RL-TR-95-164, Vol II (of five)
Final Technical Report
September 1995

HIGH-LEVEL ADAPTIVE SIGNAL PROCESSING ARCHITECTURE WITH APPLICATIONS TO RADAR NON-GAUSSIAN CLUTTER, A New Technique for Distribution Approximation of Random Data

University of Massachusetts at Amherst

Rajiv R. Shah (Syracuse University)

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Rome Laboratory
Air Force Materiel Command
Griffiss Air Force Base, New York

This report has been reviewed by the Rome Laboratory Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RL-TR-95-164, Vol II (of five), has been reviewed and is approved for publication.

APPROVED: 

DR. VINCENT C. VANNICOLA
Project Engineer

FOR THE COMMANDER: 

DONALD W. HANSON
Director of Surveillance & Photonics

If your address has changed or if you wish to be removed from the Rome Laboratory mailing list, or if the addressee is no longer employed by your organization, please notify RL (OCSS) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1995	Final Apr 91 - Jun 94

4. TITLE AND SUBTITLE
HIGH-LEVEL ADAPTIVE SIGNAL PROCESSING ARCHITECTURE WITH APPLICATIONS TO RADAR NON-GAUSSIAN CLUTTER , A New Technique for Distribution Approximation of Random Data

5. FUNDING NUMBERS
C - F30602-91-C-0038
PE - 62702F
PR - 4506
TA - 11
WU - 1B

6. AUTHOR(S)
Rajiv R. Shah (Syracuse University)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts at Amherst
Department of Computer Science
Lederle Graduate Research Center
Amherst MA 01003

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Rome Laboratory (OCSS)
26 Electronic Pky
Griffiss AFB NY 13441-4514

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
RL-TR-95-164, Vol II (of five)

11. SUPPLEMENTARY NOTES
Rome Laboratory Project Engineer: Dr. Vincent C. Vannicola/OCSS/(315) 330-2861

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
We describe the accomplishments of a collaborative effort carried out by a team of researchers from the University of Massachusetts at Amherst (headed by V. Lesser), Boston University (headed by H. Nawab), and Syracuse University (headed by D. Weiner) on the development of a high-level signal processing architecture called IPUS (Integrated Processing and Understanding of Signals). Based on the IPUS generic testbed architecture and "radar and sound understanding" (RESUN) architectures, we have been able to transfer IPUS technology from a LISP environment to a C++ environment for use in an IPUS Radar Clutter Analysis Testbed. Though not as well-developed as the sound understanding application because of its newness, this radar testbed has still clearly demonstrated the potential of IPUS-like technologies for CFAR processing of radar returns. There has also been significant development of knowledge for weak signal detection. This knowledge has involved the application of the Ozturk algorithm to hypothesize the distribution of data in a clutter patch based on a small amount of data. Also, techniques have been developed for partitioning the radar surveillance volume into background noise and clutter patches, for weak signal detection in K-distributed clutter, and the efficient use of Rejection Theorem for Weibull clutter generation. Though the effort on the application of IPUS to communications was given less priority, we still did some interesting theoretical work on (see reverse)

14. SUBJECT TERMS
Radar, Signal processing, Artificial intelligence detection.
Clutter, Non-Gaussian

15. NUMBER OF PAGES
102

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

13. (Cont'd)

weak signal detection in communication systems subject to spherically invariant random processes (SIRP) interference.

Acknowledgements

This work was supported by the Rome Laboratory, U.S. Air Force, under contract number F30602-91-C-0038, which I gratefully acknowledge. I would like to thank my parents for their patience and support during the course of this effort. I would like to thank my advisor, Dr. Donald.D.Weiner, for his patience, encouragement, support and thought provoking suggestions which contributed significantly to this effort. Thanks are also due to Lisa Slaski, Dr. Prakash Chakravarthi, Dr. Muralidhar Rangaswamy and Mohammed Slamani for their help, suggestions and many interesting and stimulating discussions. I would also like to thank the Department of Electrical and Computer Engineering and Academic Computing Services of Syracuse University for the use of their excellent computing and other resources that made this effort possible.

Abstract

This thesis deals with the analysis of random data. Two approaches are discussed. The first approach is a Goodness of Fit test to determine whether or not random data samples are statistically consistent with a prespecified probability distribution. The well-known Kolmogorov-Smirnov test, Chi-Square test, Q-Q Plots and P-P plots are reviewed and illustrated by means of several examples. A new algorithm, the Ozturk Algorithm, is introduced.

The second approach deals with approximation of the underlying probability density function of random data samples. The previously mentioned well-known tests are not suitable for this task. However, the Ozturk Algorithm provides a powerful solution for this problem with a nice graphical interpretation. Finally, computer simulated results obtained with the Ozturk Algorithm are presented and discussed.

Contents

List of Figures

List of Tables

Chapter 1

Literature Review

1.1 Introduction

In the analysis of random data, we encounter situations where there may be various statistical models or "hypotheses" that need to be checked against the data. The usual situation is that one has a particular probability distribution in mind to be tested or checked for consistency in representing data from a certain experiment. The hypothesis that this distribution is the right one is called the *null hypothesis*, often denoted by H_0. This hypothesis may have emerged from long experience associated with an experiment and it is desired to see whether the hypothesis is still correct when there has been some change in circumstances that call it into question. Alternatively, the hypothesis may be the result of a theoretical analysis or a logical argument and the theory needs to be verified.

A null hypothesis is ordinarily taken to be quite specific. In particular, location, scale and shape parameters associated with the probability density function are specified along with the type of distribution. For example, the probability density function of the Weibull distribution changes as its shape parameter is changed. Therefore, Weibull distribution having a different shape parameter from the Weibull distribution of the null hypothesis are assumed to be different. All the other distributions, taken together, define what is referred to as the alternative hypothesis, denoted by H_1. Therefore, another question that arises in the analysis of random data is "If the null hypothesis is not true, what are suitable approximations to the underlying distribution of the data?"

To answer these questions, several tests have been proposed and used. Each of these

tests have their own strengths and weaknesses. Some may work well for a particular set of density functions, but poorly for others. We focus our attention on four of the most frequently used tests for analyzing random data. Detailed discussions of these tests follow.

1.2 The Kolmogorov-Smirnov Test

This test is based on the idea of a "sample distribution function", a statistic that is the sample version of the population distribution function.

Given a sample $(X_1, X_2, \ldots, X_n)$ of size n, the *sample distribution function* is the cumulative distribution function (cdf) of a discrete probability density function where the random variable assumes the values $X_1, X_2, \ldots, X_n$ with probability $1/n$. Consequently, the cdf increases in steps of size $1/n$ at each sample value, rising from 0 to the left of the smallest X_i to 1 at the largest X_i.

1.2.1 Example[13]

Consider a sample of the following 5 observations.

$$2.22, -0.83, 0.18, 1.18, 2.05.$$

The sample distribution function is easily constructed after the sample values are marked on the x-axis: Starting at height zero for the values less than -0.83, the cdf is increased successively by steps of height 1/5 at the ordered sample value -0.83, 0.18, 1.18, 2.05, and 2.22. The result is shown in fig.1.1.

The Kolmogorov-Smirnov test statistic is defined as the maximum absolute vertical deviation D_n, of the sample distribution function, $F_n(\mathrm{x})$, from the cdf, $F_0(\mathrm{x})$, specified by the null hypothesis H_0. If the fit is good, D_n is expected to have a small value. On the other hand, if the underlying distribution has a cdf significantly different from $F_0(\mathrm{x})$, it is expected that the fit will be poor and D_n will be large. Consequently, if values of D_n exceed a pre-specified value, called the acceptance limit, H_0 is rejected. Fortunately, the distribution of the statistic D_n depends only on the sample size and not on the shape of the distribution being tested. The distribution of D_n has been computed under the assumption that the null hypothesis holds. Results of the acceptance limits are given in table 1.1 [12] and [13] for different sample sizes and for various pre-selected values of

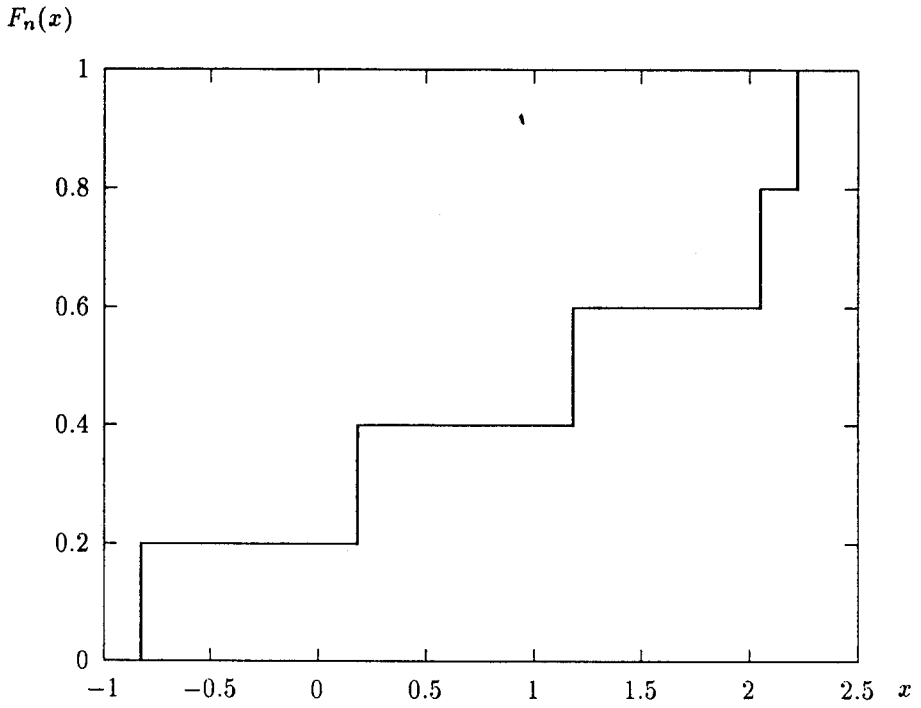

Figure 1.1: Sample Distribution function for example 1.1.1

$$\alpha = \Pr(\text{reject } H_0 | H_0 \text{ true}) \tag{1.1}$$

where α is called the significance level. For large values of n, asymptotic formulae are given for the acceptance limits. In summary, the test consists of the following steps:

1) Plot F_n(x) and F_0(x) in the co-ordinate axes.

2) By inspection, determine the maximum vertical absolute deviation, given by,

$$D_n = \overset{max}{x} |F_n(x) - F_0(x)|. \tag{1.2}$$

3) Select a significance level

$$\alpha = \Pr(\text{reject } H_0 | H_0 \text{ true}). \tag{1.3}$$

4) Accept H_0 if $D_n \leq K$ and reject otherwise.

Note that

$$1 - \alpha = \Pr(\text{accept } H_0 | H_0 \text{ true}). \tag{1.4}$$

Let the cdf of D_n be denoted by $F_{D_n}(x|H_0)$. It follows that

$$Pr(D_n > K|H_0) = 1 - F_{Dn}(K|H_0) = \alpha. \tag{1.5}$$

Consequently, K is the $100 \times (1 - \alpha)$ percentile of $F_{D_n}(x|H_0)$.

1.2.2 Example[13]

Let the null hypothesis F_0(x) be Gaussian with mean = 32 and standard deviation = 1.8. Consider the 10 observations: 31.0, 31.4, 33.3, 33.4, 33.4, 33.5, 33.7, 34.4, 34.9, 36.2. The corresponding sample distribution function F_n(x), is sketched in fig.1.2. along with the normal distribution whose mean is 32 and whose standard deviation is 1.8. Assume the significance level is chosen to be $\alpha = 0.05$. From fig.1.2 it is determined that the maximum deviation D_n, between the two curves is 0.56. From table 1.1 the acceptance limit is $K = 0.409$. Since $D_n > K$, H_0 is rejected.

Although the Kolmogorov-Smirnov test is found to perform quite well even for small sample sizes, it has two principal disadvantages.

1. To perform the test it is necessary to have a priori knowledge about the data in order to be able to specify meaningful null hypotheses.
2. When the null distribution is rejected. no information is provided about which distributions are suitable for approximating the underlying distribution of the data.

1.3 The Chi-Square Test

The chi-square test was originally developed for discrete random variables. It is applied to the case of continuous random variables by making a discrete approximation to the continuous probability density function. Since the distribution of the statistic used becomes tractable only as the sample size becomes infinite, the chi-square test should be employed only for large sample sizes.

Consider a null hypothesis with probability density function $f_0(x)$ and distribution function $F_0(x)$, as shown in fig.1.3. Divide the x-axis into k contiguous intervals

Sample size	**Significance level**				
(n)	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.829
4	0.494	0.525	0.564	0.624	0.734
5	0.446	0.474	0.510	0.563	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.409	0.486
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.375
16	0.258	0.274	0.295	0.328	0.391
17	0.250	0.266	0.286	0.318	0.380
18	0.244	0.259	0.278	0.309	0.270
19	0.237	0.252	0.272	0.301	0.361
20	0.231	0.246	0.264	0.294	0.352
25	0.21	0.22	0.24	0.264	0.32
30	0.19	0.20	0.22	0.242	0.29
35	0.18	0.19	0.21	0.23	0.27
40				0.21	0.25
50				0.19	0.23
60				0.17	0.21
70				0.16	0.19
80				0.15	0.18
90				0.14	
100				0.14	
Asymptotic formula:	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Table 1.1: Acceptance limits for the Kolmogorov-Smirnov test

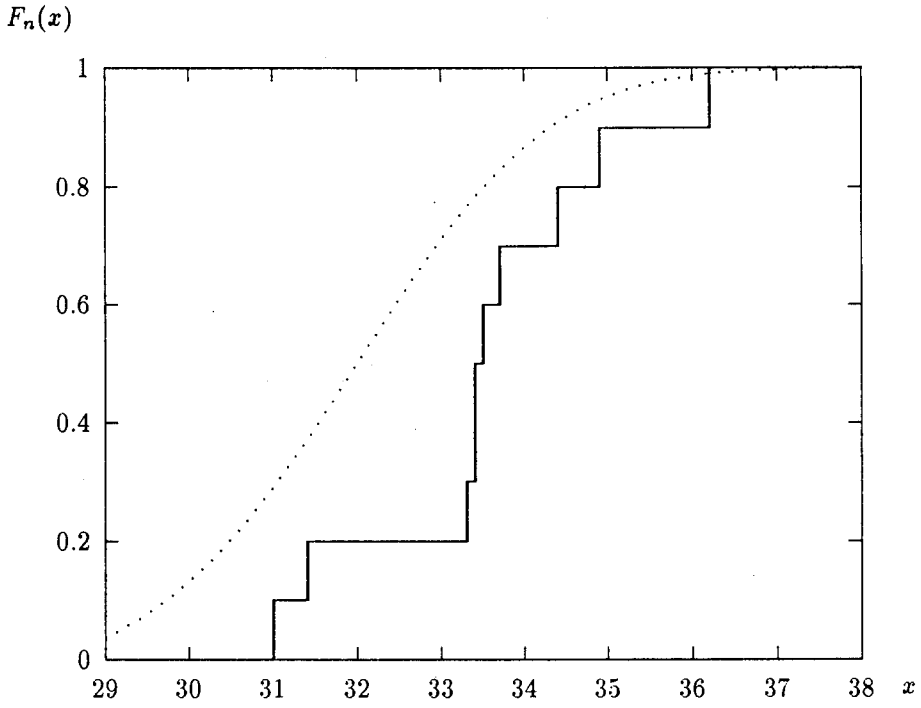

Figure 1.2: The Distribution function for example 1.2.2.

$E_1, E_2, \ldots, E_k$ from left to right. Note that the i^{th} interval, E_i, consists of the set of points such that

$$a_{i-1} \leq x \leq a_i$$

where $a_0 = -\infty$ and $a_k = +\infty$. Consequently, a_{i-1} and a_i are the end points of the i^{th} interval, E_i.

Define the probabilities

$$\begin{aligned} p_i &= Pr(X \in E_i) = Pr(a_{i-1} \leq X \leq a_{i-1}) \\ &= (F_0(a_i) - F_0(a_{i-1})); \quad i = 1, 2, \ldots, k \end{aligned} \tag{1.6}$$

Observe that p_i is the area under $f_0(\mathrm{x})$ between $\mathrm{x} = a_{i-1}$ and $\mathrm{x} = a_i$. Also,

$$\sum_{i=1}^{k} p_i = 1. \tag{1.7}$$

Now consider a random experiment consisting of n independent trials. Define

$$f_i = (\text{number of outcomes in } E_i).$$

According to the relative frequency concept,

$$p_i = \lim_{n \to \infty} \frac{f_i}{n}. \tag{1.8}$$

Note that

$$\sum_{i=1}^{k} f_i = n. \tag{1.9}$$

To test whether the null hypothesis is statistically consistent with the data, the statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \tag{1.10}$$

is evaluated. The null hypothesis is rejected when χ^2 exceeds a critical level M. To determine M, the significance level

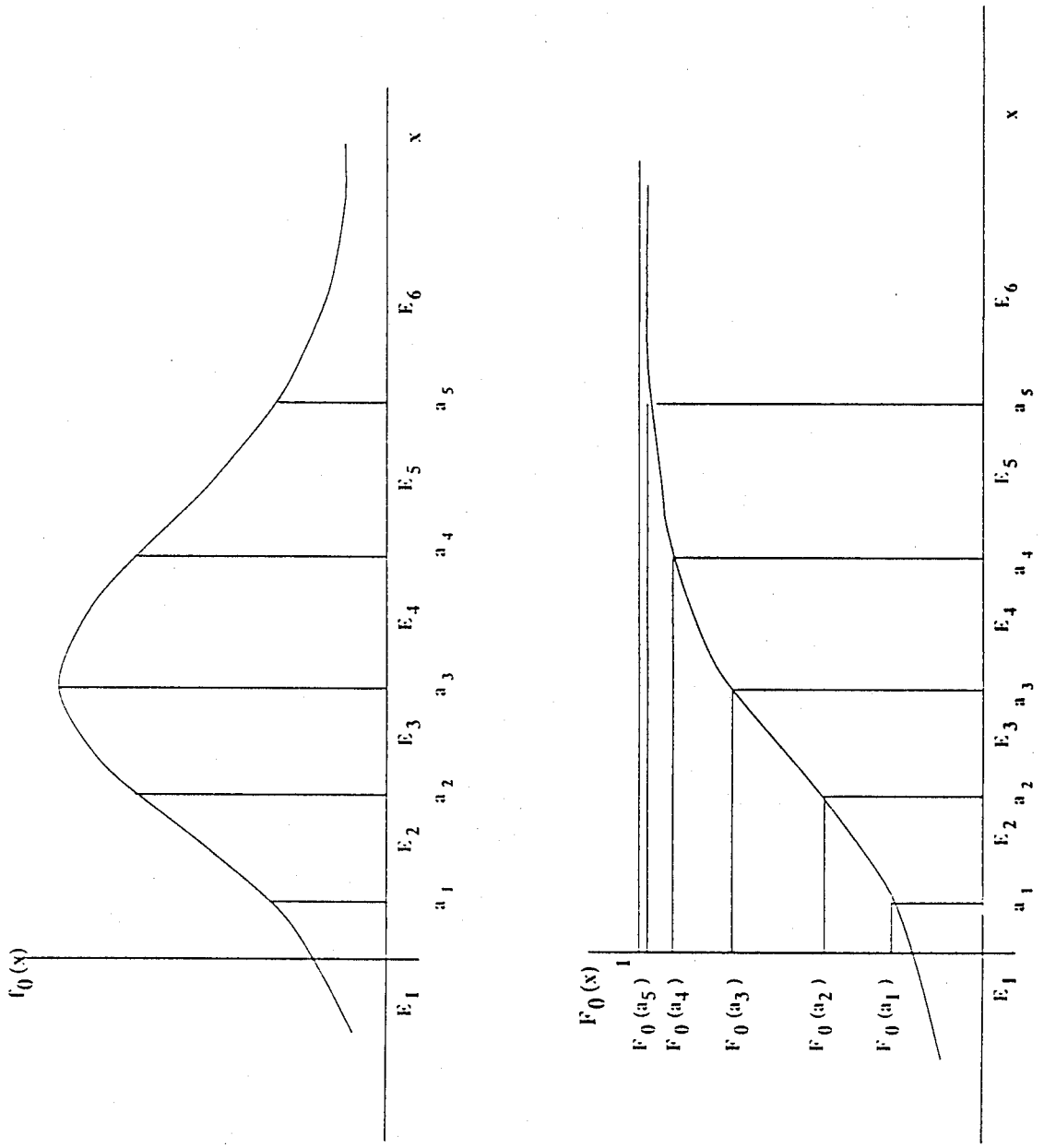

Figure 1.3: The Distribution function and PDF divided into intervals.

$$\alpha = \Pr(\text{reject } H_0 | H_0 \text{ true}) \tag{1.11}$$

is specified. Observe that

$$\alpha = Pr(\chi^2 > M|H_0) = (1 - F_{\chi^2}(M)) \tag{1.12}$$

where $F_{\chi 2}(.)$ denotes the distribution function of the χ^2 statistic. Since

$$F_{\chi^2}(M) = 1 - \alpha \tag{1.13}$$

M is the $100 \times (1 - \alpha)$ percentile of the χ^2 statistic. M is also referred to as the rejection limit. Values of M are tabulated in table 1.2 [12] and [13].

The exact distribution of the statistic χ^2 is not simple and depends on the p_i's in the null distribution (H_0) and the distribution under test. However, remarkably, it was found that these difficulties disappear as the sample size n, becomes large [12]. The distribution of χ^2 under H_0 for large n is approximately one of the family of chi-square distributions, depending on the number of intervals, k, but not on the distribution under test. This family of distributions is characterized by the *number of degrees of freedom*, defined to be k-1. Various percentiles of the chi-square distribution, for selected numbers of degrees of freedom, are given in table 1.2

1.3.1 Example[13]

Two hundred measurements of viscosity are given in table 1.3. The table gives frequencies f_i corresponding to the ten intervals. Data given in the table can be used to test the hypothesis that the probability density function (PDF) from which they come is normal with mean 32 and standard deviation 1.8. We select the significance level α to be 0.05, so that M equals the 95^{th} percentile of the χ^2 distribution. Table 1.3 shows that k, the number of intervals is 10. Consequently, the *number of degrees of freedom* given by $k - 1$ equals 9. By extracting the 95^{th} percentile of the χ^2 distribution with 9 degrees of freedom from table 1.2 it is seen that $M = 16.9$.

