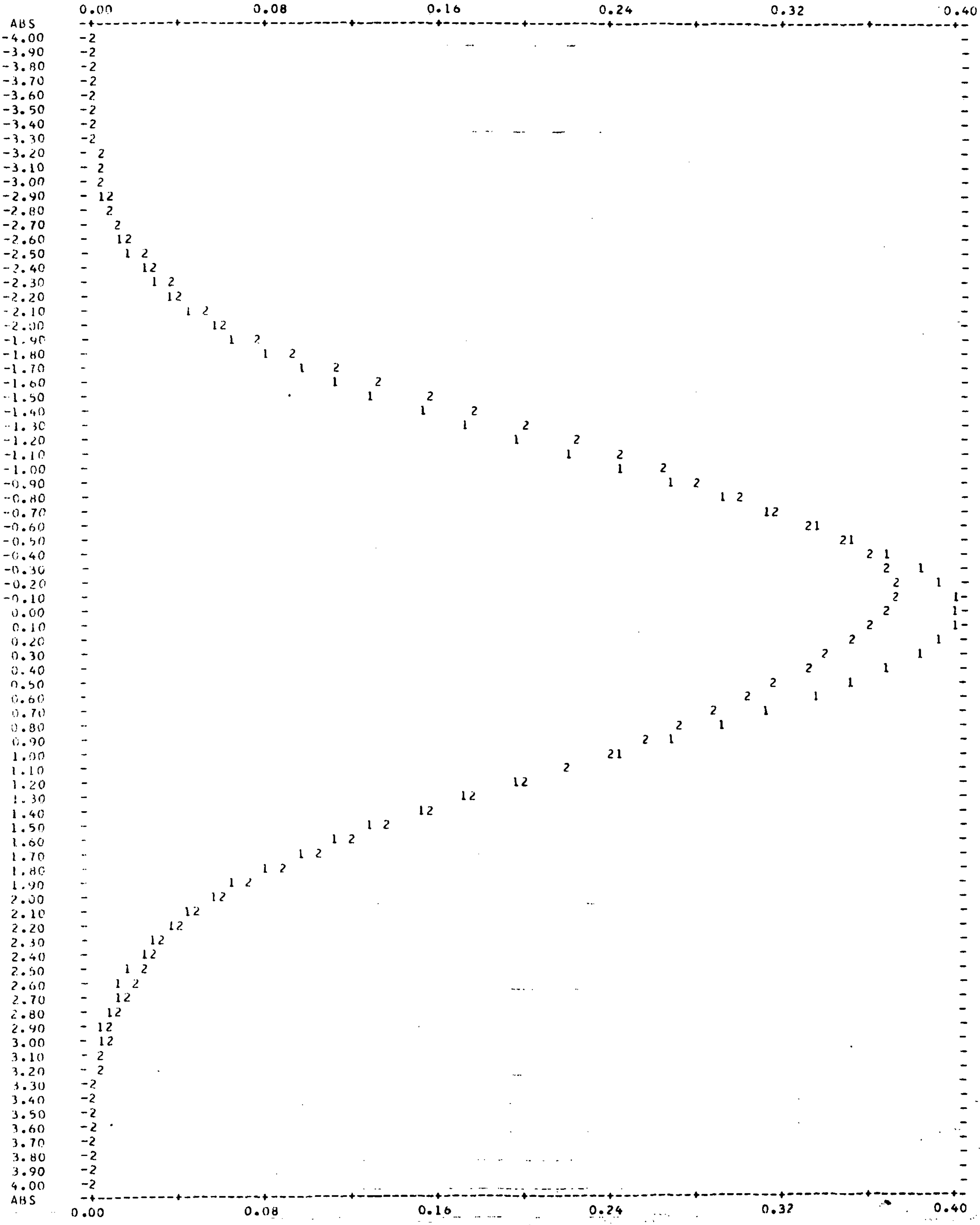
GRAPH 13

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 500 S = 0.07



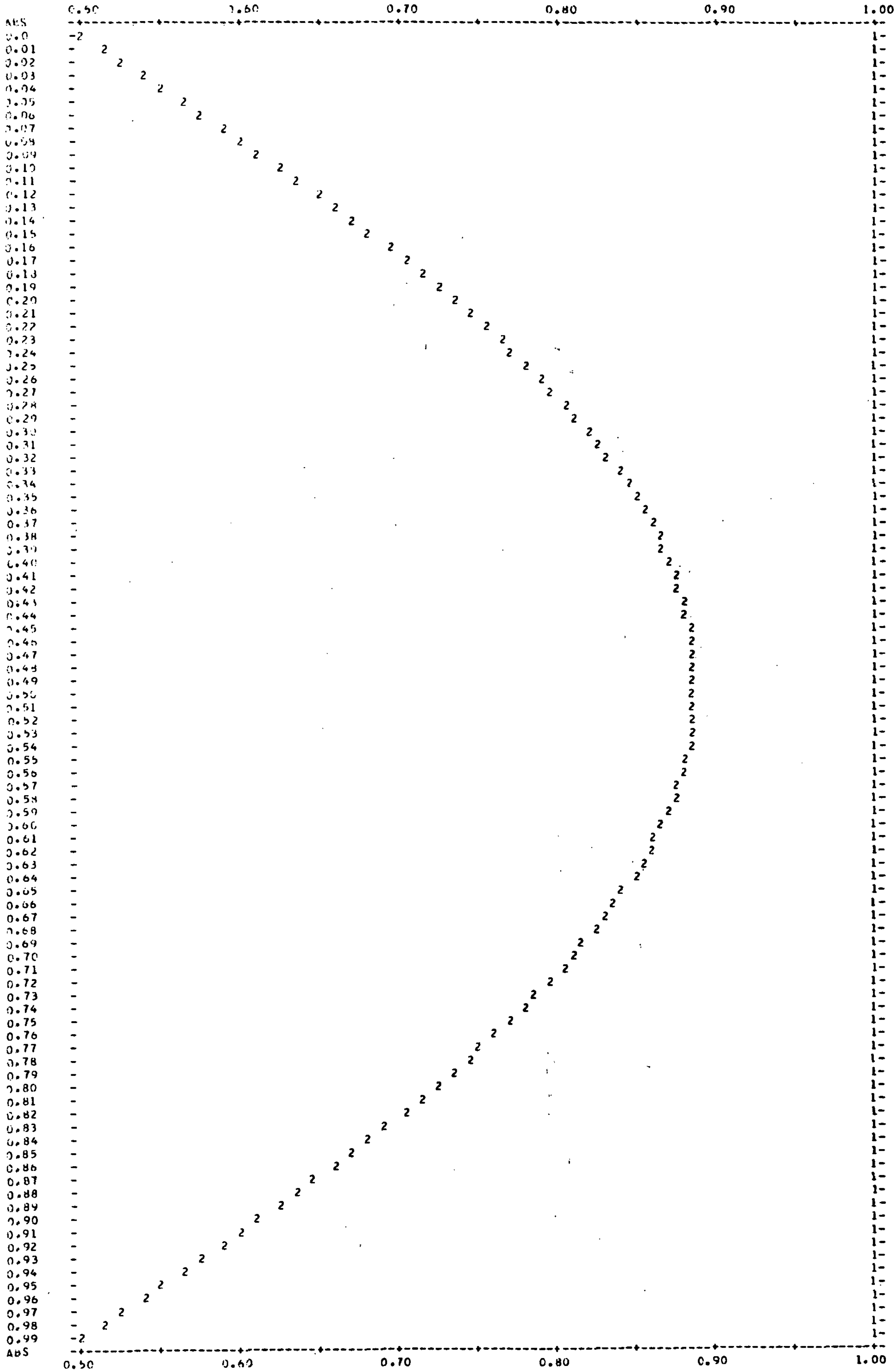
GRAPH 14

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE N(0,1) , 2 DENOTES ESTIMATED
N = 1000 S = 0.07



Graph 15

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE U(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.10

