Multivariate Nichemetrics

by

Ruey-Pyng Lu

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

APPROVED:

Eric P. Smith, Chairman

Irving J. Good

Golde I. Holtzman

Marvin Lentner

Robert S. Schulman

September, 1986

Blacksburg, Virginia

Multivariate Nichemetrics

by

Ruey-Pyng Lu

Eric P. Smith, Chairman

Statistics

(ABSTRACT)

In the study of ecological community structure, the multivariate niche model has always been the assumed structural model. This model is closely connected to the multivariate two-sample problem. Important to the understanding of species interactions in a community is the measurement of the degree to which the niches of two species overlap, or to measure the similarity between the resource use distributions of the species. Discriminant analysis is the tool used most often to analyze the similarity. In this study, we discuss the most commonly used similarity measures, and develop measures that are less dependent on the assumptions of the usual discriminant analysis. Specifically measures are derived assuming normal distributions with heterogeneous variance-covariance matrices are derived.

The problem of estimating the measures and their precision and accuracy is investigated. Two methods, the jackknife and the bootstrap, are described for estimating the bias and variance of an estimated measure. The performance of these methods was evaluated using simulation. When the number of variables involved in the model is large, the estimates of these measures may be severely biased, and the bias is consistently negative. By collecting larger samples the bias can be reasonably adjusted. Two potentially important factors affecting results are the disparity in the means and the heterogeneity of the variance-covariance matices. It is shown that when the mean separation is small, the heterogeneity of the covariance matrices has a moderate effect on the bias, but the effect is diminished when the mean separation becomes larger. The variance of the similarity estimates is also related to the value of the measure and is a quadratic function of the similarity. The logarithmic transformation of the similarity is seen to linearize the variance of the similarity estimate.

The jackknife method gives good adjustment of the bias of the estimated measures. Generally, the bootstrap method performs worse than the jackknife method. In some cases, especially when there are many redundant variables neither method gives reliable results.

Acknowledgements

I would like to express my sincere appreciation to my major advisor, Dr. E. P. Smith, for his ideas, guidance, and encouragement during this study. I also thank Dr. G. I. Holtzman for his assistance and suggestions in writing this dissertation. My advisory committee was exceedingly helpful and I thank Dr. I. J. Good, Dr. G. I. Holtzman, Dr. R. S. Schulman and Dr. B. O. Skarpness for their advice and support. I am grateful to Dr. M. Lentner for his willingness on short notice to participate in the oral examination.

I am grateful to for the financial support during my graduate study at this department.

A very special note of thanks is also expressed to my parents and my wife for their love, encouragement and support which has made it possible for me to accomplish this study.

Table of Contents

List of Illustrations

Chapter I

INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

Multivariate analysis is popular among ecologists in a variety of ecological investigations. It has been applied to geographical ecology, social behavior, niche structure, organism morphology and physiology. In such studies, many observable variables are involved, and multivariate methods are utilized to detect and describe subtle patterns among these variables, and to facilitate interpretation. The methods are not without pitfalls, however, so researchers must carefully study the results of an analysis (Williams [47]). Recently, much interest has focused on the multivariate niche model, or the measurement of overlap between the niches of two species, and on the problems inherent in some of the measures

of overlaps and similarity (Carnes and Slade [2], Van Horne and Ford [46], Porter and Dueser [37], Dueser and Shugart [6], Green [13,14]).

The ecological community can be thought of as a large n-dimensional hyperspace, within which each species population evolves to occupy its own region of the available space. The position of the species and its response to factors of the community hyperspace defines its niche. Each species thus occupies a vaguely outlined, diffuse volume that differs from but perhaps overlaps with, those of other species in the community. Hutchinson [19] gave a concrete meaning to the multivariate niche model: A niche is an n-dimensional hypervolume, expressed as the range and combination of environmental factors that permit a species to persist in a community. The value of Hutchinson's approach is that it directed ecologists to test and develop the hypotheses about structure of communities.

It has often been suggested that the key to understanding species interactions in a community is to measure the degree to which niches of two species overlap, rather than to try to describe the niches of all species (Ricklefs [39]). Such overlap is usually measured in terms of utilization of resources such as food and habitat, the important and easy to measure factors. Niche overlap is thus described as overlap of utilization between two adjacent species on a resource gradient. Typically, "niche overlap" is the degree of similarity between the niches of two species, and the "measure of overlap" is the similarity between the resource use distributions of the species. The multivariate niche model together with advances in statistical computing have allowed modellers to analyze complex data sets and

extract relevant information about community structure. For example, Porter and Dueser [37] used the multivariate niche model to test the hypothesis of Pianka [36] about the degree of niche overlap and the intensity of competition. As pointed out by Porter and Dueser, it is important to have reliable estimates of niche overlap if the tests are to be valid.

The subject of this research is the assessment of several methods of estimating multivariate niche overlap and the development of methods for estimating and assessing overlap when assumptions are violated. The focus will be on the methods associated with discriminant analysis [Harner and Whitmore, 16, Porter and Dueser, 37]. Some of the methodology will also be useful for other niche metrics such as niche breadth. In addition, the results are applicable to studies in physical anthropology, pattern recognition, and geology.

1.2 LITERATURE REVIEW

There are several methods for estimating multivariate niche overlap. The most commonly used methods resulted from the works of Shugart and Patten [42] and Green [13,14]. Shugart and Patten [44] developed a number of measures based on generalized distances and discriminant functions, while Green [13,14] suggested using the percent overlap of the 50% probability ellipses. The methods of Green were criticized by Dueser and Shugart [7] because the ellipses were computed only based on the assumption of equal covariance matrices, and because the methods are sample size dependent. Dueser and Shugart [6,7] extended

these ideas and estimated overlap as the area that overlaps between concentration ellipsoids, relative to the total area of the ellipsoids. The concentration ellipsoids are probability ellipses for observations, not means, and hence are relatively independent of sample sizes. However, Carnes and Slade [2] criticize the above measure as being dependent on the assumption of multivariate normality. In addition, the evaluation of the measure as suggested by Dueser and Shugart [7] implies uniform rather than bell-shaped usage density within the ellipsoid. The computation of overlap as planar area does not weight by the probability of usage: the probability density is assumed uniform over the ellipsoid. Carnes and Slade [2] note that departures from uniform usage could result in misleading estimates of niche overlap. Moreover, the methods of computing the confidence intervals for true parameters are only approximations.

Another approach is to estimate multivariate overlap based on the extensions of the univariate measures of overlap (MacArthur and Levins [25], Hurlbert [18]). Harner and Whitmore [16] gave the formulas for the estimation of overlap using the MacArthur-Levins [25] asymmetric measure of overlap

$$\alpha_{ij} = \frac{\int f_i(\mathbf{x}) f_j(\mathbf{x}) d\mathbf{x}}{\int f_i^{\,2}(\mathbf{x}) d\mathbf{x}}$$

where f_i and f_j are multivariate normal densities $N_p(\mathbf{u}_i, \Sigma)$ and $N_p(\mathbf{u}_j, \Sigma)$ respectively, describing the usage of a habitat by species i and j, and where the integral is p-fold. The measure is asymmetric, because α_{ji} has the denominator $\int f_j^{\,2}(\mathbf{x}) d\mathbf{x}$. Another common measure is the proportional similarity measure

$$PS = \int \min\ [f_1(x), f_2(x)\ ]dx.$$

Based on the assumptions of multivariate normality, equal covariance matrices and independence of observations, Harner and Whitmore showed that the multivariate MacArthur-Levins index is given by

$$\alpha = \exp[\ -\frac{1}{4}(\mathbf{u}_2 - \mathbf{u}_1)'\Sigma^{-1}(\mathbf{u}_2 - \mathbf{u}_1)]$$

which is a function of the Mahalanobis generalized distance. They also showed that the proportional similarity measure can be obtained by computing the univariate measure after projecting the multivariate densities onto the discriminant axis. The approach of Harner and Whitmore [16] is similar to that of Dueser and Shugart [7]. If the proportional similarity measure is used, the measure can be interpreted as the volume in common between the probability distributions. This volume is what Dueser and Shugart [7] are measuring approximately in a uniform sense. The proportional similarity measure is computed using the normality of the data and in most cases is more accurate.

The measures were computed by Harner and Whitmore only under the assumption of equal covariance matrices. Maurer [32] extended the results of Harner and Whitmore [16] to the univariate case with unequal variance, and provided confidence intervals for the univariate MacArthur-Levins index. Maurer suggested using the procedures when the multivariate data may be reduced to a single dimension using principal components or two group discriminant analysis with equal covariances. But it would be better to extend the

methods to the more complex situation with unequal covariance matrices, so that they would be applicable to real ecological data (Green [13], Dueser and Shugart [7]). Green [15] commented on the assumptions of multivariate analysis in ecological studies. Nonnormality does not appear to be a serious violation for the multivariate case, as it is not for the univariate case. Heterogeneity of variance-covariance matrices is a more serious problem. Significance arising from differences among covariances can be given a meaningful biological interpretation just as can significance arising from differences among means. Also, as Maurer indicated, confidence intervals based on the noncentral t distribution may require moderate sample sizes.

When the covariance matrices differ, the formulas given in Harner and Whitmore [16] and the formulas for Morisita's measure (Zaret and Smith [48]) ought to be modified. Furthermore, the two MacArthur-Levins measures α_{ij} and α_{ji} can be made extremely dissimilar by making the covariances matrices different and there are cases where the interpretations given by α have no resemblance to the interpretations given by α_{ij} and α_{ji}. Note also that Pianka's measure, $\alpha^* = \sqrt{\alpha_{ij}\,\alpha_{ji}}$, and Matusita's measure can be in considerable error, if the assumption of equal covariance matrices is not valid. In general, as with other measures, the errors will of course depend on the dimensionality of the problem, the disparity of the covariance matrices, and the magnitude of the overlap.

In community structure studies, the distribution of the variables in a population is naturally described by a multivariate probability distribution function $f(x)$. Assessing the similarity between two species populations is a problem of

comparing two probability functions $f_1(x)$ and $f_2(x)$. The general notion is to measure the overlap. The greater the overlap, the more similar are two populations. Thus, the niche overlap problem is closely connected with the two-sample multivariate problem. Ito and Schull [21], and Carter, Khatri and Srivastava [3] investigated the consequences of unequal variances in testing the hypothesis of equal means. For equal sample sizes, moderate violations are of little consequence, and as the sample sizes increase the test remains well behaved. For unequal sample sizes, on the other hand, moderate violations can seriously distort the level of significance and the power of the test.

In two-sample discriminant analysis, Marks and Dunn [27] compared the performance of three discriminant functions: the quadratic, the best linear and Fisher's linear discriminant function. They found that when the disparity of variance is large, the quadratic function is asymptotically better than the Fisher function. For small samples the quadratic performs worse than the Fisher as the disparity of variance is small, and this tendency increases with the number of parameters. The performance of the best linear function is in between that of the other two.

1.3 OBJECTIVES

The goal of this research is to compare and develop measures of multivariate niche overlap that are less dependent on the assumptions of the usual

discriminant function analysis and to investigate procedures for estimating the sampling variability of these measures.

The objectives are the following:

1. Investigate the accuracy and precision of currently used measures of niche overlap, and study the effects of sample sizes, of disparity in the metrics, of dimensionality and of unequal variances and covariances.

2. Extend the measures of overlap to provide more accurate estimates when the model assumptions are not valid.

3. Investigate the applicability of the jackknife and bootstrap methods to confidence interval estimation of overlap measures.

1.4 EDITORIAL REMARKS

In Chapter 2, Mahalanobis' generalized distance measure, Matusita's distance and affinity measure, Morisista's affinity measure, the MacArthur-Levins measure of overlap, Pianka's measure of overlap and the proportional similarity measure are defined, and some of their properties are discussed.

In Chapter 3, the explicit forms of the measures introduced in Chapter 2 are derived under the assumption of multivariate normal distributions with the "heterogeneous" variance-covariance matrices. It is shown that these forms reduce to those derived by Harner and Whitmore when the variance-covariance

matrices are homogeneous. Then, the connection of the affinity and the multivariate one-sample, two-sample problems are investigated. This relationship showing that a distance measure is the fundamental tool of multivariate analysis.

In Chapter 4, the relationships between the correlation coefficient and the affinity measures are established in the bivariate normal model. The dependency of the variables is identified through the application of distance measures. This comparison exposes the advantages and the disadvantages of distance measures for testing independence in the bivariate density. Finally, the relationships between the correlation coefficient and divergence measures are studied. It is shown that the divergence measures describe the independence between the variables in terms of the affinity measures.

We are concerned with a theoretical investigation of the estimators of various measures. The mathematical difficulties in deriving the properties of the estimators are formidable, and consequently the properties are evaluated mainly by Monte Carlo methods. In fact, it would appear there is no article of a theoretical nature on the estimation of measures for the most general case of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ all unknown, the case which is the most likely to occur in practice. In Chapter 5, the two-sample jackknife method and the two-sample bootstrap method are introduced. In Chapter 6, the simulation results are presented, comparing the effects of sample sizes, of the dimensionality (the number of variables), and of the disparity of variance-covariance matrices on the performance of the procedures.

In conclusion, several suggestions for applications the currently used overlap measures are presented. To get a reliable estimate, the jackknife method is appropriate.

Chapter II

DISTANCE, AFFINITY AND MEASURES OF SIMILARITY

In the context of multivariate statistics, the Mahalanobis generalized distance is predominant. In statistical ecology, on the other hand, Matusita's distance, Morisita's affinity measure, the MacArthur-Levins measure of overlap, Pianka's measure of overlap, and the proportional similarity measure are more commonly used to evaluate the community similarity between populations. In this chapter, the properties of the Mahalanobis generalized distance are reviewed and similar results are derived for the measures favored by ecologists.

2.1 MAHALANOBIS GENERALIZED DISTANCE MEASURE

Let the p -component random vector $\mathbf{X}_i$ follow the multivariate normal distribution $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with mean $\boldsymbol{\mu}_i$ and symmetric positive definite covariance matrix, i = 1,2.

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two independent samples of size N_1 and N_2 of $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively. Mahalanobis [26] gave the following definition.

DEFINITION 2.1.: A measure of the generalized distance between the two populations (i.e. distributions) is $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. The **estimated Mahalanobis generalized distance** is $D^2 = \mathbf{d}'\mathbf{S}^{-1}\mathbf{d}$ where $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, $\mathbf{S} = \dfrac{(n_1\mathbf{S}_1 + n_2\mathbf{S}_2)}{n}$, $n = n_1 + n_2$ and where $\bar{\mathbf{x}}_i$ is the p -dimensional column vector of sample means and $\mathbf{S}_i$ is a $p \times p$ sample covariance matrix with $n_i = N_i - 1$ degree of freedom, $i = 1, 2$.

The statistic D^2 is proportional to the two-sample version of Hotelling's T^2. It is important in discriminant analysis, cluster analysis, profile analysis, and other multivariate methods. In ecology, D^2 is closely related to the analysis of affinity, similarity, and niche overlap. We now introduce some measures of similarity that are popular in ecological work.

2.2 MATUSITA'S DISTANCE AND AFFINITY MEASURES

Let F_1 and F_2 be two distribution functions admitting probability densities f_1 and f_2, respectively, with respect to some probability measure m.

DEFINITION 2.2.: The **Hellinger distance** r_p (Hellinger [17]) between F_1 and F_2 is defined, for $p \geq 1$, by

$$r_p = \|F_1 - F_2\| = \left| \int [f_1^{1/p}(x) - f_2^{1/p}(x)]^p dm(x) \right|^{1/p}. \qquad [2.1]$$

Matusita's distance (Matusita [28]) measure is obtained when $p = 2$, that is:

$$d_2(F_1, F_2) = \{\int [f_1{}^{1/2}(x) - f_2{}^{1/2}(x)]^2 dm(x)\}^{1/2} \qquad [2.2]$$

and **Matusita's affinity** is defined as:

$$\rho(F_1, F_2) = \int [f_1(x) f_2(x)]^{1/2} dm(x). \qquad [2.3]$$

Thus, $d_2{}^2(F_1, F_2) = 2[1 - \rho(F_1, F_2)]$. Here $\rho(F_1, F_2)$ is the affinity of distributions and represents the "closeness" of the distributions. This affinity measure was investigated previously by Bhattacharyya [1].

As is easily seen, $\rho(F_1, F_2)$ has the following properties:

(i) $0 \leq \rho(F_1, F_2) \leq 1$.

(ii) $\rho(F_1, F_2) = 1$ if and only if $F_1 = F_2$.

(iii) For a sequence of distributions $\{F_n\}$, $\rho(F_n, F_0) \to 1$, i.e., for any measurable set E, $F_n(E) \to F_0(E)$ uniformly in E, where $F_n(E)$ and $F_0(E)$ denote the probabilities of E according to F_n and F_0 respectively.

The quantity ρ quantifies the closeness between distributions in the sense that the larger $\rho(F_1, F_2)$ is, the closer F_1 and F_2 lie. Matusita [28,29] discussed the distance measure and affinity measure in the context of the one sample and two-sample goodness-of-fit problem. Formulation of the statistical decision problem in terms of distance functions in general terminology is presented by Matusita [29].

2.3 MORISITA'S AFFINITY MEASURE

Suppose we have two independent populations governed by distribution functions F_1 and F_2. Let m be a measure with respect to which F_1 and F_2 admit square integrable densities f_1 and f_2 , respectively. The following measure of affinity was used implicitly by Morisita [34].

DEFINITION 2.3.: Morisita's affinity measure $\lambda(F_1, F_2)$ is defined by:

$$\lambda(F_1, F_2) = \frac{2\int f_1(x) f_2(x) dm(x)}{\int f_1{}^2(x) dm(x) + \int f_2{}^2(x) dm(x)} \qquad [2.4]$$

whenever the integrals are defined.

It can be shown, $\lambda(F_1, F_2)$ has the following properties:

(i) $0 \leq \lambda(F_1, F_2) \leq 1$.

(ii) $\lambda(F_1, F_2) = 1$ if and only if $F_1 = F_2$.

Remark: The quantity $D_i = \int f_i^2(x)dm(x)$, i = 1,2, which appears as a part of $\lambda(F_1, F_2)$, is well known to ecologists as a measure of "clumping", and $1 - D_i$, $i = 1,2$, is often used as a measure of the "diversity" of a population. See Van Belle and Ahmad [45]. In this context, $\lambda(F_1, F_2)$ measures the similarity of two populations in units of the clumping of each of the two populations.

2.4 MACARTHUR-LEVINS MEASURE OF OVERLAP

Assume that the random vector **x** has probability density functions $f_1(x)$ and $f_2(x)$ for species 1 and 2, respectively. The densities $f_i(x)$ are often called "usage distributions" since they specify the probabilities of the species using the available resources. Thus, the probability density is also the species biological density with respect to the resources. MacArthur and Levins [25] suggested the following definition.

DEFINITION 2.4.: The **MacArthur-Levins overlap** measure is defined by

$$\alpha_{ij} = \frac{\int f_i(x) f_j(x)dx}{\int f_i^2(x)dx} \qquad [2.5]$$

i,j = 1,2, whenever the integrals are defined.

The following properties obtain:

(i) $0 \leq \alpha_{ij} < \infty$.

(ii) If $f_i = f_j$ then $\alpha_{ij} = 1$. The converse is not true.

Remark: In general, $\alpha_{12} \neq \alpha_{21}$. A sufficient condition for equality is that $f_i(x)$ and $f_j(x)$ differ only by a location parameter. Apparently, α_{ij} measures the relative probability of the simultaneous presence of species i and species j compared to the species density of species i. In an intuitive sense, the numerator is the probabilities of all possible events when the two species use resources "simultaneously," while the denominator represents what the probability would be if species j were equivalent to species i in resource use.

2.5 PIANKA'S MEASURE OF OVERLAP

Following the idea of the MacArthur-Levins measure of overlap, Pianka [36] proposed the following measure.

DEFINITION 2.5. Pianka's overlap measure is defined by

$$\alpha = \sqrt{\alpha_{ij}\,\alpha_{ji}} = \frac{\int f_i(x) f_j(x) dx}{[\int f_i{}^2(x)dx]^{1/2}[\int f_j{}^2(x)dx]^{1/2}}. \qquad [2.6]$$

The properties of α are:

(i) $0 \leq \alpha \leq 1$ and

(ii) $\alpha = 1$ if and only if $f_i = f_j$.

Note that α is essentially a correlation coefficient, and intuitively has a similar interpretation to that given for λ , except that the denominator represents a "geometric average" of the cases where both species are equivalent to species 1 and both are equivalent to species 2.

2.6 PROPORTIONAL SIMILARITY MEASURE

The idea of the geometric representation of a probability distribution leads to analytic consideration of similarity. Thus we have

DEFINITION 2.6.: The **proportional similarity measure** PS is defined by

$$\text{PS} = \int \min\ [f_1(x), f_2(x)\ ]dx \qquad [2.7]$$

whenever the integral is defined.

Note that $0 \leq PS \leq 1$ and that PS is symmetric for any pair of density functions. The PS measure has another formulation, given by

$$\text{PS} = 1 - \frac{1}{2}\int \mid f_1(x) - f_2(x) \mid dx.$$

The equivalence of the formulas is verified by first noting that (see also Smith [42])

when $f_1(x) < f_2(x)$,

$$\min\ [f_1(x), f_2(x)\ ] = f_1(x) = f_2(x) - \ |\ f_1(x) - f_2(x)\ |$$

and when $f_2(x) \leq f_1(x)$,

$$\min\ [f_1(x), f_2(x)\ ] = f_2(x) = f_1(x) - \ |\ f_1(x) - f_2(x)\ |\ .$$

The equivalence follows by writing

$$\int 2 \min\ [f_1(x), f_2(x)\ ]dx = \int_{[x|f_1(x) < f_2(x)]} f_1(x)dx + \int_{[x|f_2(x) \leq f_1(x)]} f_2(x)dx$$

$$+ \int_{[x|f_2(x) \leq f_1(x)]} (f_1(x) - |f_1(x) - f_2(x)|)dx$$

$$+ \int_{[x|f_1(x) < f_2(x)]} (f_2(x) - |f_1(x) - f_2(x)|)dx$$

$$= 2 - \int |\ f_1(x) - f_2(x)\ |\ dx.$$

The PS measure can be interpreted as the probability that two individuals from different species simultaneously attempt to use the same resources, relative to the probability that an individual from the more likely species tries to use the same resources.

Chapter III

RESULTS UNDER MULTIVARIATE NORMAL DISTRIBUTIONS

The distance measures, the affinity measures, and the measures of overlap are generally difficult to formulate explicitly. However, for the important special case considered in this sequel, we will assume that $f_i(x)$ and $f_j(x)$ are multivariate normal densities. Thus the integrals of f_i and f_j can be evaluated. Since two multivariate normal populations are involved, we will first derive the explicit forms of the measures assuming equal variance-covariance matrices, then extend the results to the case with heterogeneous variance-covariance matrices.

Let F_1 and F_2 be nonsingular p-dimensional normal distributions with density functions

$$f_1(\mathbf{x}) = \frac{1}{|\Sigma_1|^{1/2}(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma_1^{-1}(\mathbf{x}-\mu_1)\right]$$

and

$$f_2(\mathbf{x}) = \frac{1}{|\Sigma_2|^{1/2}(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_2)'\Sigma_2^{-1}(\mathbf{x}-\mu_2)\right]$$

respectively.

3.1 EQUAL VARIANCE-COVARIANCE MATRICES

When $\Sigma_1 = \Sigma_2 = \Sigma$, we have the following results:

I. Matusita's affinity measure is (see Matusita [29])

$$\rho = \int [f_i(\mathbf{x}) f_j(\mathbf{x})]^{1/2} d\mathbf{x}$$

$$= \exp\left[-\frac{1}{8}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\right]. \qquad [3.1]$$

II. Morisita's affinity measure is (see Morisita [34])

$$\lambda = \frac{2\int f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x}}{\int f_1^{\,2}(\mathbf{x})d\mathbf{x} + \int f_2^{\,2}(\mathbf{x})d\mathbf{x}}$$

$$= \exp\left[-\frac{1}{4}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\right]. \qquad [3.2]$$

III. The MacArthur-Levins measure of overlap is (see MacArthur & Levin [25])

$$\alpha_{12} = \frac{\int f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x}}{\int f_1{}^2(\mathbf{x}) d\mathbf{x}} = \alpha_{21} = \frac{\int f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x}}{\int f_2{}^2(\mathbf{x}) d\mathbf{x}}$$

$$= \exp\ [-\frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]. \qquad [3.3]$$

IV. Pianka's measure of overlap is (see Pianka [36])

$$\alpha = \frac{\int f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x}}{[\int f_1{}^2(\mathbf{x}) d\mathbf{x}]^{1/2} [\int f_2{}^2(\mathbf{x}) d\mathbf{x}]^{1/2}}$$

$$= \exp\ [-\frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]. \qquad [3.4]$$

V. The proportional similarity measure cannot be evaluated explicitly for multivariate normal densities. We discuss this measure only for the univariate normal densities. Then, Harner and Whitmore [16] showed that

$$PS = \int_c^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} \exp[-\frac{1}{2\sigma^2}(x - \mu_1)^2] dx$$

$$+ \int_{-\infty}^{c} \frac{1}{(2\pi)^{1/2}\sigma} \exp[-\frac{1}{2\sigma^2}(x - \mu_2)^2] dx \qquad [3.5]$$

$$= 2\int_c^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu_1)^2\right]dx \qquad \text{for} \quad \mu_1 < \mu_2$$

and defined analogously for $\mu_2 < \mu_1$ and is 1 if $\mu_1 = \mu_2$,

where c is determined such that $f_1(c) = f_2(c)$. The multivariate normal version is similar to the above expression except that the integral is p -fold with the region of integration given by the $p - 1$ dimensional hyperplane on which the densities are equal. Because the PS is different from other measures, we will not consider it any further in this study.