To illustrate the computations in table 1.3, the entry for p_4 is obtained from the normal distribution function by means of eq. 1.6. Hence,

Degrees of freedom	p									
	.01	.025	.05	.10	.70	.80	.90	.95	.975	.99
1	.000	.001	.004	.016	1.07	1.64	2.71	3.84	5.02	6.63
2	.020	.051	.103	.211	2.41	3.22	4.61	5.99	7.38	9.21
3	.115	.216	.352	.584	3.66	4.64	6.25	7.81	9.35	11.3
4	.297	.484	.711	1.06	4.88	5.99	7.78	9.49	11.1	13.3
5	.554	.831	1.15	1.61	6.06	7.29	9.24	11.1	12.8	15.1
6	.872	1.24	1.64	2.20	7.23	8.56	10.6	12.6	14.4	16.8
7	1.24	1.69	2.17	2.83	8.38	9.80	12.0	14.1	16.0	18.5
8	1.65	2.18	2.73	3.49	9.52	11.0	13.4	15.5	17.5	20.1
9	2.09	2.70	3.33	4.17	10.7	12.2	14.7	16.9	19.0	21.7
10	2.56	3.25	3.94	4.87	11.8	13.4	16.0	18.3	20.5	23.2
11	3.05	3.82	4.57	5.58	12.9	14.6	17.3	19.7	21.9	24.7
12	3.57	4.40	5.23	6.30	14.0	15.8	18.5	21.0	23.3	26.2
13	4.11	5.01	5.89	7.04	15.1	17.0	19.8	22.4	24.7	27.7
14	4.66	5.63	6.57	7.79	16.2	18.2	21.1	23.7	26.1	29.1
15	5.23	6.26	7.26	8.55	17.3	19.3	22.3	25.0	27.5	30.6
16	5.81	6.91	7.96	9.31	18.4	20.5	23.5	26.3	28.8	32.0
17	6.41	7.56	8.67	10.1	19.5	21.6	24.8	27.6	30.2	33.4
18	7.01	8.23	9.39	10.9	20.6	22.8	26.0	28.9	31.5	34.8
19	7.63	8.91	10.1	11.7	21.7	23.9	27.2	30.1	32.9	36.2
20	8.26	9.59	10.9	12.4	22.8	25.0	28.4	31.4	34.2	37.6
21	8.90	10.3	11.6	13.2	23.9	26.2	29.6	32.7	35.5	38.9
22	9.54	11.0	12.3	14.0	24.9	27.3	30.8	33.9	36.8	40.3
23	10.2	11.7	13.1	14.8	26.0	28.4	32.0	35.2	38.1	41.6
24	10.9	12.4	13.8	15.7	27.1	29.6	33.2	36.4	39.4	43.0
25	11.5	13.1	14.6	16.5	28.2	30.7	34.4	37.7	40.6	44.3
26	12.2	13.8	15.4	17.3	29.2	31.8	35.6	38.9	41.9	45.6
27	12.9	14.6	16.2	18.1	30.3	32.9	36.7	40.1	43.2	47.0
28	13.6	15.3	16.9	18.9	31.4	34.0	37.9	41.3	44.5	48.3
29	14.3	16.0	17.7	19.8	32.5	35.1	39.1	42.6	45.7	49.6
30	15.0	16.8	18.5	20.6	33.5	36.2	40.3	43.8	47.0	50.9
40	22.1	24.4	26.5	29.0	44.2	47.3	51.8	55.8	59.3	63.7
50	29.7	32.3	34.8	37.7	54.7	58.2	63.2	67.5	71.4	76.2
60	37.5	40.5	43.2	46.5	65.2	69.0	74.4	79.1	83.3	88.4

Table 1.2: Percentiles of the Chi-Squared Distribution

i	interval	p_i	$200p_i$	f_i	$(f_i - 200p_i)^2 = z_i$	$z_i/(200p_i)$
1	<27.85	0.0105	2.10	3	0.81	0.3857
2	27.85-28.95	0.0346	6.92	7	0.0064	0.0009
3	28.95-30.05	0.0943	18.86	25	37.65	1.1996
4	30.05-31.15	0.1791	35.82	42	38.10	1.06
5	31.15-32.25	0.2388	47.76	56	67.90	1.422
6	32.25-33.35	0.2161	43.22	30	174.50	4.037
7	33.35-34.45	0.1399	27.98	22	35.80	1.279
8	34.45-35.55	0.0624	12.48	11	2.19	0.175
9	35.55-36.65	0.0195	3.90	3	0.81	0.208
10	>36.65	0.0048	0.96	1	0.0016	0.0017

Table 1.3: Measurements of viscosity for example 1.3.1

$$\begin{aligned} Pr(30.05 < X < 31.15) &= \Phi(\frac{31.15 - 32}{1.8}) - \Phi(\frac{30.05 - 32}{1.8}) \\ &= 0.1791 \end{aligned} \tag{1.14}$$

where $F_0(x) = \Phi(\frac{x-32}{1.8})$ and

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = Pr(X \leq x). \tag{1.15}$$

Values of $\Phi(x)$ are obtained from a table for the standard normal distribution function.

The statistic χ^2 is obtained by summing the entries in the last column, with the result $\chi^2 = 10.16$. This does not exceed the 95^{th} percentile of the chi-square distribution with 9 "degrees of freedom", i.e. $\chi^2 < M$. The chi-square test, therefore, calls for accepting the null distribution on the basis of the given data.

Notice that the chi-square test does not test F_0(x) but only the p_i's. In particular, the natural order of the intervals does not enter the test. Moreover, F_0 is not the only distribution function having the p_i's obtained from F_0. Despite these minor objections, the chi-square test is frequently used in testing a continuous distribution.

The chi-square test suffers from the same disadvantages as mentioned earlier for the Kolmogorov-Smirnov test. In addition, it has one more disadvantage, viz., it requires a large sample size to give accurate results.

1.4 Q-Q (Quantile-Quantile) Plot

A Q-Q plot is a special plot or graphical technique which can be performed to assess the marginal distribution of the sample observations. Consider a set of data of size n given by $x_1, x_2, \ldots, x_n$. Let the data be rank ordered such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. For the j^{th} ordered sample $x_{(j)}$, define

$$p_{(j)} = \frac{j - 1/2}{n} \tag{1.16}$$

where the 1/2 is introduced as a "continuity correction" [14]. Let $F_X(\mathrm{x})$ denote the cumulative distribution function of the data. For large enough sample values of n, it then follows that

$$F_X(x_{(j)}) = Pr(X \leq x_{(j)}) \approx p_{(j)}. \tag{1.17}$$

Denote the cdf of the null distribution by $F_0(\mathrm{z})$. The quantile of $F_0(\mathrm{z})$, denoted by $q_{(j)}$, is related to $p_{(j)}$ by

$$F_0(q_{(j)}) = Pr(Z \leq q_{(j)}) = p_{(j)}. \tag{1.18}$$

If the data comes from the same distribution as the null distribution, then

$$x_{(j)} \approx q_{(j)} \tag{1.19}$$

and $x_{(j)}$ can be interpreted as an estimate of the sample quantile.

A Q-Q plot is generated using the following steps:

1) Collect n data points $x_1, x_2, \ldots, x_n$.

2) Rank order the data such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

3) Define

$$p_{(j)} = \frac{j - 1/2}{n}; \quad j = 1, 2, \ldots, n.$$

4) Evaluate the quantile $q_{(j)}$ defined by

$$F_0(q_{(j)}) = p_{(j)}; \quad j = 1, 2, \ldots, n.$$

j	Ordered observations $x_{(j)}$	Probability levels $p_{(j)} = (j - 1/2)/n$	Standard normal quantiles $q_{(j)}$
1	-1.00	0.05	-0.826
2	-0.10	0.15	-0.235
3	0.16	0.25	0.116
4	0.41	0.35	0.396
5	0.62	0.45	0.649
6	0.80	0.55	0.891
7	1.26	0.65	1.144
8	1.54	0.75	1.424
9	1.71	0.85	1.755
10	2.30	0.95	2.366

Table 1.4: Observation Table for Example 1.4.1

5) Plot the pair of points

$$(q_{(j)}, x_{(j)}); \quad j = 1, 2, \ldots, n.$$

When the data comes from the null distribution, the Q-Q plot is likely to approximate a straight line through the origin at 45°.

1.4.1 Example[14]

A sample of $n = 10$ observations gives the values tabulated in the 2^{nd} column of the table 1.4. The sample mean and the sample variance are $\hat{m} = 0.77$ and $\hat{\sigma}^2 = 0.9414$ respectively. The values of $p_{(j)}$ are computed in the 3^{rd} column. Finally, taking the normal distribution with mean $\hat{m}$ and variance $\hat{\sigma}^2$ as the null distribution, the corresponding quantiles $q_{(j)}$ are evaluated in the 4^{th} column. For example, corresponding to $p_{(9)} = 0.85$

$$F_0(q_{(9)}) = Pr(Z \leq 1.775) = \int_{-\infty}^{1.775} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-0.77)^2}{1.8828}} dz = 0.85. \tag{1.20}$$

Consequently, $q_{(9)} = 1.775$.

The Q-Q plot for the above data, which is a plot of the ordered data $x_{(j)}$ against the normal quantiles $q_{(j)}$, is shown in fig. 1.4. The pair of points $(q_{(j)}, x_{(j)})$ lie very nearly along a straight line at 45° and we accept the notion that these are normally distributed with mean = 0.77 and variance = 0.9414.

The straightness of the Q-Q plot can be evaluated by calculating the correlation coefficient of the points in the plot. The correlation coefficient for the Q-Q plot is

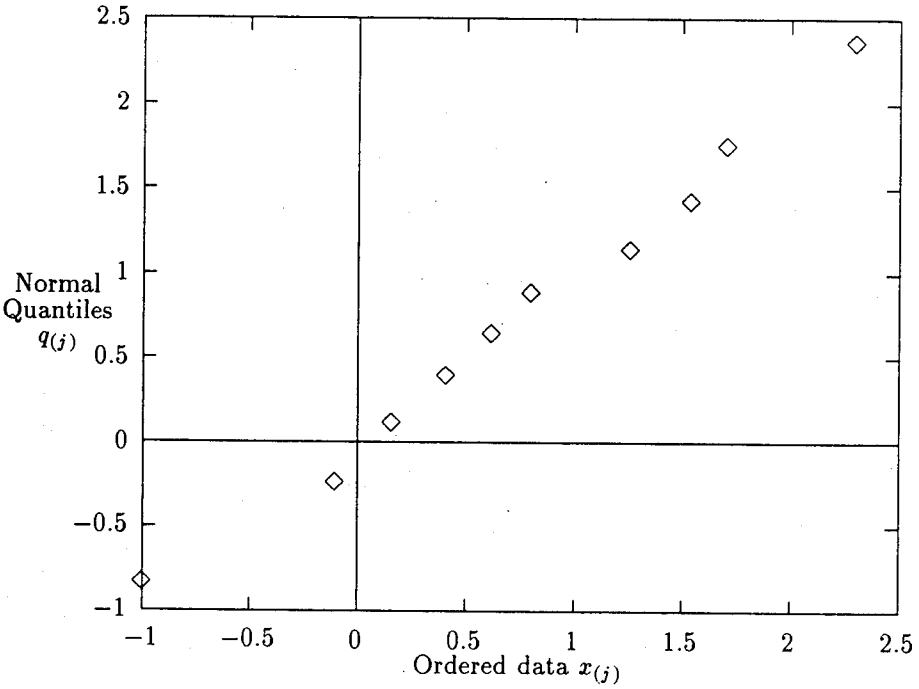

Figure 1.4: The Q-Q Plot for Example 1.4.1.

Sample size	Significance levels α		
n	0.01	0.05	0.10
5	0.8299	0.8788	0.9032
10	0.8801	0.9198	0.9351
15	0.9126	0.9389	0.9503
20	0.9269	0.9508	0.9604
25	0.9410	0.9591	0.9665
30	0.9479	0.9652	0.9715
35	0.9538	0.9682	0.9740
40	0.9599	0.9726	0.9771
45	0.9632	0.9749	0.9792
50	0.9671	0.9768	0.9809
55	0.9695	0.9787	0.9822
60	0.9720	0.9801	0.9836
75	0.9771	0.9838	0.9866
100	0.9822	0.9873	0.9895
150	0.9879	0.9913	0.9928
200	0.9905	0.9931	0.9942
300	0.9935	0.9953	0.9960

Table 1.5: Critical Points for the Q-Q Plot Correlation Coefficient Test for Normality

approximated by

$$\hat{r}_Q = \frac{\sum_{j=1}^{n}(x_{(j)} - \hat{m})(q_{(j)} - \hat{q})}{\sqrt{\sum_{j=1}^{n}(x_{(j)} - \hat{m})^2}\sqrt{\sum_{j=1}^{n}(q_{(j)} - \hat{q})^2}} \tag{1.21}$$

where $\hat{q}$ is the sample mean of the quantiles $q_{(j)}$; $j = 1, 2, \ldots, n$. Formally, we select the null hypothesis at a significance level α if $\hat{r}_Q$ exceeds a critical value denoted by M [14]. The values of M have been evaluated for the normal distribution and tabulated in table 1.5 [14] for different sample sizes and significance levels.

For the above example we select $\alpha = 0.10$. Also, using the information from table 1.4, we find that the mean of the sample quantiles and standard normal quantiles are, respectively, $\hat{m} = 0.77$ and $\hat{q} = 0$. Using eq.(1.21), we find that the correlation coefficient, $\hat{r}_Q$, is found to be 0.9943. Referring to table 1.5, we find that corresponding to $n = 10$ and $\alpha = 0.10$, the critical point M, for the Q-Q plot correlation coefficient test for normality is 0.9351. Since $\hat{r}_Q > 0.9351$, we accept the hypothesis of normality.

1.5 P-P (Probability-Probability) Plot

The P-P plot is another graphical technique which is performed for random data analysis. Just as with the Q-Q plot we consider a set of data of size n given by

$x_1, x_2, \ldots, x_n$. The data is rank ordered such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Proceeding as we did with the Q-Q plot, define for the j^{th} ordered sample, $x_{(j)}$,

$$p_{(j)} = \frac{j - 1/2}{n} \tag{1.22}$$

where the 1/2 is introduced as a "continuity correction" [14]. Let $F_X(\mathrm{x})$ denote the cumulative distribution function of the data. From the Q-Q plot we know that the $x_{(j)}$'s are the sample quantiles. Denote the cdf of the null distribution by $F_0(\mathrm{z})$. Then $p_{x_{(j)}}$ is defined to be the probability such that

$$F_0(x_{(j)}) = Pr(Z \leq x_{(j)}) = p_{x_{(j)}}. \tag{1.23}$$

If the data comes from the same distribution as the null distribution, it is likely that

$$p_{x_{(j)}} \approx p_{(j)} \tag{1.24}$$

and $p_{(j)}$ can be interpreted as an estimate of the probability $p_{x_{(j)}}$.

A P-P plot is generated using the following steps:

1) Collect n data points $x_1, x_2, \ldots, x_n$.

2) Rank order the data such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

3) Define

$$p_{(j)} = \frac{j - 1/2}{n}; \quad j = 1, 2, \ldots, n.$$

4) Evaluate the probability $p_{x_{(j)}}$ defined by

$$F_0(x_{(j)}) = p_{x_{(j)}}; \quad j = 1, 2, \ldots, n.$$

5) Plot the pair of points

$$(p_{x_{(j)}}, p_{(j)}); \quad j = 1, 2, \ldots, n.$$

When the data comes from the null distribution, the P-P plot is likely to approximate a straight line through the origin at 45°.

j	Ordered observations $x_{(j)}$	Probability levels $p_{(j)} = (j - 1/2)/n$	Standard normal probabilities $p_{x_{(j)}}$
1	-1.00	0.05	0.0342
2	-0.10	0.15	0.1853
3	0.16	0.25	0.2647
4	0.41	0.35	0.3553
5	0.62	0.45	0.4384
6	0.80	0.55	0.5124
7	1.26	0.65	0.6932
8	1.54	0.75	0.7864
9	1.71	0.85	0.8336
10	2.30	0.95	0.9424

Table 1.6: Observation Table for Example 1.5.1

1.5.1 Example

We take the example used with the Q-Q plot and find the P-P plot of the given data. The observations are tabulated in the 2^{nd} column of table 1.6. Values of $p_{(j)}$ are computed in the 2^{nd} column. The sample mean and the sample variance are $\hat{m} = 0.77$ and $\hat{\sigma}^2 = 0.9414$. Finally, taking the normal distribution with mean $\hat{m}$ and variance $\hat{\sigma}^2$ as the null distribution, the corresponding probabilities $p_{x_{(j)}}$ are evaluated in the 4^{th} column. For example, corresponding to $x_{(7)} = 1.26$

$$F_0(x_{(7)}) = Pr(Z \leq 1.26) = \int_{-\infty}^{1.26} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-0.77)^2}{1.8828}} dz = 0.6932. \tag{1.25}$$

Consequently, $p_{x_{(7)}} = 0.6932$

The P-P plot for the above data, which is a plot of the values $p_{(j)}$ against the normal probabilities $p_{x_{(j)}}$, is shown in fig. 1.5. The pair of points $(p_{(j)}, p_{x_{(j)}})$ lie very nearly along a straight line at 45° and we accept the notion that these are normally distributed with mean = 0.77 and variance = 0.9414.

The straightness of the P-P plot can be evaluated by approximating the correlation coefficient

$$\hat{r}_P = \frac{\sum_{j=1}^{n}(p_{(j)} - \hat{p})(p_{x_{(j)}} - \hat{p_x})}{\sqrt{\sum_{j=1}^{n}(p_{(j)} - \hat{p})^2}\sqrt{\sum_{j=1}^{n}(p_{x_{(j)}} - \hat{p_x})^2}} \tag{1.26}$$

where $\hat{p}$ and $\hat{p_x}$ are the sample means of $p_{(j)}$ and $p_{x_{(j)}}$, respectively, with $j = 1, 2, \ldots, n$. Unfortunately, a table for the critical value M for different values of the significance

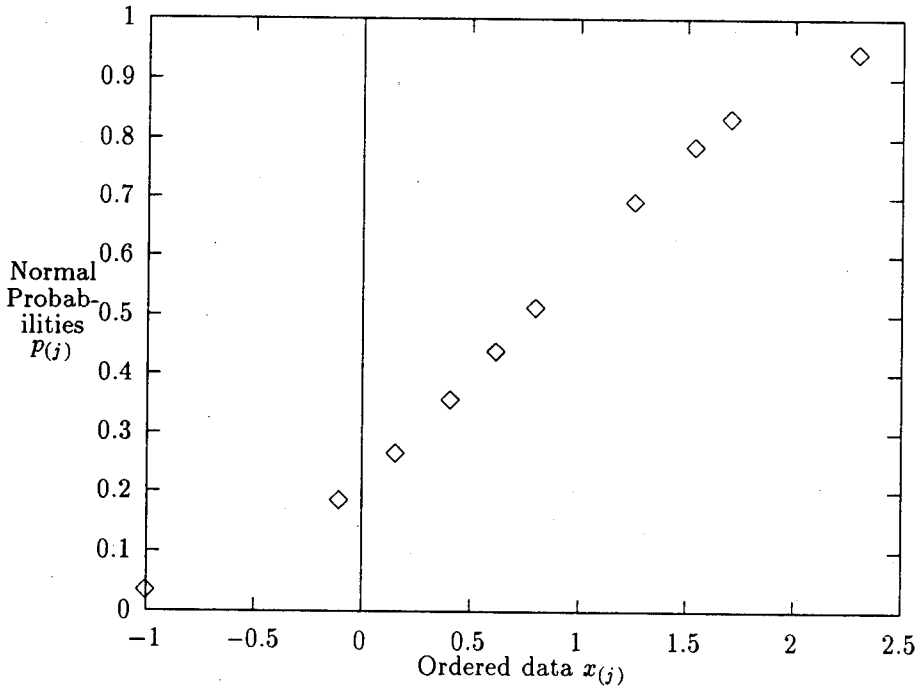

Figure 1.5: The P-P Plot for Example 1.5.1.

level, α, was not found in the literature. Nevertheless, if $\hat{r}_P$ is close to unity, $p_{(j)}$ and $p_{x_{(j)}}$ are highly correlated and although a significance level cannot be specified, it is likely that the data can be approximated with the null distribution. For this example,

$$\hat{r}_P = 0.9960. \tag{1.27}$$

Since $\hat{r}_P \approx 1$, it is concluded that the data is statistically consistent with the normal distribution having mean $\hat{m} = 0.77$ and variance $\hat{\sigma^2} = 0.9414$

An attractive property of the Q-Q plot is that it is invariant to a linear transformation. Specifically, the Q-Q plot of a linear function of $x_{(j)}$ is again a straight line at 45°. However, this time the line need not pass through the origin. P-P plots do not have this property. The main drawback of these plots is their weak performance for small and moderate sample sizes. Also, generalization of the Q-Q plot to multivariate distributions is not straightforward. On the other hand P-P plots can be applied to the multivariate situation. Although a statistic exists for evaluating the straightness of the Q-Q plot when the null distribution is standard normal, this statistic is not readily available for other distributions. Consequently, the Q-Q plots and the P-P plots do not readily offer a quantative Goodness of Fit test and the decision is mostly made on a subjective basis.

Chapter 2

The Ozturk Algorithm

2.1 Introduction

In testing a null hypothesis for a distributional assumption against an unspecified alternative there is generally no uniformly most powerful or optimal test [2]. Due to this, various test procedures have been developed for assessing these distributional assumptions. Under certain conditions, (i.e. for a specified null hypothesis, a specified sample size, and a pre-determined level of significance) one test procedure may be shown to be more powerful than the other existing procedures. Besides the power consideration of a given test, computational simplicity, desirable distributional properties of the test statistic and the generality of the test procedure are some of the important properties to be considered.

Chapter 1 gave a brief overview of some of these tests. The χ^2 test has been widely used for assesing the distributional assumptions because of its generality and its computational simplicity [2]. However, the choice of class intervals for computing the test statistic is arbitrary and the procedure can be used only for large sample sizes. Q-Q plots and P-P plots are among the most widely used graphical procedures for making asessment about the random data. But their performance is weak for small and moderate sample sizes. Also, generalizations of Q-Q plots to the multivariate distributions are not simple [5], [6]. As described in [9], [12] and [13], the Kolmogorov-Smrinov test, which is based on the empirical distribution function of the sample and the null distribution, is widely used too. In fact, comparative studies have shown that the Kolmogorov-Smirnov statistic has higher power than the χ^2 statistic for many alternatives [2]. There are many other tests such as the W test (by Shapiro

and Wilk), Anderson's A test [2], etc.

All these tests are Goodness of Fit tests.To within a certain confidence level, these tests provide information about whether a set of random data is statistically consistent with a specified null distribution. However, if the specified distribution is rejected, these tests give no clue about the alternative underlying distribution of the data. Thus, we need to have a priori knowledge about the random data to be able to use these tests. In practice, a lot of times we have no a priori knowledge about the random signals. For example, the clutter PDF encountered in radar signal processing is not known a priori. Moreover, a lot of these tests require a large number of observations to give accurate results. To get these many observations may prove costly in a real world situation. Consequently, a scheme is necessary, that not only performs the Goodness of Fit test but also approximates the PDF for small number of observations.