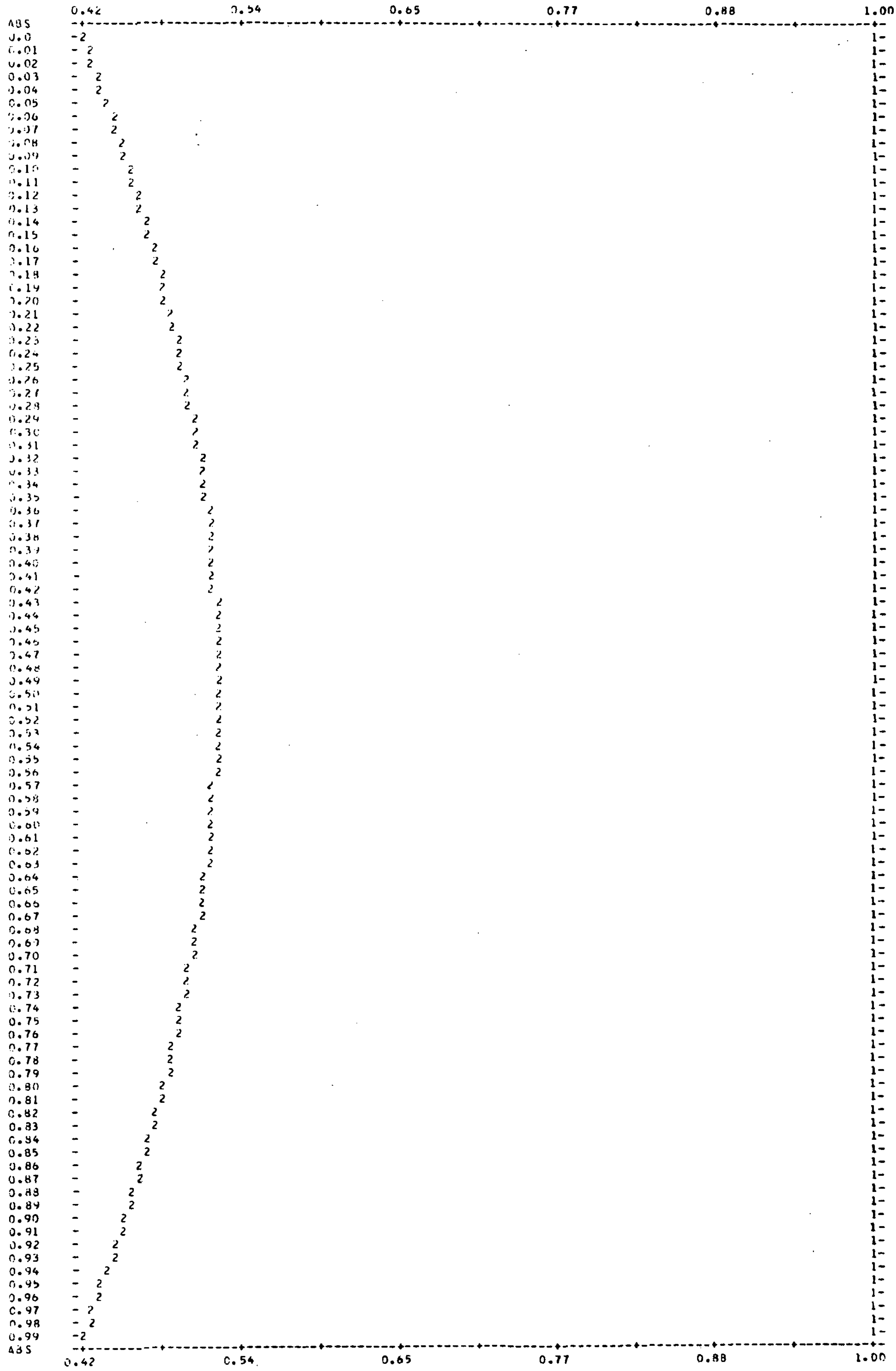


IX= 204186593

THE MAXIMUM ABSOLUTE VALUE OF THE DIFFERENCES OF FN AND FNX IS 0.495 FOR N = 7000

Graph 16

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE U(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.50