3.2 HETEROGENEOUS VARIANCE-COVARIANCE MATRICES

For the general case $\Sigma_1 \neq \Sigma_2$, we have

THEOREM 3.1. Matusita's measure

$$\rho^* = \int [f_i(\mathbf{x}) f_j(\mathbf{x})]^{1/2} d\mathbf{x}$$

$$= \frac{|\Sigma_1\Sigma_2|^{1/4}}{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|^{1/2}} \exp\left[-\frac{1}{4}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)\right]. \qquad [3.6]$$

Proof: By definition, we have

$$f_1^{1/2}(\mathbf{x}) f_2^{1/2}(\mathbf{x}) =$$

$$\frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/4}|\Sigma_2|^{1/4}} \exp\{-\frac{1}{2}[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)]\}.$$

Completing the square, the above expression may be written as

$$(2\pi)^{-p/2}|\Sigma_1|^{-1/4}|\Sigma_2|^{-1/4} \exp\{-\frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) + c]\}$$

where by matching the coefficients,

$$\mathbf{A} = \frac{1}{2}\Sigma_1^{-1} + \frac{1}{2}\Sigma_2^{-1}$$

$$\mathbf{A}\boldsymbol{\mu} = \frac{1}{2}\Sigma_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\Sigma_2^{-1}\boldsymbol{\mu}_2$$

and

$$\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + c = \frac{1}{2}\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2 .$$

Moreover,

$$\int f_1^{1/2}(\mathbf{x}) f_2^{1/2}(\mathbf{x}) d\mathbf{x} = |\mathbf{A}|^{-1/2}|\Sigma_1|^{-1/4}|\Sigma_2|^{-1/4} \exp\{\frac{-c}{2}\}.$$

Thus

$$c = \frac{1}{4}(\mu_1 - \mu_2)' \left[\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2\right]^{-1}(\mu_1 - \mu_2)$$

because $\Sigma_1 A \Sigma_2 = \frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2$.

The determinant of A is $|A| = \dfrac{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|}{|\Sigma_1\Sigma_2|}$.

Hence,

$$\int f_1^{1/2}(\mathbf{x}) f_2^{1/2}(\mathbf{x}) d\mathbf{x} = \frac{|A|^{-1/2}}{|\Sigma_1|^{1/4}|\Sigma_2|^{1/4}} \exp\{-\frac{1}{8}(\mu_1 - \mu_2)' \left[\frac{1}{2}(\Sigma_1 + \Sigma_2)\right]^{-1}(\mu_1 - \mu_2)\}$$

$$= \frac{|\Sigma_1\Sigma_2|^{1/4}}{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|^{1/2}} \exp\left[-\frac{1}{4}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)\right].$$

COROLLARY 3.1: When $\Sigma_1 = \Sigma_2 = \Sigma$, we have the equal variance-covariance case,

$$\rho^* = \exp\left[-\frac{1}{8}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\right]. \qquad [3.7]$$

This is the same result as in the previous section.

COROLLARY 3.2: When $\mu_1 = \mu_2$, we have

$$\rho^* = \frac{|\Sigma_1\Sigma_2|^{1/4}}{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|^{1/2}}, \qquad [3.8]$$

by an argument similar to that by which Matusita's affinity measure was obtained. Morisita's affinity measure can be obtained.

THEOREM 3.2.

$$\lambda^* = \frac{2\int f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x}}{\int f_1{}^2(\mathbf{x})d\mathbf{x} + \int f_2{}^2(\mathbf{x})d\mathbf{x}}$$

$$= \frac{2|\Sigma_1\Sigma_2|^{1/2}}{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|^{1/2}(|\Sigma_1|^{1/2} + |\Sigma_2|^{1/2})}$$

$$\cdot \exp[-\frac{1}{2}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)]. \qquad [3.9]$$

Proof:

$$f_1(\mathbf{x})f_2(\mathbf{x}) =$$

$$\frac{1}{(2\pi)^p|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}} \exp\{-\frac{1}{2}[(\mathbf{x} - \mu_1)'\Sigma_1^{-1}(\mathbf{x} - \mu_1) + (\mathbf{x} - \mu_2)'\Sigma_2^{-1}(\mathbf{x} - \mu_2)]\}.$$

Completing the square, the above expression may be written as

$$(2\pi)^{-p}|\Sigma_1|^{-1/2}|\Sigma_2|^{-1/2} \exp\{-\frac{1}{2}[(\mathbf{x} - \mu)'A(\mathbf{x} - \mu) + c]\}$$

where by matching the coefficients,

$$\mathbf{A} = \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1} \quad , \quad \mathbf{A}\boldsymbol{\mu} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

and

$$\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + c = \boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 .$$

Thus

$$c = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\left[\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) .$$

Since

$$\boldsymbol{\Sigma}_1\mathbf{A}\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma},$$

the determinant of $\mathbf{A}$ is $|\mathbf{A}| = \dfrac{|(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)|}{|\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2|}$.

Multiplying $f_1{}^2(\mathbf{x})$, $f_2{}^2(\mathbf{x})$ and $f_1(\mathbf{x})f_2(\mathbf{x})$ by $(2\pi)^{p/2}$, we have

$$(2\pi)^{p/2}\int f_1{}^2(\mathbf{x})d\mathbf{x} = 2^{-p/2}\,|\,\boldsymbol{\Sigma}_1\,|^{-1/2} \quad , \quad (2\pi)^{p/2}\int f_2{}^2(\mathbf{x})d\mathbf{x} = 2^{-p/2}\,|\,\boldsymbol{\Sigma}_2\,|^{-1/2}$$

and

$$(2\pi)^{p/2}\int f_1(\mathbf{x})f_2(\mathbf{x})d\mathbf{x} = \frac{|\mathbf{A}|^{-1/2}}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}}\exp\left[-\frac{1}{2}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right].$$

Hence

$$\lambda^* = \frac{2|\mathbf{A}|^{-1/2}}{|\Sigma_1|^{1/2}+|\Sigma_2|^{1/2}}\exp\{-\frac{1}{2}(\mu_1-\mu_2)'[(\Sigma_1+\Sigma_2)]^{-1}(\mu_1-\mu_2)\}$$

$$= \frac{2|\Sigma_1\Sigma_2|^{1/2}}{|\frac{1}{2}(\Sigma_1+\Sigma_2)|^{1/2}(|\Sigma_1|^{1/2}+|\Sigma_2|^{1/2})}\exp\left[-\frac{1}{2}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right].$$

COROLLARY 3.3: When $\Sigma_1 = \Sigma_2 = \Sigma$, we have the equal variance-covariance case, and

$$\lambda^* = \exp\left[-\frac{1}{4}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\right]. \qquad [3.10]$$

This is the same result as in the preceeding section.

COROLLARY 3.4: When $\mu_1 = \mu_2$, we have

$$\lambda^* = \frac{2|\Sigma_1\Sigma_2|^{1/2}}{|\frac{1}{2}(\Sigma_1+\Sigma_2)|^{1/2}(|\Sigma_1|^{1/2}+|\Sigma_2|^{1/2})}. \qquad [3.11]$$

THEOREM 3.3. For i , j = 1,2, $i \neq j$, the MacArthur-Levins measure is given by

$$\alpha_{ij}^{*} = \frac{\int f_i(\mathbf{x}) f_j(\mathbf{x}) d\mathbf{x}}{\int f_i^{\,2}(\mathbf{x}) d\mathbf{x}}$$

$$= \frac{|\Sigma_i|^{1/2}}{|\frac{1}{2}(\Sigma_i + \Sigma_j)|^{1/2}} \exp\left[-\frac{1}{2}(\mu_i - \mu_j)' (\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)\right], \qquad [3.12]$$

where f_i and f_j are the normal densities $N_p(\mu_i, \Sigma_i)$ and $N_p(\mu_j, \Sigma_j)$ respectively, and the integral is p -fold.

Proof: By the formulation of THEOREM 3.2.

Multiplying $f_i^{\,2}(\mathbf{x})$, $f_j^{\,2}(\mathbf{x})$ and $f_i(\mathbf{x})f_j(\mathbf{x})$ by $(2\pi)^{p/2}$, and integrating, we have

$$(2\pi)^{p/2}\int f_i^{\,2}(\mathbf{x}) d\mathbf{x} = 2^{-p/2} |\Sigma_i|^{-1/2},$$

$$(2\pi)^{p/2}\int f_j^{\,2}(\mathbf{x}) d\mathbf{x} = 2^{-p/2} |\Sigma_j|^{-1/2},$$

and

$$(2\pi)^{p/2}\int f_i(\mathbf{x}) f_j(\mathbf{x}) d\mathbf{x} = \frac{1}{|\Sigma_i + \Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(\mu_i - \mu_j)' (\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)\right].$$

Then

$$\alpha_{ij}^{*} = \frac{\int f_i(\mathbf{x}) f_j(\mathbf{x}) d\mathbf{x}}{\int f_i^{\,2}(\mathbf{x}) d\mathbf{x}}$$

$$= \frac{|\Sigma_i|^{1/2}}{|\frac{1}{2}(\Sigma_i + \Sigma_j)|^{1/2}} \exp\left[-\frac{1}{2}(\mu_i - \mu_j)' (\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)\right]. \qquad [3.12]$$

Note here that if $\Sigma_1 \neq \Sigma_2$, then $\alpha_{12} \neq \alpha_{21}$.

COROLLARY 3.5: When $\Sigma_1 = \Sigma_2 = \Sigma$, we have the equal variance-covariance case, and

$$\alpha_{12}^{*} = \alpha_{21}^{*} = \exp\left[-\frac{1}{4}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)\right]. \qquad [3.13]$$

This is the same result as in the preceeding section.

COROLLARY 3.6: When $\mu_1 = \mu_2$, we have

$$\alpha_{12}^{*} = \frac{|\Sigma_1|^{1/2}}{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|^{1/2}} \qquad [3.14]$$

and

$$\alpha_{21}^{*} = \frac{|\Sigma_2|^{1/2}}{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|^{1/2}}. \qquad [3.15]$$

THEOREM 3.4. Pianka's measure of overlap is

$$\alpha^{*} = \frac{\int f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x}}{[\int f_1^{\,2}(\mathbf{x}) d\mathbf{x}]^{1/2} [\int f_2^{\,2}(\mathbf{x}) d\mathbf{x}]^{1/2}}$$

$$= \frac{|\Sigma_1 \Sigma_2|^{1/4}}{|\frac{1}{2}(\Sigma_1 + \Sigma_2)|^{1/2}} \exp[-\frac{1}{2}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)]. \qquad [3.16]$$

Proof: By the formation of THEOREM 3.2.

Multiplying $f_i^{\,2}(\mathbf{x}), f_j^{\,2}(\mathbf{x})$ and $f_i(\mathbf{x}) f_j(\mathbf{x})$ by $(2\pi)^{p/2}$, and integrating we get

$$(2\pi)^{p/2} \int f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x} = \frac{1}{|\Sigma_1 + \Sigma_2|^{1/2}} \exp[-\frac{1}{2}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)],$$

$$[(2\pi)^{p/2} \int f_1^{\,2}(\mathbf{x}) d\mathbf{x}]^{1/2} = [2^{-p/2} |\Sigma_1|^{-1/2}]^{1/2} = 2^{p/4} |\Sigma_1|^{-1/4}, \text{ and}$$

$$[(2\pi)^{p/2} \int f_2^{\,2}(\mathbf{x}) d\mathbf{x}]^{1/2} = [2^{-p/2} |\Sigma_2|^{-1/2}]^{1/2} = 2^{p/4} |\Sigma_2|^{-1/4}.$$

Then α^{*} can be obtained by putting the above expressions in the numerator and the denominator respectively.

COROLLARY 3.7: When $\Sigma_1 = \Sigma_2 = \Sigma$, we have the equal variance-covariance case, and

$$\alpha^* = \exp\left[-\frac{1}{4}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\right]. \qquad [3.17]$$

This is the same result as in the preceeding section.

COROLLARY 3.8: When $\mu_1 = \mu_2$, we have

$$\alpha^* = \frac{|\Sigma_1\Sigma_2|^{1/4}}{\left|\frac{1}{2}(\Sigma_1 + \Sigma_2)\right|^{1/2}}. \qquad [3.18]$$

When $\Sigma_1 = \Sigma_2 = \Sigma$, these are the exponential forms of certain functions of Mahalanobis generalized distance $(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$. However, when $\Sigma_1 \neq \Sigma_2$, these measures are represented as the product of two terms. One term (the exponent) may be regarded mainly as the Mahalanobis generalized distance corresponding to the weighted average of Σ_1 and Σ_2. This term essentially measures how "far" apart the means are. The second term is essentially a measure of the information contributed by the differences between the covariance matrices Σ_1 and Σ_2. Chernoff [5] discussed a measure, similar to ρ, as the information for discriminating between multivariate normal distributions with unequal covariance matrices.

3.3 TEST STATISTICS AND AFFINITY

The measures discussed in the preceeding sections are functions of the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$. One can identify the affinity measure as the sole parameter between the population distributions. But how should the sample data be used to estimate the parameter? How should the estimator be used to test hypotheses concerning the parameters? First we shall consider the one-sample problem using the affinity and overlap measures.

3.3.1 THE ONE-SAMPLE PROBLEM: Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a sample if n (> 1) observations of a p dimensional random vector $\mathbf{x}$ with a nonsingular normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For those $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we define, as usual, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})$ and we consider the normal distribution $N(\bar{\mathbf{x}}, \mathbf{S})$ determined by $\bar{\mathbf{x}}$ and $\mathbf{S}$ where for $n > p$, $\mathbf{S}$ is nonsingular with probability 1. Thus the affinity and overlap measures between $F = N(\mathbf{a}, \mathbf{A})$ (where $\mathbf{a}$ and $\mathbf{A}$ are known) and $S_n = N(\bar{\mathbf{x}}, \mathbf{S})$ are

$$\rho(F, S_n) = \frac{|\mathbf{AS}|^{1/4}}{|\frac{1}{2}(\mathbf{A} + \mathbf{S})|^{1/2}} \exp\left[-\frac{1}{4}(\mathbf{a} - \bar{\mathbf{x}})'(\mathbf{A} + \mathbf{S})^{-1}(\mathbf{a} - \bar{\mathbf{x}})\right]$$

$$\lambda(F, S_n) = \frac{2|\mathbf{A} + \mathbf{S}|^{1/2}}{|\frac{1}{2}(\mathbf{A} + \mathbf{S})|^{1/2}(|\mathbf{A}|^{1/2} + |\mathbf{S}|^{1/2})} \exp\left[-\frac{1}{2}(\mathbf{a} - \bar{\mathbf{x}})'(\mathbf{A} + \mathbf{S})^{-1}(\mathbf{a} - \bar{\mathbf{x}})\right]$$

and

$$\alpha(F, S_n) = \frac{|\mathbf{AS}|^{1/4}}{|\frac{1}{2}(\mathbf{A} + \mathbf{S})|^{1/2}} \exp[-\frac{1}{2}(\mathbf{a} - \bar{\mathbf{x}})'(\mathbf{A} + \mathbf{S})^{-1}(\mathbf{a}_1 - \bar{\mathbf{x}})] .$$

With these statistics, we can make inferences concerning $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $N(\mathbf{a}, \mathbf{A})$, when the former is unknown.

I. When it is known beforehand that $\boldsymbol{\Sigma} = \mathbf{A}$, although $\boldsymbol{\Sigma}$ itself is unknown, we consider

$$\rho_1(F, S_n) = \exp[-\frac{1}{8}(\mathbf{a} - \bar{\mathbf{x}})' \mathbf{S}^{-1}(\mathbf{a} - \bar{\mathbf{x}})]$$

and

$$\lambda_1(F, S_n) = \alpha_1(F, S_n) = \exp[-\frac{1}{4}(\mathbf{a} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{a} - \bar{\mathbf{x}})],$$

which are the affinities and overlaps between $N(\mathbf{a}, \mathbf{S})$ and $N(\bar{\mathbf{x}}, \mathbf{S})$. Further, when $\boldsymbol{\Sigma}$ is known, we take

$$\rho_2(F, S_n) = \exp[-\frac{1}{8}(\mathbf{a} - \bar{\mathbf{x}})' \boldsymbol{\Sigma}^{-1}(\mathbf{a} - \bar{\mathbf{x}})]$$

and

$$\lambda_2(F, S_n) = \alpha_2(F, S_n) = \exp[-\frac{1}{4}(\mathbf{a} - \bar{\mathbf{x}})' \boldsymbol{\Sigma}^{-1}(\mathbf{a} - \bar{\mathbf{x}})] .$$

For these ρ_i, λ_i, and α_i $(i = 1, 2)$, we have

$$-8 \log \rho_1(F, S_n) = -4 \log \lambda_1(F, S_n) = -4 \log \alpha_1(F, S_n) = (\mathbf{a} - \bar{\mathbf{x}})' \mathbf{S}^{-1}(\mathbf{a} - \bar{\mathbf{x}})$$

and

$$-8 \log \rho_2(F, S_n) = -4 \log \lambda_2(F, S_n) = -4 \log \alpha_2(F, S_n) = (\mathbf{a} - \bar{\mathbf{x}})' \boldsymbol{\Sigma}^{-1}(\mathbf{a} - \bar{\mathbf{x}}) .$$

These are familiar expressions in multivariate analysis, since $n(\mathbf{a} - \bar{\mathbf{x}})\boldsymbol{\Sigma}^{-1}(\mathbf{a} - \bar{\mathbf{x}})$ is the one-sample generalized T^2.

Let S_n be the distribution determined by n observations of a random variable with $F = N(\mathbf{a}, \boldsymbol{\Sigma})$. Then

$$-8n \log \rho_1(F, S_n) = -4n \log \lambda_1(F, S_n) = -4n \log \alpha_1(F, S_n)$$

$$= n[(\bar{\mathbf{x}} - \mathbf{a})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \mathbf{a})]$$

is distributed according to a noncentral F distribution, and

$$-8n \log \rho_2(F, S_n) = -4n \log \lambda_2(F, S_n) = -4n \log \alpha_2(F, S_n)$$

$$= n[(\bar{\mathbf{x}} - \mathbf{a})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \mathbf{a})]$$

follows a χ^2 distribution with n degrees of freedom.

II. When the problem is concerned only with the covariance matrix, we consider

$$\rho_3(F, S_n) = \alpha_3(F, S_n) = \frac{|\mathbf{AS}|^{1/4}}{|\frac{1}{2}(\mathbf{A} + \mathbf{S})|^{1/2}}$$

and

$$\lambda_3(F, S_n) = \frac{2|\mathbf{A} + \mathbf{S}|^{1/2}}{|\frac{1}{2}(\mathbf{A} + \mathbf{S})|^{1/2}(|\mathbf{A}|^{1/2} + |\mathbf{S}|^{1/2})},$$

which are the affinities between $N(\mathbf{b}, \mathbf{A})$ and $N(\mathbf{b}, \mathbf{S})$, $\mathbf{b}$ being any vector.

We note that these statistics are invariant under any nonsingular transformation. In fact, for a nonsingular transformation $\mathbf{y} = \mathbf{Gx} + \mathbf{b}$, the distribution of $\mathbf{y}$ is $F_y = N(\mathbf{Ga} + \mathbf{b}, \mathbf{G\Sigma G'})$, and the sample mean and sample variance-covariance matrix become $\bar{\mathbf{y}} = \mathbf{G}\bar{\mathbf{x}} + \mathbf{b}$ and $S_y = \mathbf{GSG'}$, respectively. Moreover,

$$\rho_3(F_y, S_{ny}) = \frac{|(\mathbf{G\Sigma G'})(\mathbf{GSG'})|^{1/4}}{|\frac{1}{2}((\mathbf{G\Sigma G'}) + (\mathbf{GSG'})|^{1/2}} = \frac{|\mathbf{\Sigma S}|^{1/4} abs.|\mathbf{G}|}{|\frac{1}{2}(\mathbf{\Sigma} + \mathbf{S})|^{1/2} abs.|\mathbf{G}|}$$

$$= \rho_3(F_x, S_{nx})$$

where $abs.|\mathbf{G}|$ is the absolute value of the determinant of $\mathbf{G}$.

Now, by assumption, $\boldsymbol{\Sigma}$ is a positive-definite symmetric matrix, so there exists an orthogonal matrix $\mathbf{T}$ such that $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}'$ has a diagonal form

$$[\eta_1, \ldots, \eta_p]$$

where $\eta_i > 0$ for all i. Let

$\mathbf{L}$ = Diagonal $[\eta_1^{1/2}, \ldots, \eta_p^{1/2}]$.

Here $\mathbf{L} = \mathbf{L}'$. Further, let $\mathbf{G} = \mathbf{L}^{-1}\mathbf{T}$. Then we have

$$\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}' = (\mathbf{L}^{-1}\mathbf{T})\boldsymbol{\Sigma}(\mathbf{L}^{-1}\mathbf{T})' = \mathbf{I}.$$

This, with the invariance of ρ_3, implies that, when we are concerned with the distribution of $\rho_3(F, S_n)$, we may assume that $F = N(\mathbf{0}, \mathbf{I})$. That is, the distribution of $\rho_3(F, S_n)$ is independent of a particular distribution. For $F = N(\mathbf{0}, \mathbf{I})$, we obtain

$$\rho_3(F, S_n) = \frac{|\mathbf{IS}|^{1/4}}{|\frac{1}{2}(\mathbf{I} + \mathbf{S})|^{1/2}} = \frac{|\mathbf{S}|^{1/4}}{|\frac{1}{2}(\mathbf{I} + \mathbf{S})|^{1/2}}.$$

Let

$$X_i' = [X_{i1}, \ldots, X_{ip}] \quad (i = 1, \ldots, n).$$

As the matrix S converges to **I** with probability 1, the distribution of $\rho_3(F, S_n)$ is asymptotically equal to that of

$$\frac{|\mathbf{W}|^{1/4}}{|\frac{1}{2}(\mathbf{I} + \mathbf{W})|^{1/2}},$$

where $\mathbf{W} = \text{Diagonal}\ [\frac{1}{n}\sum_{i=1}^{n} X_{i1}^{\ 2}\ ,\ \ \ldots\ ,\ \ \frac{1}{n}\sum_{i=1}^{n} X_{ip}^{\ 2}]$.

If $Z_1 = \frac{1}{n}\sum_{i=1}^{n} X_{i1}^{\ 2}\ ,\ \ \ldots\ ,\ \ Z_p = \frac{1}{n}\sum_{i=1}^{n} X_{i1}^{\ 2}\ ,$

then

$$\frac{|\mathbf{W}|^{1/4}}{|\frac{1}{2}(\mathbf{I} + \mathbf{W})|^{1/2}} = \frac{(\prod_{i=1}^{p} Z_i)^{1/4}}{(\prod_{i=1}^{p} \frac{1}{2}(1 + Z_i))^{1/2}} = [\prod_{i=1}^{p} \frac{4Z_i}{(1 + Z_i)^2}]^{1/4},$$

and $nZ_1, \ldots, nZ_p$ are independent random variables having the χ^2 distribution with n degrees of freedom.

For the more general case,

$$\frac{|\mathbf{\Sigma S}|^{1/4}}{|\frac{1}{2}(\mathbf{\Sigma} + \mathbf{S})|^{1/2}} = \left[\prod_{i=1}^{p} \frac{4\eta_i Z_i}{(\eta_i + Z_i)^2}\right]^{1/4}.$$

Similarly, with the invariance of λ_3, for $F = N(\mathbf{0}, \mathbf{I})$ we obtain

$$\lambda_3(F, S_n) = \frac{2|\mathbf{I} + \mathbf{S}|^{1/2}}{|\frac{1}{2}(\mathbf{I} + \mathbf{S})|^{1/2}(|\mathbf{I}|^{1/2} + |\mathbf{S}|^{1/2})}.$$

As the matrix S converges to **I** with probability 1, the distribution of $\lambda_3(F, S_n)$ converges to that of

$$\frac{2|\mathbf{I} + \mathbf{W}|^{1/2}}{|\frac{1}{2}(\mathbf{I} + \mathbf{W})|^{1/2}(|\mathbf{I}|^{1/2} + |\mathbf{W}|^{1/2})} = \frac{2\prod_{i=1}^{p}(1 + Z_i)^{1/2}}{(\prod_{i=1}^{p}\frac{1}{2}(1 + Z_i))^{1/2}(1 + \prod_{i=1}^{p} Z_i^{1/2})}$$

$$= 2^{\frac{(p+2)}{2}}[\prod_{i=1}^{p}(1 + Z_i)^{1/2}]^{-1}.$$

For the more general case,

$$\frac{2|\mathbf{\Sigma} + \mathbf{S}|^{1/2}}{|\frac{1}{2}(\mathbf{\Sigma} + \mathbf{S})|^{1/2}(|\mathbf{\Sigma}|^{1/2} + |\mathbf{S}|^{1/2})} = 2^{\frac{(p+2)}{2}}[\prod_{i=1}^{p}(\eta_i)^{1/2} + \prod_{i=1}^{p}(Z_i)^{1/2}]^{-1}.$$

To test the hypotheses $H_o : \mathbf{A} = \mathbf{\Sigma}$ vs $H_a : \mathbf{A} \neq \mathbf{\Sigma}$, using ρ_3, it can be shown that

$$\left[\prod_{i=1}^{p}\frac{4\eta_i\frac{\chi^2_{n,1-\alpha}}{n}}{(\eta_i + \frac{\chi^2_{n,1-\alpha}}{n})^2}\right]^{1/4}$$

is the critical value of the α level test based on the criterion $\left[\prod_{i=1}^{p}\frac{4\eta_i Z_i}{(\eta_i + Z_i)^2}\right]^{1/4}$.