A new algorithm based on sample order statistics has been developed in [1] through [3] and has been reported in [10] for univariate distribution approximation. This algorithm has two modes of operation. In the first mode, the algorithm performs a Goodness of Fit test. Specifically, the test determines, to a desired confidence level, whether the random data is statistically consistent with a specified probability distribution. In the second mode of operation, the algorithm approximates the PDF underlying the random data. In particular, by analyzing the random data and without any a priori knowledge, the algorithm identifies from a stored library of PDFs that density function which best approximates the data. Estimates of the location, scale, and shape parameters of the PDF are provided by the algorithm. The algorithm is typically found to work well for observation sizes of the order of 75-100.

In this chapter we present the Ozturk algorithm. It will be demonstrated through examples that the algorithm can be used to test for any distributional assumption (not limited to location-scale family) including univariate and multivariate random variables.

2.2 Definitions

Let $F_Y(\mathrm{y})$ denote the PDF of a random variable Y. Consider the linear transformation defined by

$$x = \beta y + \alpha. \tag{2.1}$$

The PDF of X is given by

$$f_X(x) = \frac{1}{|\beta|} f_Y(\frac{x-\alpha}{\beta}) \tag{2.2}$$

where α and β are defined to be the location and scale parameters of $f_X(\mathrm{x})$, respectively. The mean μ_x and the variance σ_x^2 of the random variable x are given by

$$\begin{aligned} \mu_x &= E[x] \\ \sigma_x^2 &= E[(x-\mu_x)^2], \end{aligned} \tag{2.3}$$

where E is the expectation operator.

Although the mean and the variance are related to the location and scale parameters, note that the location parameter is not the mean value and the scale parameter is not the square root of the variance, in general. However, for a standard Gaussian PDF $f_Y(\mathrm{y})$, for which the mean is zero and variance unity, the location parameter is the mean of X and the scale parameter is the standard deviation (square root of the variance) of X.

The coefficient of skewness α_3, and the coefficient of kurtosis α_4, of X are defined to be

$$\begin{aligned} \alpha_3 &= \frac{E[(x-\mu_x)^3]}{\sigma_x^3} \\ \alpha_4 &= \frac{E[(x-\mu_x)^4]}{\sigma_x^4} \end{aligned} \tag{2.4}$$

It is readily shown that α_3 and α_4 are invariant to the values of μ_x and σ_x. For any PDF that is symmetric about the mean, $\alpha_3 = 0$. For the case of the Gaussian distribution, $\alpha_3 = 0$ and $\alpha_4 = 3$.

2.3 The Ozturk Algorithm

Any distribution, or a family of distributions, can be represented by a single point or by a region on an $\alpha_3 - \alpha_4$ plane, respectively, where α_3 is the coefficient of skewness and α_4 is the coefficient of kurtosis (see for example [15], p.14). A set of random data can also be represented by a point whose co-ordinates are given by the sample

estimates of α_3 and α_4. Then the best candidate for the underlying true distribution can be identified tobe the nearest neighbour distribution on the chart. Although such a chart, based on the coefficient of skewness and kurtosis, provides a useful way of characterizing the distributions, its use is limited by the fact that the moments of some distributions do not exist. Other drawbacks of this approach are

1. Estimates of α_3 and α_4 are highly sensitive to extreme observations.
2. Estimates of these moments are highly biased for small sample sizes [4].
3. The moment estimators are greatly affected by outliers.

In this chapter we introduce the Ozturk Algorithm, a general graphical technique which works in two specific modes.

1. In the first mode it performs a formal Goodness of Fit test for a specified null distribution.
2. In its second mode it provides a graphical representation that gives insight into what distribution best approximates the data set and thus provides a way of characterizing the data.

2.3.1 Goodness of Fit Test

The Goodness of Fit test is a complex algorithm which determines whether or not the set of data samples provided to the algorithm is statistically consistent with a specified distribution (the null hypothesis). Using the standard normal distribution with zero mean and unit variance as the *reference distribution*, the standardized sample order statistics are represented by a system of linked vectors. The terminal point of the linked vectors, as well as the shape of their trajectories, are used in determining whether or not to accept the null hypothesis. In its present form the algorithm uses the standard Gaussian distribution as the *reference distribution.* However, any other distribution could be used as the *reference distribution.* The null hypothesis is the distribution against which the sample data is to be tested. Note that the reference distribution need not be the same as the null distribution.

We begin by introducing several sample order statistics used in the algorithm and then proceed to explain the Goodness of Fit test procedure

Consider the following three sets of data size n:

1. A sample data set

$$X_1, X_2, X_3, \ldots, X_n.$$

with mean and standard deviation given by μ_x and σ_x.

2. A null hypothesis data set

$$Z_1, Z_2, Z_3, \ldots, Z_n$$

generated from any available distribution against which the sample set will be tested. The mean and standard deviation of this data set are defined to be μ_z and σ_z, respectively.

3. A reference distribution data set

$$W_1, W_2, W_3, \ldots, W_n$$

generated from the standardized Gaussian.

Let $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ denote the ordered set of samples obtained by ordering $X_i; i = 1, 2, \ldots, n$, where $X_{1:n}$ is the smallest data sample. Similiarly, the other two data sets are ordered resulting in the three ordered data sets

$$\begin{aligned} &X_{1:n}, X_{2:n}, X_{3:n}, \ldots, X_{n:n} \\ &Z_{1:n}, Z_{2:n}, Z_{3:n}, \ldots, Z_{n:n} \\ &W_{1:n}, W_{2:n}, W_{3:n}, \ldots, W_{n:n}. \end{aligned} \tag{2.5}$$

Define

$$Y_{i:n} = \frac{X_{i:n} - \hat{\mu_x}}{\hat{\sigma_x}} \;\; ; \;\; i = 1, 2, \ldots, n \tag{2.6}$$

where $\hat{\mu_x} = \sum X_i/n$ is the sample mean and $\hat{\sigma_x} = \sum[(X_i - \mu_x)^2/(n-1)]^{1/2}$ is the sample standard deviation. These are the standardized order statistics of the sample data. For the null hypothesis, a Monte Carlo simulation consisting of 2,000 trials is utilized. The estimate of the expected value of the standardized i^{th} order statistic is defined as

$$\hat{T}_{i:n} = \frac{1}{2000}\sum_{k=1}^{2000}\frac{(Z_{i:n})_k - \hat{\mu_z}}{\hat{\sigma_z}}; \quad i = 1, 2, \ldots, n. \tag{2.7}$$

where $(Z_{i:n})_k$ denotes the i^{th} order statistic from the k^{th} Monte Carlo trial, and $\hat{\mu_z}$ and $\hat{\sigma_z}$ denote the sample mean and sample standard deviation. Also, $\hat{m}_{i:n}$ is defined as the estimate of the expected value of the i^{th} order statistic of the reference distribution, the standardized Gaussian. Using 2,000 Monte Carlo trials,

$$\hat{m}_{i:n} = \frac{1}{2000}\sum_{k=1}^{2000}(W_{i:n})_k; \quad i = 1, 2, \ldots, n \tag{2.8}$$

where $(W_{i:n})_k$ denotes the i^{th} order statistic from the k^{th} Monte Carlo trial of the reference distribution. When the null hypothesis is the reference distribution, the standardized Gaussian, then

$$\hat{T}_{i:n} \approx \hat{m}_{i:n}. \tag{2.9}$$

The Goodness of Fit test proceeds by joining together two sets of n linked vectors, one for the sample data and one for the null hypothesis. The i^{th} linked vector in each set is characterized by its length and orientation with respect to the horizontal axis. For the sample data, the length of the i^{th} vector, a_i, is obtained from the magnitude of the i^{th} standardized order statistic of the data, while its angle or orientation, θ_i, is related to $\hat{m}_{i:n}$. More specifically, for the sample data

$$a_i = \frac{|Y_{i:n}|}{n}$$
$$\theta_i = \pi\phi(\hat{m}_{i:n}) \quad ; \quad \phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} exp(-\frac{t^2}{2})dt. \tag{2.10}$$

ϕ_x is the cumulative distribution of the standard Gaussian distribution. We define the sample points Q_k in a two dimensional plane (U, V) by

$$Q_k = (U_k, V_k)\ ;\ k = 1, 2, \ldots, n \tag{2.11}$$

where $U_0 = V_0 = 0$ and

$$\begin{aligned} U_k &= \frac{1}{k}\sum_{i=1}^{k} |Y_{i:n}| \cos(\theta_i) \\ V_k &= \frac{1}{k}\sum_{i=1}^{k} |Y_{i:n}| \sin(\theta_i) \\ & k = 1, 2, \ldots, n. \end{aligned} \tag{2.12}$$

Similiarly, for the null hypothesis the length of the i^{th} vector, b_i, is obtained from the magnitude of the i^{th} standardized order statistic of the null data set. Specifically, for the null data

$$b_i = \frac{|\hat{T}_{i:n}|}{n}, \qquad \theta_i = \pi\phi(m_{i:n}). \tag{2.13}$$

Using the same two dimensional plane, we plot the sample points for the null distribution defined by

$$Q_{0k} = (U_{0k}, V_{0k})\ ;\ k = 1, 2, \ldots, n \tag{2.14}$$

where, $U_{00} = V_{00} = 0$ and

$$\begin{aligned} U_{0k} &= \frac{1}{k}\sum_{i=1}^{k} |\hat{T}_{i:n}| \cos(\theta_i) \\ V_{0k} &= \frac{1}{k}\sum_{i=1}^{k} |\hat{T}_{i:n}| \sin(\theta_i) \\ & k = 1, 2, \ldots, n. \end{aligned} \tag{2.15}$$

Note that the angle θ remains the same for both sets of linked vectors. However, the magnitude of the linked vector for the sample data is a_i whereas it is b_i for the null distribution. The angle θ_i is solely dependent on the reference distribution while the magnitudes $|Y_{i:n}|$ and $|\hat{T}_{i:n}|$ are solely dependent on the sample data and null data

sets, respectively. In particular, for the i^{th} sample linked vector, a_i is dependent on the standardized i^{th} order statistic of the sample data set whereas for the i^{th} linked vector of the null hypothesis, b_i is dependent on the estimate of the expected value of the i^{th} standardized order statistic of the Monte Carlo simulation of the null distribution.

Although $Y_{i:n}$ and $\hat{T}_{i:n}$ are ordered statistics from the smallest to the largest, note that the magnitudes of $Y_{i:n}$ and $\hat{T}_{i:n}$ are not. In fact, with increasing i, $|Y_{i:n}|$ and $|\hat{T}_{i:n}|$ would begin large, decrease to approximately zero and then increase again.

The i^{th} sample and null linked vectors, respectively, are drawn by joining the points (Q_i, Q_{i-1}) and $(Q_{0i}, Q_{0(i-1)})$. It should be noted that the Q_n and Q_{0n} given in equations (2.11) and (2.14) represent the terminal point, respectively, of the linked vectors defined above. Fig. 2.1 shows the two sets of linked vectors obtained when both the sample and null data sets are obtained from the the Gaussian distribution with $n = 6$ and $n = 50$. The solid curves in fig.2.1 show the linked vector for the sample distribution while the dashed curves show the ideal linked vector for the null distribution. When the length n, of the data set, is large (on the order of 50 points), then the linked vector is a smooth arc, as seen in fig.2.1.

For a typical set of ordered data samples drawn from the null distribution, it is reasonable to expect that the sample linked vectors would follow the null linked vectors closely. If the ordered set of samples is not from the null distribution, then the sample linked vectors are not expected to follow the path of the null linked vectors closely. Hence, the procedure provides visual information about how well the ordered set of data fit the null distribution. However this is not an *ad hoc* statistical procedure. As shall be seen later on, we do construct test statistics to present a formal way of performing the Goodness of Fit test to determine whether the data set is statistically consistent with the null hypothesis.

2.3.1.1 Properties of the test statistic Q_{0n}

An important property of the Q_{0n} statistic is that it is invariant under linear transformation. In particular, we consider the standardization used in eq.(2.6). Let $S_i = cX_i + d$ where c and d are constants. Let μ_s and σ_s denote the mean and standard deviation of the samples, S_i, respectively. Then, it is readily shown that $|\frac{X_i - \mu_x}{\sigma_x}| = |\frac{S_i - \mu_s}{\sigma_s}|$. The invariance property follows as a consequence. The advantage of this property is that the PDF of $Q_{0n} = (U_{0n}, V_{0n})$ for a given sample set and reference

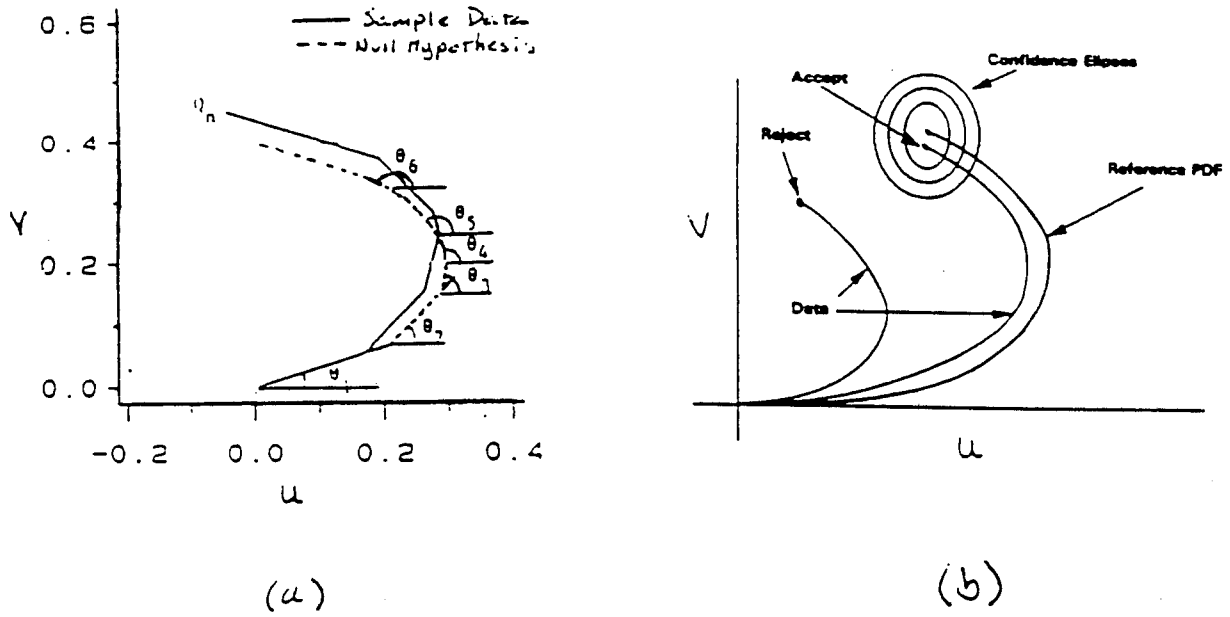

Figure 2.1: The Linked Vectors:Dashed lines P_0 = Null Linked Vectors, Solid Lines P_1 = Sample Linked Vectors

distribution depends only on the sample size n and is unaffected by the location and scale parameters. The distributional properties of this statistic for testing normality is studied by Ozturk and Dudewicz in [3].

The exact sampling distribution of Q_{0n} is usually difficult to obtain. However, the empirical distribution of the test statistic Q_{0n} was obtained via Monte Carlo experimentation by Ozturk and Dudewicz in [3] . Using the means, variances and coefficients of skewness and kurtosis of U_{0n} and V_{0n} based on 50,000 samples for values of n from 3 till 100, they found that the distributions of U_{0n} and V_{0n} approach the normal distribution even for moderate sample sizes. The distributional properties of the statistic $Q_{0n} = (U_{0n}, V_{0n})$ for testing normality is studied by Ozturk and Dudewicz in [3]. Some of the empirical results obtained by them for the statistic Q_{0n} for standard normal distribution for $3 \leq n \leq 100$ are given below.

$$E(U_{0n}) = 0 \tag{2.16}$$

$$E(V_{0n}) = \mu_v \approx 0.326601 + \frac{0.412921}{n}$$

$$E(U_{0n}V_{0n}) \approx 0$$

$$Var(U_{0n}) = \sigma_u^2 \approx \frac{0.02123}{n} + \frac{0.01765}{n^2} \tag{2.17}$$

$$Var(V_{0n}) = \sigma_v^2 \approx \frac{0.04427}{n} - \frac{0.0951}{n^2}.$$

Also, it was found empirically, for $n > 10$, that, U_{0n} and V_{0n} are approximately bivariate normal.

An interesting property of this algorithm is that any one of the points $Q_{0k}; k = 1, 2, \ldots, n$, or a selected group of these points can be used as a test statistic to establish a formal test. The algorithm in its present form proposes the general statistic Q_{0n} as the test statistic for testing the null hypothesis.

2.3.1.2 Basic Concept of the Confidence Contours

The algorithm provides quantitative information as to how consistent the sample data set is with the null hypothesis distribution by the use of the confidence contours. An example of these contours is shown in fig.2.1. If the end point of the sample data linked vector curve falls within one or more of these contours, then the sample data set is said to be statistically consistent with the null hypothesis at a confidence level

based on the confidence contours. If the sample data set is truly consistent with the null hypothesis, note that the sample linked vector is likely to closely follow the null linked vector.

Now consider the linked vector for the null hypothesis which is based on the standardized expected values of the order statistic, Z, for 2,000 Monte Carlo simulations. The test statistic Q_{0n}, found by computing the expected value of 2,000 end points of the 2,000 linked vectors provided by the Monte Carlo simulation, is random. The coordinates of Q_{0n}, U_{0n} and V_{0n}, may or may not be bivariate Gaussian.

When U_{0n} and V_{0n} are bivariate Gaussian, the confidence contours of the null hypothesis are readily determined. A three dimensional bell shaped bivariate Gaussian curve is fitted to the 2,000 end points arising from the Monte Carlo simulation. The elliptical contours of this distribution are plotted for various parameters of the significance level α (eg. 0.01, 0.05, 0.1) where α is defined as the conditional probability that Q_{0n} falls outside the specified ellipse given that the data comes from the null distribution. $(1-\alpha)$ is called the confidence level and the corresponding contour is called the $100 \times (1-\alpha)$ percent confidence contour. Note that $(1-\alpha)$ is the conditional probability tha Q_{0n} falls inside the specified ellipse given that the data comes from the null distribution.

This could be done for any of the n points of the ordered statistic, Z, along the null linked vector. Thus, more than one set of confidence contours could be created if there are more than one test statistic. Then, if the sample data is truly consistent with the null hypothesis, the sample data linked vector is likely to pass through a series of confidence contours determined from the distributions of the test statistics. However, it was found to be unnecessary to clutter up the graphics with so many contours, as the human eye can readily detect whether or not the linked vectors are closely following the same trajectory. The option of using more than one test statistic is provided in the algorithm.

Note that the average value of the test statistic, Q_{0n}, of the null distribution is at the center of the contours. Thus, the closer the end point of the sample data linked vector is to the center of the confidence contour, the more likely it is that the sample data is coming from the null hypothesis. As the significance level decreases, the confidence level increases and the probabillity that Q_{0n} will fall within the corresponding ellipse

will also increase. This results in the fact that the size of the confidence contours increase as the confidence level increases.

For a given sample size, n, the i^{th} angle of any linked vector depends solely on the reference distribution which remains unchanged throughout. Consequently, for a given value of sample size, n, and for a given null hypothesis, values for the magnitude and angle of the points (U_{0k}, V_{0k}) on the null linked vector, $k = 1, 2, \ldots, n$, may be tabulated. This table, which is dependent on n and the null hypothesis, could be stored and recalled when desired. This can significantly reduce the computational requirements.

2.3.1.3 Determining Confidence Contours

As described earlier, the confidence contours are contours from the bivariate probability density function of the end point coordinates used to determine the test statistic Q_{0n}. These 2,000 end points are obtained from the Monte Carlo simulation. Plotting confidence contours is usually not easy when the joint distribution is not bivariate normal. Further, in order to analytically determine the confidence contours, the joint PDF of U_{0n} and V_{0n} must be known [4]. However, it is difficult to analytically determine this joint PDF. Consequently, a normality transformation is made on the end point coordinates U_{0n} and V_{0n} to obtain statistics $r_{0u} = \psi_1(U_{0n})$ and $r_{0v} = \psi_2(V_{0n})$ where $\psi_1(.)$ and $\psi_2(.)$ are functions operating on U_{0n} and V_{0n}, respectively. A family of distributions called the *Johnson System* is used to perform the transformation on U_{0n} and V_{0n} so as to obtain a bivariate normal distribution.

The Johnson system of distributions is a flexible family of distributions having four parameters. This system is used to summarize a set of data by means of a mathematical function which will fit the data. The system proposed by Johnson contains three families of distributions which are obtained by transformations of the form

$$R = \gamma + \eta f_i(G; \lambda, \epsilon) \; ; i = 1, 2, 3 \tag{2.18}$$

where R is a standard normal variable and G is the random variable on which the transformation is performed. γ, η, λ, and ϵ are four parameters of the Johnson system of distributions. In particular, let

$$\begin{aligned} R_{0u} &= \gamma_1 + \eta_1 f_i(U_{0n}; \lambda_1, \epsilon_1) \\ R_{0v} &= \gamma_2 + \eta_2 f_i(V_{0n}; \lambda_2, \epsilon_2) \\ i &= 1, 2, 3 \end{aligned} \tag{2.19}$$

where f_i ; $i = 1, 2, 3$ represent the following three functions suggested by Johnson:

$$f_1(g; \lambda, \epsilon) = \sinh^{-1}(\frac{g - \epsilon}{\lambda}) \tag{2.20}$$

denotes the S_U distribution,

$$f_2(g; \lambda, \epsilon) = \ln(\frac{g - \epsilon}{\lambda + \epsilon - g}), \quad \epsilon \leq g \leq \epsilon + \lambda \tag{2.21}$$

denotes the S_B distribution, and

$$f_3(g; \lambda, \epsilon) = \ln(\frac{g - \epsilon}{\lambda}), \quad g > \epsilon \tag{2.22}$$

denotes the S_L distribution.

Note that $f_i(g; \lambda, \epsilon), i = 1, 2, \ldots, 3$, are single- valued monotonically increasing functions for the allowed ranges of g. S_L is, in essence, a three parameter distribution since the parameter λ can be eliminated by letting $\gamma^* = \gamma - \eta \ln \lambda$ so that $r = \gamma^* + \eta \ln(g - \epsilon)$. S_B is a distribution bounded on $(\epsilon, \epsilon + \lambda)$ and the S_U is an unbounded distribution. In a plot of the third and fourth order standardized moments where $\sqrt{\alpha_3}$ is plotted versus α_4, the chosen functions are such that the S_L distributions form a curve dividing the $(\sqrt{\alpha_3}, \alpha_4)$ plane in two regions. The S_B distributions lie in one of the regions and the S_U lie in the other.

In using this system of transformations, the first step is to determine which of the three families should be used for performing the normality transformation. A possible procedure is to compute the sample estimate of the standardized moments, viz., the coefficients of skewness and kurtosis, and choose the distribution according to which of the two regions contains the computed point. However, as described at the beginning of the chapter, this method has major drawbacks. Consequently, another procedure is used to determine the family of distributions to be used to perform the

transformations. It is a simple selection rule which is a function of four percentiles to select one of the three families and to give estimates of the parameters for all the families. It was developed by Slifker J. and Shapiro S. in [4].