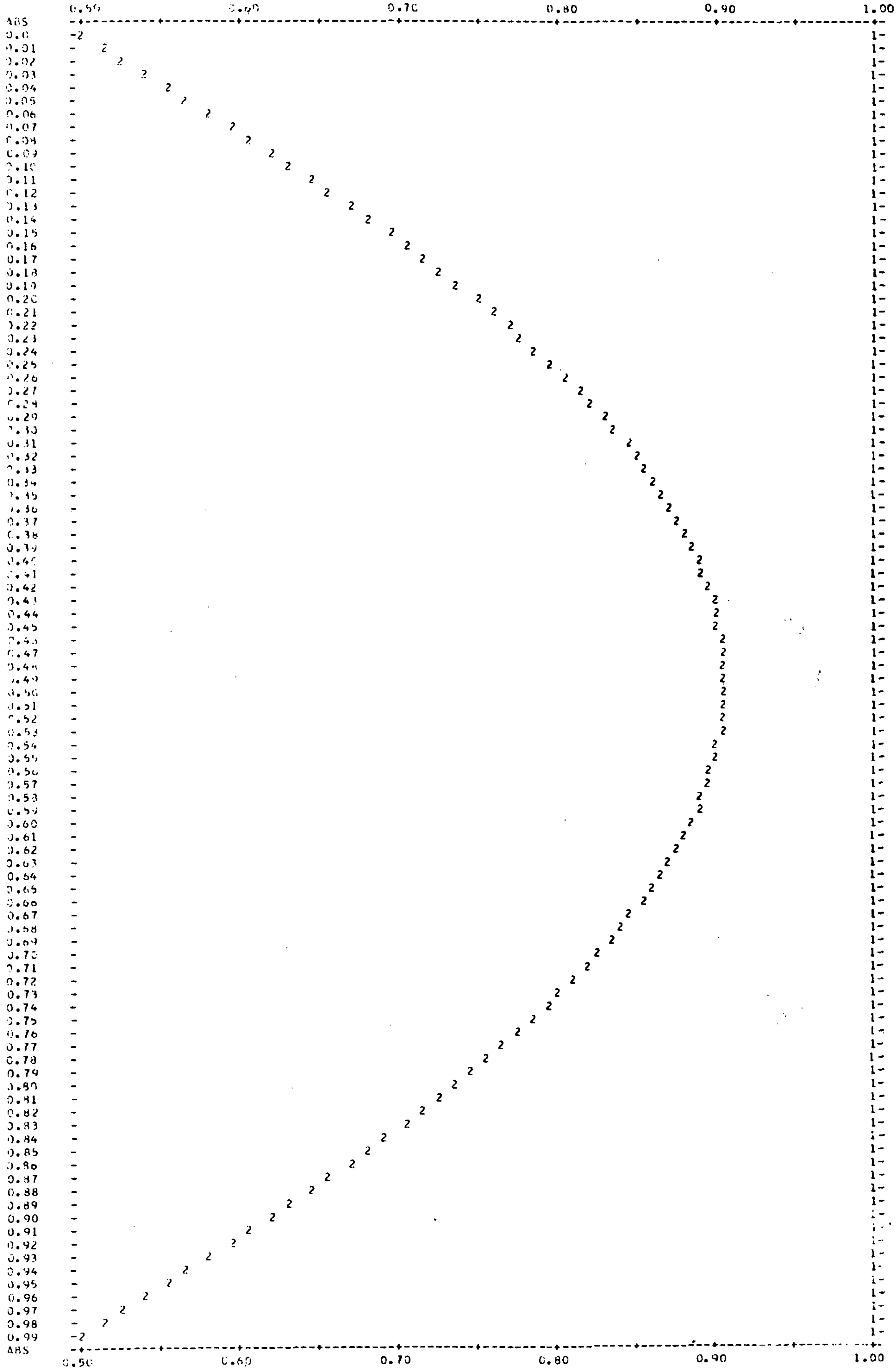


IX= 204186593

THE MAXIMUM ABSOLUTE VALUE OF THE DIFFERENCES OF FN AND FNX IS 0.496 FOR N = 7000

Graph 17

PLOTS OF TRUE AND ESTIMATED DENSITY FUNCTIONS
1 DENOTES TRUE U(0,1) , 2 DENOTES ESTIMATED
N = 7000 S = 0.09



IX= 204186593