Alternatively, using λ_3 , it can be shown that

$$2^{\frac{(p+2)}{2}}\left[\prod_{i=1}^{p}(\eta_i)^{1/2} + \prod_{i=1}^{p}\left(\frac{\chi^2_{n,1-\alpha}}{n}\right)^{1/2}\right]^{-1}$$

is the critical value of the α level test based on the test statistic $2^{\frac{(p+2)}{2}}\left[\prod_{i=1}^{p}(\eta_i)^{1/2} + \prod_{i=1}^{p}(Z_i)^{1/2}\right]^{-1}$.

These give the test criteria for testing hypotheses about the covariance matrices using the affinity measures.

3.3.2 THE TWO-SAMPLE PROBLEM: Suppose we want to make inferences about $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, the difference between the population-mean vectors.

With a few tentative assumptions we are able to make such inferences through the distance, affinity or overlap measures.

We make the following assumptions concerning the structure of the data.

1. The sample $\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$ is a random sample of size n_1 from a p -variate population with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$.

2. The sample $\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$ is a random sample of size n_2 from a p -variate population with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$.

3. Also, $\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$ are independent of $\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$.

Further assumptions are required when n_1 and n_2 are small:

1. Both populations are multivariate normal.

2. $\Sigma_1 = \Sigma_2$.

From the samples $\mathbf{x}_{1j}$, $(j = 1, \ldots, n_1)$ and $\mathbf{x}_{2j}$, $(j = 1, \ldots, n_2)$, we calculate $\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}$, $\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}$,

$$\mathbf{S}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$$

and

$$\mathbf{S}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'.$$

When $\Sigma_1 = \Sigma_2 = \Sigma$, $\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$ is an estimate of $(n_1 - 1)\Sigma$ and $\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'$ is an estimate of $(n_2 - 1)\Sigma$. Consequently, we can pool the two samples in order to estimate the common covariance Σ. We obtain

$$\mathbf{S}_p = \frac{\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)' + \sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'}{n_1 + n_2 - 2}$$

$$= \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}.$$

To test the hypothesis that $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$, a specified vector. We proceed as follows. The likelihood ratio test is based on the squared statistical distance T^2, and

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)'[(\frac{1}{n_1} + \frac{1}{n_2})\mathbf{S}_p]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)$$

is distributed as $\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}$. However, when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, we are unable to find a "distance" measure like T^2 whose distribution does not depend on the unknowns $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. If $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ were known, the test statistic would be

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))'[\frac{1}{n_1}\boldsymbol{\Sigma}_1 + \frac{1}{n_2}\boldsymbol{\Sigma}_2]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)),$$

which has an approximate χ^2_p distribution. When n_1 and n_2 are large, with high probability, $\mathbf{S}_1$ will be close to $\boldsymbol{\Sigma}_1$, and $\mathbf{S}_2$ will be close to $\boldsymbol{\Sigma}_2$, with high probability. Consequently,

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))'[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$$

will have an approximate χ^2_p distribution (Johnson and Wichern [23]).

The performance of the sample test on the heterogeneous variance-covariance matrices depends on the sample sizes n_1 and n_2, the dimension p, and the degree of the similarity of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

Chapter IV

CORRELATION AND DISTANCE IN NORMAL DISTRIBUTIONS

Let (X_1, X_2) be a two-dimensional random vector with distribution F over R^2, and let Ω be the set of two-dimensional distributions over R^2 each of which is a direct product of one-dimensional distributions over R. Then, X_1 and X_2 are independent of each other if and only if $F \in \Omega$. Therefore, we say that when F is "near" Ω , the random variables X_1 and X_2 are "nearly" independent. The affinity between distributions may be interpreted as the "closeness" of the distributions. As discussed in the previous chapters, several affinity measures and overlap measures are closely related to the distance between distributions. On the other hand, the correlation coefficient also represents a relation between two variates. In this chapter, we will present the relation between the correlation co-

efficient and the affinities in term of the joint distribution or the marginal distributions. Here, the study is confined to the normal distribution.

4.1 CORRELATION COEFFICIENT AND THE MATUSITA AFFINITY

First, consider the distribution G and assume that $G \in \Omega$. The Matusita affinity of F and Ω is defined as

$$\rho_1 = \max_{G \in \Omega} \rho(F, G). \qquad [4.1]$$

Here ρ_1 represents the degree of association between X_1 and X_2 , where $\rho(F, G)$ denotes the affinity between F and G . See, for example, Matusita [31]. However, when $F \in \Omega$, F becomes the direct product of the marginal distributions of X_1 and X_2. Therefore, we can also consider

$$\rho_{11} = \rho(F, F_1 \times F_2) \qquad [4.2]$$

as a measure of association between X_1 and X_2, where F_1 and F_2 are the marginal distributions of X_1 and X_2. We shall give the relations between ρ_1 or ρ_{11} and the correlation coefficient r of X_1 and X_2 , then examine the relation between the concepts of distance and independence.

Let (X_1, X_2) be a two-dimensional random vector with normal distribution $F = N(\mathbf{a}, \Sigma)$, where

$$\mathbf{a} = (a_1, a_2), \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \qquad \sigma_{12} = \sigma_{21}.$$

Now, let Ω be the set of normal distributions with mean $\mathbf{a}$ and covariance matrix

$$\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \quad \text{where} \quad x, y > 0.$$

Let G be a distribution of Ω . Then Matusita's affinity between F and G is calculated to be

$$\rho = 2\frac{\left|\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}\right|^{-1/4}}{\left|\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} + \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}^{-1}\right|^{1/2}}$$

$$= 2\frac{\left|\begin{matrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{matrix}\right|^{1/4}\left|\begin{matrix} x & 0 \\ 0 & y \end{matrix}\right|^{1/4}}{\left|\begin{matrix} x + \sigma_{11} & \sigma_{12} \\ \sigma_{21} & y + \sigma_{22} \end{matrix}\right|^{1/2}}. \qquad [4.3]$$

To find ρ_1 , we take the derivatives of the above expression with respect to x and y. The maximum of ρ obtains when

$x = [(\sigma_{11}/\sigma_{22})(\sigma_{11}\sigma_{22} - \sigma_{12}^2)]^{1/2}$ and

$$y = [(\sigma_{22}/\sigma_{11})(\sigma_{11}\sigma_{22} - \sigma_{12}^2)]^{1/2}, \qquad [4.4]$$

or, equivalently, when

$$x = \sigma_{11}[(1 - r^2)]^{1/2} \quad \text{and}$$

$$y = \sigma_{22}[(1 - r^2)]^{1/2} \qquad [4.5]$$

where r denotes the correlation coefficient $\sigma_{12}/[(\sigma_{11}\sigma_{22})^{1/2}]$. Thus we obtain

$$\rho_1 = \max_{x,y>0} \rho(F, G)$$

$$= \frac{\sqrt{2}(1 - r^2)^{1/4}}{[1 + (1 - r^2)^{1/2}]^{1/2}}. \qquad [4.6]$$

For ρ_{11} , we have

$$\rho_{11} = 2\frac{\left|\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}\right|^{-1/4}}{\left|\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}^{-1} + \begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}^{-1}\right|^{1/2}}$$

$$= \frac{2(1 - r^2)^{1/4}}{(4 - r^2)^{1/2}}. \qquad [4.7]$$

From these relations it can be seen that (figure 1)

1. ρ_1 and ρ_{11} are equal to each other if and only if r = 0, that is, if and only if X_1 and X_2 are mutually independent.

2. ρ_1 and ρ_{11} are both monotone-decreasing functions of r^2. When r = 0, $\rho_1 = \rho_{11} = 1$, and when $r = \pm 1$, $\rho_1 = \rho_{11} = 0$.

If F admits density $f(\mathbf{x})$ and G admits density $g(\mathbf{x})$, then

$$1 - \rho_1 = \frac{1}{2} d_2^2(F, G).$$

Furthermore, if $F_1 \times F_2$ admits density $f_1(x)f_2(x)$, then

$$1 - \rho_{11} = \frac{1}{2} d_2^2(F, F_1 \times F_2).$$

These are the expressions in terms of Matusita's affinity.

Although ρ_1 or ρ_{11} are represented as measures of association between X_1 and X_2, $1 - \rho_1$ or $1 - \rho_{11}$ can be taken suitably as a coefficient of association between X_1 and X_2 , since they are monotone-increasing functions of r^2 ; when r = 0, they vanish, and when $r = \pm 1$, they are 1.

Note that the distribution in Ω which is the closest to F is not

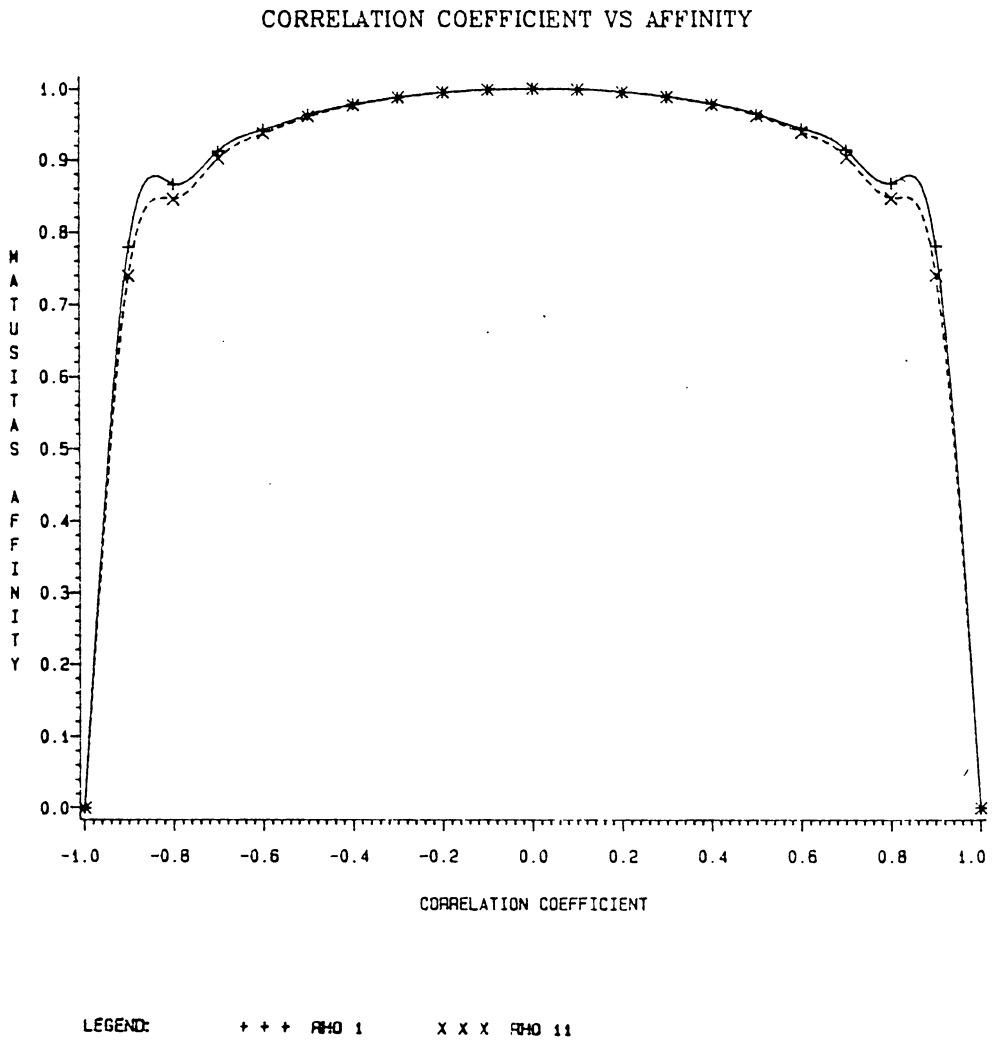

Figure 1. Correlation and Matusita's Affinity: The relation between correlation coefficient and ρ_1, ρ_{11}.

$$N\ (\mathbf{a}, \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix})$$

but

$$N\ (\mathbf{a}, \begin{bmatrix} \sigma_{11}(1 - r^2)^{1/2} & 0 \\ 0 & \sigma_{22}(1 - r^2)^{1/2} \end{bmatrix}).$$

4.2 CORRELATION COEFFICIENT AND THE MORISITA AFFINITY

Let (X_1, X_2) be a two-dimensional random vector with normal distribution $F = N(\mathbf{a}, \Sigma)$, where

$$\mathbf{a} = (a_1, a_2), \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \qquad \text{and } \sigma_{12} = \sigma_{21}\ .$$

Now, let Ω be the set of normal distributions with mean $\mathbf{a}$ and covariance matrix

$$\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \qquad \text{where} \quad x, y > 0.$$

Let G be a distribution of Ω . Then Morisita's affinity between F and G is calculated to be

$$\lambda = 2\frac{\left|\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\begin{bmatrix}x & 0\\ 0 & y\end{bmatrix}\right|^{1/2}}{\left|\frac{1}{2}\left(\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}+\begin{bmatrix}x & 0\\ 0 & y\end{bmatrix}\right)\right|^{1/2}\left(\begin{vmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{vmatrix}^{1/2}+\begin{vmatrix}x & 0\\ 0 & y\end{vmatrix}^{1/2}\right)}.$$

To find λ_1 , take the derivatives of the above expression with respect to x and y, this λ attaining its maximum when

$$x = [(\sigma_{11}/\sigma_{22})(\sigma_{11}\sigma_{22} - \sigma_{12}^2)]^{1/2} = \sigma_{11}[(1 - r^2)]^{1/2} \quad \text{and}$$

$$y = [(\sigma_{22}/\sigma_{11})(\sigma_{11}\sigma_{22} - \sigma_{12}^2)]^{1/2} = \sigma_{22}[(1 - r^2)]^{1/2}. \qquad [4.8]$$

Thus we obtain

$$\lambda_1 = \max_{x,y>0} \lambda(F, G)$$

$$= \frac{\sqrt{2}\,(1 - r^2)^{1/4}}{[1 + (1 - r^2)^{1/2}]^{1/2}}. \qquad [4.9]$$

For λ_{11}, we have

$$\lambda_{11} = 2\frac{\left|\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}\right|^{1/2}}{\left|\frac{1}{2}\left(\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}+\begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}\right)\right|^{1/2}\left(\begin{vmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{vmatrix}^{1/2}+\begin{vmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{vmatrix}^{1/2}\right)}$$

$$= \frac{4(1 - r^2)^{1/2}}{(4 - r^2)^{1/2}[1 + (1 - r^2)^{1/2}]}. \qquad [4.10]$$

From the above relations it can be seen that (figure 2)

1. λ_1 and λ_{11} are equal to each other if and only if r = 0, that is, X_1 and X_2 are mutually independent.

2. λ_1 and λ_{11} are both monotone-decreasing functions of r^2. When r = 0, $\lambda_1 = \lambda_{11} = 1$, and when $r = \pm 1, \lambda_1 = \lambda_{11} = 0$.

If F admits density $f(\mathbf{x})$ and G admits density $g(\mathbf{x})$, then

$$1 - \lambda_1 = \frac{\int(f(\mathbf{x}) - g(\mathbf{x}))^2 d\mathbf{x}}{\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x}}.$$

Furthermore, if $F_1 \times F_2$ admits density $f_1(x)f_2(x)$, then

$$1 - \lambda_{11} = \frac{\int(f(\mathbf{x}) - f_1(x)f_2(x))^2 d\mathbf{x}}{\int f^2(\mathbf{x})d\mathbf{x} + \int f_1^2(x)f_2^2(x)dx}.$$

These are the expressions in terms of the Morisita's affinity.

Although λ_1 or λ_{11} are represented as a measure of association between X_1 and X_2, $1 - \lambda_1$ or $1 - \lambda_{11}$ can be taken suitably as a coefficient of association

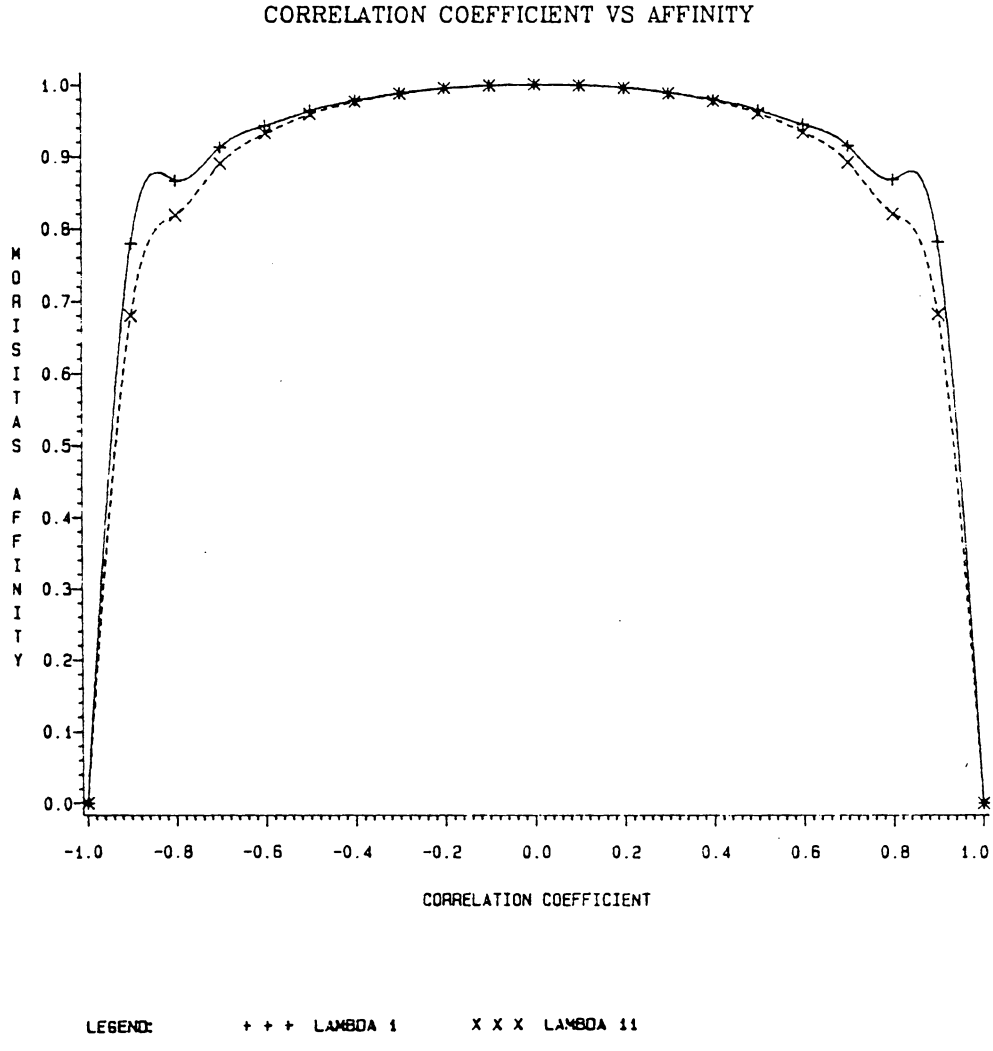

Figure 2. Correlation and Morisita's Affinity: The relation between correlation coefficient and λ_1, λ_{11}.

between X_1 and X_2 , since they are monotone-increasing functions of r^2 ; when r = 0, they vanish, and when $r = \pm 1$, they are 1.

It is interesting to notice that the distribution in Ω which is the closest to F is not

$$N(\mathbf{a}, \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix})$$

but

$$N(\mathbf{a}, \begin{bmatrix} \sigma_{11}(1 - r^2)^{1/2} & 0 \\ 0 & \sigma_{22}(1 - r^2)^{1/2} \end{bmatrix}).$$

Now, there is relationship between the measures ρ_{11} and λ_{11}, that is, $\rho_{11} \geq \lambda_{11}$ when $-1 \leq r \leq 1$, and $\rho_{11} = \lambda_{11}$ only if $r = \pm 1$. Figure 3 shows that the affinities approach the same independent concept as the correlation coefficient, but one cannot use ρ_1, ρ_{11} or λ_{11} as criteria for testing the independence of X_1 and X_2, because they have less variation than r around 0. However, taking the logarithm of $1 - \rho_1$, $1 - \rho_{11}$ and $1 - \lambda_{11}$, as the coefficients of association between X_1 and X_2 (figure 4), one finds that as r approaches 0, these "logarithmic" coefficients of association approach infinity. Thus they may serve as criterion for testing the independence of X_1 and X_2.

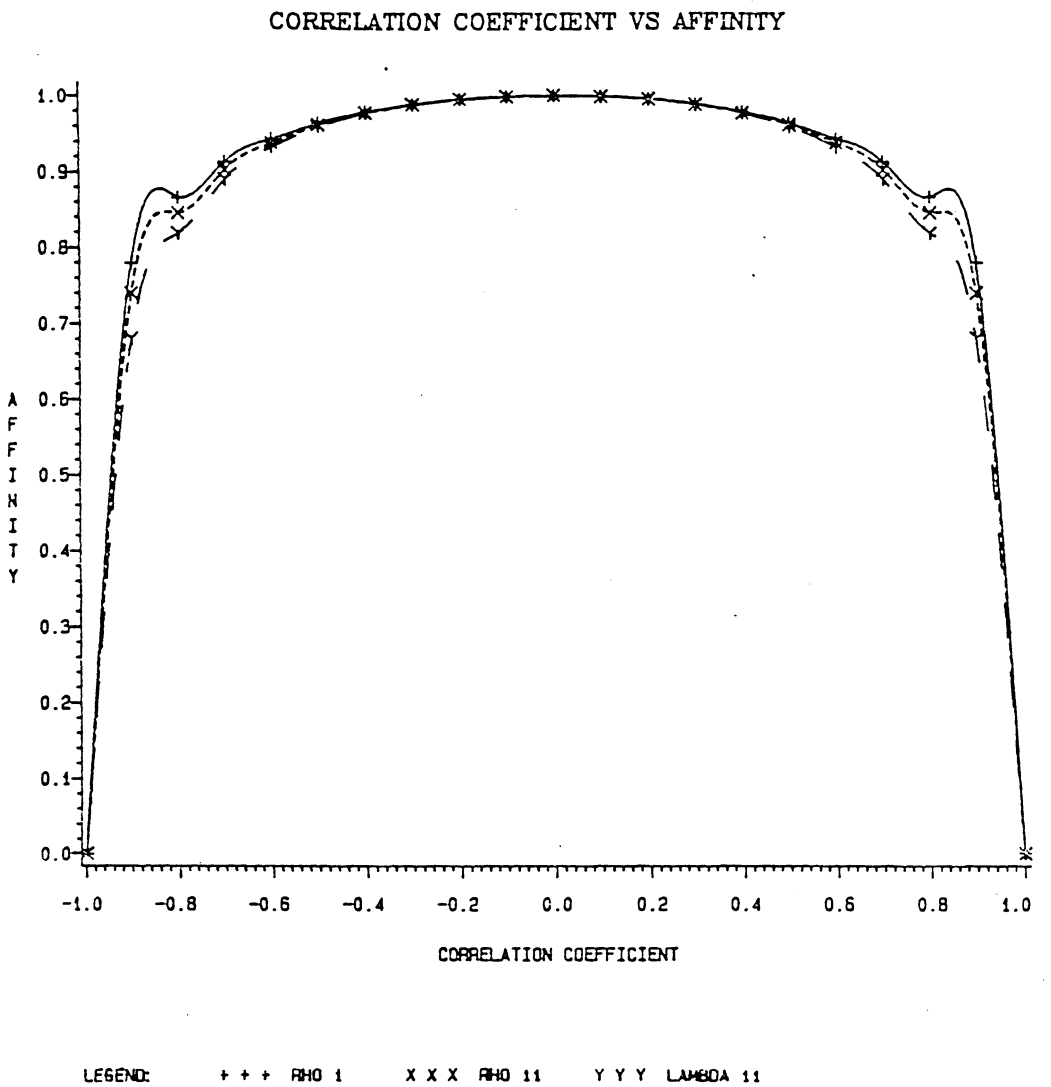

Figure 3. Correlation and Affinity: The relation between correlation coefficient and ρ_1, ρ_{11}, λ_1 and λ_{11}.