The idea of the selection rule is to try and find a property of the transformation given in eq.(2.18) and use it to select an appropriate member of the Johnson family to approximate a set of data. It was heuristically felt by Slifker J. and Shapiro S. that there must be some relationship concerning the distances in the tails vs. distances in the central portion of the distribution which could be used to distinguish between the bounded and unbounded cases. This led to the following formalization.

Consider any one of the transformations described by eq.(2.18). Choose any fixed value of $r > 0$ of a standard normal variate. Then the points $\pm r$ and $\pm 3r$ divide a horizontal axis into three intervals of equal length given by $(-3r, -r), (-r, r), (r, 3r)$. Let g_{3r}, g_r, g_{-r} and g_{-3r} be the values corresponding to $3r, r, -r$ and $-3r$,under the transformation given in eq.(2.18), respectively. Let

$$\begin{aligned} m &= g_{3r} - g_r \\ l &= g_{-r} - g_{-3r} \\ p &= g_r - g_{-r}. \end{aligned} \tag{2.23}$$

Since $f_i(g; \lambda, \epsilon), i = 1, 2, \ldots, 3$, are single-valued monotonically increasing functions for the allowed ranges of g, it is readily seen that m, l and p are all greater than 0. For a bounded symmetrical Johnson distribution, it was hypothesized that the distances m and l between each of the outer and inner points would be smaller than the distance p between the two inner points. The converse would be true for the unbounded case. This led to the following more general result:

$$\begin{aligned} &(i) \quad \tfrac{ml}{p^2} > 1 \quad \textit{for any } S_U \textit{ distribution;} \\ &(ii) \quad \tfrac{ml}{p^2} < 1 \quad \textit{for any } S_B \textit{ distribution;} \\ &(iii) \quad \tfrac{ml}{p^2} = 1 \quad \textit{for any } S_L \textit{ distribution.} \end{aligned} \tag{2.24}$$

These properties are proven in Appendix A and can be used to discriminate among

the three families.

Selection Procedure

The selection procedure consists of the following steps:

1. Choose a fixed value of $r > 0$. This choice should be motivated by the number of data points. In general, for moderate sized data sets, a value of r less than 1 should be chosen [4]. Any choice of r greater than 1 would make it difficult to estimate the percentile of $\pm 3r$. A *typical* choice is to use a value of r close to 0.5 such as 0.524. This would make $3r = 1.572$ and these points correspond to the 70^{th} and the 94.2^{th} percentiles of the standard normal distribution, respectively. However, the larger the number of data points, the larger the value of r that can be selected. In the *Ozturk algorithm* r is chosen to be 0.775449.

2. Determine from a table for the normal distribution the probability $P_a = Pr(R \leq a)$, where a is taken to be either $3r, r, -r$ or $-3r$. For example, if $r = 0.5$ then $P_{0.5} = Pr(R \leq 0.5) = 0.6915$.

3. Determine integer values of k_a such that

$$P_a \approx \frac{k_a - \frac{1}{2}}{n} \tag{2.25}$$

where

$$k_a = [nP_a + \frac{1}{2}], \tag{2.26}$$

[.] denotes the closest integer, and $a = 3r, r, -r, -3r$.

4. Obtain n observations of the random variable G, where G is related to the random variable R through eq.(2.18). Order these observations from the smallest to the largest and denote the k^{th} ordered observation by g^k.

5. Let

$$g_a = g^{k_a} \tag{2.27}$$

where $a = 3r, r, -r, -3r$. The connections between g_a, k_a, and P_a are explained in Appendix B.

6. From the values of g_a obtained in step 5, compute the distances m, l, and p according to eq.(2.23).

7. Use the criteria in eq.(2.24) to select the appropriate member of the family of distributions.

Since the g^i's are continuous random variables, the probability is zero that $(ml/p^2) = 1$. Thus, choice of the S_L distribution requires that ml/p^2 fall within some small prespecified tolerance interval around 1.

After completion of the selection process, the next step is to estimate the parameters of the distribution selected. Estimation of the parameters is accomplished by using the formulae given. These allow the estimates to be simply calculated by means of a scientific hand calculator. The formulae for the estimates are given in terms of the chosen values of r and the computed values of m, l and p. Derivations of these formulae are provided in Appendix A.

Note that the following formulae express the parameter values as functions of m, l and p which in turn are functions of g_{3r}, g_r, g_{-r} and g_{-3r}. In practice, the corresponding parameter estimates are computed based on the ordered sample values, $g^{k_a}; a = 3r, r, -r, -3r$.

(i) Johnson S_U Distribution

$$r = \gamma + \eta \sinh^{-1}\left(\frac{g - \epsilon}{\lambda}\right) \tag{2.28}$$

Parameter Estimates for Johnson S_U Distribution

$$\begin{aligned} \eta &= \frac{2r}{\cosh^{-1}[\frac{1}{2}(\frac{m}{p} + \frac{l}{p})]}; \ (\eta > 0) \\ \gamma &= \eta \sinh^{-1}[\frac{\frac{l}{p} - \frac{m}{p}}{2(\frac{m}{p}\frac{l}{p} - 1)^{1/2}}]; \\ \lambda &= \frac{2p(\frac{m}{p}\frac{l}{p} - 1)^{1/2}}{(\frac{m}{p} + \frac{l}{p} - 2)(\frac{m}{p}\frac{l}{p} + 2)^{1/2}}; \ (\lambda > 0) \end{aligned} \tag{2.29}$$

$$\epsilon = \frac{g_r + g_{-r}}{2} + \frac{p(\frac{l}{p} - \frac{m}{p})}{2(\frac{m}{p} + \frac{l}{p} - 2)}.$$

(ii) Johnson S_B Distribution

$$r = \gamma + \eta \ln(\frac{g - \epsilon}{\lambda + \epsilon - g}). \tag{2.30}$$

Parameter Estimates for Johnson S_B Distribution

$$\begin{aligned}
\eta &= \frac{r}{\cosh^{-1}(\frac{1}{2}[(1 + \frac{p}{m}(1 + \frac{p}{l})]^{1/2})}; \quad (\eta > 0) \\
\gamma &= \eta \sinh^{-1}[\frac{(\frac{p}{l} - \frac{p}{m})\{(1 + \frac{p}{m})(1 + \frac{p}{l}) - 4\}^{1/2}}{2(\frac{p}{m}\frac{p}{l} - 1)}]; \\
\lambda &= \frac{p[\{(1 + \frac{p}{m})(1 + \frac{p}{l}) - 2\}^2 - 4]^{1/2}}{\frac{p}{m}\frac{p}{l} - 1}; \quad (\eta > 0) \\
\epsilon &= \frac{g_r - g_{-r}}{2} - \frac{\lambda}{2} + \frac{p(\frac{p}{l} - \frac{p}{m})}{2(\frac{p}{m}\frac{p}{l} - 1)}.
\end{aligned} \tag{2.31}$$

(iii) Johnson S_L Distribution

$$r = \gamma^* + \eta \ln(g - \epsilon). \tag{2.32}$$

Parameter Estimates for Johnson S_L Distribution

$$\begin{aligned}
\eta &= \frac{2r}{\ln \frac{m}{p}} \\
\gamma^* &= \eta \ln[\frac{\frac{m}{p} - 1}{p(\frac{m}{p})^{1/2}}] \\
\epsilon &= \frac{g_r + g_{-r}}{2} - \frac{p}{2}\frac{\frac{m}{p} + 1}{\frac{m}{p} - 1}.
\end{aligned} \tag{2.33}$$

Note that the values of the parameters above are presented in such a way as to emphasize their dependence on the ratios m/p and l/p for the S_U distribution and on p/m and p/l for the S_B distribution . For the S_L distribution, we see from eq.(2.24) that $(l/p) = (m/p)^{-1}$. Thus, the formulae for the S_L distribution parameters are given solely in terms of the single ratio m/p.

Resistor Interval (in kΩ)	Observed Frequencies
<9.25	-
9.25-9.75	1
9.75-10.25	7
10.25-10.75	18
10.75-11.25	36
11.25-11.75	70
11.75-12.25	115
12.25-12.75	199
12.75-13.25	437
13.25-13.75	929
13.75-14.25	1787
14.25-14.75	2294
14.75-15.25	2082
15.25-15.75	1129
15.75-16.25	275
16.25-16.75	55
16.75-17.25	6
>17.25	-
Total	9440

Table 2.1: Resistor Values

Example:

We consider a set of data representing the resistor values in a VLSI circuit. The data and the observed frequencies are shown in table 2.3.1.3. We choose the value of r to be 1. Thus, the four values assumed by a are +3, 1, -1 and -3. From the table for the normal distribution, the probabilities, $P_a = Pr(R \leq a)$, for $a = 3, 1, -1$, and -3 are found to be $0.9986, 0.8413, 0.1587$ and 0.0014, respectively.

First, consider $a = 3$, for which $P_3 = 0.9986$. The value of the order number k_3 is given by

$$k_3 = [nP_3 + \frac{1}{2}] = [(9440)(0.9986) + 0.5] = 9427 \tag{2.34}$$

where [.] denotes the closest integer value. If the raw data were available, we would simply let g_3 equal to the 9427^{th} ordered sample, g^{9427}. However, because the raw data has been grouped into the intervals tabulated in table 2.3.1.3, the value of g^{9427} is unknown. Consequently, interpolation is used to estimate a value for g^{9427}.

Note that the 9427^{th} ordered observation falls in the interval $(16.25, 16.75)$. The probabilities that the resistor values are less than or equal to $16.25k\Omega$ and $16.75k\Omega$,

respectively, are given by

$$\begin{aligned} Pr(G \leq 16.25) &= \frac{9440 - 6 - 55}{9440} = 0.9935 \\ Pr(G \leq 16.75) &= \frac{9440 - 6}{9440} = 0.9994. \end{aligned} \tag{2.35}$$

Thus, by the method of interpolation,

$$\frac{16.75 - 16.25}{0.9994 - 0.9935} = \frac{g^{9427} - 16.25}{0.9986 - 0.9935} \tag{2.36}$$

The value of g^{9427} is found to be $16.25 + 0.439 = 16.689$. Setting g_3 equal to g^{9427}, it follows that $g_3 = 16.689$. Values of g_1, g_{-1}, and g_{-3} are found in a similiar manner. In summary,

$$\begin{aligned} g_3 &= 16.689 \\ g_1 &= 15.242 \\ g_{-1} &= 13.581 \\ g_{-3} &= 10.409. \end{aligned} \tag{2.37}$$

Consequently,

$$\begin{aligned} p &= g_1 - g_{-1} = 1.661 \\ m &= g_3 - g_1 = 1.447 \\ l &= g_{-1} - g_{-3} = 3.172 \end{aligned} \tag{2.38}$$

yielding

$$\frac{ml}{p^2} = 1.664. \tag{2.39}$$

Since the value of $\frac{ml}{p^2}$ is found to be significantly greater than 1, it is decided from eq.(2.24) that an S_U distribution is appropriate for transformation. The formulae given in eq.(2.29) are used to obtain the parameter values. Thus,

$$
\begin{aligned}
\eta &= \frac{2(1)}{\cosh^{-1}[\frac{1}{2}(0.871+1.910)]} = 2.333 \\
\gamma &= 2.333\sinh^{-1}[\frac{1.910-0.871}{2\sqrt{1.910(0.871)-1}}] = 1.402 \\
\lambda &= \frac{2(1.661)\sqrt{(1.910)(0.871)-1}}{(0.871+1.910-2)\sqrt{0.871+1.910+2}} = 1.585 \\
\epsilon &= \frac{15.242+13.581}{2} + \frac{1.661(1.910-0.871)}{2(1.910+0.871-2)} = 15.516.
\end{aligned}
\tag{2.40}
$$

The transformation equation, therefore, becomes

$$
r = 1.402 + 2.333\sinh^{-1}(\frac{g-15.516}{1.585}) \tag{2.41}
$$

where r is a standard normal variable and g is a random variable corresponding to the resistors values.

Once the transformation equations have been obtained for the end point coordinates U_{0n} and V_{0n}, they are applied to the end point data arising from the 2,000 Monte Carlo simulations to generate the standard bivariate normal random variables R_{0u} and R_{0v}, respectively. If a type j transformation, $j = 1, 2, 3$, is used, the original data is said to have a Johnson type j distribution. In practice, U_{0n} and V_{0n} need not have the same distributions (i.e., U_{0n} may be of type i whereas V_{0n} may be of type j and $i \neq j$). An estimate of the correlation coefficient between R_{0u} and R_{0v} is given by

$$
\hat{\rho} = \frac{1}{1999}\sum_{i=1}^{2000}[\frac{(R_{0u_i}-\hat{\mu}_{r_{0u}})(R_{0v_i}-\hat{\mu}_{r_{0v}})}{\hat{\sigma}_{r_{0u}}\hat{\sigma}_{r_{0v}}}] \tag{2.42}
$$

where $\hat{\mu}_{r_{0u}}$, $\hat{\mu}_{r_{0v}}$ and $\hat{\sigma}_{r_{0u}}$, $\hat{\sigma}_{r_{0v}}$ are the sample means and variances of the 2,000 transformed statistics R_{0u} and R_{0v}, respectively.

Since R_{0u} and R_{0v} are bivariate standard normal random variables, their joint PDF can be written as

$$
f_{R_{0u},R_{0v}}(r_{0u}, r_{0v}) = \frac{1}{2\pi\sqrt{1-\hat{\rho}^2}}\exp(-\frac{t}{2}) \tag{2.43}
$$

where

$$t = \frac{1}{1-\hat{\rho}^2}(r_{0u}^2 + r_{0v}^2 - 2\hat{\rho} r_{0u} r_{0v}). \tag{2.44}$$

Let $t = t_0$. Then the equation

$$t_0 = \frac{1}{1-\hat{\rho}^2}(r_{0u}^2 + r_{0v}^2 - 2\hat{\rho} r_{0u} r_{0v}) \tag{2.45}$$

is that of an ellipse in the r_{0u}, r_{0v} plane for which

$$f_{R_{0u},R_{0v}}(r_{0u}, r_{0v}) = \frac{1}{2\pi\sqrt{1-\hat{\rho}^2}} \exp(-\frac{t_0}{2}). \tag{2.46}$$

Points that fall within the ellipse correspond to those points in the r_{0u}, r_{0v} plane for which

$$f_{R_{0u},R_{0v}}(r_{0u}, r_{0v}) > \frac{1}{2\pi\sqrt{1-\hat{\rho}^2}} \exp(-\frac{t_0}{2}). \tag{2.47}$$

Let α be defined as the probability that r_{0u} and r_{0v} fall outside the ellipse given that the data comes from the null hypothesis. It follows that

$$\alpha = Pr(T > t_0). \tag{2.48}$$

Note that the bivariate normal distribution is a special case of the spherically invariant random vector (SIRV) where the characteristic PDF is given by [11].

$$f_S(s) = \delta(s-1). \tag{2.49}$$

The PDF of an N-dimensional SIRV involves the same quadratic form t that arises in the N-dimensional multivariate Gaussian PDF. For an SIRV, Rangaswamy [11] shows that the PDF of t is

$$f_T(t) = \frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})} t^{\frac{N}{2}-1} h_N(t); \quad (0 \le t \le \infty) \tag{2.50}$$

where $h_N(t)$ is a monotonically decreasing function given by

$$h_N(t) = \int_0^\infty s^{-N} \exp -\frac{t}{2s^2} f_S(s) ds. \tag{2.51}$$

Substituting eqn.(2.49) into eq.(2.51), there results

$$h_N(t) = \int_0^\infty s^{-N} \exp(-\frac{t}{2s^2})\delta(s-1)ds \qquad (2.52)$$
$$= \exp(-\frac{t}{2}).$$

For the bivariate case, $N = 2$. Consequently, eq.(2.50) reduces to

$$f_T(t) = \frac{1}{2}\exp(-\frac{t}{2}); \quad 0 \leq t \leq \infty. \qquad (2.53)$$

Hence,

$$\alpha = Pr(T > t_0) = \int_{t_0}^\infty \frac{1}{2}\exp(-\frac{t}{2})dt = \exp(-\frac{t_0}{2}). \qquad (2.54)$$

Consequently, $t_0 = -2\ln(\alpha)$. Thus eq.(2.45) becomes

$$\frac{1}{1-\hat{\rho}^2}(r_{0u}^2 + r_{0v}^2 - 2\hat{\rho}r_{0u}r_{0v}) = -2\ln(\alpha). \qquad (2.55)$$

The contour equation

$$r_{0u}^2 - 2\hat{\rho}r_{0u}r_{0v} + r_{0v}^2 = -2(1-\hat{\rho}^2)\ln(\alpha), \qquad (2.56)$$

which is the equation of an ellipse, is used to determine the $100 \times (1-\alpha)\%$ confidence contour. This is also shown in [15] and [16]. When the statistics R_{0u} and R_{0v} are uncorrelated, the correlation coefficient is 0 and eq.(2.56) becomes

$$r_{0u}^2 + r_{0v}^2 = -2\ln(\alpha), \qquad (2.57)$$

which is the equation of a circle. Also, eq.(2.56) degenerates into a line as the correlation coefficient approaches ± 1.

In the *Ozturk Algorithm*, an *inverse* Johnson Transformation is applied to the points for the confidence ellipses. The locus of the resulting points obtained is then plotted to obtain the corresponding confidence contours in the $U - V$ plane. Consequently, these confidence contours are not necessarily ellipsoidal.

The confidence contours are plotted for a given sample size n. These are then used to make a visual as well as computational test of the null hypothesis. If the terminal

point, Q_n, of the sample data, falls inside the contour, the data is declared as being consistent with the null hypothesis with confidence level $(1-\alpha)$. Otherwise the null hypothesis is rejected with a significance level α. Fig.2.2 shows the linked vectors and the confidence contours when the null distribution is standard normal and the sample data size is 100. From the figure it is seen that the sample data is statistically consistent with the null hypothesis at confidence level of 90%. Fig.2.3 shows a case where the sample data is not consistent with the null hypothesis at a siginificance level of 1%.

2.3.2 Distribution Approximation

The *distribution approximation* procedure is simply an extension of the Goodness of Fit test. Following a similiar approach to that outlined in the section for the Goodness of Fit test, random samples are generated from a library of different univariate probability distributions. In the Goodness of Fit test, the statistic $Q_{0n} = (U_{0n}, V_{0n})$ given by eq.(2.15) was obtained for the null hypothesis and for a specified n. For the distribution approximation we go one step further and for each distribution taken from a library of distributions, we obtain the end point statistic Q_n from eq.(2.12) for a given n and for various choices of the shape parameter. Thus, depending on whether it has a shape parameter or not, each distribution is represented by a point or a trajectory in a two dimensional plane whose coordinates are U_n and V_n. Fig. 2.4 shows an example of such a representation. The distributions which are plotted on the distribution approximation chart are (1) Gaussian, (2) Uniform, (3) Exponential, (4) Laplace, (5) Logistic, (6) Cauchy, (7) Extreme Value, (8) Gumbel type-2, (9) Gamma, (10) Pareto, (11) Weibull, (12) Lognormal, (13) Student-T, (14) K-distributed, (15) Beta, and (16) Su-Johnson. Tables 2.2 and 2.3 give the standard and the general form respectively, of these distributions.

vspace*8in

vspace*8in

As mentioned before, the points on the linked vectors for various distributions are computed using eq.(2.12). The magnitude for each point on the linked vectors is computed from values averaged over 2,000 Monte Carlo simulations of the ordered statistic, $Y_{i:n}$, obtained from eq.(2.6) while the angles are computed from the *reference distribution* (standard Gaussian). The confidence ellipses are computed only for the

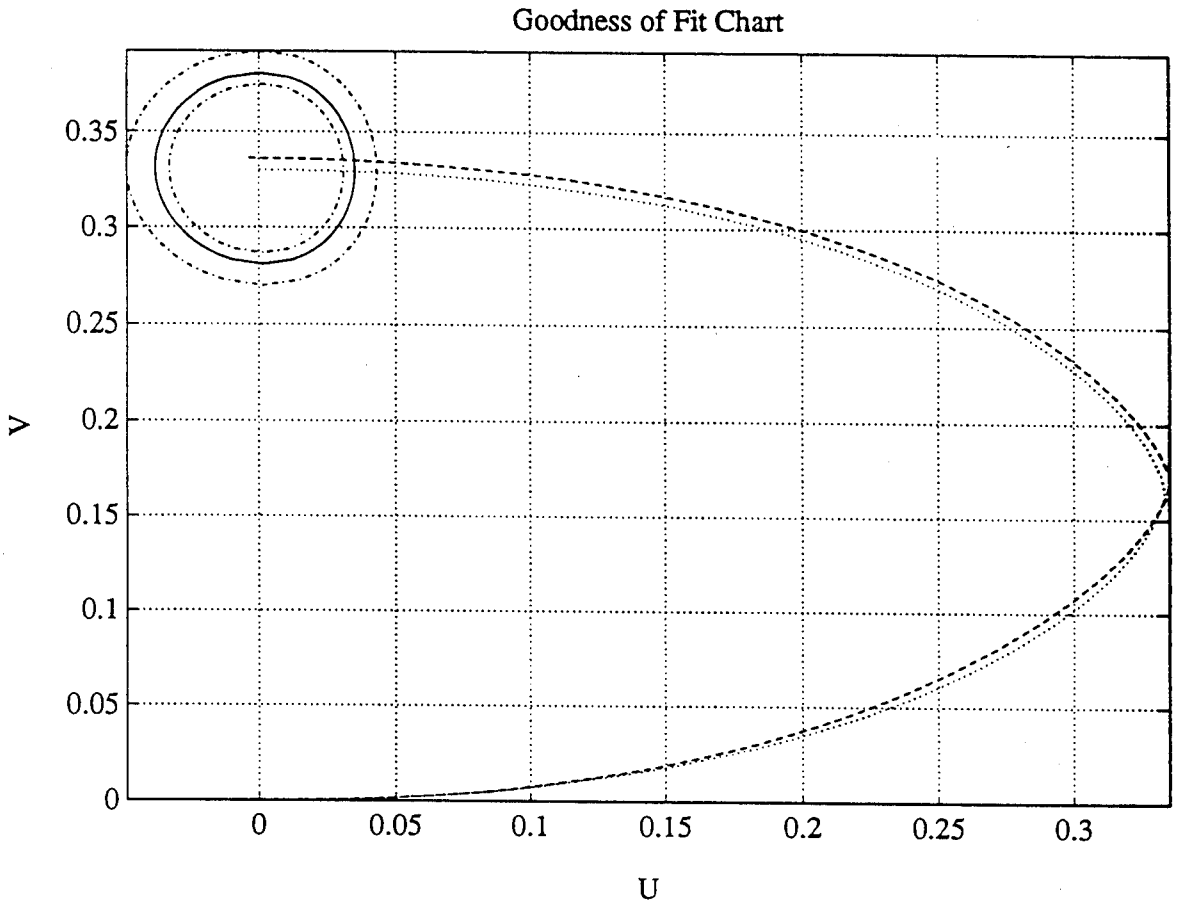

Figure 2.2: The Confidence Contours and the linked vectors with standard normal as null. Dotted Line = Null Distribution Pattern, Dashed Line = Sample Distribution Pattern. 90, 95, 99% contours from the innermost to the outermost respectively.

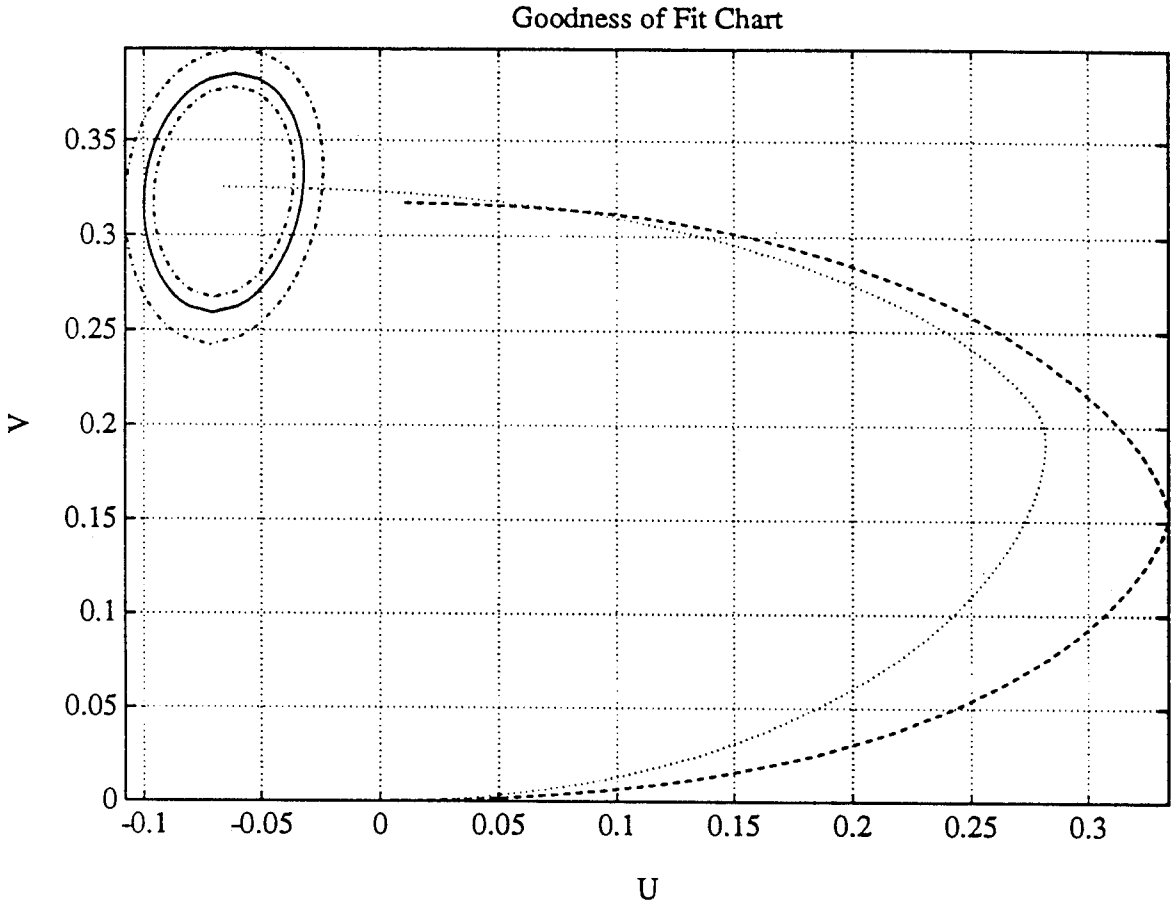

Figure 2.3: The Sample Data is not consistent with the null hypothesis. Dotted line = Null Distribution Pattern, Dashed Line = Sample Distribution Pattern.

Distribution	Standard Form $f_Y(y)$
Gaussian	$(\sqrt{2\pi}\beta)^{-1}\exp(-\frac{y^2}{2}) \quad -\infty < y < \infty$
Uniform	$1 \quad 0 < y < 1$
Exponential	$\exp(-y) \quad 0 < y < \infty$
Laplace	$0.5\exp(-\lvert y\rvert) \quad -\infty < y < \infty$
Logistic	$\exp(-y)[1+\exp(-y)]^{-2} \quad -\infty < y < \infty$
Cauchy	$\frac{1}{\pi(1+y^2)} \quad -\infty < y < \infty$
Extreme Value (Type 1)	$\exp(-y)\exp[-\exp(-y)] \quad -\infty < y < \infty$
Gumbel (Type 2)	$\gamma y\exp(-\gamma-1)\exp(-y^{-\gamma}) \quad \infty < y < \infty$
Gamma	$\frac{1}{\Gamma(\alpha)}\exp(-y)y^{\gamma-1} \quad \alpha < y < \infty$
Pareto	$\frac{\gamma}{y^{\gamma+1}} \quad y > 1, \gamma > 0$
Weibull	$\gamma y^{\gamma-1}\exp(-y^{\gamma}) \quad y > 0$
Lognormal	$\frac{\gamma}{\sqrt{2\pi}y}\exp[-\frac{\{\gamma\log(y)\}^2}{2}] \quad y > 0$
K-Distribution	$\frac{2}{\Gamma(\gamma)}(\frac{y}{2})^{\gamma}K_{\gamma-1}y \quad y > 0$
Beta	$\frac{1}{B(\gamma,\delta)}y^{\gamma-1}(1-y)^{\delta-1} \quad 0 < y < 1$
Johnson-SU	$\frac{\exp(\frac{(\sinh^{-1}(y)-\gamma)^2}{2\delta^2})}{\sqrt{2\pi}\delta\sqrt{1+y^2}} \quad -\infty < y < \infty$

Table 2.2: Standard Forms of the PDF's used in the Approximation Chart

Distribution	General Form $f_X(x)$
Gaussian	$(\sqrt{2\pi}\beta)^{-1}\exp(-\frac{(x-\alpha)^2}{2\beta^2} \quad -\infty < x < \infty$
Uniform	$\frac{1}{\beta} \quad \alpha < x < \alpha+\beta$
Exponential	$\frac{1}{\beta}\exp(-\frac{(x-\alpha)}{\beta}) \quad \alpha < x < \infty$
Laplace	$\frac{0.5}{\beta}\exp[-\lvert\frac{(x-\alpha)}{\beta}\rvert] \quad -\infty < x < \infty$
Logistic	$\frac{1}{\beta}\exp[-\frac{x-\alpha)}{\beta}][1+\exp(-\frac{(x-\alpha)}{\beta})]^{-2} \quad -\infty < x < \infty$
Cauchy	$\frac{1}{\pi\beta[1+\frac{(x-\alpha)^2}{\beta_2}]} \quad -\infty < x < \infty$
Extreme Value (Type 1)	$\frac{1}{\beta}\exp[-\frac{(x-\alpha)}{\beta}]\exp[-\exp\{-\frac{(x-\alpha)}{\beta}\}] \quad -\infty < x < \infty$
Gumbel (Type 2)	$\frac{\gamma}{\beta}\frac{(x-\alpha)}{\beta}\exp(-\gamma-1)\exp[-\frac{(x-\alpha)^{-\gamma}}{\beta\gamma}] \quad -\infty < x < \infty$
Gamma	$\frac{1}{\beta\Gamma(\alpha)}\exp[-\frac{(x-\alpha)}{\beta}](\frac{x-\alpha}{\beta})^{\gamma-1} \quad \alpha < x < \infty$
Pareto	$\frac{\gamma}{\beta}\frac{1}{(\frac{x-\alpha}{\beta})^{\gamma+1}} \quad x > \alpha+\beta, \gamma > 0$
Weibull	$\frac{\gamma}{\beta}(\frac{x-\alpha}{\beta})^{\gamma-1}\exp[-(\frac{x-\alpha}{\beta})^{\gamma}] \quad x > \alpha$
Lognormal	$\frac{\gamma}{\sqrt{2\pi}\beta(\frac{x-\alpha}{\beta})}\exp[-\frac{\{\gamma\log(\frac{x-\alpha}{\beta})\}^2}{2}] \quad x > \alpha$
K-Distribution	$\frac{2}{\beta\Gamma(\gamma)}[\frac{x-\alpha}{2\beta}]^{\gamma}K_{\gamma-1}[\frac{x-\alpha}{\beta}] \quad x > \alpha$
Beta	$\frac{1}{\beta B(\gamma,\delta)}(\frac{x-\alpha}{\beta})^{\gamma-1}[1-(\frac{x-\alpha}{\beta})]^{\delta-1} \quad \alpha < x < \alpha+\beta$
Johnson-SU	$\frac{1}{\beta}\frac{\exp(\frac{[\sinh^{-1}(\frac{x-\alpha}{\beta})-\gamma]^2}{2\delta^2})}{\sqrt{2\pi}\delta\sqrt{1+(\frac{x-\alpha}{\beta})^2}} \quad -\infty < x < \infty$

Table 2.3: General Forms of the PDF's used in the Approximation Chart

null hypothesis used in the prior Goodness of Fit test. Only the end point coordinates Q_n of the linked vectors are provided in the approximation chart. This is due to the fact that the plot would become too cluttered to properly interpret the data if all the linked vectors for these various distributions were provided in the graphics. Also, meaningful information from the linked vectors is contained in the location of their end points. Therefore, only the end points of all the linked vectors are provided in the approximation chart, along with the confidence ellipses for the selected null distribution.

For each distribution, such as Gaussian, that is uniquely specified by its mean and variance (no shape parameters), there exists a single end point on the approximation chart corresponding to the single unique linked vector.

For distributions dependent on a single shape parameter, such as Weibull, the end point of the of the linked vector will also be dependent on the shape parameter. Therefore, a sequence of linked vectors must be computed in order to obtain the trajectory on which the end point travels for varying shape parameter. In a sense, the trajectory represents a family of PDFs having the same distribution but with different shape parameter values. For example, the trajectory for the Weibull distribution is obtained by joining the end points for which the shape parameters are 0.3, 0.4, 0.5, 0.6, 0.8, 1.1, 1.5, 2.0, 3.0, 5.0. As the shape parameter increases, note that the Weibull distribution approaches the Gaussian distribution. This is shown in fig. 2.4. The representation of fig. 2.4 is called an approximation chart.

Similiarly, for a distribution dependent on two shape parameters, such as Beta, a sequence of linked vectors must be computed in order to plot the trajectories on which the end point travels for varying shape parameters. This is performed by holding the first shape parameter constant and varying the second shape parameter to generate a trajectory, then changing the first shape parameter and again holding it constant while varying the second shape parameter, etc... until a family of trajectories is produced that characterizes the distribution.

Thus, an approximation chart such as that in fig. 2.4 can be produced. It is apparent that this approximation chart provides a one to one graphical representation for each PDF for a given n. Therefore, every point in the approximation chart correponds to a specific distribution. Thus, if the null hypothesis in the Goodness of

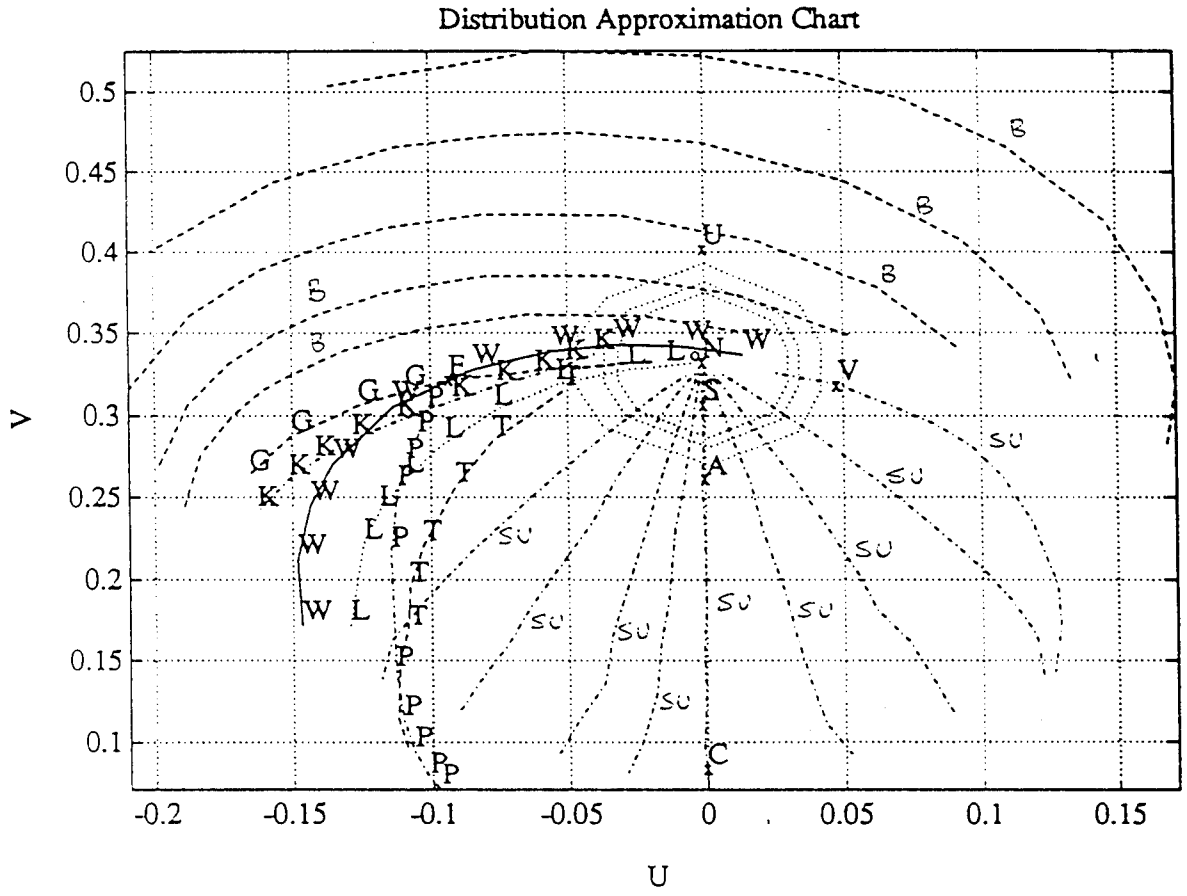

Figure 2.4: The Approximation Chart. 1) N = Normal, 2) U = Uniform, 3) C = Cauchy, 4) L = Lognormal, 5) S = Logistic, 6) A = Laplace, 7) V = Extreme Value, 8) T = T2-Gumbel, 9) G = Gamma, 10) E = -ve Exponential, 11) P = Pareto, 12) K = K-Distributed, 13) W = Weibull, 14) B = Beta, 15) SU = SU-Johnson.

Fit test is rejected, then the distribution which approximates the underlying PDF of the set of random data can be obtained by comparing Q_n obtained for the samples with the existing trajectories in the chart. The end point or trajectory closest to the Q_n of the sample data is chosen as an approximation to the PDF underlying the random data. This closest point or trajectory is determined by projecting the sample point Q_n to neighbouring points or trajectories on the chart and selecting that point or trajectory whose perpendicular distance from the sample point is the smallest. For example consider the situation of fig. 2.5. Let $Q_n = (u_n, v_n)$ denote the coordinates of the sample point. Let (x_1, y_1) and (x_2, y_2) denote the coordinates of the points A and B on the trajectory shown in fig. 2.5. The segment of the trajectory between points A and B is assumed to be linear. Let (x_0, y_0) denote the coordinates of the point of intersection of the straight line between A and B and the projection of $Q_n = (u_n, v_n)$ onto this straight line. The equation of the straight line between the points A and B can be written as

$$y - y_1 = m(x - x_1) \tag{2.58}$$

where $m = \frac{y_2 - y_1}{x_2 - x_1}$ and (x, y) is a point on the line. Also, the equation of the straight line joining (x_0, y_0) and (u_n, v_n) is

$$y - v_n = -\frac{1}{m}(x - u_n) \tag{2.59}$$

where (x, y) is a point on the perpendicular. The coordinates (x_0, y_0) result from letting $x = x_0$ and $y = y_0$ in eqs.(2.58) and (2.59). Their solution yields

$$\begin{aligned} x_0 &= \frac{1}{m^2+1}[m^2 - my_1 + u_n + mv_n] \\ y_0 &= \frac{1}{m^2+1}[y_1 - mx_1 + m_2 v_n + mu_n]. \end{aligned} \tag{2.60}$$

Finally, the perpendicular distance from the sample point onto the trajectory between points A and B is

$$D = \sqrt{\frac{1}{m^2+1}[m^2\psi_1^2 - 2m\psi_1\psi_2 + \psi_2^2]} \tag{2.61}$$

where

$$\begin{aligned} \psi_1 &= u_n - x_1, \\ \psi_2 &= v_n - y_1. \end{aligned} \tag{2.62}$$

The complete distribution approximation algorithm is summarized as follows.

1. Sort the samples $X_1, X_2, \ldots, X_n$ in increasing order.
2. Obtain the standardized order statistic $Y_{i:n}$.
3. Compute U_n and V_n from eq.(2.12) for the library of PDFs mentioned.
4. Obtain an approximation chart based on the sample size n and plot the sample point Q_n on this chart.
5. Compute the distance, D between the sample point Q_n and each of the end points on the chart. Choose the PDF corresponding to the point or trajectory that results in the smallest value for D as an approximation to the PDF of the samples.

The approximation to the underlying PDF of the set of random data can be improved by including as many distributions as possible in the approximation chart so as to fill as much of the space as possible with candidate distributions. It is emphasized, however, that this procedure does not identify the underlying PDF. It merely gives the best approximation to the distribution underlying the PDF of the data from those available in the chart.

2.3.3 Parameter Estimation

Once the probability distribution of the samples is approximated, the next step is to estimate its parameters. The method of distribution approximation discussed in 2.3.2 lends itself for estimating the parameters of the approximated distribution. We present the estimation procedure for the location, scale and shape parameters in this section.

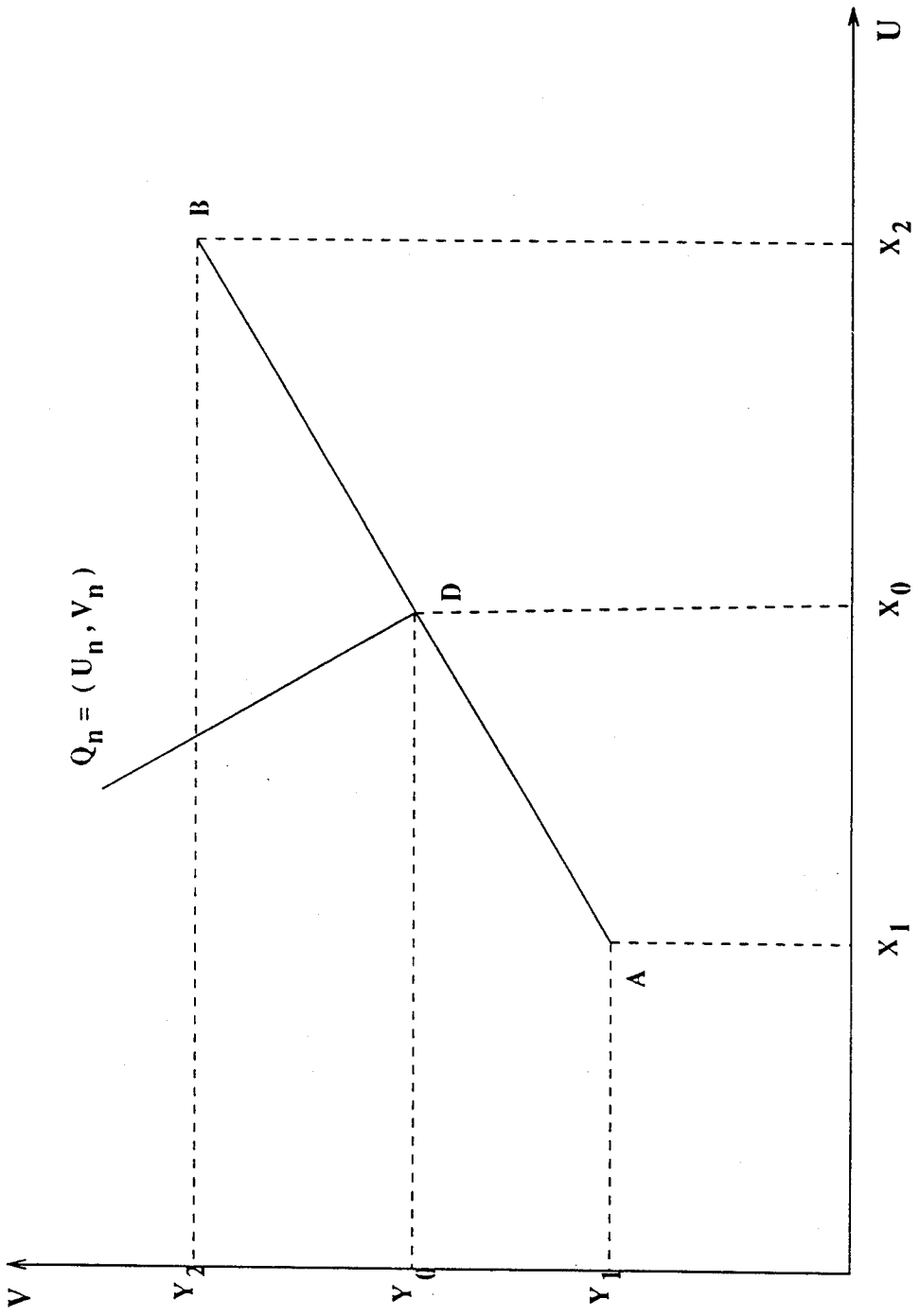

Figure 2.5: Distance Computation

2.3.3.1 Estimation of Location and Scale Parameters

Let $f(x : \alpha, \beta)$ denote a known distribution which approximates the PDF of the set of random data, where α and β are the location parameter and scale parameter, respectively, of the approximating PDF. Let $X_{i:n}$ denote the ordered statistics of X from a sample of size n. Let $S_{i:n}$ be defined by

$$S_{i:n} = \frac{X_{i:n} - \alpha}{\beta}. \tag{2.63}$$

Also, let

$$\mu_{i:n} = E[S_{i:n}]. \tag{2.64}$$

Then

$$E[X_{i:n}] = \beta\mu_{i:n} + \alpha. \tag{2.65}$$

We consider the following statistics

$$\begin{aligned} T_1 &= \sum_{i=1}^{n} \cos(\theta_i) X_{i:n} \\ T_2 &= \sum_{i=1}^{n} \sin(\theta_i) X_{i:n} \end{aligned} \tag{2.66}$$

where θ_i is the angle defined in eq.(2.10). The expected values of T_1 and T_2 are

$$\begin{aligned} E[T_1] &= \sum_{i=1}^{n} \cos(\theta_i)[\beta\mu_{i:n} + \alpha] \\ E[T_2] &= \sum_{i=1}^{n} \sin(\theta_i)[\beta\mu_{i:n} + \alpha]. \end{aligned} \tag{2.67}$$

These can be written as

$$\begin{aligned} E[T_1] &= a\alpha + b\beta \\ E[T_2] &= c\alpha + d\beta \end{aligned} \tag{2.68}$$

where

$$\begin{aligned} a &= \sum_{i=1}^{n} \cos(\theta_i) \\ b &= \sum_{i=1}^{n} \mu_{i:n} \cos(\theta_i) \\ c &= \sum_{i=1}^{n} \sin(\theta_i) \\ d &= \sum_{i=1}^{n} \mu_{i:n} \sin(\theta_i). \end{aligned} \tag{2.69}$$

Because the standardized Gaussian distribution is used as the reference distribution for θ_i, it can be shown that $a = 0$ [1]. The estimates for β and α are then given by

$$\begin{aligned} \hat{\beta} &= \frac{\widehat{E[T_1]}}{\hat{b}} \\ \hat{\alpha} &= \frac{\widehat{E[T_2]} - \hat{d}\hat{\beta}}{\hat{c}}. \end{aligned} \tag{2.70}$$

For sufficiently large n (i.e., $n > 50$), suitable estimates for $E[T_1]$ and $E[T_2]$ are

$$\begin{aligned} \widehat{E[T_1]} &\approx T_1 \\ \widehat{E[T_2]} &\approx T_2. \end{aligned} \tag{2.71}$$

Estimates for b and d rely upon an estimate of $\mu_{i:n}$. $\hat{\mu}_{i:n}$ is obtained from a Monte Carlo simulation of $S_{i:n}$ where $S_{i:n}$ is generated from the known approximating distribution $f(x; 0, 1)$ having zero location and unity scale parameters. $\hat{\mu}_{i:n}$ is the sample mean of $S_{i:n}$ based upon 2,000 Monte Carlo trials. Having $\hat{\mu}_{i:n}$, the estimates for b and d are given by

$$\begin{aligned} \hat{b} &= \sum_{i=1}^{n} \hat{\mu}_{i:n} \cos(\theta_i) \\ \hat{d} &= \sum_{i=1}^{n} \hat{\mu}_{i:n} \sin(\theta_i). \end{aligned} \tag{2.72}$$

The scale and location parameters are then estimated by application of eqs.(2.70) and (2.71).