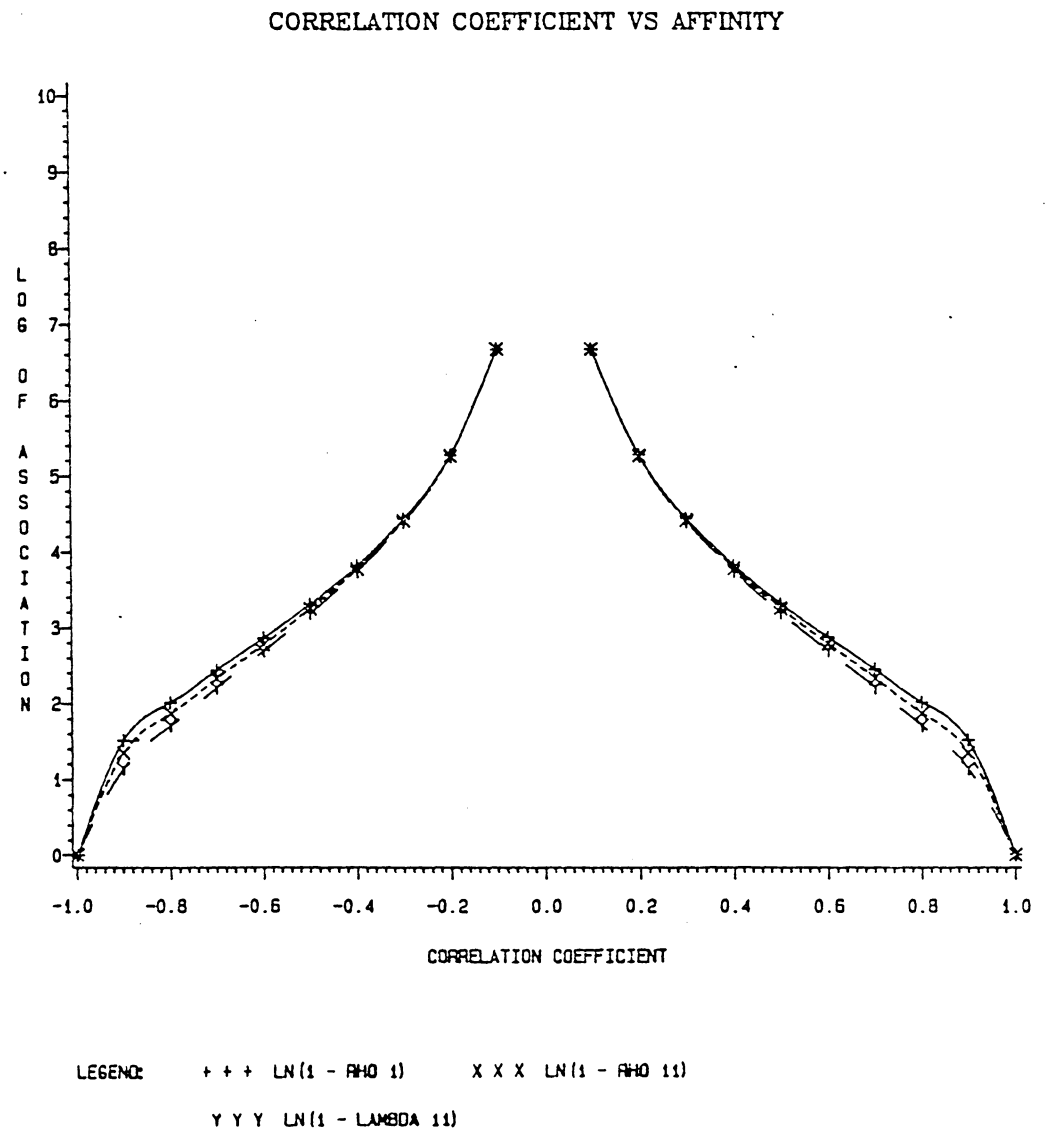

Figure 4. Correlation and Log of association: The relation between correlation coefficient and log of association of ρ and λ .

4.3 CORRELATION COEFFICIENT AND DIVERGENCE

There are other measures that represent the discrepancy among the probability distributions: the Kullback-Leibler information number, and the divergence. Here, the relationship between the correlation coefficient and the divergence will be discussed in terms of the joint distribution on the marginal distributions.

The following definition is due to Kullback-Leibler [24]:

DEFINITION 4.1. Let $g_1(x)$ and $g_2(x)$ be two absolutely continuous probability density functions. The Kullback-Leibler information number $I(g_1, g_2)$ is defined by

$$I(g_1, g_2) = \int g_1(x) \log \frac{g_1(x)}{g_2(x)} dx.$$

The divergence J of g_1 and g_2 is defined by

$$J(g_1, g_2) = \int (g_1(x) - g_2(x)) \log \frac{g_1(x)}{g_2(x)} dx.$$

This was first introduced by Jeffrey [22].

The above definition can easily be extended to the multivariate case. If $g_1(\mathbf{x}) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ and $g_2(\mathbf{x}) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$, then

$$I(g_1, g_2) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \frac{1}{2} \mathrm{tr} \boldsymbol{\Sigma}_1 (\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})$$

$$+ \frac{1}{2}\mathrm{tr}\Sigma_2^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)'. \qquad [4.11]$$

$$J = \frac{1}{2}\mathrm{tr}(\Sigma_1 - \Sigma_2)(\Sigma_2^{-1} - \Sigma_1^{-1}) + \frac{1}{2}\mathrm{tr}(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)(\mu_1 - \mu_2)'. \qquad [4.12]$$

Assuming equal population covariance matrices, $\Sigma_1 = \Sigma_2 = \Sigma$, then [4.11] and [4.12] become,

$$I(g_1, g_2) = \frac{1}{2}\mathrm{tr}\Sigma^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)' \qquad [4.13]$$

$$J(g_1, g_2) = \mathrm{tr}\Sigma^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)'. \qquad [4.14]$$

Both are constant multiples of the Mahalanobis generalized distance.

Assuming equal population means, $\mu_1 = \mu_2$, then [4.11] and [4.12] become, (see Kullback [22])

$$I(g_1, g_2) = \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{p}{2} + \frac{1}{2}\mathrm{tr}\Sigma_1\Sigma_2^{-1} \qquad [4.15]$$

$$J(g_1, g_2) = \frac{1}{2}\mathrm{tr}\Sigma_1\Sigma_2^{-1} + \frac{1}{2}\mathrm{tr}\Sigma_2\Sigma_1^{-1} - p. \qquad [4.16]$$

As in the discussion of the previous sections, let the distribution $G \in \Omega$, and define the divergence measure of F and Ω as

$$J_1 = \min_{G\in\Omega} J(F, G). \qquad [4.17]$$

Here J_1 represents the closeness between X_1 and X_2. However, when $F \in \Omega$, F becomes the direct product of the marginal distributions of X_1 and X_2. Therefore, we can also consider

$$J_{11} = J(F, F_1 \times F_2). \qquad [4.18]$$

Let (X_1, X_2) be a two-dimensional random vector with normal distribution $F = N(\mathbf{a}, \Sigma)$, where

$$\mathbf{a} = (a_1, a_2) \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \text{ and } \quad \sigma_{12} = \sigma_{21}.$$

Now, let Ω be the set of normal distributions with mean **a** and covariance matrix

$$\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \quad \text{where} \quad x, y > 0.$$

Let G be a distribution of Ω . Then divergence measure between F and G is calculated to be

$$J = \frac{1}{2}\left(\frac{\sigma_{11}}{x} + \frac{\sigma_2}{y}\right) + \frac{1}{2}\frac{x\sigma_{22} + y\sigma_{11}}{\sigma_{11}\sigma_{22}(1 - r^2)} - 2. \qquad [4.19]$$

To find J_1, we take the derivatives of the above expression with respect to x and y, and ρ attains its minimum when

$$x = [(\sigma_{11}/\sigma_{22})(\sigma_{11}\sigma_{22} - \sigma_{12}^2)]^{1/2}, \quad y = [(\sigma_{22}/\sigma_{11})(\sigma_{11}\sigma_{22} - \sigma_{12}^2)]^{1/2} \qquad [4.20]$$

or, equivalently, when

$$x = \sigma_{11}[(1 - r^2)]^{1/2}, \quad y = \sigma_{22}[(1 - r^2)]^{1/2}, \qquad [4.21]$$

where r denotes the correlation coefficient $\sigma_{12}/[(\sigma_{11}\sigma_{22})^{1/2}]$. Thus we obtain

$$J_1 = \frac{2}{(1 - r^2)^{1/2}} - 2. \qquad [4.22]$$

For J_{11}, we have

$$J_{11} = \frac{1}{2}\mathrm{tr}\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}^{-1} + \frac{1}{2}\mathrm{tr}\begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}^{-1} - 2$$

$$= \frac{r^2}{1 - r^2}. \qquad [4.23]$$

From these relations, it can be seen that (figure 5)

1. $J_1 = J_{11}$ if and only if r = 0, that is, if and only if X_1 and X_2 are mutually independent.

2. J_1 and J_{11} are both monotone-increasing functions of r^2 for $o \leq r^2 \leq 1$. When r = 0, $J_1 = J_{11} = 1$ and when $r = \pm 1$, $J_1 = J_2 = \infty$.

These can be taken as coefficients of association between X_1 and X_2 since they are monotone-increasing functions of r^2 .

It is interesting to notice, whichever measure (Matusita affinity, Morisita affinity, divergence) is applied, the distribution in Ω which is the closest to F is not

$$N(\mathbf{a}, \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix})$$

but

$$N(\mathbf{a}, \begin{bmatrix} \sigma_{11}(1 - r^2)^{1/2} & 0 \\ 0 & \sigma_{22}(1 - r^2)^{1/2} \end{bmatrix}).$$

Since ρ_1 and λ_1 are measures of affinity between X_1 and X_2, one can take $-\ln(\rho_1)$ and $-\ln(\lambda_1)$ as the measures of association to clarify the independence concept between X_1 and X_2. In figure 3, J_1, $-\ln(\rho_1)$ and $-\ln(\lambda_1)$ are compared with the correlation coefficient. It turns out that J_1 moves away from zero faster

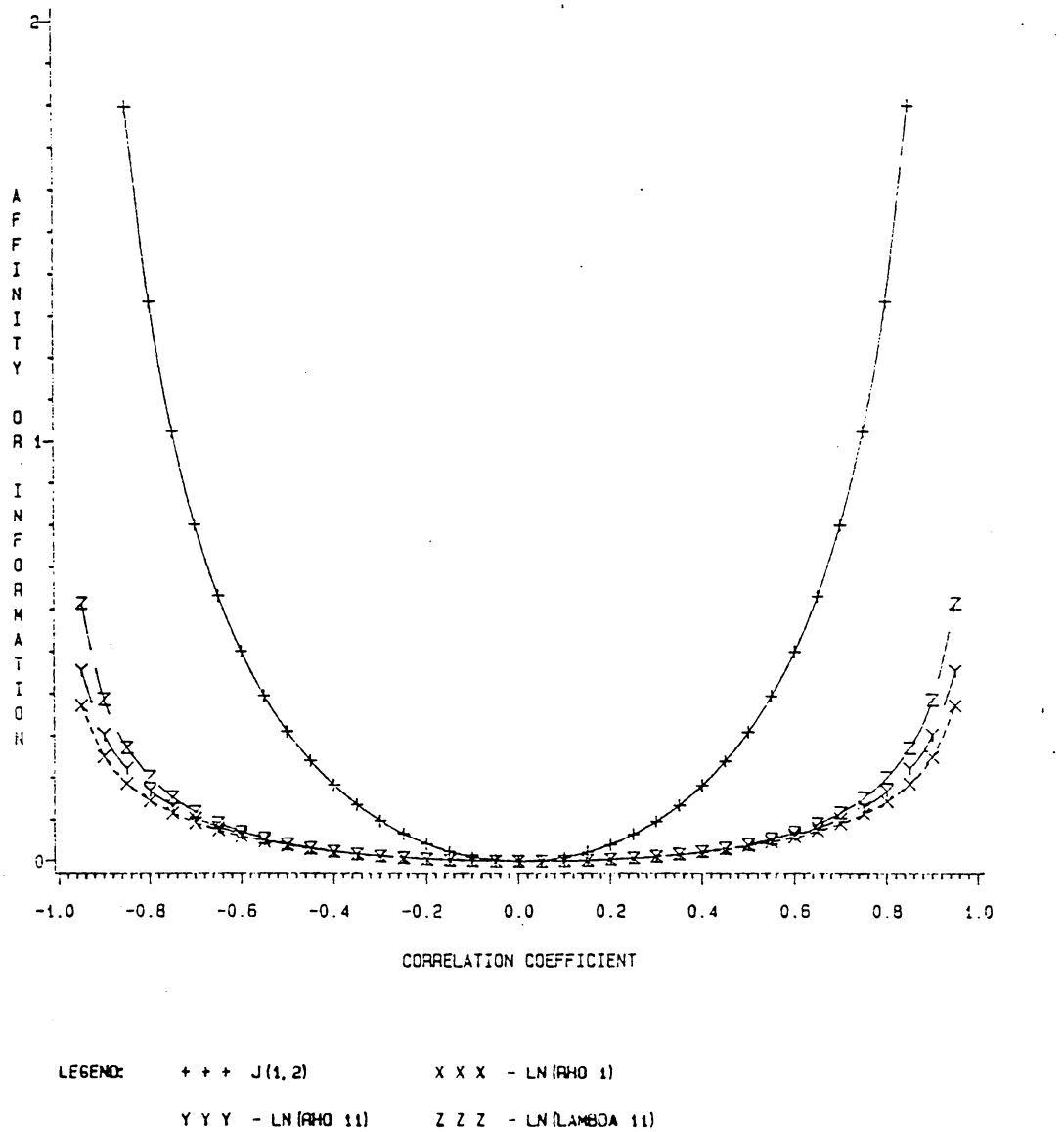

Figure 5. Correlation and Divergence: The relation between correlation coefficient and - log $(1 - \rho)$, - log $(1 - \lambda)$ and divergence.

than the other measures of association, when the correlation coefficient is away from zero.

Chapter V

ESTIMATION PROCEDURES

In this chapter, we describe two nonparametric procedures, the jackknife and the bootstrap, for the estimation of the measures introduced in the previous chapters. The main idea is to assess the usefulness, the reliability and the precision of the currently used measures of overlap. In practice, it is important to obtain accurate estimates of the variance of the measures. Several factors affecting the reliability of sampled data, such as the number of variables (the dimensionality), the sample sizes, the inequality of the variance-covariance matrices, the differences of means, are investigated. The task is focused on the biases and the adjustments for biases, the variances of measures, the precision of the jackknife and the bootstrap methods and the coverages of the confidence intervals. Monte Carlo simulation techniques are used to generate the sample data from multivariate normal distributions. Results are for the two sample case as this is the most common situation.

5.1 TWO-SAMPLE JACKKNIFE METHOD

The jackknife method provides direct numerical approximations of both bias and standard error, and can give reasonably reliable confidence limits. It is a relative of various nonparametric methods (Miller [33]). The one-sample jackknife method has been applied to numerous problems, and its description can be found in many texts. Here, the measures of affinity and overlap are based on two populations, so the two-sample version of the jackknife method is introduced.

Assume that the data consists of two groups of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$. Let the affinity measure or the overlap measure be given by $\gamma = \gamma(\mathbf{x}, \mathbf{y})$. Then the jackknife procedure can be performed as follows:

1. Compute $\hat{\gamma}(n_1, n_2)$ from the samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$.

2. Remove an observation, say $\mathbf{x}_i$.

3. Compute $\hat{\gamma}$ without $\mathbf{x}_i$, refer to this as $\hat{\gamma}_{-i}(n_1 - 1, n_2)$.

4. Repeat 2 and 3 for all n_1 observations of $\mathbf{x}$.

5. Repeat 2-4 for all n_2 observations in the second set, removing $\mathbf{y}_j$.

There are now $n_1 + n_2$ estimates of the measure.

These estimates can be used to compute the variance of the estimator and also to adjust for its bias. Approximate symmetric confidence intervals can then be constructed using a normal approximation (Miller [33]).

The exact expression of the sampled jackknife estimator and the sample variance are as follows. Compute, for $i = 1, \ldots, n_1$,

$$\hat{\gamma}_i(n_1 - 1, n_2) = (n_1 - 1/2)\hat{\gamma}(n_1, n_2) - (n_1 - 1)\hat{\gamma}_{-i}(n_1 - 1, n_2), \qquad [5.1]$$

where $\hat{\gamma}_{-i}(n_1 - 1, n_2)$ is the estimated overlap with the ith individual removed from sample **x** , and is calculated from a sample of $n_1 - 1$ individuals. Similarly, for $j = 1, \ldots, n_2$,

$$\hat{\gamma}_j(n_1, n_2 - 1) = (n_2 - 1/2)\hat{\gamma}(n_1, n_2) - (n_2 - 1)\hat{\gamma}_{-j}(n_1, n_2 - 1). \qquad [5.2]$$

These are the *pseudovalues* of the jackknife estimator for the sample set **x** and **y**. Let

$$\bar{\gamma}_1(n_1 - 1, n_2) = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{\gamma}_i(n_1 - 1, n_2) \qquad [5.3]$$

and

$$\bar{\gamma}_2(n_1, n_2 - 1) = \frac{1}{n_2} \sum_{i=1}^{n_2} \hat{\gamma}_i(n_1, n_2 - 1). \qquad [5.4]$$

The two-sample jackknife estimator is then

$$\gamma_J(n_1, n_2) = \bar{\gamma}_1(n_1 - 1, n_2) + \bar{\gamma}_2(n_1, n_2 - 1). \qquad [5.5]$$

The sample variance of $\bar{\gamma}_1(n_1 - 1, n_2)$ is

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} [\hat{\gamma}_i(n_1 - 1, n_2) - \bar{\gamma}_1(n_1 - 1, n_2)]^2 \qquad [5.6]$$

and the sample variance of $\bar{\gamma}_2(n_1, n_2 - 1)$ is

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} [\hat{\gamma}_j(n_1, n_2 - 1) - \bar{\gamma}_2(n_1, n_2 - 1)]^2. \qquad [5.7]$$

Tukey conjectured that the psuedo-values obtained by the jackknife method can be treated as approximately independent, identically distributed, normal random variables; thus the approximate variance of $\gamma_J(n_1, n_2)$ and $\gamma(n_1, n_2)$ is calculated as S_J^2 , which is

$$S_J^2 = \frac{1}{n_1} S_1^2 + \frac{1}{n_2} S_2^2. \qquad [5.8]$$

If $E[\hat{\gamma}(n_1, n_2)] = \gamma + \frac{a}{n_1} + \frac{b}{n_2} + \frac{c}{n_1 n_2} + O(\frac{1}{n_1^2}, \frac{1}{n_2^2})$,

then the above jackknife procedure removes the bias associated with $\frac{1}{n_1}$ and $\frac{1}{n_2}$.

Confidence intervals based on jackknife estimates of the affinity measure or the overlap measure are calculated by assuming that the sampling distribution

of γ is approximately normal. The limits of a 95% confidence interval calculated for jackknife estimates are thereby

$$\gamma_J \pm (t_{n_1 + n_2 - 2, 0.975})S_J, \qquad [5.9]$$

where n_1, n_2 are the sample sizes of the original samples **x** and **y**.

5.2 TWO-SAMPLE BOOTSTRAP METHOD

The bootstrap method (Efron [8]) is similar to the jackknife but can be used to obtain estimates of confidence intervals that do not require a normality assumption. The basic idea is to simulate the properties of the statistic γ by sampling from an empirical estimate of the underlying probability distribution. Assume that the data consists of two groups of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$. Let the affinity measure or the overlap measure be given by $\gamma = \gamma(\mathbf{x}, \mathbf{y})$. The method for the two-sample case is

1. Compute $\hat{\gamma}$ from the samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$.

2. Take a random sample with replacement of size n_1 from the **x** data and a random sample with replacement of size n_2 from the **y** data.

3. Compute the *lth bootstrap replicate* $\hat{\gamma}_l$, using this new set of samples.

4. Repeat the process k times, i.e. for $l = 1, \ldots, k$, to obtain k bootstrap replicates $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$.

5. Compute the mean $\gamma_B = \bar{\gamma}_k = \frac{1}{k}\sum_{l=1}^{k}\hat{\gamma}_l$ and the variance $S_B^2 = \frac{1}{k-1}\sum_{l=1}^{k}(\hat{\gamma}_l - \gamma_B)^2$.

These are the bootstrap estimates.

Let γ denote the unknown affinity measure or the unknown overlap measure for the populations under consideration. If $\hat{\gamma}$ is a biased estimate of γ, then γ_B will generally be biased because it estimates $\hat{\gamma}$, not γ. The bias of $\hat{\gamma}$ is defined as

$$bias(\hat{\gamma}) = E(\hat{\gamma}) - \gamma,$$

which can be estimated by $\widehat{bias}(\hat{\gamma}) = \gamma_B - \hat{\gamma}$. Therefore, a bias-adjusted bootstrap estimate of γ can be calculated as follows:

$$\gamma_{B,adj} = \hat{\gamma} - (\gamma_B - \hat{\gamma}) = 2\hat{\gamma} - \gamma_B .$$

Because

$$bias(\gamma_{B,adj}) = E(\gamma_{B,adj}) - \gamma = E(\hat{\gamma} - (\gamma_B - \hat{\gamma})) - \gamma$$

$$= (E(\hat{\gamma}) - \gamma) - E(\gamma_B - \hat{\gamma}) = 0.$$

Confidence intervals surrounding bootstrap estimates of the affinity measure or the overlap measure are calculated by assuming that the sampling distribution

of γ would be approximately normal, so that limits of a 95% confidence interval calculated for the bias-adjusted bootstrap estimates are

$$\gamma_{B,adj} \pm (t_{n_1 + n_2 - 2, 0.975})S_B, \qquad [5.10]$$

where n_1, n_2 are the sample sizes in the original samples **x** and **y**.

The percentile method also may be applied (Efron [8]), but is not considered here though because of the expense in computing.

Chapter VI

SIMULATION RESULTS FOR SAMPLE MEASURES

We are concerned with a theoretical investigation of the estimators of various measures. The mathematical difficulties in deriving the properties of the estimators are formidable, and consequently the properties must be evaluated primarily by Monte Carlo methods. In fact, it would appear there is no article of a theoretical nature on the estimation of measures for the most general case of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ all unknown, the case which is the most likely to occur in practice.

This chapter describes the results of a simulation study to assess the sampling properties of the measures and the estimation methods. The simulation program generated samples of various size from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, using IMSL subroutines [20], computed the measures, and the process repeated 100 times. The jackknife procedure was used at each step to estimate the variance and adjust for

bias, and the bootstrap procedure was replicated 100 times at each step to estimate the variance and adjust for bias. Symmetric 95% confidence intervals were formed using the estimated variances. The simulation study focused on

1. Bias and the adjustment for bias.

2. Variance of measures and precision of the jackknife and bootstrap methods.

3. Coverage of the confidence intervals.

Parameters that were varied in all of the studies include the covariance matrices, Σ_1, Σ_2, the number of variables, p, the differences between the means, $\mu_1 - \mu_2$, and the sample sizes, N_1, N_2.

For two random vectors **X** and **Y**, we start the simulation with two bivariate normal densities by setting 5 different values for $\mu_1 - \mu_2$ as (0,0); (0.90,0.90); (1.36,1.36); (1.82,1.82) and (2.41,2.41). The variance-covariance matrices are paired as

$$(i)\quad \Sigma_1 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

$$(ii)\quad \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

$$(iii)\quad \Sigma_1 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

(*iv*) $\Sigma_1 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

The ratio of generalized variance (RD) is the determinant of Σ_1 over the determinant of Σ_2; the four RDs are 1.0, 2.08, 2.67 and 1.28 respectively. We also looked at the case with $\sigma_{x1} = 1$, 3 and 5. For example, when $\sigma_{x1} = 3$, (iii) is changed to

$$\Sigma_1 = \begin{bmatrix} 9 & 0.6 \\ 0.6 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

and RD = 24. There are thus two ways used to investigate covariance effects. The sample sizes are combinations of 10, 20 and 30. When the dimension is 4, we add two more uncorrelated variables into the above bivariate densities. These variables have same mean for the extraneous variables, for example, $\mu_1 - \mu_2 =$ (1.36,1.36,0.0,0.0). We have a similar setting when the dimension is 8. The effects of extraneous variables is of great interest as many researchers measure many variables. If unimportant variables are included, how will this affect bias and variance estimates? In the first four sections, Matusita's measure is primarily used to demonstrate the effects of those factors. The results for other measures are quite similar. Some of the differences are described in the later section. We will present the results in three parts based on the three criteria: bias and adjustment of bias, the estimated variance and the coverage of 95% confidence intervals.

6.1 BIAS AND ADJUSTMENT OF BIAS

Our interest in this section is the degree to which the estimator of ρ^* (EQ. [3.6]) is biased and how well the jackknife and the bootstrap procedures adjust for the bias. Because the measure ρ^* is bounded between 0 and 1, bias is expected, especially near 1. As will be seen, the bias may be as large as 0.50. In all the simulations, the bias was negative indicating underestimation of the measure. Thus bias becomes an important consideration in the use of this measure. It is also important to investigate the influence of the parameters Σ_1, Σ_2, the mean separation, $\mu_1 - \mu_2$ and the sample sizes N_1, N_2 on bias.