2.3.3.2 Shape Parameter Estimation

In this section we present the approximate method used for estimating the shape parameter of the approximating PDF. We first consider distributions with only one shape parameter. Let γ denote the shape parameter of the approximating PDF. Since U_n and V_n are location and scale invariant, the point Q_n depends only on the sample size n and the shape parameter γ.

Recall that a point on the trajectories of the approximation chart is obtained by averaging for a specified value of the shape parameter the results from a large number of trials for U_n and V_n. Consequently, for given values of n and γ the coordinates of the corresponding point along the trajectory for a specified distribution can be characterized by

$$\begin{aligned} E(U_n) &= \phi_1(n,\gamma) \\ E(V_n) &= \phi_2(n,\gamma) \end{aligned} \tag{2.73}$$

where the complete trajectory is obtained by repeating the large number of trials for U_n and V_n over a suitable range of γ. On a given trial involving the random data it is likely that the coordinates U_n and V_n obtained for the samples will not coincide with any of the trajectories on the chart. The PDF underlying the random data is approximated by selecting the distribution corresponding to the point in the trajectory that falls closest to the sample point Q_n. The situation is illustrated in fig. 2.6. Q_n appears in the figure with coordinates (U_n, V_n). The straight line $\hat{T}_r$ denotes an approximation to a segment of the nearest trajectory which, in general, is a curved segment between points A and B. A is that point on the actual trajectory corresponding to the shape parameter γ_A. Its coordinates are (u_A, v_A). Similiarly, B is the point on the actual trajectory corresponding to the shape parameter γ_B. Its coordinates are (u_B, v_B). The slope of the straight line between points A and B is

$$m = \frac{v_B - v_A}{u_B - u_A}. \tag{2.74}$$

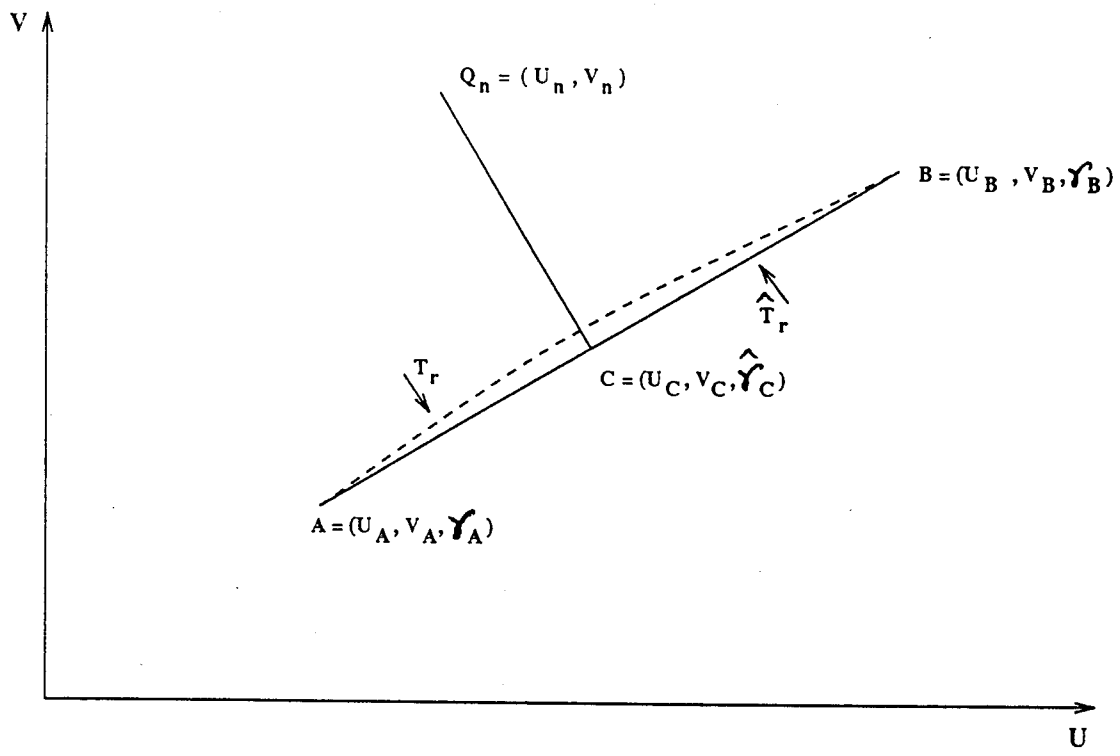

Figure 2.6: Shape Parameter Estimation

The equation for the straight line $\hat{T}_r$ is

$$v = v_A + m(u - u_A). \tag{2.75}$$

Point C, with coordinates (u_c, v_c), is the perpendicular projection of Q_n onto $\hat{T}_r$. The straight line linking Q_n and C has a slope equal to $-\frac{1}{m}$ and an equation of the form

$$v = V_n - \frac{1}{m}(u - U_n). \tag{2.76}$$

Since C is a point common to both straight lines, it follows from eqs.(2.75) and (2.76) that

$$v_A + m(u_C - u_A) = V_n - \frac{1}{m}(u_C - U_n). \tag{2.77}$$

Solution for u_C results in

$$u_C = \frac{m(V_n - v_A) + mu_A + U_n}{m^2 + 1}. \tag{2.78}$$

Let γ_C denote the shape parameter corresponding to the point on the actual trajectory closest to Q_n. An approximation to γ_C is then obtained by linear interpolation on $\hat{T}_r$. The result is

$$\hat{\gamma_C} = \gamma_A + \frac{(\gamma_B - \gamma_A)(u_C - u_A)}{(u_B - u_A)}. \tag{2.79}$$

The accuracy of the procedure can be improved by employing a non-linear interpolation method. It must be emphasized that the location, scale and shape parameter estimation procedures presented in this section are approximate procedures.

The proposed estimation procedure can also be extended to the two -shape parameter case. In this case one needs to choose at least three points $(u_1, v_1), (u_2, v_2)$ and (u_3, v_3) and let the shape parameter values corresponding to these three points be γ_1, γ_2 and γ_3, respectively. The points are chosen in such a way that they form the three vertices of a triangle inside which falls the sample point Q_n [1]. Again, by using a linear interpolation in the plane, an approximate solution can be obtained for the parameter estimates.

Chapter 3

Simulation Results of the Ozturk Algorithm

For univariate cases, the power of the Ozturk Algorithm has been studied for various distributions in [1], [2] and [3]. It was noted in [1] through [3] that the power of the algorithm depends on the sample size n, type of the standardized statistic and the null distribution. This algorithm has been found to compare favorably against all the well known tests. Also, the algorithm has been put to use to test its performance against different known distributions. Random data were generated using computer simulations as given in [18] and [19]. The Goodness of Fit test as well as the Distribution Approximation test was performed on these data using this Algorithm. In this chapter, a brief summary of some of the results obtained is presented.

Data was generated from four different null distributions, viz.,

- Univariate Gaussian
- Weibull (Shape Parameter 1)
- Gamma (Shape Parameter 1)
- Lognormal (Shape Parameter 1).

The Goodness of Fit test results are tabulated and presented first. The results of the Distribution Approximation are not easy to tabulate. We shall, therefore, present the result of a single case for the purpose of illustration.

3.1 Goodness of Fit Test Results

3.1.1 The Univariate Gaussian Case:

Data was generated from a Gaussian pseudo random number generator using computer simulations. The data set represented a zero mean and unit variance normal PDF. The following observations were noted.

- It was observed that a sample size of less than 40 is not advisable for the Goodness of Fit test as it almost always shows that the data is statistically consistent with any null. This is due to the fact that such a small sample size could be used to represent any PDF.
- A sample size between 75 and 100 is found to be good enough to accurately perform the Goodness of Fit test.
- For a sample size greater than 75 and when the null specified was Gaussian, the Goodness of Fit test showed that the data was statistically consistent with the null in almost all the cases . For other null hypotheses which are not close to Gaussian in the approximation chart, the Goodness of Fit test always showed that the data was statistically inconsistent with the nulls. But for null hypotheses which are close to Gaussian in the approximation chart, such as logistic, the Goodness of Fit test comes up with statistical consistency almost always. This vindicates the fact that the logistic PDF curve is very similiar to the Gaussian PDF curve.

Table 3.1 shows the results obtained for this case.

3.1.2 The Weibull Case:

Data was generated from the Weibull PDF with shape parameter 1 and the Goodness of Fit test was performed on it. The following observations were noted.

- The Goodness of Fit test worked well for a sample size between 75 and 100. For a smaller sample size, the Goodness of Fit test is not advisable since the results obtained were not accurate.

Sample Size (n)	Data Generated From	Null Distribution	No. of Cases of SC/ Total No.of Cases
5	Gaussian	Gaussian	3/3
25			3/3
40			4/7
50			4/7
75			7/7
100			7/8
125			8/8
150			8/8
25		Uniform	2/5
40			2/5
50			1/5
75			0/5
100			0/5
150			0/5
75		Exponential	0/8
100			0/8
150			0/8
75		Laplace	4/8
100			2/8
150			1/8
75		Logistic	7/8
100			7/8

Table 3.1: Results of the Ozturk Algorithm when the data generated was Gaussian. SC indicates Statistical Consistency

Sample Size (n)	Data Generated From	Null Distribution	No. of Cases of SC/ Total No.of Cases
5	Weibull (Sh.1)	Weibull (Sh.1)	7/8
25			7/9
40			4/8
50			5/8
75			6/8
100			8/8
150			8/8
5	Weibull (Sh.1)	Weibull (Sh.2)	7/8
25			3/8
40			2/8
50			0/8
75			0/9
100			0/8
150			0/8

Table 3.2: Ozturk Algorithm Results when data generated was Weibull with Shape Parameter 1. SC indicates Statistical Consistency and Sh. indicates Shape Parameter.

- When the null specified was Weibull with shape parameter 2, the Goodness of Fit test showed that the data was statistically inconsistent with the null for a sample size greater than 75 in all the cases. This is due to the fact that Weibull (Shape Parameter 1) is far away from Weibull (Shape Parameter 2) on the approximation chart.

Table 3.2 shows the results obtained for this case.

3.1.3 The Gamma Case:

Data was generated from the Gamma PDF with shape parameter 1 and the Goodness of Fit test was performed on it. Observations noted were almost the same as those for the Weibull case. Again a sample size between 75 and 100 was observed to have performed well in this case. The results for this case are tabulated in table 3.3.

3.1.4 The Lognormal Case:

A Lognormal pseudo random number generator was used to generate random data representing the Lognormal PDF with a shape parameter of 1. Observations noted for the Goodness of Fit test performance on this data were noted.

- A sample size between 50 and 75 was found to be sufficient to perform the Goodness of Fit test accurately.

Sample Size (n)	Data Generated From	Null Distribution	No. of Cases of SC/ Total No.of Cases
40	Gamma (Sh.1)	Gamma (Sh.1)	5/7
50			7/8
75			7/9
100			7/8
150			8/8
50	Gamma (Sh.1)	Gamma (Sh.5)	1/8
75			0/8
100			1/8

Table 3.3: Ozturk Algorithm Results when data generated was Gamma with Shape Parameter 1. SC indicates Statistical Consistency and Sh. indicates Shape Parameter.

Sample Size (n)	Data Generated From	Null Distribution	No. of Cases of SC/ Total No.of Cases
40	Lognormal (Sh.1)	Lognormal (Sh.1)	5/8
50			7/8
75			8/8
100			8/8
150			8/8
40	Lognormal (Sh.1)	Lognormal (Sh.0.5)	0/9
50			0/8
75			0/8
100			0/8

Table 3.4: Ozturk Algorithm Results when data generated was Lognormal with Shape Parameter 1. SC indicates Statistical Consistency and Sh. indicates Shape Parameter.

- This sample size gave very good results as far as the distribution approximation was concerned. About 30 times out of 40 the Lognormal PDF showed up in the five closest distributions that could be approximated.

Table 3.4 shows the results of the Goodness of Fit test for this case.

In general, the Goodness of Fit test seemed to perform well for a sample size of 100. The confidence contours grow smaller when the sample size is increased. In effect, we could hypothesize that for an infinite sample size the contours would become a point in the two dimensional (U, V) plane. This is intuitively satisfying.

Distribution No.	Distance	Rank
21	0.23728E-07	1
22	0.89802E-05	2
20	0.23473E-04	3
23	0.28047E-04	4
5	0.47542E-04	5

Table 3.5: Five closest PDF's given by distribution approximation test for a standard Gaussian data set

Distribution No.	Location Parameter	Scale Parameter	Shape (1) Parameter	Shape (2) Parameter
21	-0.23435	1.9774	2.4099	-0.2
22	-0.14976	1.9392	2.3697	-0.1
20	-0.39697	1.9747	2.4310	-0.4
23	-0.058871	1.9437	2.3717	0.0
5	-0.06215	0.49795	0.0	0.0

Table 3.6: Estimates of the Parameters of the five closest distributions chosen by the distribution approximation test for a standard Gaussian data

3.2 Distribution Approximation Test Results

The distribution approximation test was performed for a number of cases. In fact, it was performed for all the cases in which the Goodness of Fit test was performed. As mentioned previously, since it is not very easy to tabulate the results of the distribution approximation test for all these cases, results for a single test case are presented below.

Data was generated from standard Gaussian distribution using a Gaussian random number generator. A single test case consisting of 100 data points was considered. Using standard Gaussian as the null distribution, the distribution approximation test was performed on the data set. The first results of this test gave the five closest PDF's that the data could approximate. This result is shown in table 3.5
Distributions 20, 21, 22 and 23 are all SU-Johnson distributions with different shape parameters, where as distribution number 5 is logistic distribution. Note that the standard Gaussian was the 11th ranked PDF with a distance of 0.47879E-03. Estimates of the location, scale and the shape parameters given by the distribution approximation test for these distributions are given in table 3.6.

The approximation chart for this test case is shown in fig. 3.1. It is obvious

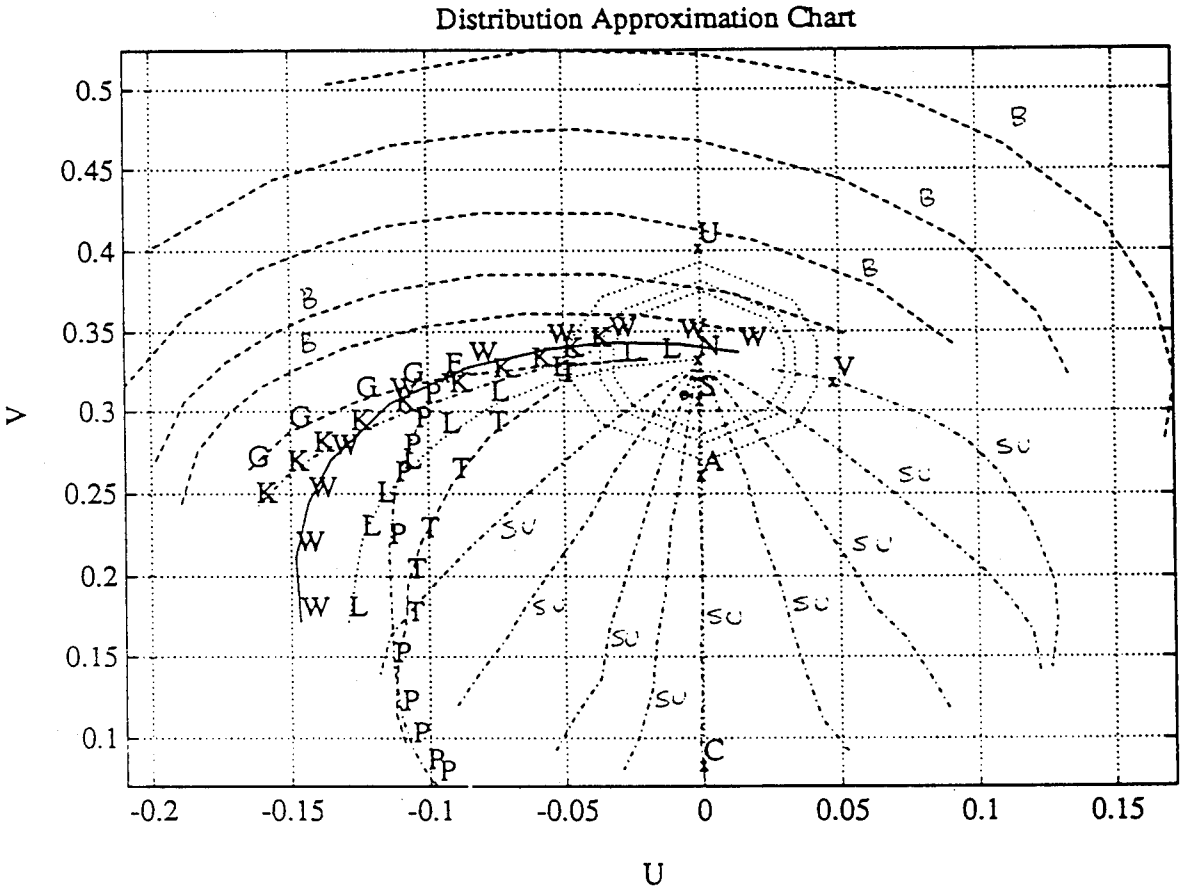

Figure 3.1: Approximation Chart for a standard Gaussian data set. 1) N = Normal, 2) U = Uniform, 3) C = Cauchy, 4) L = Lognormal, 5) S = Logistic, 6) A = Laplace, 7) V = Extreme Value, 8) T = T2-Gumbel, 9) G = Gamma, 10) E = -ve Exponential, 11) P= Pareto, 12) K = K-Distributed, 13) W = Weibull, 14) B = Beta, 15) SU = SU-Johnson.

from the approximation chart that the PDF's identified for this case are very close to the Gaussian PDF. It is therefore concluded that even though this data set has passed the Goodness of Fit test with standard Gaussian as the null, they could also be approximated by the set of 5 PDF's identified in table 3.5. In fact these are better approximations than standard Gaussian. This is shown by histogram plots shown in figures 3.2, 3.3 and 3.4. In these plots the histogram of the data is plotted along with the null hypothesis, which is the standard Gaussian, and one of the five distributions, given by the distribution approximation test, on the same coordinate axes. As is obvious from the figures there is very little to choose amongst the five PDF's approximated.

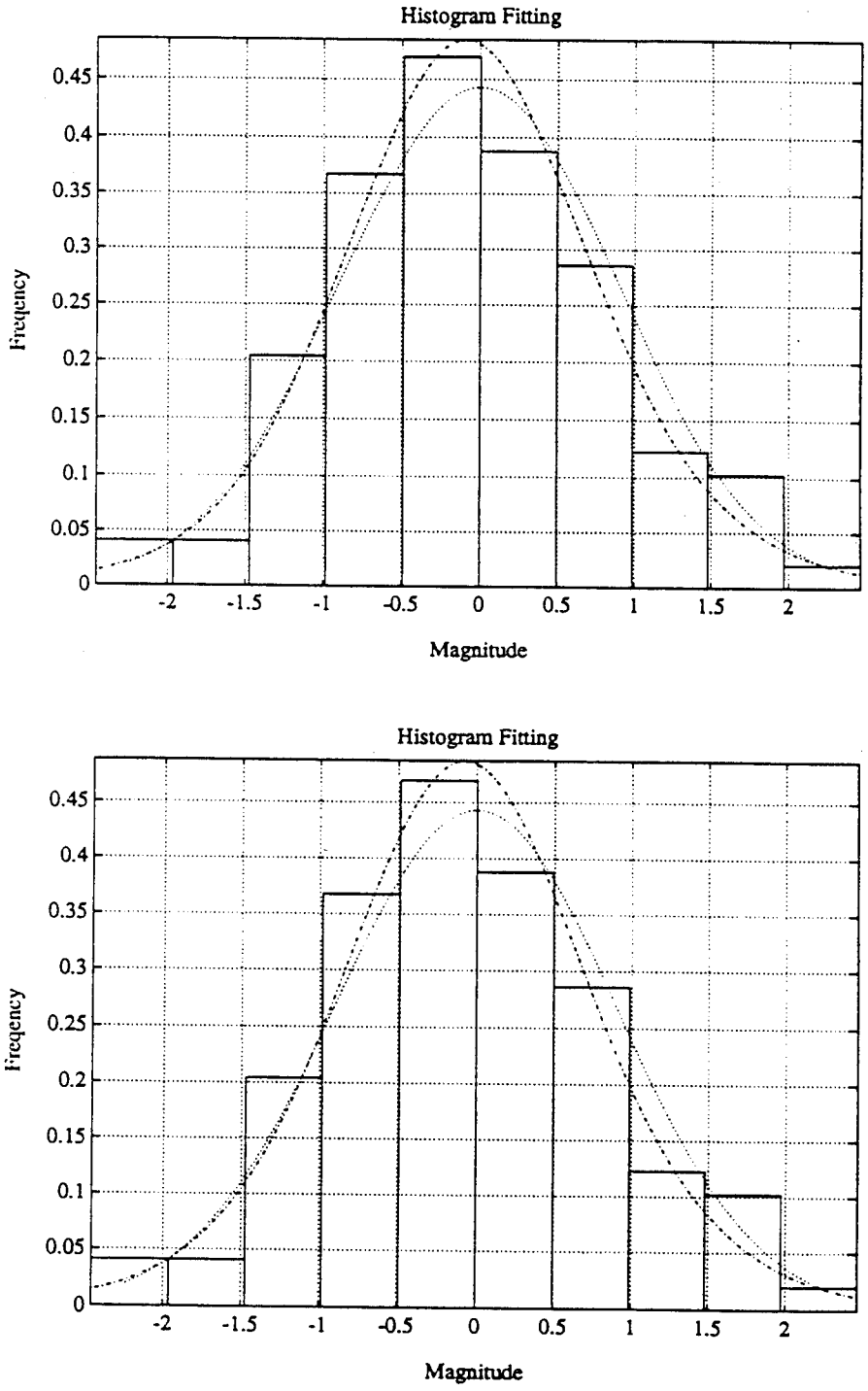

Figure 3.2: Histogram Plot: 1)Histogram plotted for the data, 2)Dotted curve is the standard Gaussian, 3)Dashed curve is PDF no. 21 for the top plot and PDF no 22 for the bottom plot. Parameters of the PDFs 21 and 22 are given in table

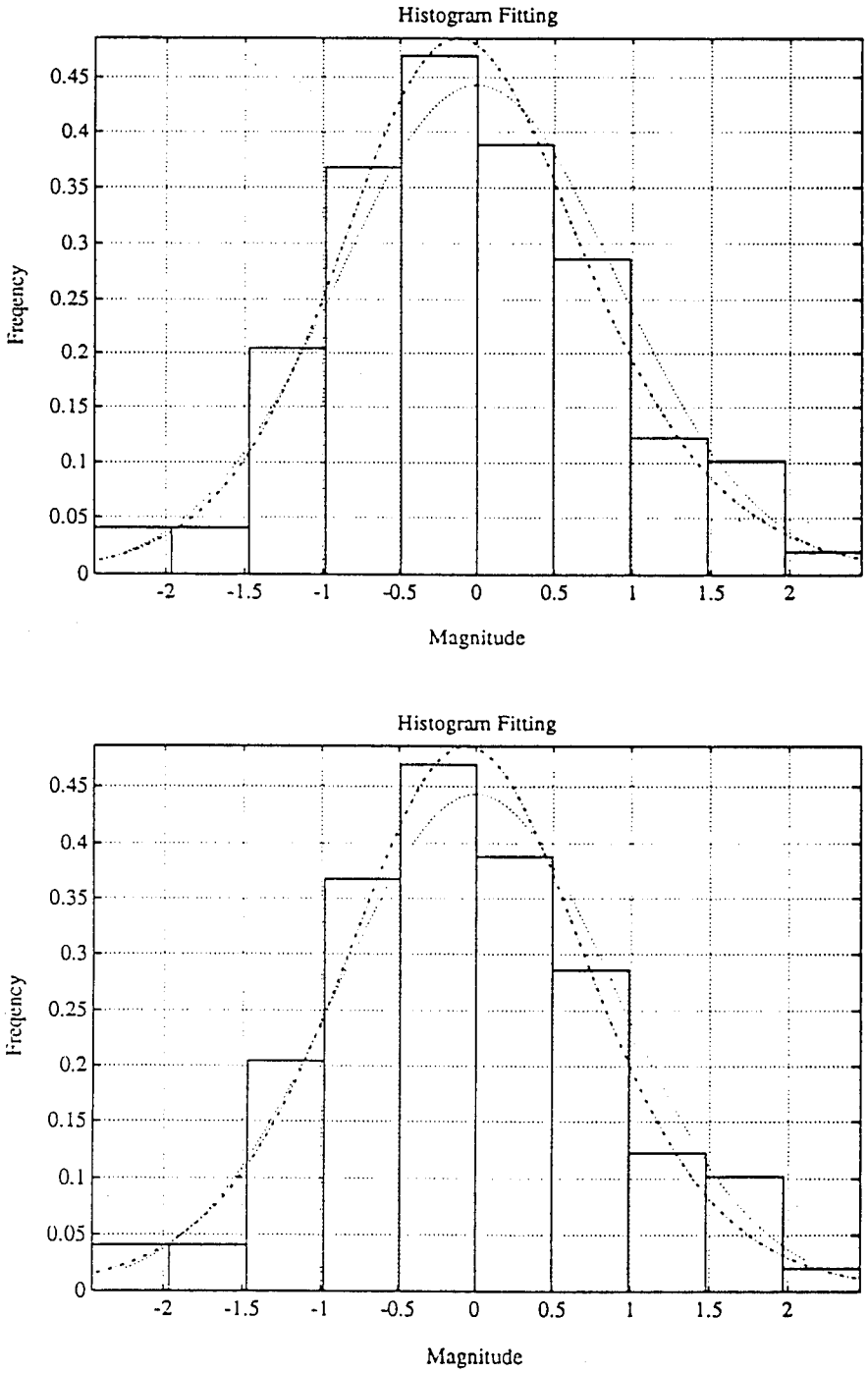

Figure 3.3: Histogram Plot: 1)Histogram plotted for the data, 2)Dotted curve is the standard Gaussian, 3)Dashed curve is PDF no. 20 for the top plot and PDF no 23 for the bottom plot. Parameters of the PDFs 20 and 23 are given in table

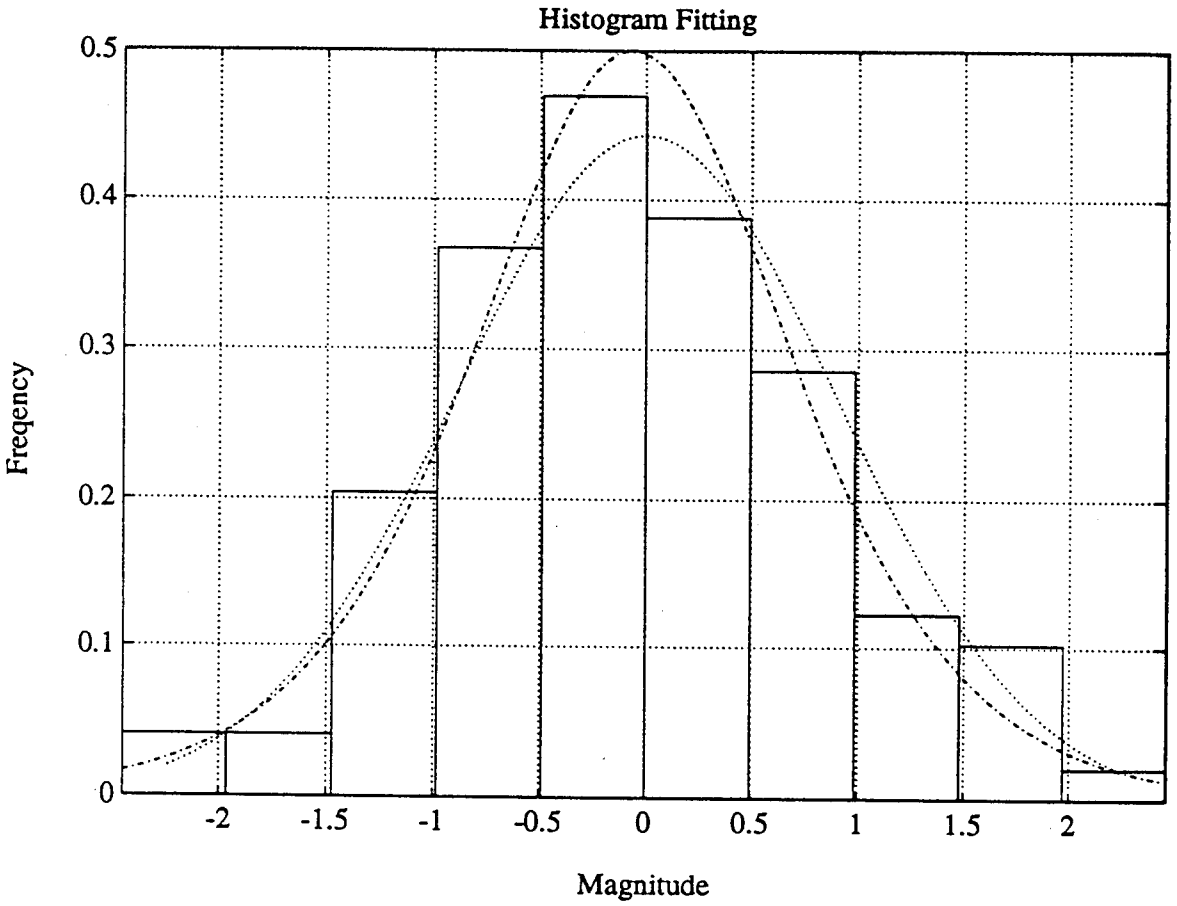

Figure 3.4: Histogram Plot: 1)Histogram plotted for the data, 2)Dotted curve is the standard Gaussian, 3)Dashed curve is PDF no. 5. Parameters of PDF 5 are given in table

Chapter 4

Conclusions and Suggestions for Future Work

4.1 Conclusions

This thesis has discussed various techniques for analyzing random data. Two areas were considered. The first area dealt with Goodness of Fit tests to determine whether or not a set of random data is statistically consistent with a prespecified probability distribution. After reviewing the Kolmogorov-Smirnov test, Chi-Square test, Q-Q Plots and P-P Plots, a new test, called the Ozturk Algorithm, was introduced. This test has an easily understood graphical presentation and works well for sample sizes as small as 100. The second area dealt with approximation of the underlying PDF of random data. Although the other tests were not applicable to this second area, the Goodness of Fit test of the Ozturk Algorithm was shown to lend itself to generation of a distribution approximation chart from which approximations to the underlying PDF of the random data can be obtained. Again, good results were observed for sample sizes as small as 100. An analysis was provided for generating confidence contours when the random data was non Gaussian. Simulated data was used to evaluate performance of the Ozturk Algorithm and some of the results were presented.

4.2 Suggestions for Future Work

Several problems remain to be explored with the Ozturk Algorithm:

1. The Ozturk Algorithm works well for continuous probability density functions. Generalizations of the Ozturk Algorithm to the

discrete case should be explored.

2. Extension of the Ozturk Algorithm from univariate to multivariate PDF's should be considered. One possibility involves utilization of quadratic forms of the data [7], [11].

3. Rangaswamy [11] demonstrated that multivariate spherically invariant random processes (SIRP's) can be approximated by means of their quadratic forms. However, a probability distribution approximation chart for SIRP's that could be utilized by the Ozturk Algorithm remains to be generated.

4. The univariate PDF's currently included in the approximation chart are unimodal. Extension to multimodal PDF's should be explored.

5. When the number of data points is much greater than 100, the Ozturk Algorithm requires considerable time to process the data. Ways should be examined for making the algorithm more efficient. This includes parallelization of the algorithm as well as processing the data in groups of 100 and averaging the results.

6. Reduction of the Ozturk Algorithm to chip form should be investigated for real-time applications.

Appendix A

Algebraic Derivations for Johnson Distributions

In this appendix, the criteria given in eq.(2.25) of chapter 2 are established and the parameter estimates given in equations (2.30), (2.32) and (2.34) are developed.

A.1 Johnson S_U Distributions

The transformation for the *Johnson S_U Distribution* is of the form

$$R = \gamma + \eta \sinh^{-1}(\frac{G-\epsilon}{\lambda}) \tag{A.1}$$

where R is a standard normal variable and ϵ is a location parameter, λ is a scale parameter, and γ and η are shape parameters for the PDF of the *S_U Distribution.* Solving eq.(A.1) for G in terms of R, there results

$$G = \epsilon + \lambda \sinh(\frac{R-\gamma}{\eta}) = \epsilon - \lambda \sinh(\frac{\gamma-R}{\eta}) \tag{A.2}$$

where we have made use of the fact that $\sinh(A)$ is an odd function of A. Define

$$\begin{aligned} m &= g_{3r} - g_r \\ l &= g_{-r} - g_r \\ p &= g_r - g_{-r} \end{aligned} \tag{A.3}$$

where $g_{\pm r}, g_{\pm 3r}$ are obtained from eq.(A.2) for $\pm r$ and $\pm 3r$, respectively.

For a fixed positive value of r, eqs.(A.2) and(A.3) give

$$
\begin{aligned}
m &= \lambda[\sinh(\frac{\gamma - r}{\eta}) - \sinh(\frac{\gamma - 3r}{\eta})] \\
&= \lambda[\sinh(\frac{\gamma - 2r + r}{\eta}) - \sinh(\frac{\gamma - 2r - r}{\eta})], \\
l &= \lambda[\sinh(\frac{\gamma + 3r}{\eta}) - \sinh(\frac{\gamma + r}{\eta})] \\
&= \lambda[\sinh(\frac{\gamma + 2r + r}{\eta}) - \sinh(\frac{\gamma + 2r - r}{\eta})], \\
p &= \lambda[\sinh(\frac{\gamma + r}{\eta}) - \sinh(\frac{\gamma - r}{\eta})].
\end{aligned}
\tag{A.4}
$$

Using the standard formula

$$
\sinh(A + B) - \sinh(A - B) = 2\cosh A \sinh B, \tag{A.5}
$$

we obtain the values of m, l and p from eq.(A.4) as

$$
\begin{aligned}
m &= 2\lambda \cosh(\frac{\gamma - 2r}{\eta}) \sinh(\frac{r}{\eta}) \\
l &= 2\lambda \cosh(\frac{\gamma + 2r}{\eta}) \sinh(\frac{r}{\eta}) \\
p &= 2\lambda \cosh(\frac{\gamma}{\eta}) \sinh(\frac{r}{\eta}).
\end{aligned}
\tag{A.6}
$$

Thus,

$$
\begin{aligned}
\frac{m}{p} &= \frac{\cosh(\frac{\gamma - 2r}{\eta})}{\cosh(\frac{\gamma}{\eta})} \\
\frac{l}{p} &= \frac{\cosh(\frac{\gamma + 2r}{\eta})}{\cosh(\frac{\gamma}{\eta})},
\end{aligned}
\tag{A.7}
$$

which gives

$$
\frac{ml}{p^2} = \frac{\cosh(\frac{\gamma - 2r}{\eta}) \cosh(\frac{\gamma + 2r}{\eta})}{\cosh^2(\frac{\gamma}{\eta})}. \tag{A.8}
$$

Using the identity

$$\cosh(A+B)\cosh(A-B) = \cosh^2 A + \cosh^2 B - 1, \tag{A.9}$$

we obtain from eq.(A.8)

$$\begin{aligned} \frac{ml}{p^2} &= \frac{\cosh^2(\frac{\gamma}{\eta}) + \cosh^2(\frac{2r}{\eta}) - 1}{\cosh^2(\frac{\gamma}{\eta})} \\ &= 1 + \frac{\cosh^2(\frac{2r}{\eta}) - 1}{\cosh^2(\frac{\gamma}{\eta})}. \end{aligned} \tag{A.10}$$

Since $\cosh^2(\frac{2r}{\eta}) > 1$, it is clear that $\frac{ml}{p^2} > 1$ for the S_U distribution.

Applying the identity,

$$\cosh(A+B) + \cosh(A-B) = 2\cosh A \cosh B, \tag{A.11}$$

to the sum of m/p and l/p in eq.(A.7), we get

$$\begin{aligned} \frac{m}{p} + \frac{l}{p} &= \frac{2\cosh(\frac{\gamma}{\eta})\cosh(\frac{2r}{\eta})}{\cosh(\frac{\gamma}{\eta})} \\ &= 2\cosh(\frac{2r}{\eta}). \end{aligned} \tag{A.12}$$

Solving eq.(A.12) for η, we get

$$\eta = \frac{2r}{\cosh^{-1}[\frac{1}{2}(\frac{m}{p} + \frac{l}{p})]} \quad \eta > 0. \tag{A.13}$$

Since $\eta > 0$ is assumed, the positive value of the inverse of cosh(.) must be chosen.

The expression for $\cosh(2r/\eta)$ in eq.(A.12) can be substituted in eq.(A.10) to give

$$\frac{ml}{p^2} = \frac{\cosh^2(\frac{\gamma}{\eta}) + (\frac{m+l}{2p})^2 - 1}{\cosh^2(\frac{\gamma}{\eta})}. \tag{A.14}$$

Solving for $\cosh^2(\gamma/\eta)$,

$$\cosh^2(\frac{\gamma}{\eta}) = \frac{(m+l)^2 - 4p^2}{4(ml - p^2)}. \tag{A.15}$$

Since $\cosh^2(A) - \sinh^2(A) = 1$, eq.(A.15) leads to

$$\sinh^2(\frac{\gamma}{\eta}) = \frac{(m-l)^2}{4(ml-p^2)} \tag{A.16}$$

Thus,

$$\sinh(\frac{\gamma}{\eta}) = \frac{[(m-l)^2]^{1/2}}{2(ml-p^2)^{1/2}}. \tag{A.17}$$

Unlike η, the parameter γ may be either positive or negative. Thus, a determination of the sign of the numerator in eq.(A.17) must be made. Taking the difference of l/p and m/p in eq.(A.7) and applying the hyperbolic identity,

$$\cosh(A+B) - \cosh(A-B) = 2\sinh A \sinh B, \tag{A.18}$$

yields

$$\frac{l-m}{p} = \frac{2\sinh(\frac{\gamma}{\eta})\sinh(\frac{2r}{\eta})}{\cosh(\frac{\gamma}{\eta})}. \tag{A.19}$$

Both $\cosh(\gamma/\eta) > 0$ and $\sinh(2r/\eta) > 0$ (since $r > 0$) and hence the sign of $\sinh(\gamma/\eta)$ is the same as that of $l-m$. It follows that

$$\sinh(\frac{\gamma}{\eta}) = \frac{l-m}{2(ml-p^2)^{1/2}} \tag{A.20}$$

and γ is given by

$$\gamma = \eta sinh^{-1}[\frac{(\frac{l}{p} - \frac{m}{p})}{2(\frac{m}{p}\frac{l}{p} - 1)^{1/2}}]. \tag{A.21}$$

As indicated earlier, the expression for p is given by eq.(A.6) as

$$p = 2\lambda\gamma\cosh(\frac{\gamma}{\eta})\sinh(\frac{r}{\eta}). \tag{A.22}$$

$\mathrm{Sinh}(r/\eta)$ can be obtained by using the relationship $\sinh^2 A = \frac{\cosh 2A - 1}{2}$ in conjunction with eq.(A.12). This gives

$$\sinh^2(\frac{r}{\eta}) = (\frac{\frac{m+l}{2p} - 1}{2}) = \frac{m+l-2p}{4p}. \tag{A.23}$$

$\cosh(\gamma/\eta)$ is known from eq.(A.15). Using eqs.(A.22), (A.15), and (A.23), we get

$$p = 2\lambda \frac{(m+l-2p)(m+l+2p)^{1/2}}{4[p(ml-p^2)]^{1/2}}. \tag{A.24}$$

Consequently,

$$\lambda = \frac{2p(\frac{ml}{p^2}-1)^{1/2}}{(\frac{m}{p}+\frac{l}{p}-2)(\frac{m}{p}+\frac{l}{p}+2)^{1/2}}. \tag{A.25}$$

Finally, consider

$$g_r + g_{-r} = 2\epsilon - \lambda[\sinh(\frac{\gamma - r}{\eta}) + \sinh(\frac{\gamma + r}{\eta})]. \tag{A.26}$$

Using the standard formula,

$$\sinh(A+B) + \sinh(A-B) = 2\sinh A \cosh B, \tag{A.27}$$

eq.(A.26) can be rewritten as

$$g_r + g_{-r} = 2\epsilon - 2\lambda \sinh(\frac{\gamma}{\eta})\cosh(\frac{r}{\eta}). \tag{A.28}$$

$\mathrm{Cosh}(r/\eta)$ in eq.(A.28) is obtained using the relation, $\cosh^2 A = \frac{1+\cosh(2A)}{2}$. It follows from eq.(A.12), that

$$\cosh^2(\frac{r}{\eta}) = \frac{1+\frac{(m+l)}{2p}}{2} = \frac{2p+m+l}{4p}. \tag{A.29}$$

Thus, using the values of $\sinh(\gamma/\eta)$ from eq.(A.20), $\cosh(r/\eta)$ from eq.(A.29) and λ from eq.(A.25), it is seen that

$$\begin{aligned} g_r + g_{-r} &= 2\epsilon - 2[\frac{2p[p(ml-p^2)]^{1/2}}{(m+l-2p)(m+l+2p)^{1/2}}][\frac{l-m}{2(ml-p^2)^{1/2}}][\frac{(2p+m+l)^{1/2}}{2p^{1/2}}] \\ &= 2\epsilon - \frac{p(l-m)}{m+l-2p}. \end{aligned} \tag{A.30}$$

Consequently, we get

$$\epsilon = \frac{1}{2}[g_r + g_{-r} + \frac{p(\frac{l}{p}-\frac{m}{p})}{(\frac{m}{p}+\frac{l}{p}-2)}]. \tag{A.31}$$

A.2 Johnson S_B Distribution

The transformation of the *Johnson S_B Distribution* is of the form

$$R = \gamma + \eta \ln(\frac{G-\epsilon}{\lambda+\epsilon-G}) \tag{A.32}$$

where R is the standard normal random variable and ϵ is a location parameter, λ is a scale paramter, and γ and η are the shape parameters of the S_B *Distribution.* Solving eq.(A.32) for G, we obtain

$$G = \epsilon + \frac{\lambda}{1+\exp\frac{\gamma-R}{\eta}}. \tag{A.33}$$

As was done in the previous section, we now proceed to find m, l and p in terms of the parameters γ, λ, η, and ϵ. Thus,

$$\begin{aligned} m &= g_{3r} - g_r \\ &= \frac{\lambda}{1+\exp(\frac{\gamma-3r}{\eta})} - \frac{\lambda}{1+\exp(\frac{\gamma-r}{\eta})}. \end{aligned} \tag{A.34}$$

Therefore,

$$m = \lambda[\frac{\exp(\frac{\gamma-r}{\eta}) - \exp(\frac{\gamma-3r}{\eta})}{1+\exp(\frac{\gamma-3r}{\eta}) + \exp(\frac{\gamma-r}{\eta}) + \exp(\frac{2\gamma-4r}{\eta})}] \tag{A.35}$$

and consequently,

$$m = \lambda\frac{\exp(\frac{\gamma}{\eta})\exp(\frac{-2r}{\eta})[\exp(\frac{r}{\eta}) - \exp(\frac{-r}{\eta})]}{1+\exp(\frac{\gamma-3r}{\eta}) + \exp(\frac{\gamma-r}{\eta}) + \exp(\frac{2\gamma-4r}{\eta})}. \tag{A.36}$$

This implies

$$m = \lambda\frac{2\exp(\frac{\gamma-2r}{\eta})\sinh(\frac{r}{\eta})}{\exp(\frac{\gamma-2r}{\eta})[\exp(-\frac{\gamma-2r}{\eta}) + \exp(\frac{-r}{\eta}) + \exp(\frac{r}{\eta}) + \exp(\frac{\gamma-2r}{\eta})]}. \tag{A.37}$$

Finally, we get

$$m = \lambda\frac{\sinh(\frac{r}{\eta})}{\cosh(\frac{\gamma-2r}{\eta}) + \cosh(\frac{r}{\eta})}. \tag{A.38}$$

Proceeding in a similiar fashion, we find

$$
\begin{aligned}
l &= g_{-r} - g_{-3r} \\
&= \lambda \frac{\sinh(\frac{r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma+2r}{\eta})}
\end{aligned}
\tag{A.39}
$$

and

$$
\begin{aligned}
p &= g_{r} - g_{-r} \\
&= \lambda \frac{\sinh(\frac{r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta})}.
\end{aligned}
\tag{A.40}
$$

In summary, the values of m, l and p are

$$
\begin{aligned}
m &= \frac{\lambda \sinh(\frac{r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma-2r}{\eta})} \\
l &= \frac{\lambda \sinh(\frac{r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma+2r}{\eta})} \\
p &= \frac{\lambda \sinh(\frac{r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta})}.
\end{aligned}
\tag{A.41}
$$

Hence,

$$
\begin{aligned}
\frac{p}{m} &= \frac{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma-2r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta})} \\
\frac{p}{l} &= \frac{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma+2r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta})},
\end{aligned}
\tag{A.42}
$$

which gives

$$
\frac{p}{m}\frac{p}{l} = \frac{A}{B} \tag{A.43}
$$

where

$$A = \cosh^2(\frac{r}{\eta}) + \cosh(\frac{\gamma-2r}{\eta})\cosh(\frac{\gamma+2r}{\eta}) + \cosh(\frac{r}{\eta})\cosh(\frac{\gamma-2r}{\eta}) + \cosh(\frac{r}{\eta})\cosh(\frac{\gamma+2r}{\eta}) \quad \text{(A.44)}$$

and

$$B = \cosh^2(\frac{r}{\eta}) + \cosh^2(\frac{\gamma}{\eta}) + 2\cosh(\frac{\gamma}{\eta})\cosh(\frac{r}{\eta}). \quad \text{(A.45)}$$

Using the identities

$$\begin{aligned} \cosh(A+B)\cosh(A-B) &= \cosh^2(A) + \cosh^2(B) - 1 \quad \text{(A.46)} \\ \cosh(A+B) + \cosh(A-B) &= 2\cosh(A)\cosh(B) \\ \cosh^2(A) - \sinh^2(A) &= 1 \end{aligned}$$

and applying them to eq.(A.45), we get

$$A = \cosh^2(\frac{r}{\eta}) + \cosh^2(\frac{\gamma}{\eta}) + 2\cosh(\frac{r}{\eta})\cosh(\frac{\gamma}{\eta})\cosh(\frac{2r}{\eta}) + \sinh^2(\frac{2r}{\eta}). \quad \text{(A.47)}$$

B is smaller than A because the cosh function assumes only positive values and since $\cosh(\frac{2r}{\eta}) > 1$ and $\sinh^2(\frac{2r}{\eta}) > 0$. This gives $\frac{p^2}{ml} > 1$ or equivalently, $\frac{ml}{p^2} < 1$.