6.1.1.The ratio of the generalized variances and different dimensions

For the sample sizes $N_1 = 20, N_2 = 20$, we compare the bias of the unadjusted estimator $\hat{\rho}^*$ from (EQ. [3.6]) to different ratios of the generalized variances (1.0, 1.28, 2.08 and 2.67) in figure 6. As indicated in figure 6, in the case of 2 or 4 variables, the slight disparity of the variance-covariance matrices has no impact on the bias of $\hat{\rho}^*$. There is a slight effect of the ratio of the generalized variances on the bias of $\hat{\rho}^*$ when 8 variables are involved. In fact, when there are 2 variables, the bias of $\hat{\rho}^*$ is around -0.05; when the number of variables increases to 8, the bias increases 5 to 10 times (dramatically). It can be seen that the mean separation influences heavily the bias of ρ^* when there are 8 variables involved, and affects slightly the bias of ρ^* when there are a few variables in the model. This suggests that for the same ratio of generalized variances, adding the extra-

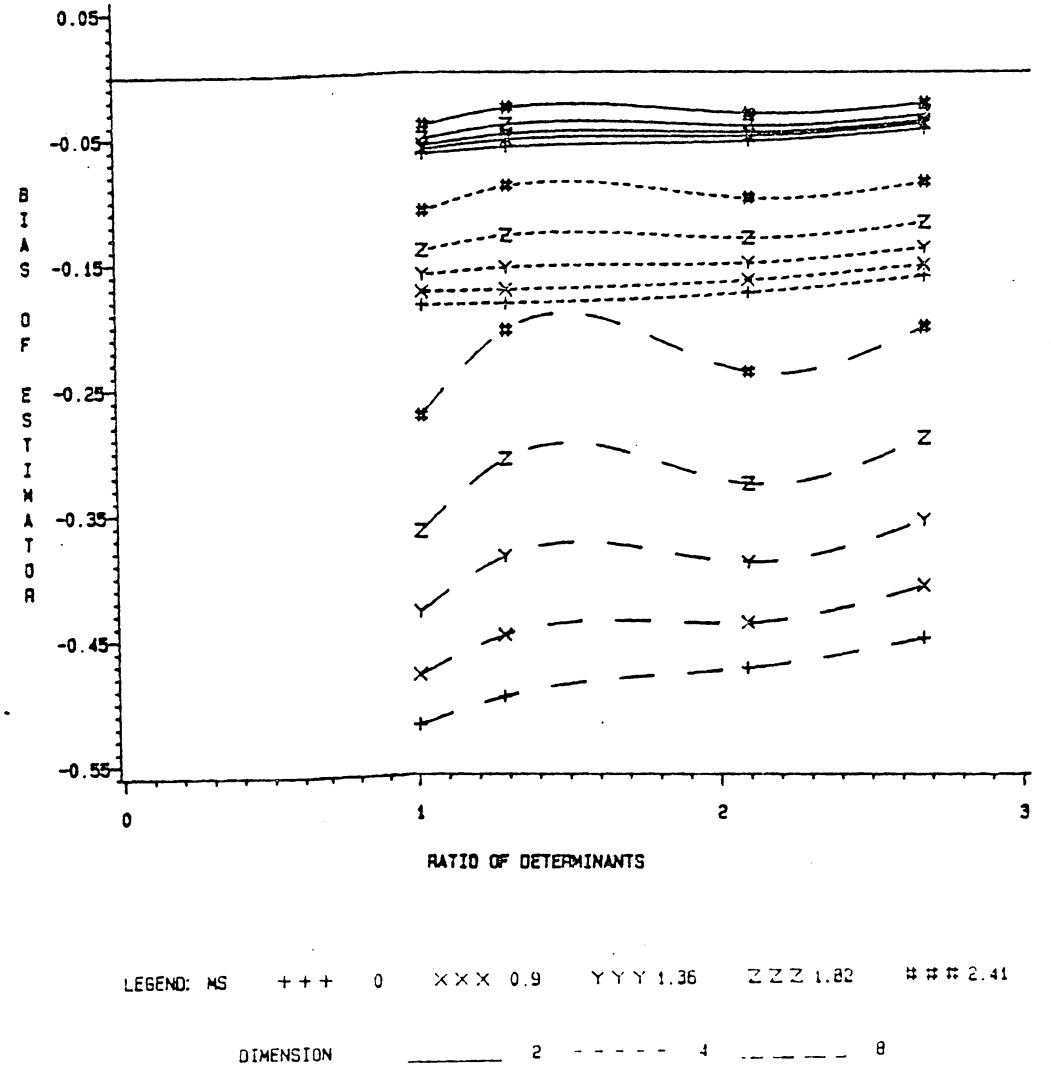

Figure 6. **Bias versus RD:** Bias of the unadjusted measure $\hat{\rho}^*$ versus the ratio of the generalized variances (RD) for different values of mean separation and dimensions, $N_1 = 20$, $N_2 = 20$.

neous uncorrelated variables to estimate ρ^* will dramatically increase the bias relative to the mean separation.

From figures 7 and 8, the use of the jackknife method to reduce the bias of $\hat{\rho}^*$, works extremely well. The bootstrap method does not work as well. The reduction in bias is good regardless of the different ratios of the generalized variances. When we consider the number of variables involved in the model, we see that the reduction in bias using the bootstrap method is not as good as the reduction using the jackknife method. Generally, the jackknife method is more advantageous than the bootstrap method in reducing the bias of $\hat{\rho}^*$, especially with the sample sizes considered in this study.

6.1.2.The theoretical measure ρ^* and unequal samples: In figure 9 we graph the bias of the estimator $\hat{\rho}^*$ versus ρ^* for sample sizes 10,10; 10,20 and 10,30 and different ratios of the generalized variances (1.0, 1.28, 2.08 and 2.67), when the number of variables is 2 (the bivariate case). When ρ^* is close to 1, the bias of $\hat{\rho}^*$ is large. As ρ^* decreases, the bias of $\hat{\rho}^*$ is decreases gradually. When the ratio of the generalized variances changes from 1.0 to 2.67, ρ^* decreases and the bias of $\hat{\rho}^*$ is reduced by 30%. When the sample size increases from 10 to 30 in one population, the bias of $\hat{\rho}^*$ is reduced by 50%. This indicates that when ρ^* is close to 1, one needs to obtain better estimates of the variance-covariance matrices and also collect larger samples.

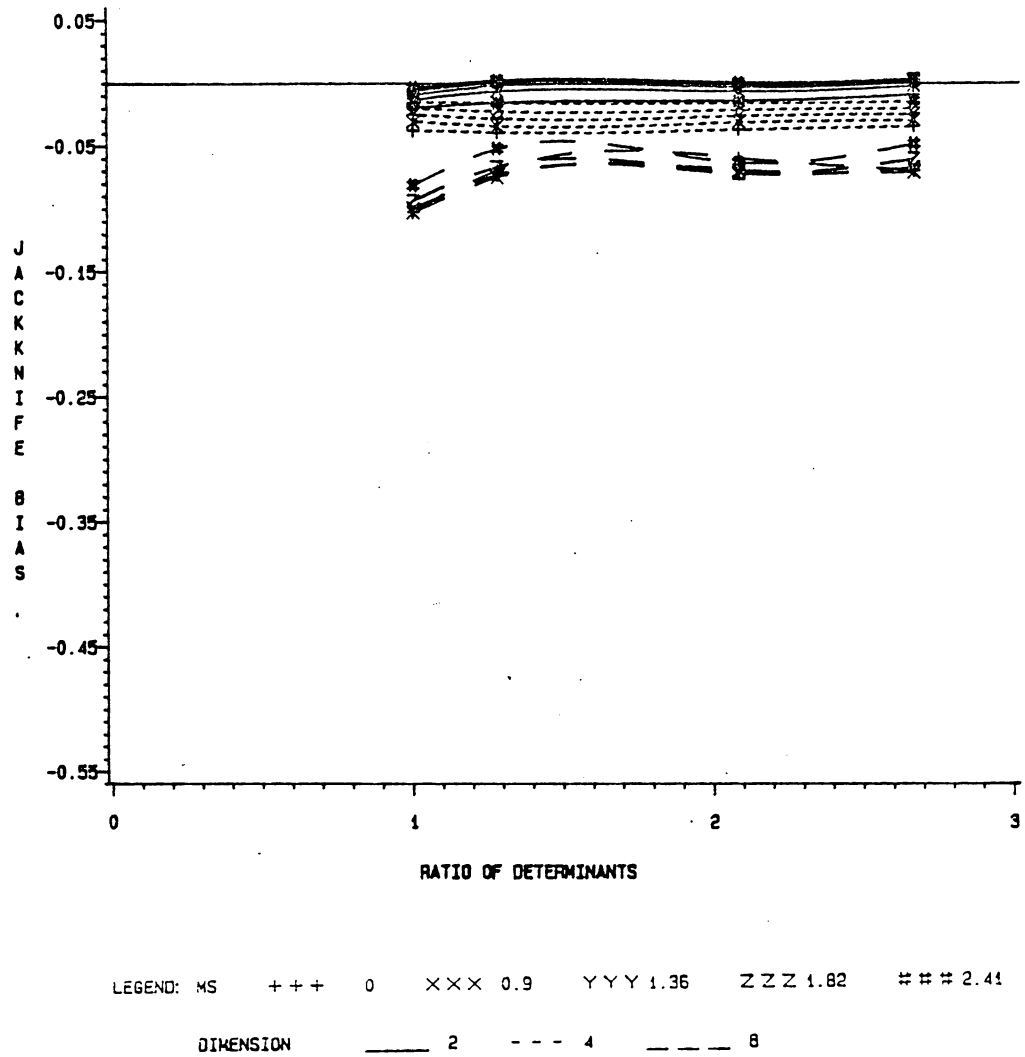

Figure 7. **Jackknife bias versus RD:** Bias of the jackknife estimate ρ_j^* versus the ratio of the generalized variance (RD) for different values of mean separation and dimensions, $N_1 = 20$, $N_2 = 20$.

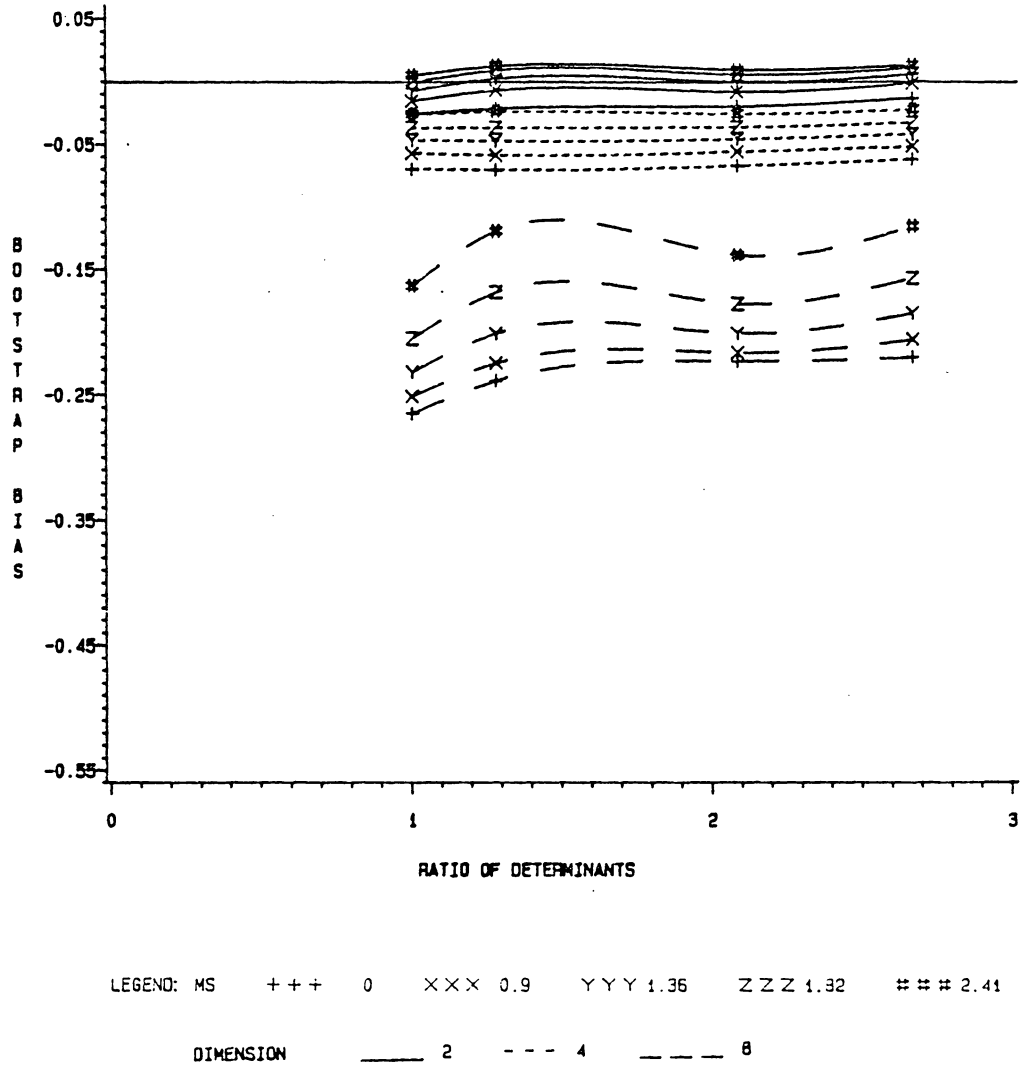

Figure 8. Bootstrap bias versus RD: Bias of the adjusted bootstrap estimate $\rho^{*}_{B,adj}$ versus the ratio of the generalized variance (RD) for different values of mean separation and dimensions, $N_1 = 20$, $N_2 = 20$.

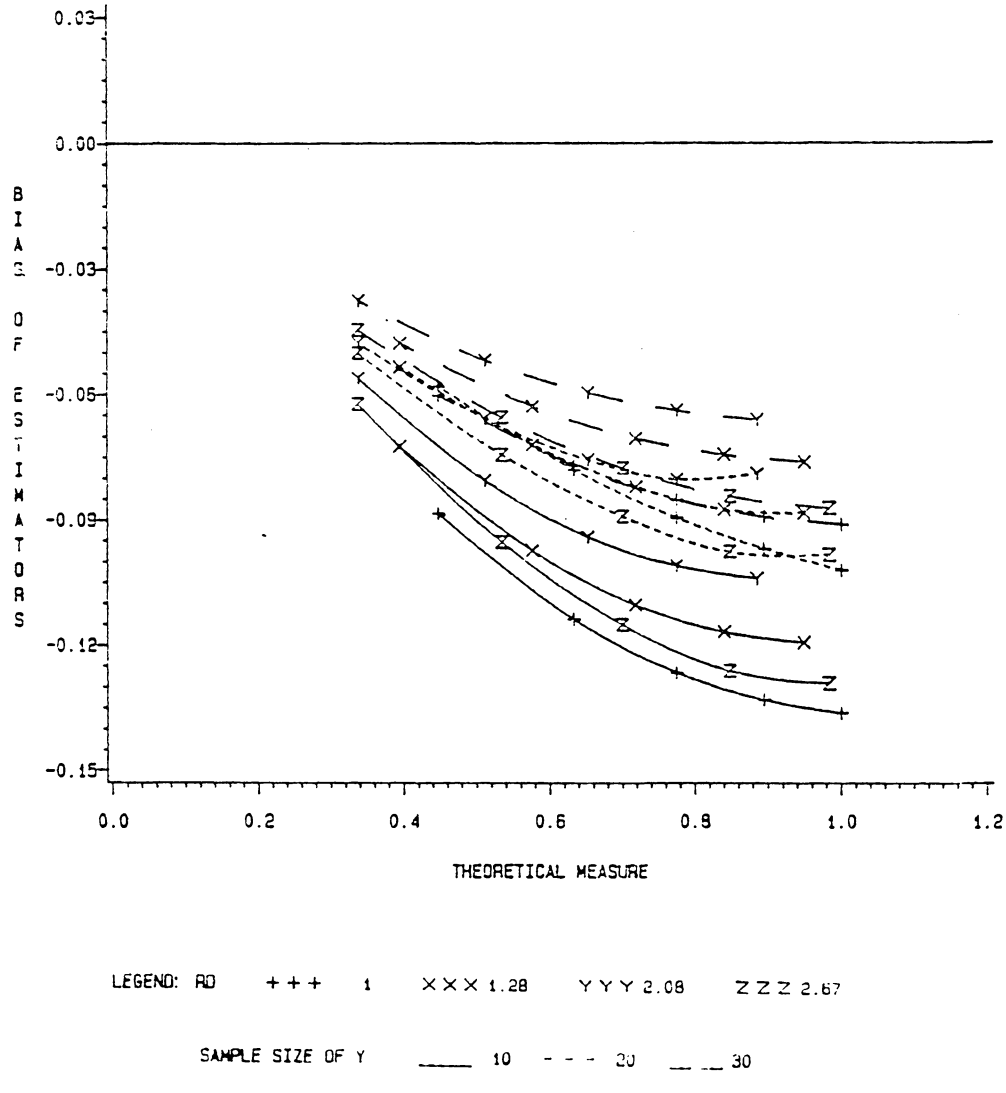

Figure 9. Bias versus Theoretical measure: Bias of the unadjusted estimate $\hat{\rho}^*$ versus ρ^* for different sample sizes of Y, with $N_2 = 10, 20, 30$, and various ratios of the generalized variances (RD). here $N_1 = 10$ and the dimension = 2.

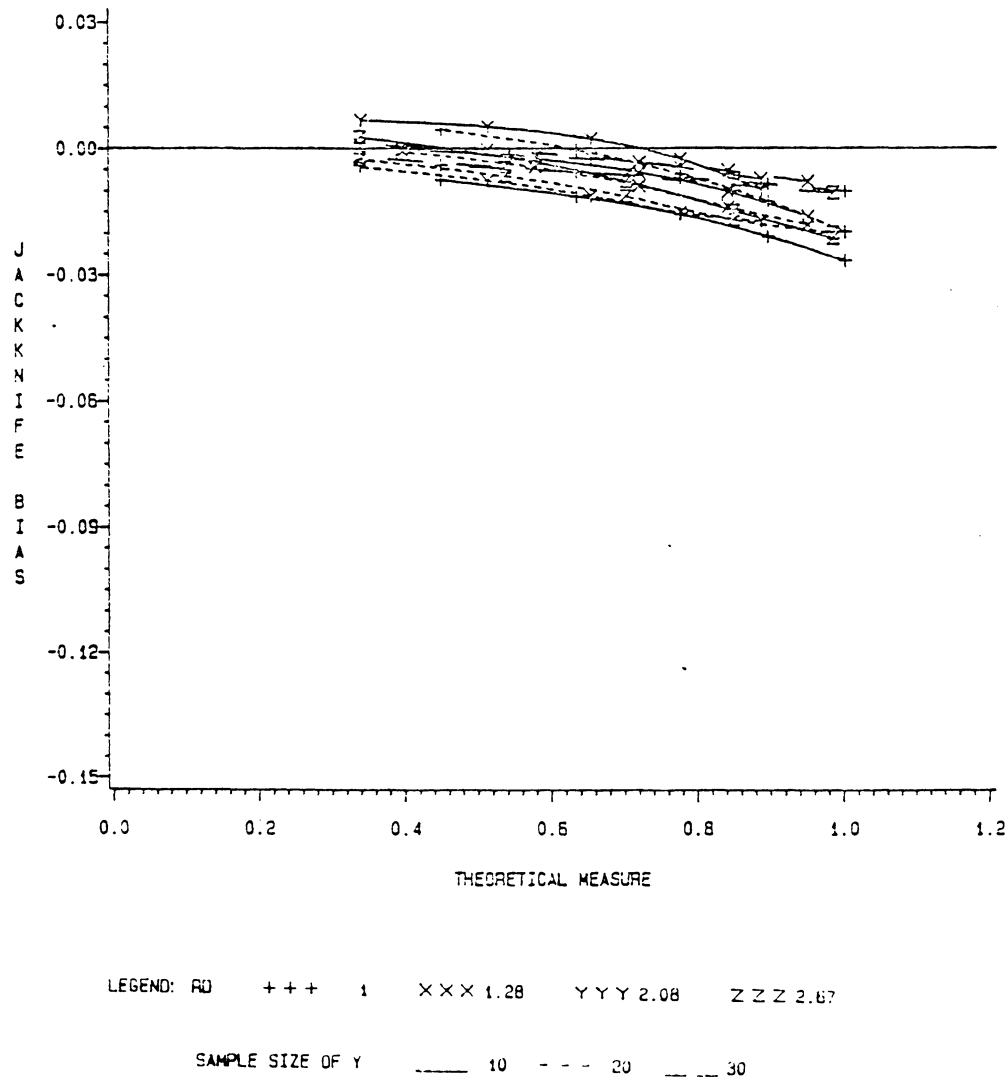

Figure 10. Jackknife bias versus Theoretical measure: Bias of the jackknife estimate ρ_J^* versus ρ^* for different sample sizes of Y, with $N_2 = 10, 20, 30$, and various ratios of the generalized variances (RD). Here $N_1 = 10$ and the dimension = 2.

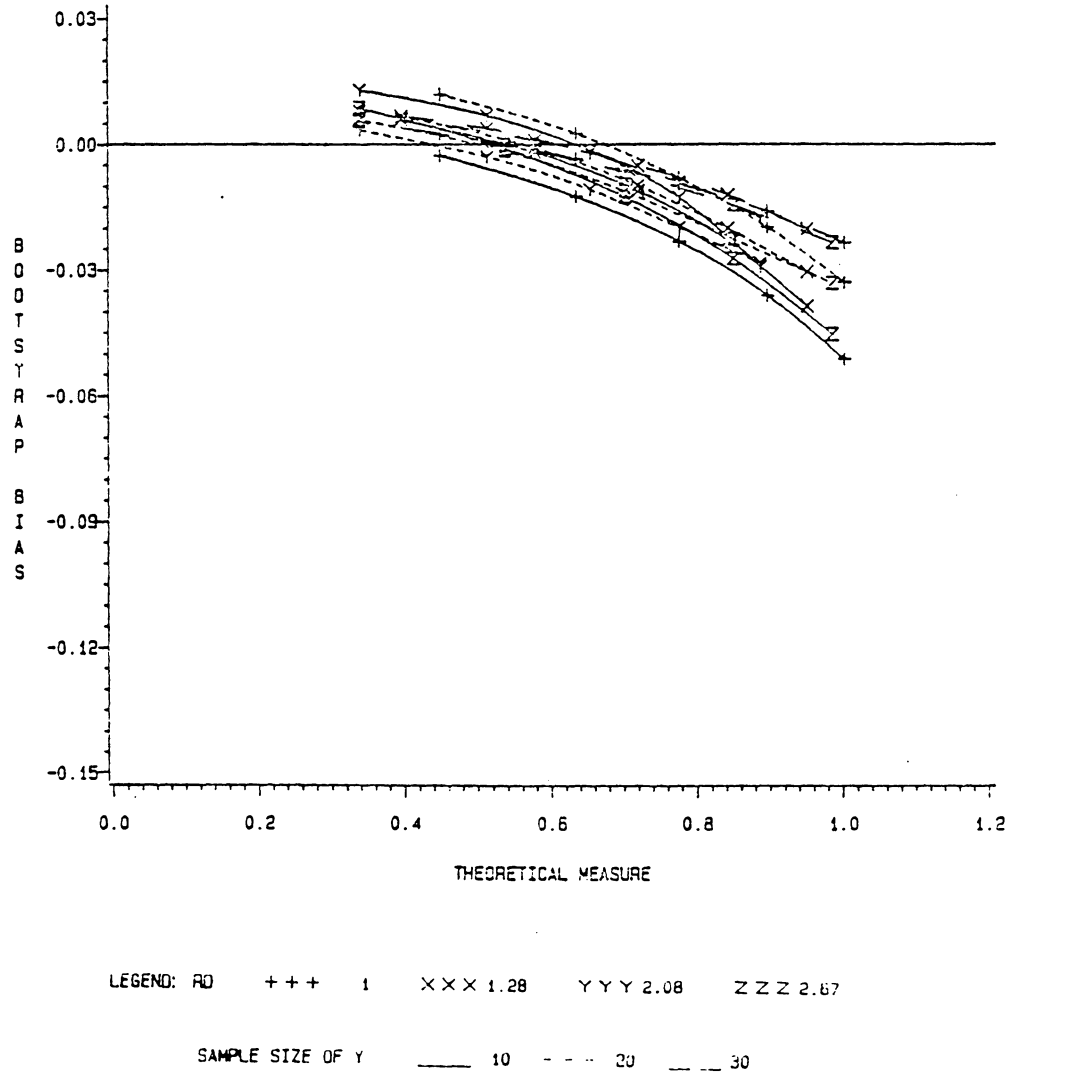

Figure 11. Bootstrap bias versus Theoretical measure (bootstrap): Bias of the adjusted bootstrap estimate $\rho^{*}_{B,adj}$ versus ρ^{*} for different sample sizes of Y, with $N_2 = 10, 20, 30$, and various ratios of the generalized variances (RD). Here $N_1 = 10$ and the dimension = 2.

From figures 10 and 11, the jackknife and the bootstrap methods are seen to satisfactorily reduce the bias of $\hat{\rho}^*$. As ρ^* becomes close to 1, the jackknife method is again seen to perform better than the bootstrap method.

6.1.3.The theoretical measure ρ^* and different variances: For the sample sizes $N_1 = 20, N_2 = 20$, we compare the bias of $\hat{\rho}^*$ for different ratios of the generalized variances (1.0, 1.28, 2.08 and 2.67) and various values of the variances of **X**, σ_{x1} = 1, 3 and 5. For the bivariate case, in figure 12, when ρ^* decreases the bias of $\hat{\rho}^*$ is reduced. As the ratio of the generalized variances changes from 1.0 to 2.67 the bias is reduced 30%. When σ_{x1} changes from 1 to 3, there is a significant reduction of the bias. As σ_{x1} gets larger, a small reduction in the bias can be obtained. The reason for this decrease in bias is that as σ_{x1} increases the value of the theoretical measure decreases and hence the bias decreases. There is thus an interaction between the variance and the theoretical measure. This is also the reason why there is a slant in the curve. Even though the mean separation, for example, for the first points in a connected curve is the same, we note they shift to the right and down as we move down the graph (as the variance σ_{x1} decreases). This slant is associated with the effect of the variance on the measure.

Comparing figures 13 and 14, the jackknife method reduces the bias better than the bootstrap method. The bootstrap method tends to overestimate the measure ρ^* especially when the ratio of the generalized variances is large and σ_{x1} is large.

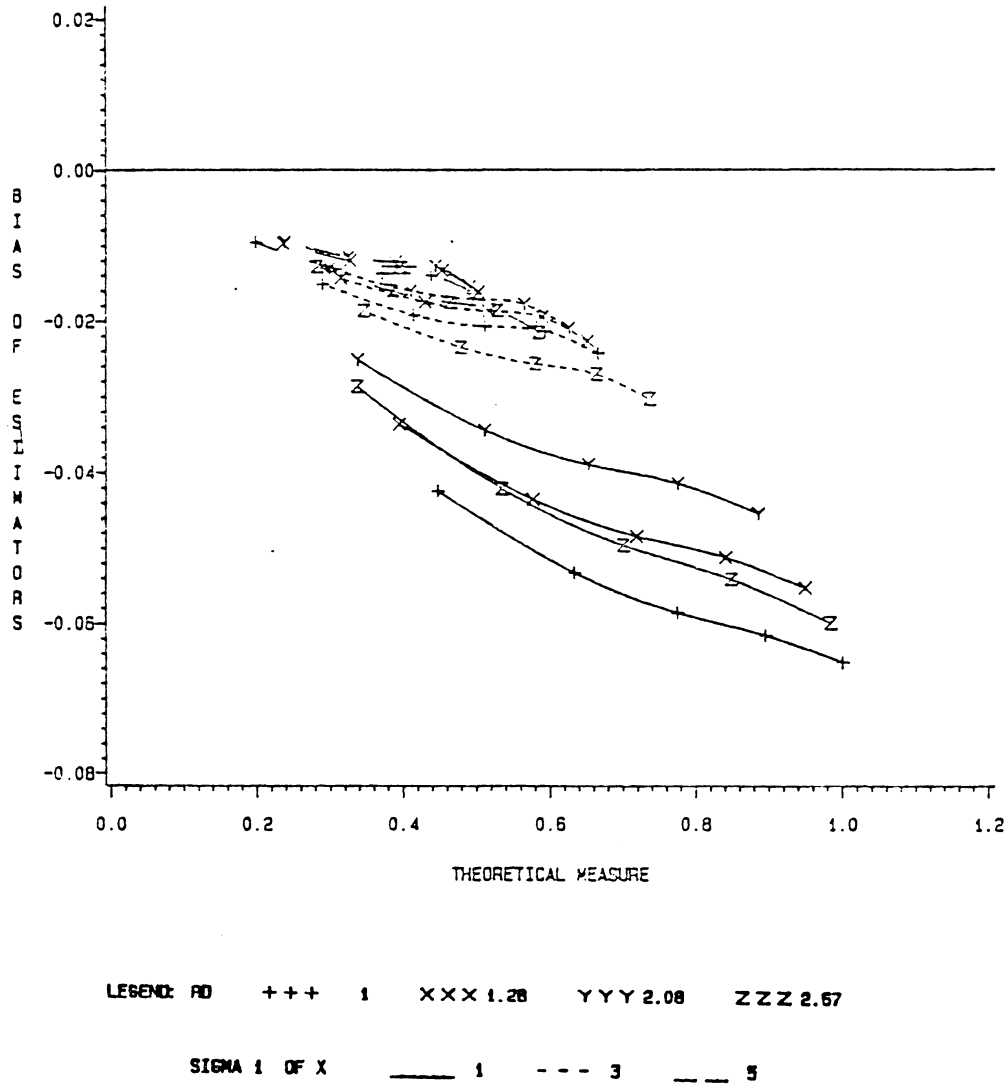

Figure 12. Bias versus Theoretical measure: Bias of the unadjusted estimate of $\hat{\rho}^*$ versus ρ^* for different variances of X, with σ_{x1} = 1 , 3 , 5 and various ratios of generalized variances. Here σ_{y1} = 1 and the dimension = 2.

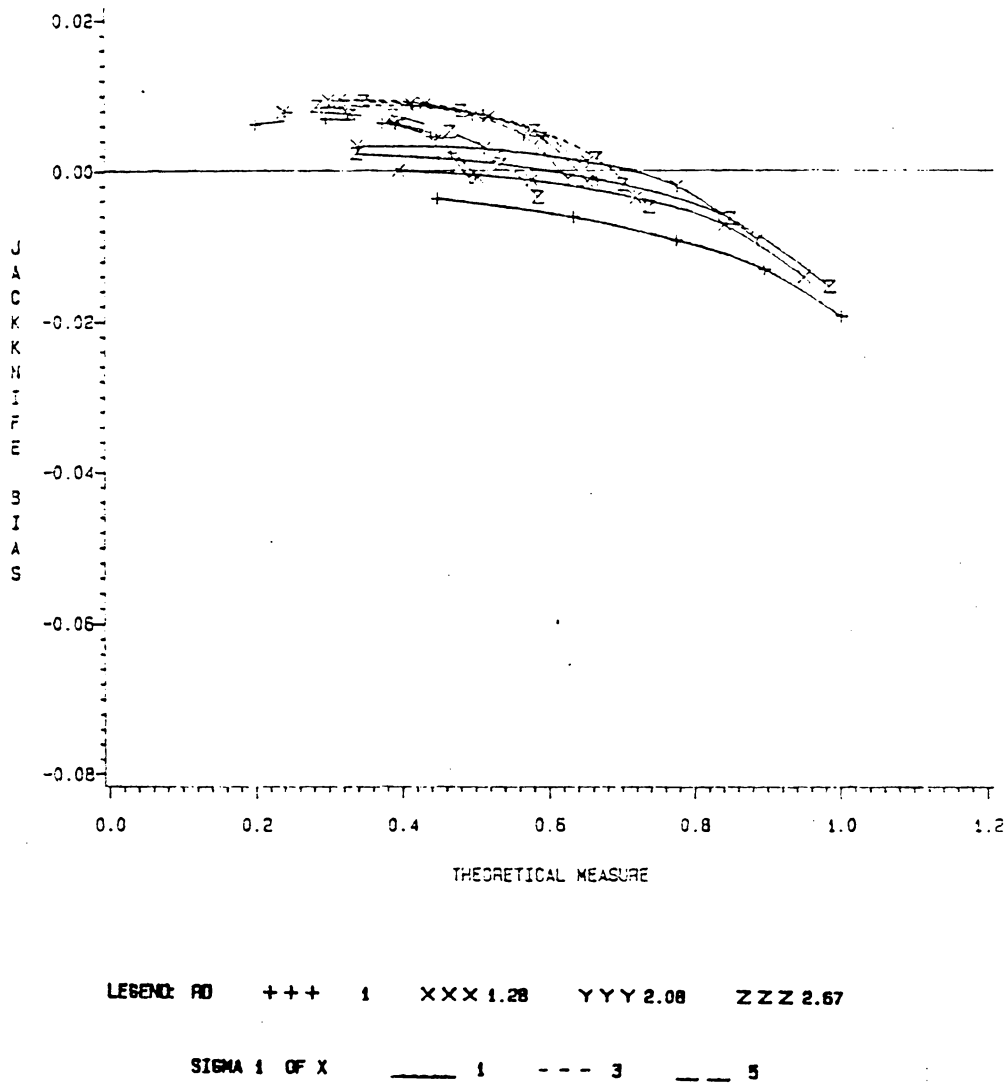

Figure 13. Jackknife bias versus Theoretical measure: Bias of the jackknife estimate ρ_j^* versus ρ^* for different variances of X, with σ_{x1} = 1 , 3 , 5 and various ratios of generalized variances. Here σ_{y1} = 1 and the dimension = 2.

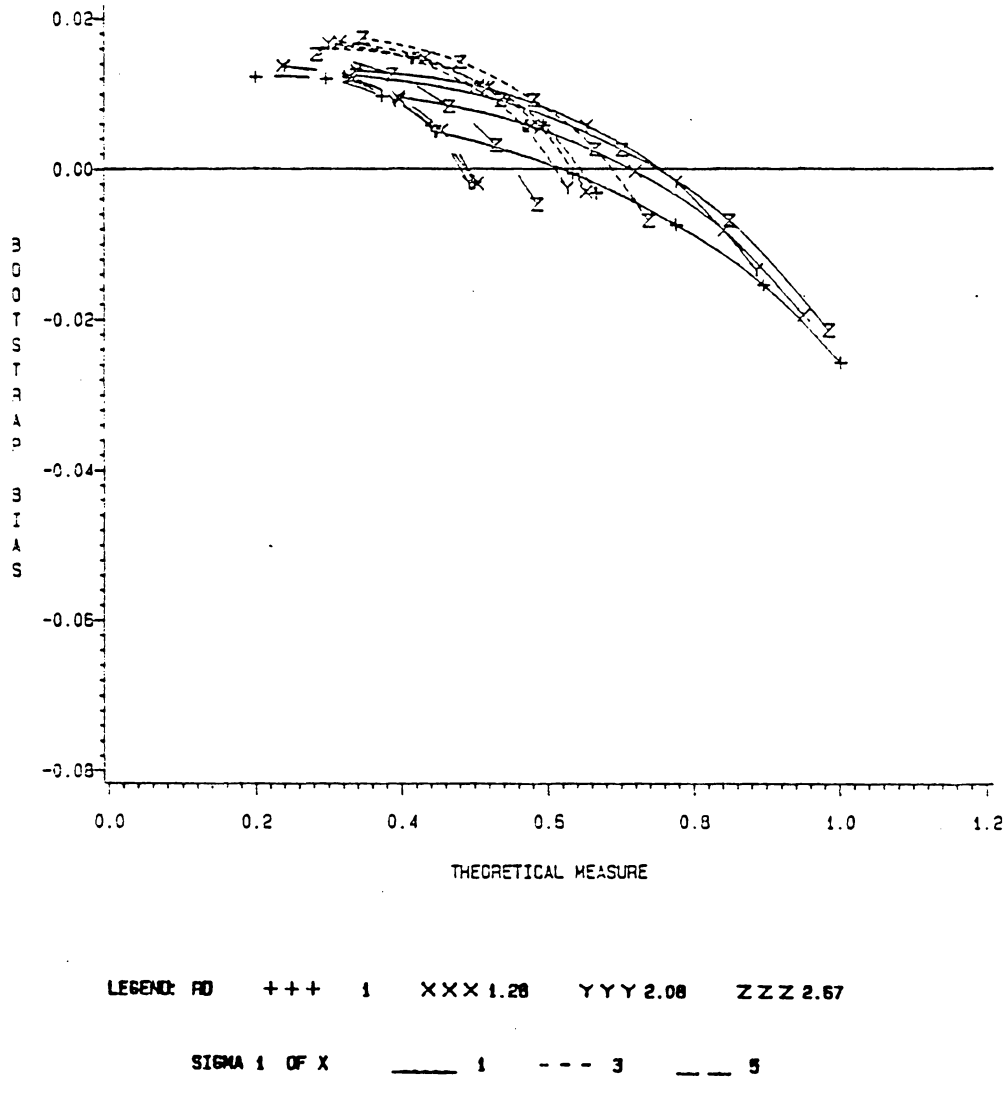

Figure 14. **Bootstrap bias versus Theoretical measure:** Bias of the adjusted bootstrap estimate $\rho^{*}_{B,adj}$ versus ρ^{*} for different variances of X, with σ_{x1} = 1 , 3 , 5 and various ratios of generalized variances. Here σ_{y1} = 1 and the dimension = 2.

6.1.4.The theoretical measure ρ^* for various dimensions: For the sample sizes $N_1 = 20, N_2 = 20$, we compare the bias of $\hat{\rho}^*$ for different numbers of variables (2, 4 and 8) in figure 15. When the number of variables increases from 2 to 4 the bias of $\hat{\rho}^*$ increases roughly 3-fold. When the number of variables increases from 4 to 8 the bias of $\hat{\rho}^*$ increases another 3 times. Also when ρ^* is close to 1, the bias of $\hat{\rho}^*$ is up to 0.50. This is a significant result. The slope and the bias (in magnitude) increases with increasing dimension. The bias of $\hat{\rho}^*$ is a linear function of ρ^* with different slopes for different dimensions. If there are more uncorrelated extraneous variables involved in estimating the measure, the bias of $\hat{\rho}^*$ gets larger regardless of the mean separation and the ratio of the generalized variances.

In figures 16 and 17, we see that the jackknife method may reduce the bias considerately when the number of variables is 2 and the sample sizes are (20,20). Using (20,20) samples to estimate ρ^* with 8 variables in the model, ρ_J^* reduces the bias of $\hat{\rho}^*$ 70%. The bootstrap estimate $\rho^*_{B,adj}$ reduces 50%, but the bias is still moderate. Thus the impact of the dimensionality is quite strong. The sample size (20,20) is sufficient to estimate the measure ρ^* accurately in the 2 variable case. In the 8 variable case, one will not obtain a reasonable estimate by utilizing (20,20) samples. One possible improvement is to use a variable selection procedure and just use the most informative variables to estimate the measure ρ^*. Although in these figures, the dimensionality effect is obtained by adding redundant variables, we note that simulations with non-redundant variables produced similar curves and the bias was not as significant.

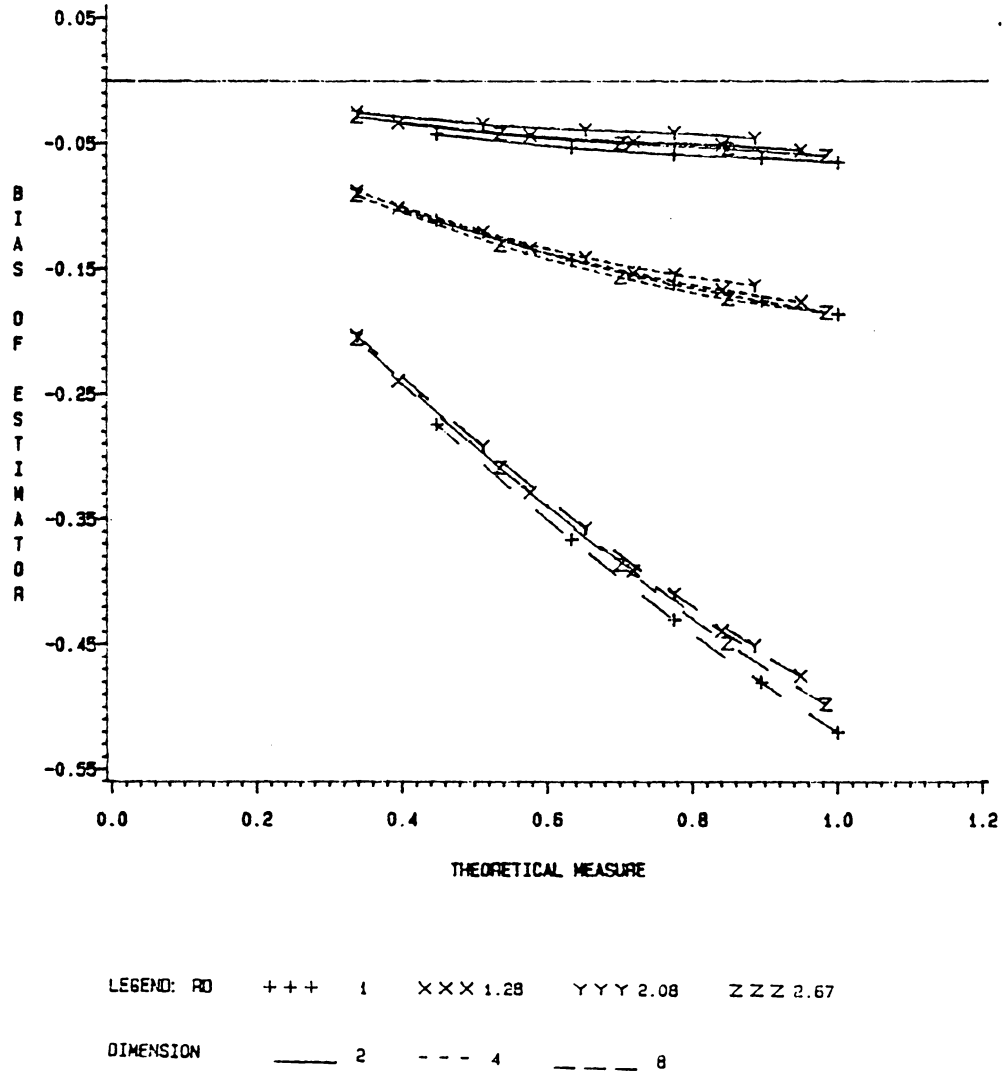

Figure 15. Bias vs Theoretical measure: Bias of the unadjusted estimate $\hat{\rho}^*$ versus ρ^* for dimensions, 2, 4, 8. Here $N_1 = 20$, $N_2 = 20$.

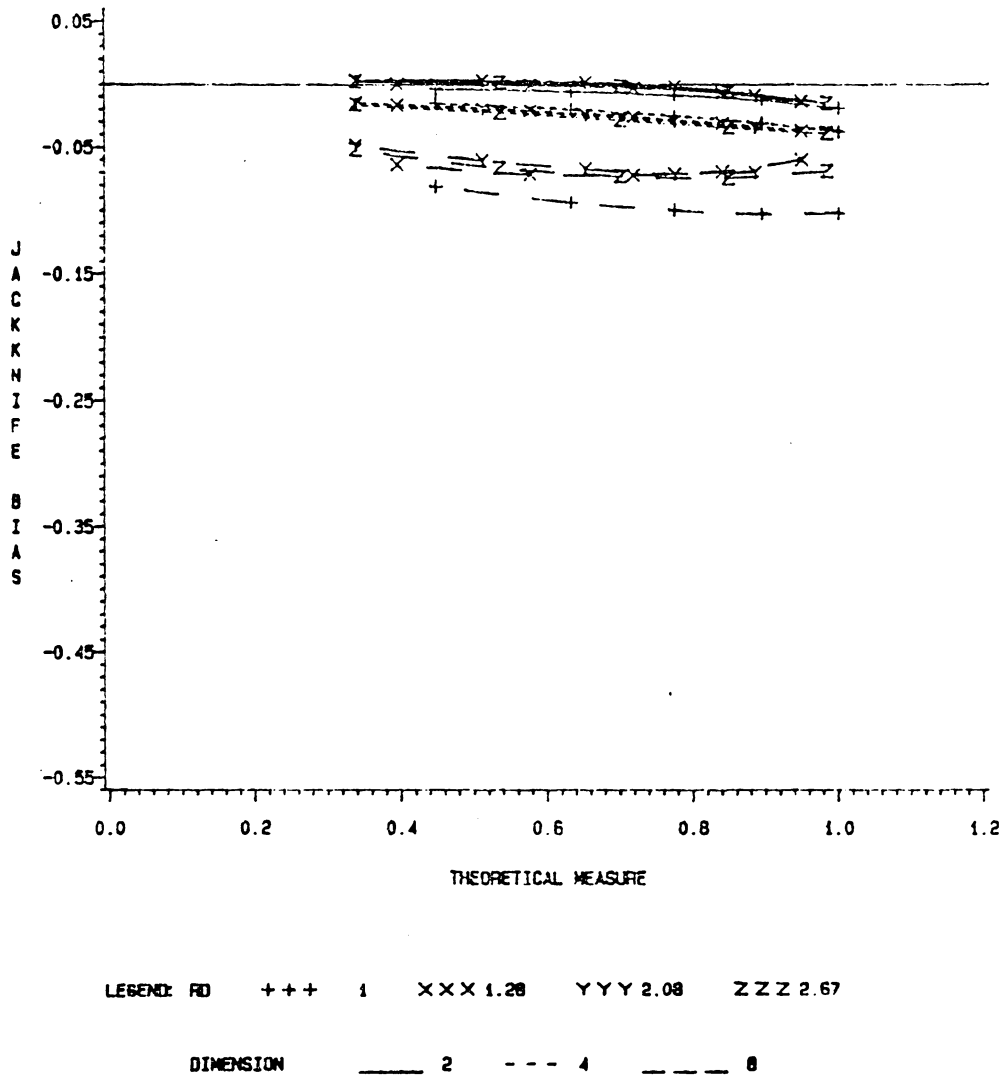

Figure 16. Jackknife bias vs Theoretical measure: Bias of the jackknife estimate $\hat{\rho}_J^*$ versus ρ^* for dimensions, 2, 4, 8. Here $N_1 = 20$, $N_2 = 20$.

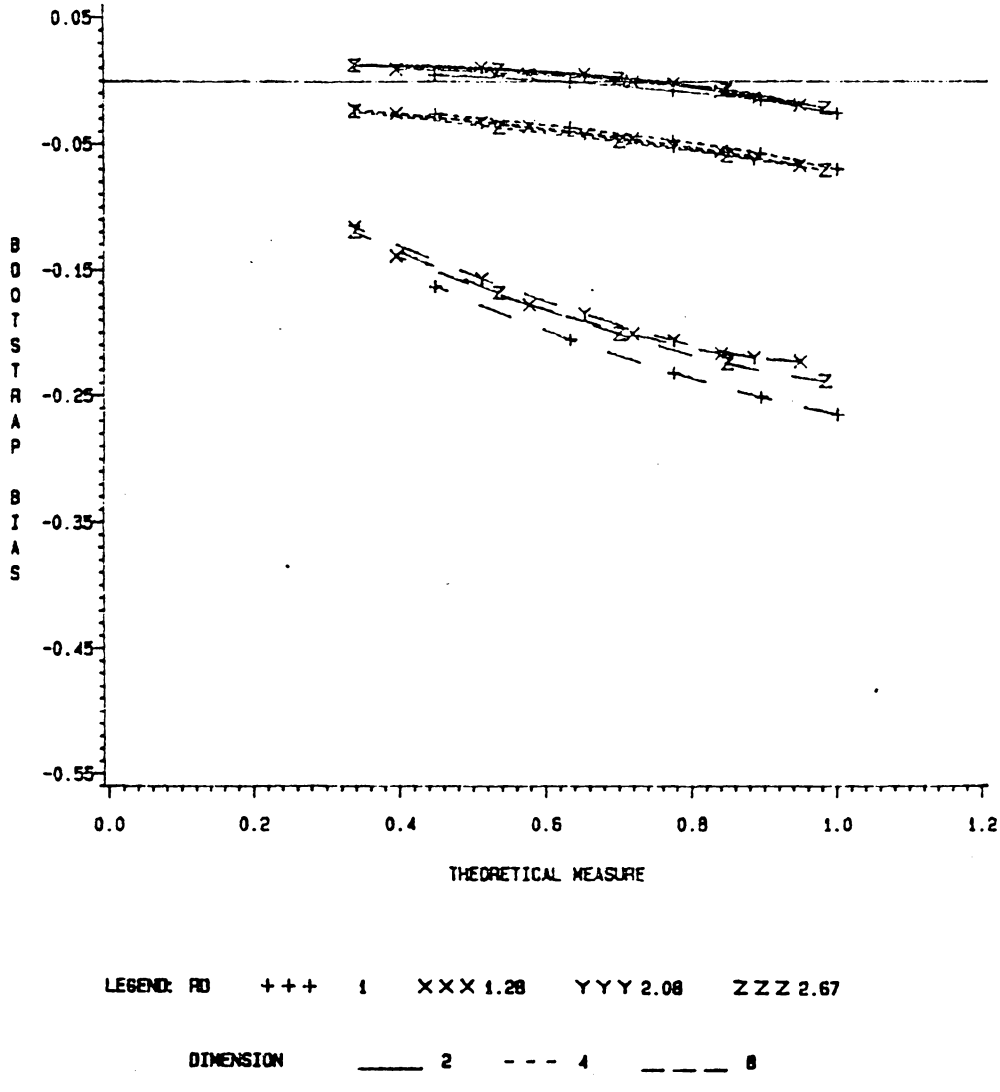

Figure 17. Bootstrap bias vs Theoretical measure: Bias of the unadjusted bootstrap estimate $\rho^*_{B,adj}$ versus ρ^* for dimensions , 2, 4, 8. Here $N_1 = 20$, $N_2 = 20$.