From eq.(A.42)

$$\begin{aligned} 1 + \frac{p}{m} &= \frac{2\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta}) + \cosh(\frac{\gamma-2r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta})} \quad \text{(A.48)} \\ 1 + \frac{p}{l} &= \frac{2\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta}) + \cosh(\frac{\gamma+2r}{\eta})}{\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta})}. \end{aligned}$$

Multiplying the expressions in eq.(A.48) and using the identities in eq.(A.46) we get

$$(1+\frac{p}{m})(1+\frac{p}{l}) = \frac{C}{D} \tag{A.49}$$

where

$$\begin{aligned} C &= 4\cosh^2(\frac{r}{\eta}) + 2\cosh^2(\frac{\gamma}{\eta})(1+\cosh(\frac{2r}{\eta}) \\ &\quad + 4\cosh(\frac{\gamma}{\eta})\cosh(\frac{r}{\eta})(\cosh(\frac{2r}{\eta})+1) + \cosh^2(\frac{2r}{\eta}) - 1 \end{aligned} \tag{A.50}$$

and

$$D = \cosh^2(\frac{r}{\eta}) + \cosh^2(\frac{\gamma}{\eta}) + 2\cosh(\frac{r}{\eta})\cosh(\frac{\gamma}{\eta}). \tag{A.51}$$

Using the property, $\cosh(2A)+1 = 2\cosh^2 A$ in eq.(A.50), we get

$$\begin{aligned} C &= 4\cosh^2(\frac{r}{\eta}) + 4\cosh^2(\frac{\gamma}{\eta})\cosh^2(\frac{r}{\eta}) \\ &\quad + 8\cosh(\frac{\gamma}{\eta})\cosh^3(\frac{r}{\eta}) + 4\cosh^4(\frac{r}{\eta}) - 4\cosh^2(\frac{r}{\eta}). \end{aligned} \tag{A.52}$$

Thus

$$C = 4\cosh^2(\frac{r}{\eta})D. \tag{A.53}$$

Finally, from eqs.(A.49) and (A.53) we get

$$(1+\frac{p}{m})(1+\frac{p}{l}) = 4\cosh^2(\frac{r}{\eta}). \tag{A.54}$$

Solution for η results in

$$\eta = \frac{r}{\cosh^{-1}[\frac{1}{2}\{(1+\frac{p}{m})(1+\frac{p}{l})\}^{1/2}]}. \tag{A.55}$$

Now consider the sum of the terms in eq.(A.42). Using the identities in eq.(A.46) on this sum we get

$$\frac{p}{m}+\frac{p}{l} = \frac{2\cosh(\frac{r}{\eta})+\cosh(\frac{\gamma+2r}{\eta})+\cosh(\frac{\gamma-2r}{\eta})}{\cosh(\frac{r}{\eta})+\cosh(\frac{\gamma}{\eta})} \tag{A.56}$$
$$= \frac{2\cosh(\frac{r}{\eta})+2\cosh(\frac{\gamma}{\eta})\cosh(\frac{2r}{\eta})}{\cosh(\frac{r}{\eta})+\cosh(\frac{\gamma}{\eta})}.$$

Solving the above equation for $\cosh(\gamma/\eta)$, we get

$$\cosh(\frac{\gamma}{\eta}) = \frac{\cosh(\frac{r}{\eta})(\frac{p}{m}+\frac{p}{l}-2)}{(\frac{p}{m}+\frac{p}{l}+2)-4\cosh^2(\frac{r}{\eta})} \tag{A.57}$$

where again we have made use of the identities in eq.(A.46). From eq.(A.54) expressions of $\cosh(r/\eta)$ and $\sinh^2(r/\eta)$ are obtained and substituted in eq.(A.57). The result is

$$\cosh(\frac{\gamma}{\eta}) = \frac{(\frac{p}{m}+\frac{p}{l}-2)[(1+\frac{p}{m})(1+\frac{p}{l})]^{1/2}}{2(1+\frac{p}{m}+\frac{p}{l}+\frac{p}{m}\frac{p}{l}-\frac{p}{m}-\frac{p}{l}-2)} \tag{A.58}$$
$$= \frac{(\frac{p}{m}+\frac{p}{l}-2)[(1+\frac{p}{m})(1+\frac{p}{l})]^{1/2}}{2(\frac{p^2}{ml}-1)}.$$

Rather than solve this for γ, it is preferable to derive $\sinh(\gamma/\eta)$ because $\sinh^{-1}(.)$ yields the correct sign of γ due to it being single valued. Again we use the identity in eq.(A.46) to get sinh(.) from cosh(.). Thus

$$\sinh(\frac{\gamma}{\eta}) = \frac{[(\frac{p}{m}-\frac{p}{l})^2]^{1/2}[(1+\frac{p}{m})(1+\frac{p}{l})-4]^{1/2}}{2(\frac{p}{m}\frac{p}{l}-1)}. \tag{A.59}$$

To see which root should be taken in the first factor in the numerator, observe from eq.(A.42) that

$$\frac{p}{l}-\frac{p}{m} = \frac{2\sinh(\frac{\gamma}{\eta})\sinh(\frac{2r}{\eta})}{\cosh(\frac{r}{\eta})+\cosh(\frac{\gamma}{\eta})}, \tag{A.60}$$

where the hyperbolic identity, $\cosh(A+B)-\cosh(A-B) = 2\sinh(A)\sinh(B)$, is used. The denominator on the right of eq.(A.60) is always positive. Also, $\sinh(2r/\eta) > 0$ since $r > 0$. Hence, it follows that the sign of $\sinh(\gamma/\eta)$ is the same as that of $p/l - p/m$. Eq.(A.59) then becomes

$$\sinh(\frac{\gamma}{\eta}) = \frac{(\frac{p}{l} - \frac{p}{m})[(1+\frac{p}{m})(1+\frac{p}{l}) - 4]^{1/2}}{2(\frac{p}{m}\frac{p}{l} - 1)}. \tag{A.61}$$

The value of γ is consequently given by

$$\gamma = \eta \sinh^{-1}[\frac{(\frac{p}{l} - \frac{p}{m})\{(1+\frac{p}{m})(1+\frac{p}{l}) - 4\}^{1/2}}{2(\frac{p}{m}\frac{p}{l} - 1)}]. \tag{A.62}$$

The parameter λ can be evaluated using the expression for p in eq.(A.41).

$$\lambda = \frac{p[\cosh(\frac{r}{\eta}) + \cosh(\frac{\gamma}{\eta})]}{\sinh(\frac{r}{\eta})}. \tag{A.63}$$

$\mathrm{Cosh}(r/\eta)$ and $\sinh(r/\eta)$ are known in terms of p, l and m from eq.(A.54) and $\cosh(\gamma/\eta)$ is obtained from eq.(A.59). Thus

$$\lambda = p\frac{\{[(1+\frac{p}{m})(1+\frac{p}{l})]^{1/2} + \frac{[(1+\frac{p}{m})(1+\frac{p}{l})]^{1/2}(\frac{p}{m}+\frac{p}{l}-2)}{\frac{p^2}{ml}-1}\}}{[(1+\frac{p}{m})(1+\frac{p}{l}) - 4]^{1/2}}. \tag{A.64}$$

Going through the algebra we get

$$\lambda = p\frac{[(1+\frac{p}{m})(1+\frac{p}{l})]^{1/2}[1 + \frac{\frac{p}{m}+\frac{p}{l}-2}{(\frac{p^2}{ml}-1)}]}{[(1+\frac{p}{m})(1+\frac{p}{l}) - 4]^{1/2}} \tag{A.65}$$

which gives

$$\lambda = p\frac{[(1+\frac{p}{m})(1+\frac{p}{l})]^{1/2}[(1+\frac{p}{m})(1+\frac{p}{l}) - 4]^{1/2}}{(\frac{p^2}{ml} - 1)}. \tag{A.66}$$

Consequently,

$$\lambda = p\frac{\{[(1+\frac{p}{m})(1+\frac{p}{l})]^2 - 4(1+\frac{p}{m})(1+\frac{p}{l})\}^{1/2}}{(\frac{p^2}{ml} - 1)}. \tag{A.67}$$

Finally the formula for λ is obtained as

$$\lambda = p\frac{\{[(1+\frac{p}{m})(1+\frac{p}{l}) - 2]^2 - 4\}^{1/2}}{(\frac{p^2}{ml} - 1)}. \tag{A.68}$$

For ϵ, eq.(A.33) is used to determine the sum

$$g_r + g_{-r} = 2\epsilon + \lambda[\frac{1}{1+\exp(\frac{\gamma-r}{\eta})} + \frac{1}{1+\exp(\frac{\gamma+r}{\eta})}]. \quad \text{(A.69)}$$

This is easily shown to be equivalent to

$$g_r + g_{-r} = 2\epsilon + \lambda[1 - \frac{\sinh(\frac{\gamma}{\eta})}{\cosh(\frac{\gamma}{\eta}) + \cosh(\frac{r}{\eta})}]. \quad \text{(A.70)}$$

From eq.(A.63), this reduces to

$$g_r + g_{-r} = 2\epsilon + \lambda - \frac{p\sinh(\frac{\gamma}{\eta})}{\sinh(\frac{r}{\eta})}. \quad \text{(A.71)}$$

Substituting the previously determined values of $\sinh(\gamma/\eta)$ and $\sinh(r/\eta)$ in terms of p, l and m from eqs.(A.61) and (A.54) respectively, quickly yields the desired result for ϵ as

$$\epsilon = \frac{g_r + g_{-r}}{2} - \frac{\lambda}{2} + \frac{p(\frac{p}{l} - \frac{p}{m})}{2(\frac{p}{m}\frac{p}{l} - 1)}. \quad \text{(A.72)}$$

A.3 Johnson S_L Distribution

The *Johnson S_L Distribution* is given by the transformation of the form

$$R = \gamma^* + \eta \ln(G - \epsilon) \quad \text{(A.73)}$$

where again R is a standard normal variable, ϵ is a location parameter, and γ and η are the shape parameters. Solving eq.(A.73) for G we get

$$G = \epsilon + \exp(\frac{R - \gamma^*}{\eta}). \quad \text{(A.74)}$$

Then

$$\begin{aligned} m &= g_{3r} - g_r \\ &= \exp(\frac{3r - \gamma^*}{\eta}) - \exp(\frac{r - \gamma^*}{\eta}) \end{aligned}$$

$$
\begin{aligned}
l &= g_{-r} - g_{-3r} \\
&= \exp(\frac{-r-\gamma^*}{\eta}) - \exp(\frac{-3r-\gamma^*}{\eta}) \qquad \text{(A.75)} \\
p &= g_r - g_{-r} \\
&= \exp(\frac{r-\gamma^*}{\eta}) - \exp(\frac{-r-\gamma^*}{\eta}).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\frac{m}{p} &= \exp(\frac{2r}{\eta}) \qquad \text{(A.76)} \\
\frac{l}{p} &= \exp(\frac{-2r}{\eta})
\end{aligned}
$$

and therefore $\frac{ml}{p^2} = 1$.

Moreover, the value of η is obtained from eq.(A.76) as

$$\eta = \frac{2r}{\ln(\frac{m}{p})}. \qquad \text{(A.77)}$$

From eqs.(A.75) and (A.76),

$$p = \exp(-\frac{\gamma^*}{\eta})[(\frac{m}{p})^{1/2} - (\frac{l}{p})^{1/2}] \qquad \text{(A.78)}$$

which yields γ^* as

$$\gamma^* = \eta \ln[\frac{\frac{m}{p}-1}{p(\frac{m}{p})^{1/2}}]. \qquad \text{(A.79)}$$

Finally,

$$g_r + g_{-r} = 2\epsilon + \exp(-\frac{\gamma^*}{\eta})[\exp(\frac{r}{\eta}) + \exp(-\frac{r}{\eta})]. \qquad \text{(A.80)}$$

Substituting the known expressions for $\exp(-\gamma^*/\eta)$ and $\exp(r/\eta)$ from eqs.(A.75) and (A.76) respectively, we get the desired result for ϵ as

$$\epsilon = \frac{g_r + g_{-r}}{2} - \frac{p}{2}\frac{\frac{m}{p}+1}{\frac{m}{p}-1}. \qquad \text{(A.81)}$$

Appendix B

Connections between g_a, k_a, P_a

According to eq.(48) of chapter 2, the Gaussian random variable R is related to the non-Gaussian random variable G by the transformation

$$R = \gamma + \eta f_i(G; \lambda, \epsilon) \tag{B.1}$$

where $f_i(g; \lambda, \epsilon)$ are single valued monotonically increasing functions, $i = 1, 2, 3$. Let r_0 and g_0 satisfy eq.(B.1). Because of the single-valued monotonically increasing nature of the transformation

$$Pr(G \leq g_0) = Pr(R \leq r_0). \tag{B.2}$$

From a relative frequency point of view

$$Pr(R \leq r_0) \approx \frac{No.\ of\ observations\ less\ than\ equal\ to\ r_0}{n} \tag{B.3}$$

where n is the total number of observations.

Now assume that n observations of the random variable G are obtained. Ordering the observations of G from the smallest to the largest, denote the k^{th} ordered observation by g^k. Then k equals the number of observations less than or equal to g^k. Corresponding to the ordered observations of the random variable G are ordered realizations of the random variable R (See eq.(2.18)). Denote the k^{th} ordered realization of R by r^k. Because the transformation in eq.(B.1) is single valued and monotonically increasing, it follows that

$$Pr(R \leq r^k) = Pr(G \leq g^k) \approx \frac{k}{n} \tag{B.4}$$

Introducing the "continuity correction", as was done with the q-q and p-p plots in chapter 1, $Pr(R \leq r^k)$ is approximated by

$$Pr(R \leq r^k) \approx \frac{k - \frac{1}{2}}{n}. \tag{B.5}$$

By definition,

$$P_a = Pr(R \leq a) \tag{B.6}$$

where $a = 3r, r, -r, -3r$. We define the integer k_a such that

$$P_a \approx \frac{k_a - \frac{1}{2}}{n} \tag{B.7}$$

where

$$k_a = [nP_a + \frac{1}{2}], \tag{B.8}$$

[.] denotes the closest integer and $a = 3r, r, -r, -3r$.

Equations (B.6) and (B.7) imply that a approximately equals the k_a^{th} ordered sample of R. Note that the k_a^{th} ordered sample of G is g^{k_a}, where r^{k_a} and g^{k_a} satisfy eq.(B.1). From equations (B.4) and (B.5)

$$Pr(G \leq g^{k_a}) = Pr(R \leq r^{k_a}) \approx Pr(R \leq a) \approx \frac{k_a - \frac{1}{2}}{n}. \tag{B.9}$$

It follows, given P_a, that one can determine k_a, and given k_a, one can determine g_a by the simple relation

$$g_a = g^{k_a}. \tag{B.10}$$

Bibliography

[1] Ozturk A., "A New Method for Distribution Identification." Submitted for publication to JASA.

[2] Ozturk A., "A General Algorithm for Univariate and Multivariate Goodnes of Fit Tests Based on Graphical Representation," Communication in Statistics, 1991.

[3] Ozturk A. and Dudewicz E.J., "A New Statistical Goodness of Fit Test Based on Graphical Representation," The Biometrical Journal, 1991.

[4] Slifker J. and Shapiro S., "Johnson System: Selection and Parameter Estimation," Technometrics, vol. 22, pp. 239-246, May 1980.

[5] Aly E.E.A.A and Ozturk A., "Hodges Lehman Quantile-Quantile Plots," Computational Statistics and Data Analysis, vol. 6, pp. 99-108.

[6] Wilk M.B. and Gnanadesikan R., "Probability Plotting Methods for the Analysis of Data," Biometrika, vol. 55, pp. 1-8, 1968.

[7] Small N.J.H., "Plotting Squared Radii", Biometrika, vol. 65, pp. 657-658, 1978.

[8] Filliben J.J., "The Probability Plot Correlation Coefficient Test for Normality," Technometrics, vol. 17, pp. 111-112, Feb. 1975.

[9] Massey F.J., "The Kolmogorov-Smironov Test for Goodness of Fit," Journal of the American Statistical Association, vol. 46, pp. 68-78, 1951.

[10] Slaski L.K., "An Introduction to Dr. Ozturk's Algorithm for PDF Approximation," Rome Laboratory Internal Report, USAF, March 1993.

[11] Rangaswamy M., "Spherically Invariant Random Processes for Radar Clutter Modelling, Simulation and Distribution Identification", Ph.D. Dissertation, Department of Electrical Engineering, Syracuse Univerisity, December 1992.

[12] Lindgren B.W., McElrath G.W. and Berry D.A., Introduction to Probability and Statistics. New York: Macmillan Publishing Co. Inc., 4th ed. 1978.

[13] Lindgren B.W. and McElrath G.W., Introduction to Probability and Statistics. New York: Macmillan Publishing Co. Inc., 3rd ed. 1969.

[14] Johnson R.A. and Wichern D.W., Applied Multivariate Statistical Analysis, New Jersey: Prentice Hall Inc., 3rd ed., 1992.

[15] Johnson N. and Kotz S., Distributions in Statistics: Continuous Univariate Distributions, New York: John Wiley and Sons Inc., 1970.

[16] Johnson N. and Kotz S., Distributions in Statistics: Continuous Multivariate Distributions, New York: John Wiley and Sons Inc., 1976.

[17] Papoulis A. Probability, Random Variables and Stochastic Processes, New York: McGraw Hill, 1984.

[18] Bratley P., Fox B. and Schrage L., A Guide to Simulation, New York: Springer Verlag, 1987.

[19] Johnson M.E., Multivariate Statistical Simulation, New York: John Wiley and Sons Inc., 1987.

Rome Laboratory

Customer Satisfaction Survey

RL-TR-_____

Please complete this survey, and mail to RL/IMPS, 26 Electronic Pky, Griffiss AFB NY 13441-4514. Your assessment and feedback regarding this technical report will allow Rome Laboratory to have a vehicle to continuously improve our methods of research, publication, and customer satisfaction. Your assistance is greatly appreciated.
Thank You

_____

_____

Organization Name:_____(Optional)

Organization POC: _____(Optional)

Address:_____

1. On a scale of 1 to 5 how would you rate the technology developed under this research?

5-Extremely Useful 1-Not Useful/Wasteful

Rating_____

Please use the space below to comment on your rating. Please suggest improvements. Use the back of this sheet if necessary.

2. Do any specific areas of the report stand out as exceptional?

Yes___ No_____

If yes, please identify the area(s), and comment on what aspects make them "stand out."

3. Do any specific areas of the report stand out as inferior?

Yes___ No___

If yes, please identify the area(s), and comment on what aspects make them "stand out."

4. Please utilize the space below to comment on any other aspects of the report. Comments on both technical content and reporting format are desired.

*U.S. GOVERNMENT PRINTING OFFICE: 1995-610-126-20051

MISSION

OF

ROME LABORATORY

Mission. The mission of Rome Laboratory is to advance the science and technologies of command, control, communications and intelligence and to transition them into systems to meet customer needs. To achieve this, Rome Lab:

a. Conducts vigorous research, development and test programs in all applicable technologies;

b. Transitions technology to current and future systems to improve operational capability, readiness, and supportability;

c. Provides a full range of technical support to Air Force Materiel Command product centers and other Air Force organizations;

d. Promotes transfer of technology to the private sector;

e. Maintains leading edge technological expertise in the areas of surveillance, communications, command and control, intelligence, reliability science, electro-magnetic technology, photonics, signal processing, and computational science.

The thrust areas of technical competence include: Surveillance, Communications, Command and Control, Intelligence, Signal Processing, Computer Science and Technology, Electromagnetic Technology, Photonics and Reliability Sciences.