6.1.5.The ratio of the generalized variances for mean separation

For the sample sizes $N_1 = 20$, $N_2 = 20$, we compare the various ratios of the generalized variances to the bias of $\hat{\rho}^*$. In the bivariate case, figure 18 indicates that when the ratio of the generalized variances increases from 1 to 9 the bias of $\hat{\rho}^*$ is reduced sharply up to 50%. The reduction is small when the ratio of the generalized variances gets larger. If the disparity of the variance-covariance matrices is severe, the bias of $\hat{\rho}^*$ tends to decrease gradually. $\hat{\rho}^*$ has smaller bias if it is away from 1 and close to 0. This means that if the variance-covariance matrices are more disparate , ρ^* decreases, and the bias of $\hat{\rho}^*$ is smaller. Whenever the ratio of the generalized variances is small, the mean separation has the larger effect on bias. This kind of impact will be limited as the ratio of the generalized variances increases.

In figure 19 and 20, the jackknife method and the bootstrap method perform equally well in reducing the bias of $\hat{\rho}^*$. They tend to slightly overestimate ρ^* as the ratio of the generalized variances gets larger. This indicates that when there is a disparity in variance-covariance matrices, it is appropriate to apply either method adjusting the bias of $\hat{\rho}^*$ for these sample sizes.

6.2 SIMULATED VARIANCES

In this section, the variance of the estimate of Matusita's measure ρ^* is simulated. The jackknife method and the bootstrap method are used to estimate the variance. Our interest is to examine the performance of these methods in esti-

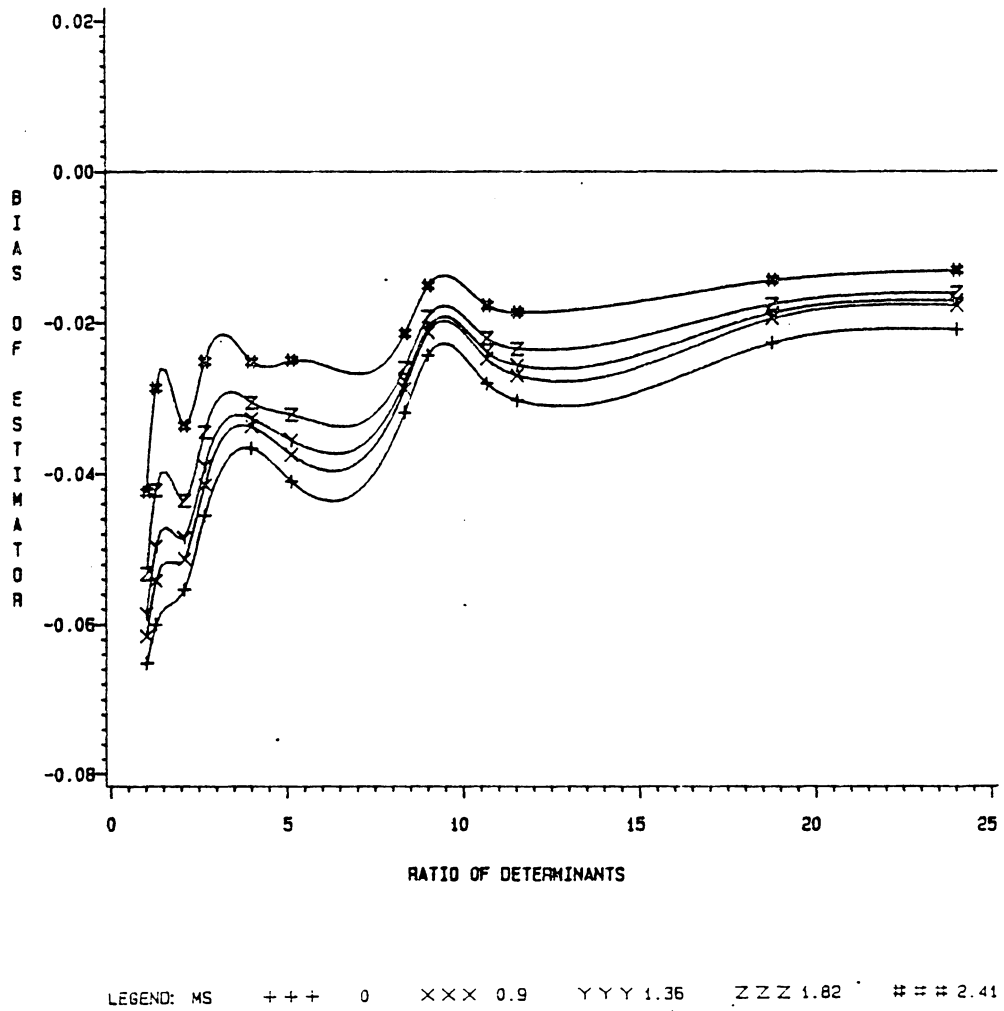

Figure 18. **Bias versus RD:** Bias of the unadjusted estimate $\hat{\rho}^*$ versus the ratio of the generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

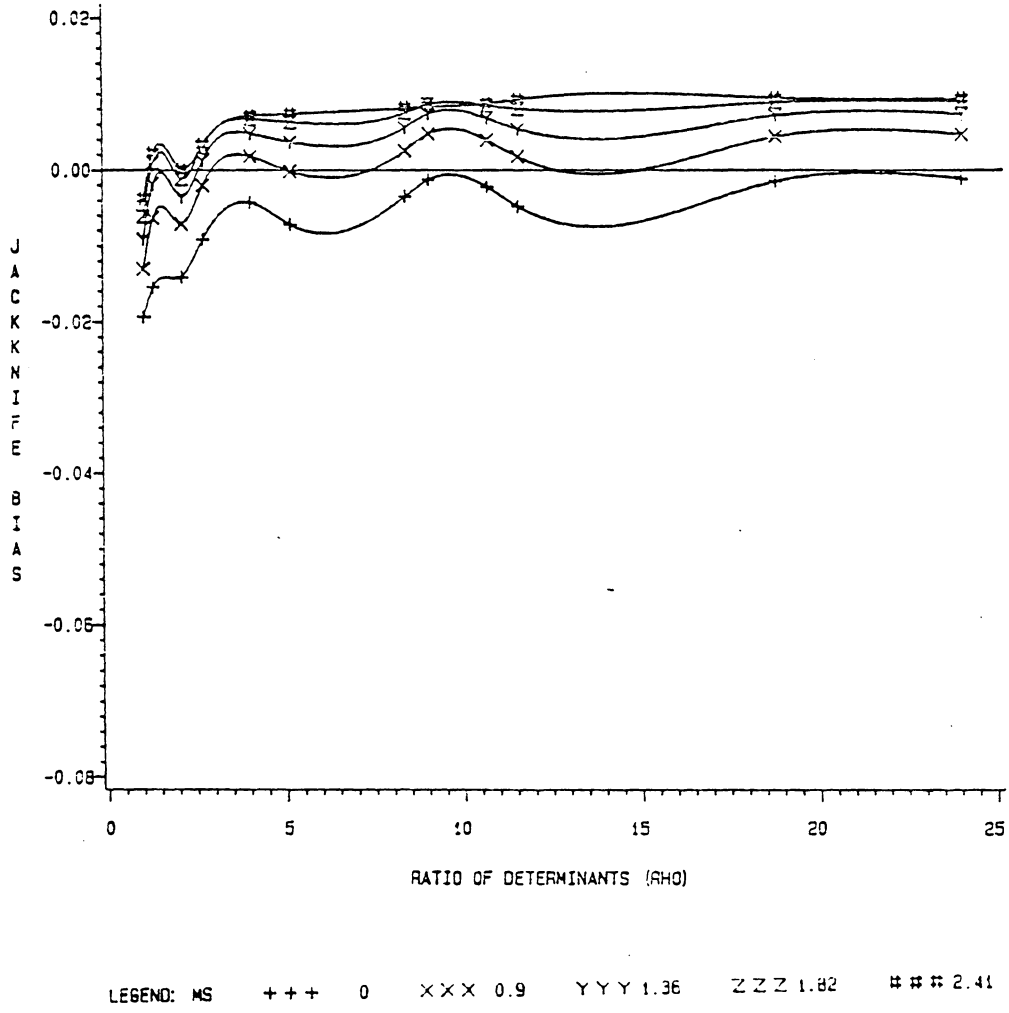

Figure 19. Jackknife bias versus RD: Bias of the jackknife estimate ρ_J^* versus the ratio of the generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

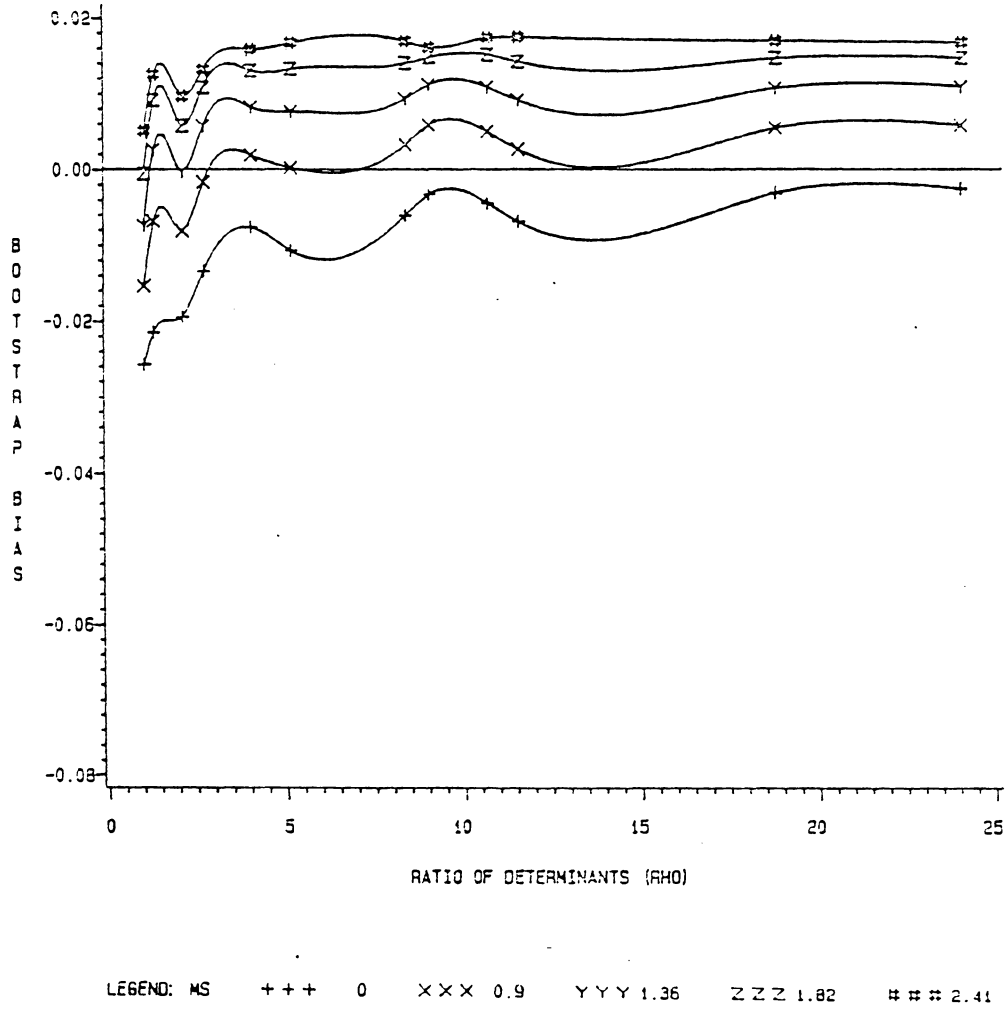

Figure 20. Bootstrap bias versus RD: Bias of the adjusted bootstrap estimate $\rho^{*}_{B,adj}$ versus the ratio of the generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

mating the variance relative to the factors of dimensionality, the ratio of the generalized variances, the mean separation and the sample sizes.

6.2.1.The ratio of the generalized variances and different dimensions

For the sample sizes $N_1 = 20$, $N_2 = 20$, we compare the simulated variance of $\hat{\rho}^*$ to various ratios of the generalized variances (1.0, 1.28, 2.08 and 2.67) for the number of variables (2, 4 and 8) and mean separation in figure 21. It can be seen that the ratio of the generalized variances has almost no effect on the simulated variance. The simulated variance is about the same for the ratio of the generalized variances ranging between 1 and 2.67. When there are 2 variables, the simulated variance of $\hat{\rho}^*$ depends mainly on the mean separation. When the number of variables increases to 8, the simulated variance of $\hat{\rho}^*$ is about the same regardless of the mean separation.

From figures 22 and 23, the use of the jackknife method or the bootstrap method to estimate the variance of $\hat{\rho}^*$ is not good when the number of variables is large. In applying the bootstrap method to estimate the variance of $\hat{\rho}^*$, we see it underestimates the variance. When either the jackknife method or the bootstrap method is adopted, the ratio of generalized variances has little effect on the variance of ρ^*. Generally, the bootstrap method tends to underestimate while the jackknife method tends to overestimate slightly.

6.2.2.The theoretical measure ρ^* for unequal samples: The effects of sample sizes 10,10; 10,20 and 10,30 for different ratios of the generalized variances (1.0, 1.28,

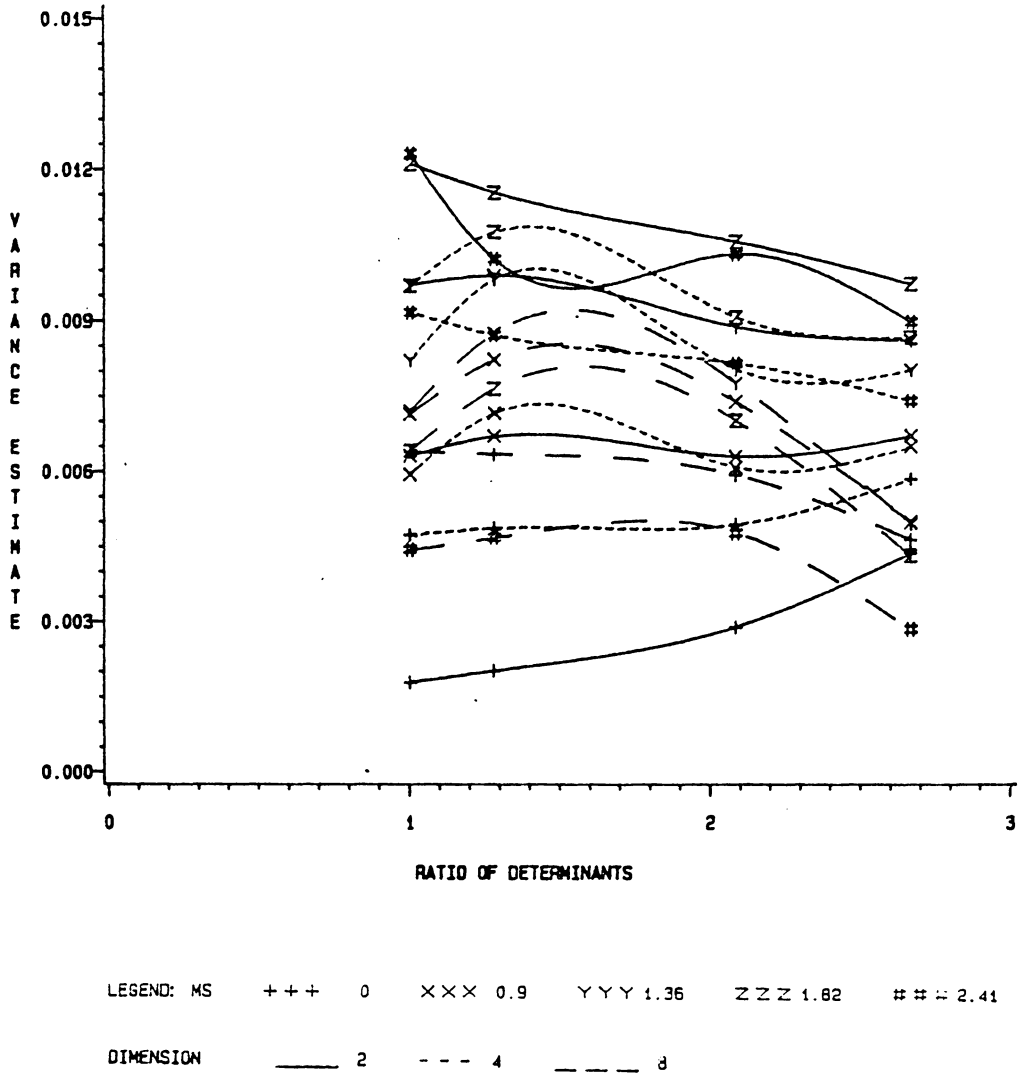

Figure 21. Simulated variance versus RD: Simulated variance of $\hat{\rho}^*$ versus the ratio of the generalized variance for different values of mean separation and dimension. Here $N_1 = 20$, $N_2 = 20$.

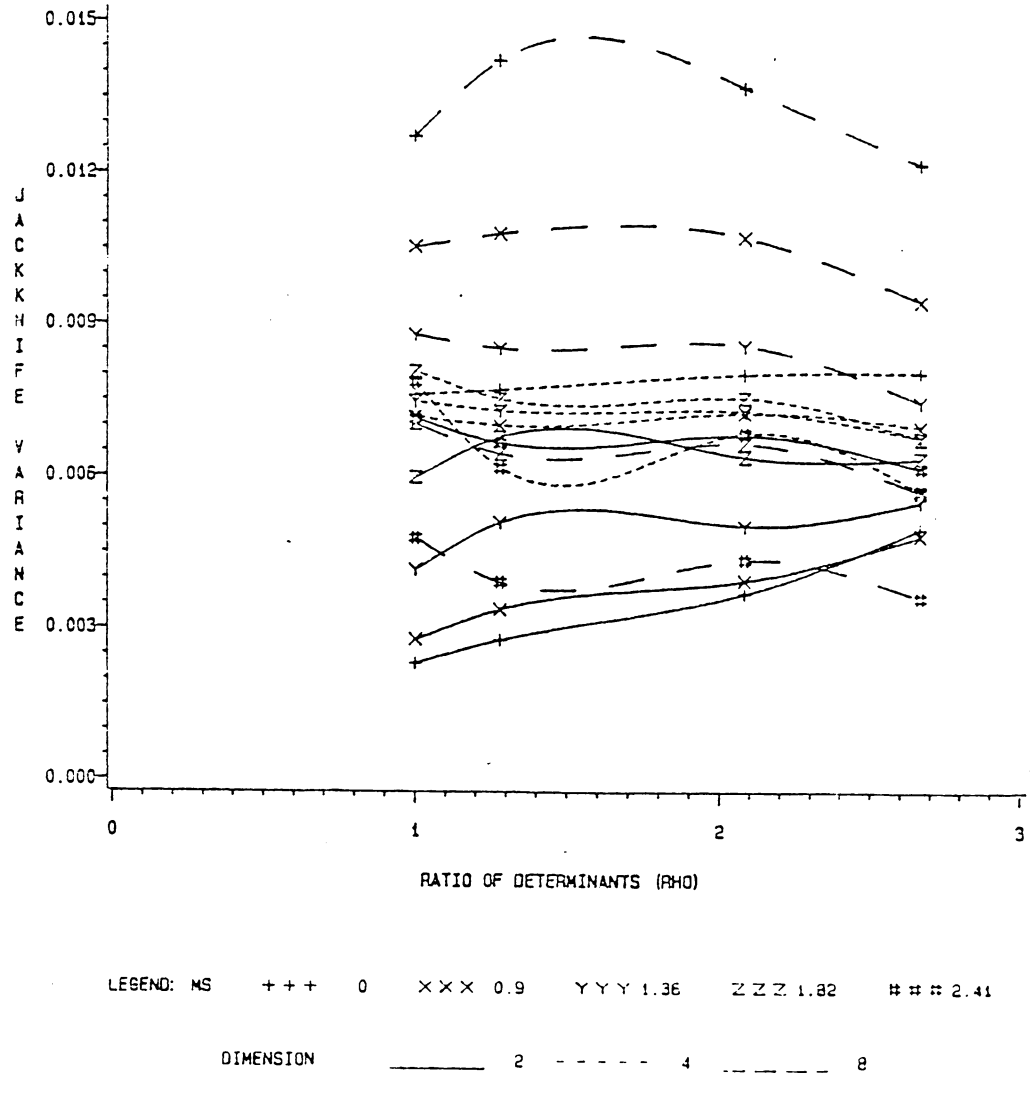

Figure 22. Jackknife variance estimate versus RD: Jackknife variance estimate of $\hat{\rho}^*$ versus the ratio of the generalized variance for different values of mean separation and dimension. Here $N_1 = 20$, $N_2 = 20$.

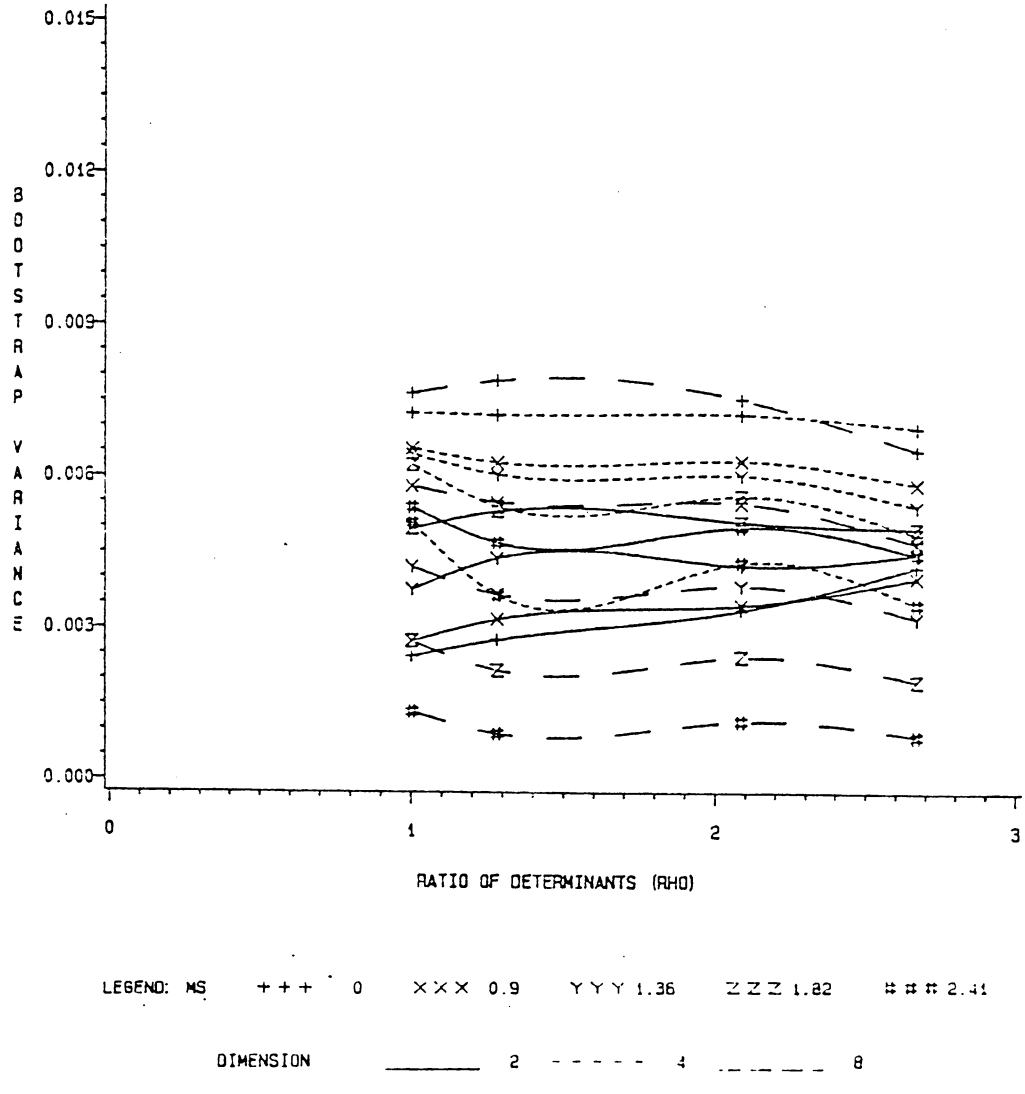

Figure 23. **Bootstrap variance estimate versus RD:** Bootstrap variance estimate of $\hat{\rho}^*$ versus the ratio of the generalized variance for different values of mean separation and dimension. Here $N_1 = 20$, $N_2 = 20$.

2.08 and 2.67) are graphed in figure 24 for the bivariate case. When the sample size of **Y** increases from 10 to 20 the simulated variance of $\hat{\rho}^*$ is reduced about 30%, and when the sample size of **Y** increases from 20 to 30 the simulated variance of $\hat{\rho}^*$ is reduced only 10%. This suggests that the simulated variance of $\hat{\rho}^*$ approaches a certain value asymptotically as the sample size for the **Y** sample increase while the sample size of the **X** sample remains fixed. It can be seen that when ρ^* is close to either 1 or 0, the variance of of $\hat{\rho}^*$ is small. The curve of the simulated variance of $\hat{\rho}^*$ is a quadratic function of ρ^* and it is concave downward between 0 and 1. When the ratio of the generalized variances increases from 1.0 to 2.67, ρ^* decreases and the simulated variance of $\hat{\rho}^*$ is reduced slightly.

Figures 25 and 26 describe the use of the jackknife method and the bootstrap methods to estimate the variance. When ρ^* is less than 0.9 both methods underestimate the variance. However, both methods overestimate the variance as ρ^* approaches 1. The shapes of the curves are similar to the simulated variance curves for smaller values of ρ^* but increase instead of decrease when ρ^* is near 1. Since the bias of $\hat{\rho}^*$ is large when ρ^* is close to 1 and the bias decreases as ρ^* decreases, we suspect that the bias is causing poor estimation in these areas.

6.2.3.The theoretical measure ρ^* and different variances: For the sample sizes $N_1 = 20, N_2 = 20$, we compare the simulated variance of $\hat{\rho}^*$ to the theoretical measure ρ^* for different ratios of the generalized variances (1.0, 1.28, 2.08 and 2.67) and different variances of X (σ_{x1} = 1, 3 and 5) for the bivariate case in figure 27. It can be seen that for a given value of ρ^* the simulated variance of $\hat{\rho}^*$

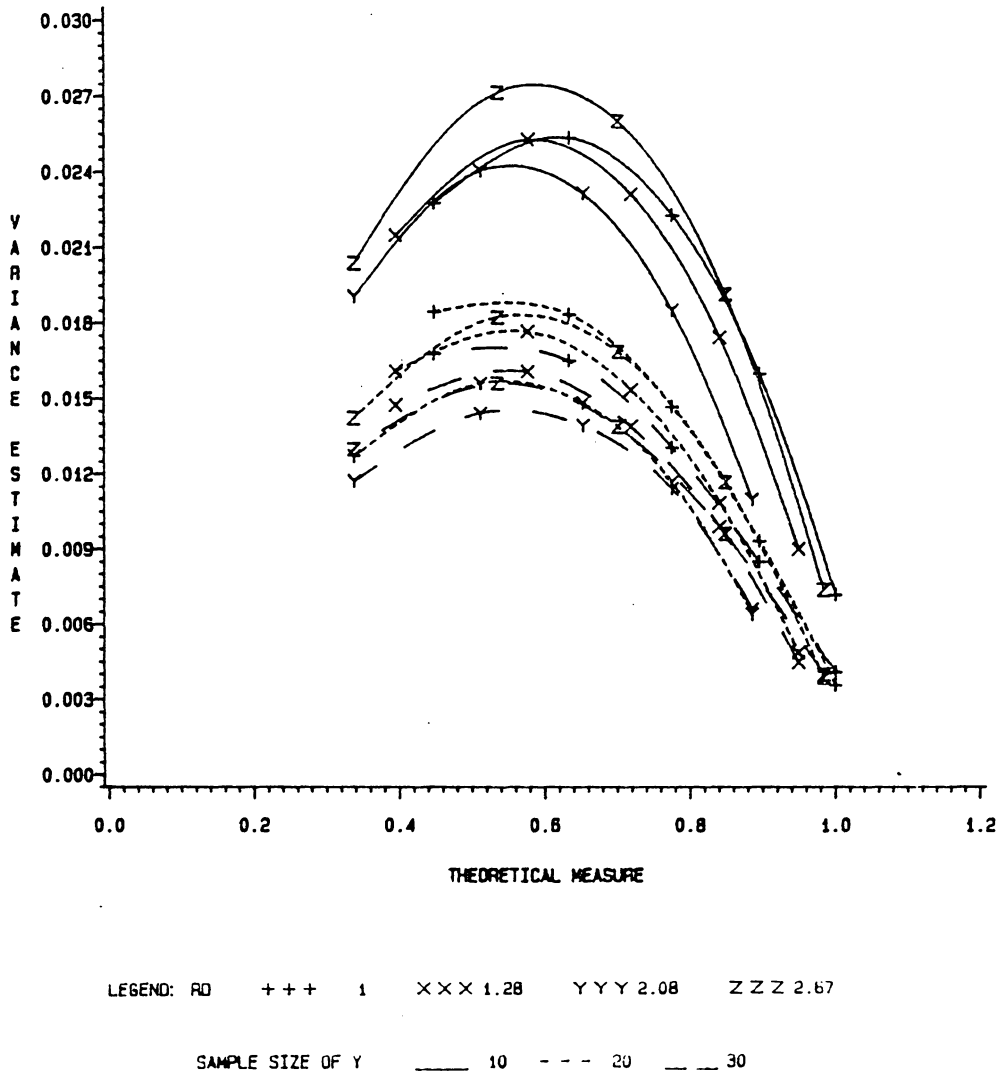

Figure 24. Simulated variance versus Theoretical measure: Simulated variance of $\hat{\rho}^*$ versus ρ^* for different sample sizes of Y, $N_2 = 10, 20, 30$ and various ratios of the generalized variances. Here $N_1 = 10$ and the dimension = 2.

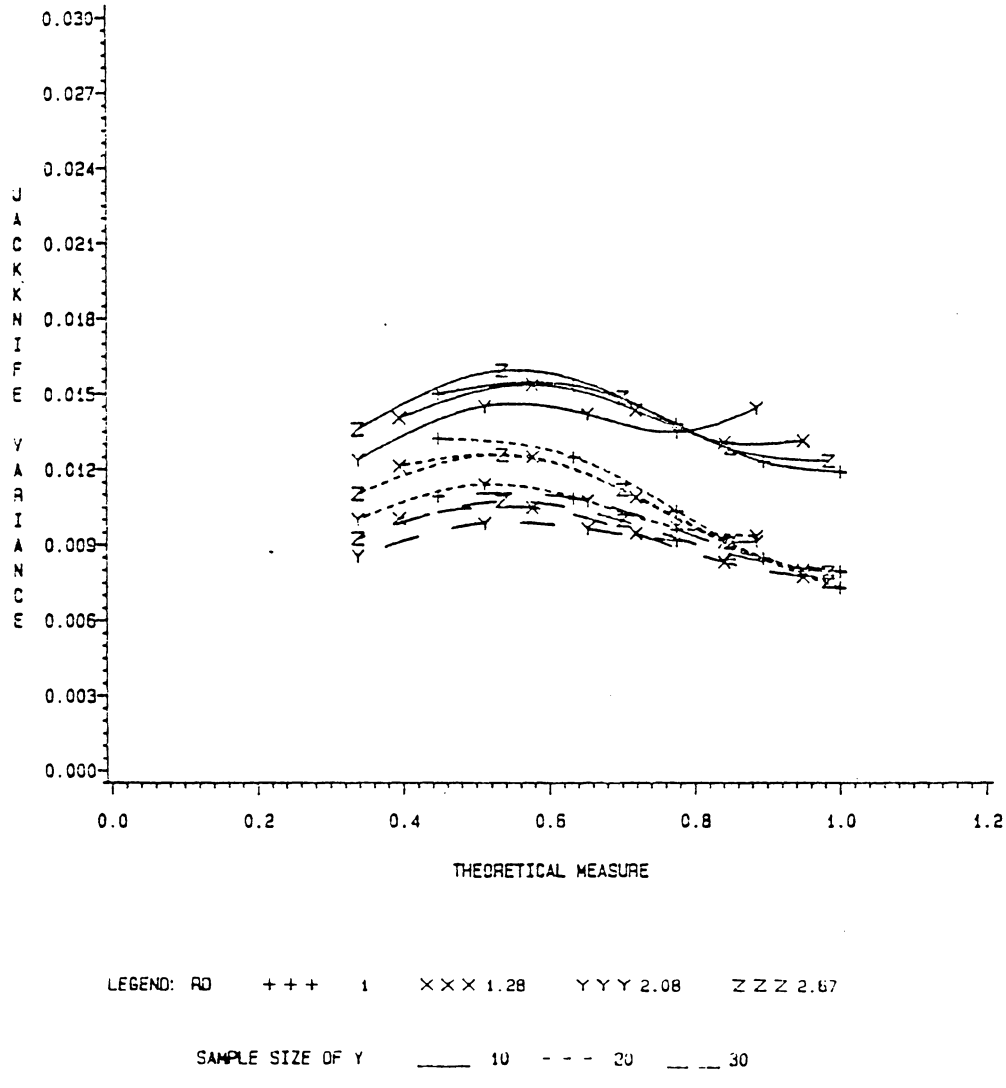

Figure 25. Jackknife variance estimate versus Theoretical measure: Jackknife variance estimate of ρ^* versus ρ^* for different sample sizes of Y, $N_2 = 10, 20, 30$ and various ratios of the generalized variances. Here $N_1 = 10$ and the dimension = 2.

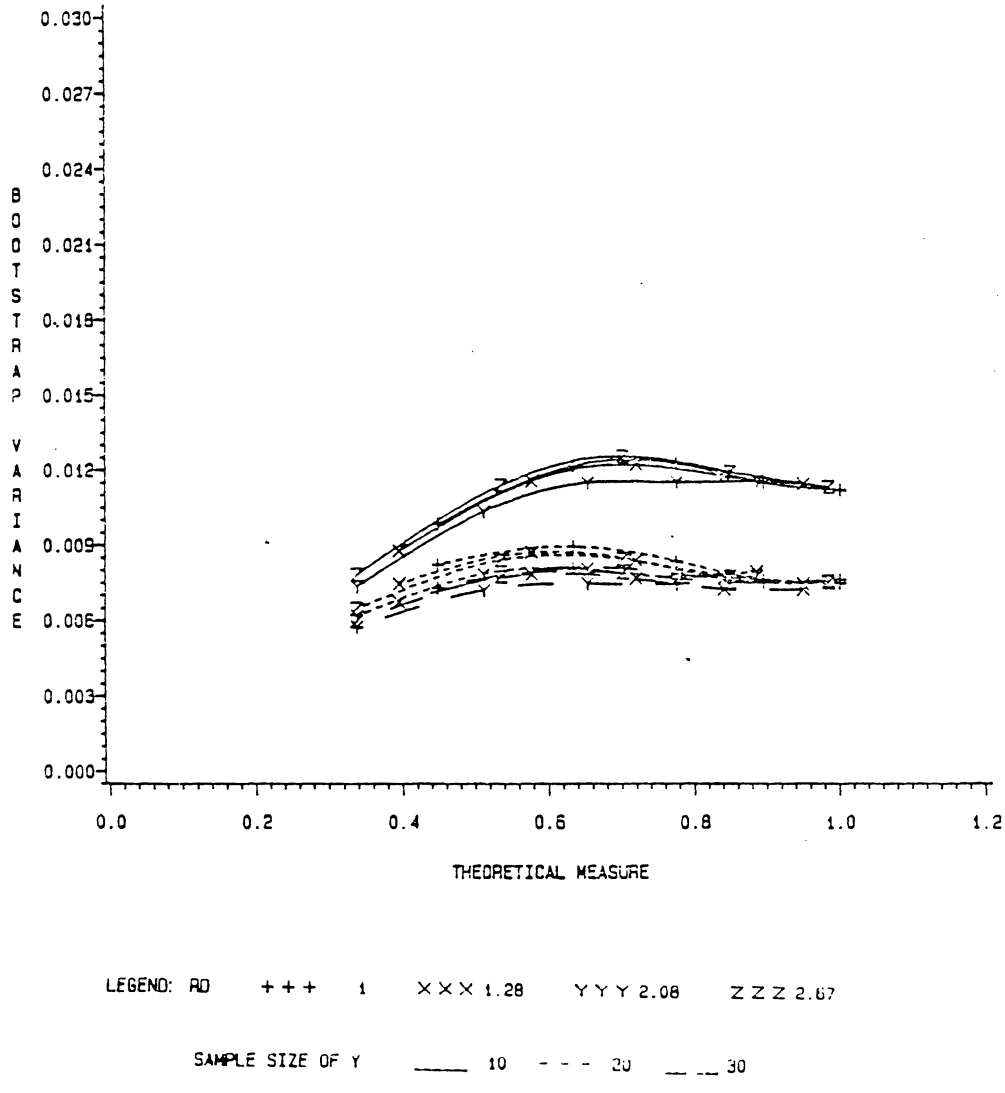

Figure 26. Bootstrap variance versus Theoretical measure: Bootstrap variance estimate of $\hat{\rho}^*$ versus ρ^* for different sample sizes of Y, $N_2 = 10, 20, 30$ and various ratios of the generalized variances. Here $N_1 = 10$ and the dimension = 2.

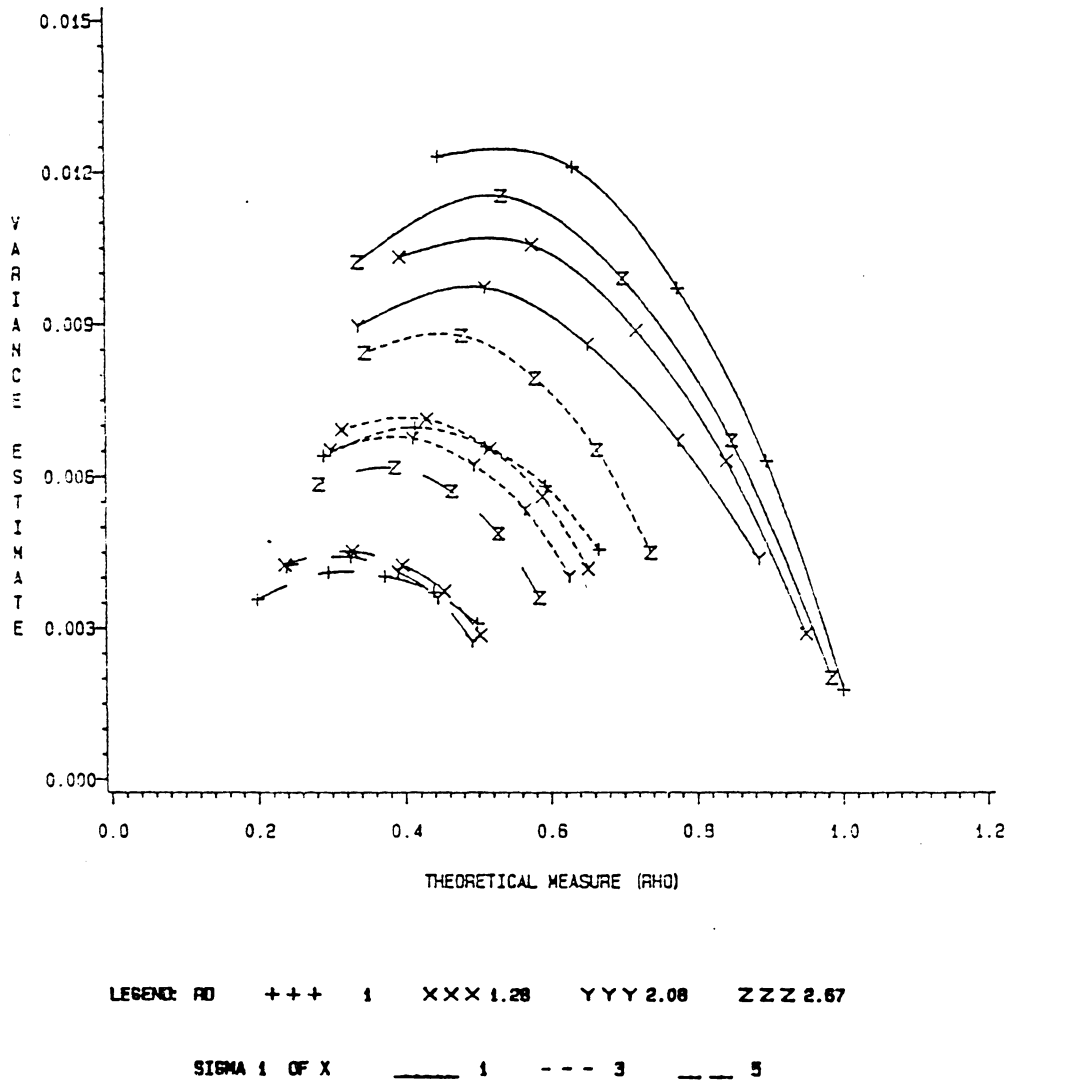

Figure 27. **Simulated variance estimate versus Theoretical measure:** Simulated variance of $\hat{\rho}^*$ versus ρ^* for different variances of X, with σ_{x1} = 1 , 3 , 5 and various ratios of the generalized variances. Here σ_{y1} = 1 and the dimension = 2.

varies and it depends heavily on the disparity of the variance-covariance matrices. When the ratio of the generalized variances increases from 1.0 to 2.67 the variance is reduced 25% at the vertex of the curve . When σ_{x1} changes from 1 to 3, the ratio of the generalized variances increases, the theoretical measure ρ^* decreases and the simulated variance of $\hat{\rho}^*$ is reduced around the vertex of the curve. This is a significant reduction of the variance. As σ_{x1} gets larger, which means the ratio of the generalized variances is larger, the simulated variance of $\hat{\rho}^*$ becomes smaller. The inequality of the variances in the variance-covariance matrices is important in determining the variance estimate of $\hat{\rho}^*$.

Comparing figures 28 and 29, when the theoretical measure ρ^* is close to 1 the jackknife method and the bootstrap method estimate the simulated variance of $\hat{\rho}^*$ well. The jackknife method underestimates the simulated variance of $\hat{\rho}^*$ and the bootstrap estimate of the simulated variance of $\hat{\rho}^*$ is slightly worse. When there is no mean separation, both methods estimate the variance quite well. As σ_{x1} increases, the degree of underestimation is increased slightly. The ratio of the generalized variances seems to have no effect on estimating the simulated variance of $\hat{\rho}^*$ using either method.

6.2.4.The theoretical measure ρ^* for various dimensions: For the sample sizes $N_1 = 20, N_2 = 20$, we compare the simulated variance of $\hat{\rho}^*$ to the theoretical measure ρ^* for the number of variables (2, 4 and 8) and various mean separations in figure 30. When the theoretical measure ρ^* is around 0.9 the simulated variance of $\hat{\rho}^*$ is around 0.006. For other values of ρ^* the simulated variance of

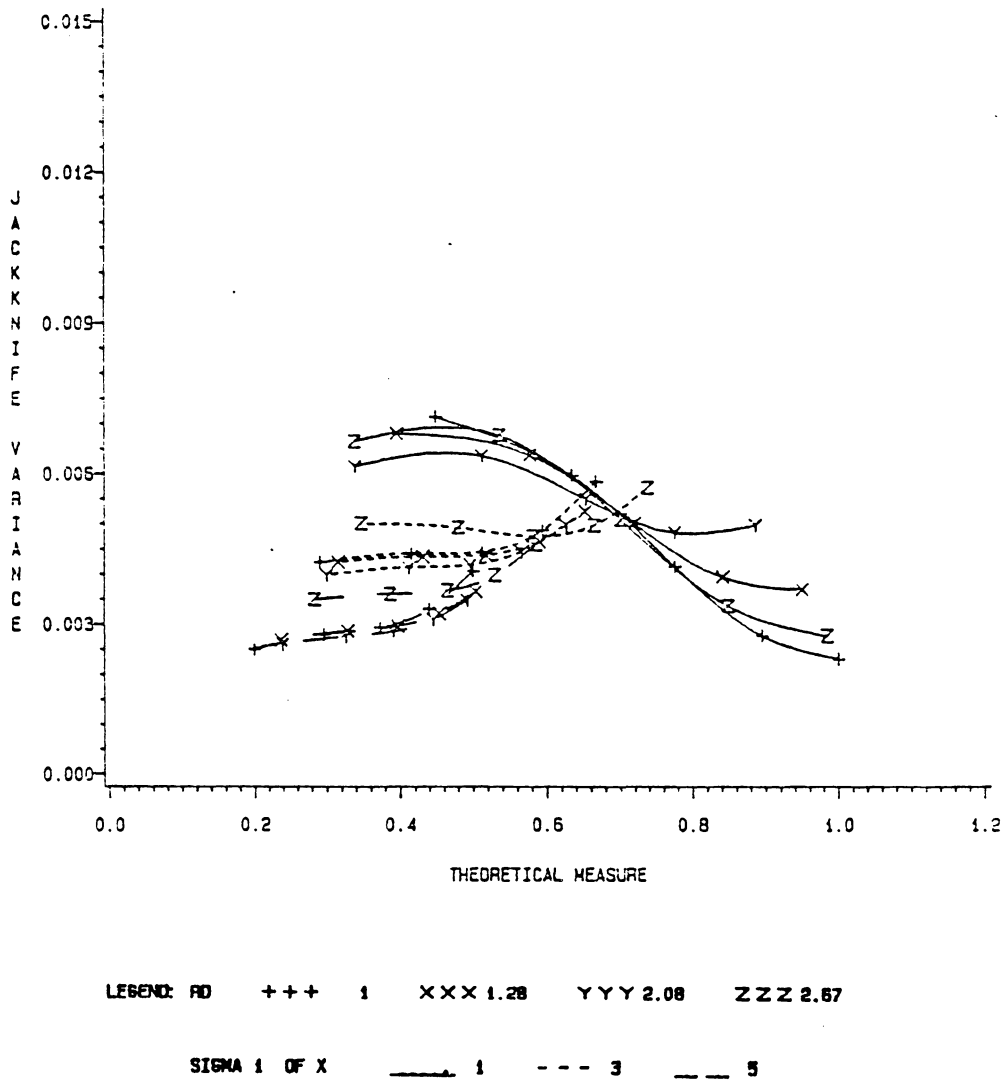

Figure 28. Jackknife variance estimate versus Theoretical measure: Jackknife variance estimate of ρ^* versus ρ^* for different variances of X, with σ_{x1} = 1 , 3 , 5 and various ratios of the generalized variances. Here σ_{y1} = 1 and the dimension = 2.

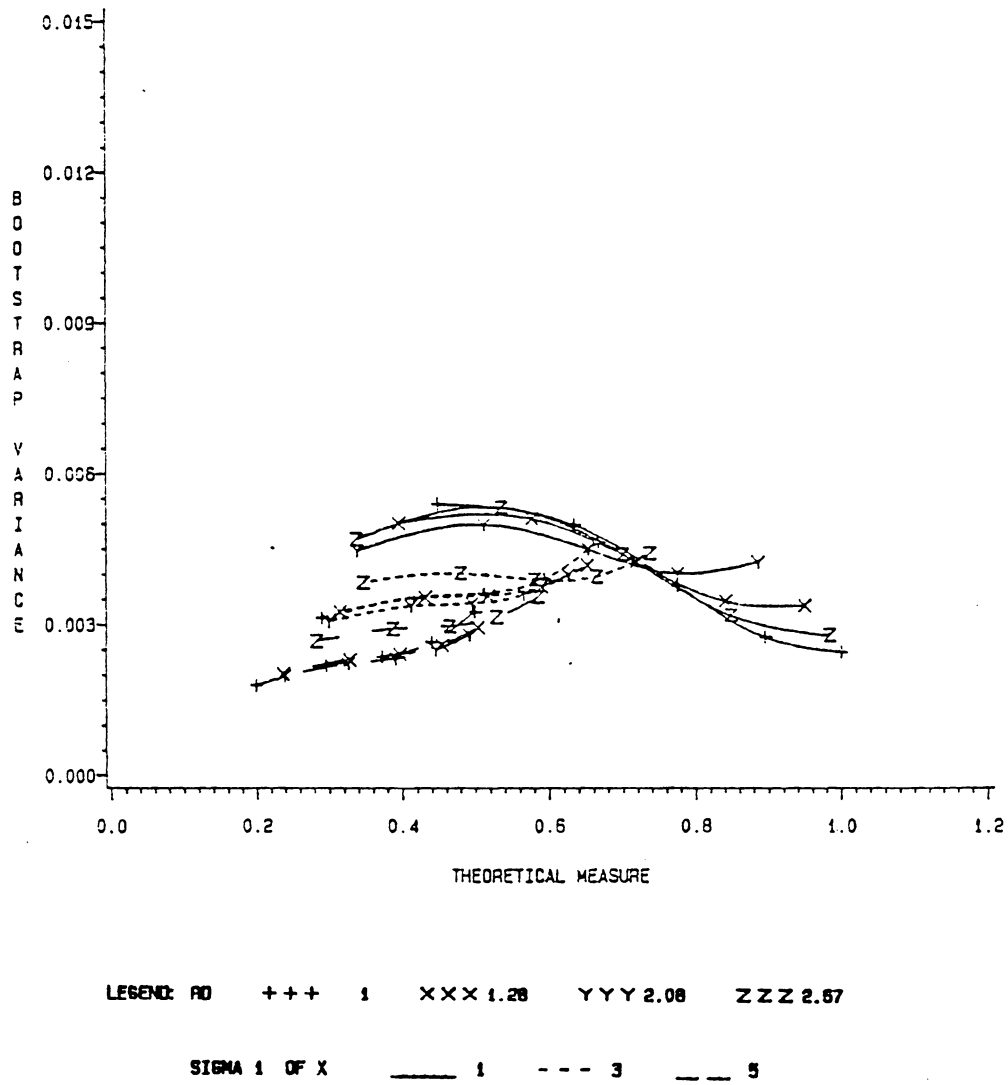

Figure 29. **Bootstrap variance estimate versus Theoretical measure:** Bootstrap variance estimate of ρ^* versus ρ^* for different variances of X, with σ_{x1} = 1 , 3 , 5 and various ratios of the generalized variances. Here σ_{y1} = 1 and the dimension = 2.

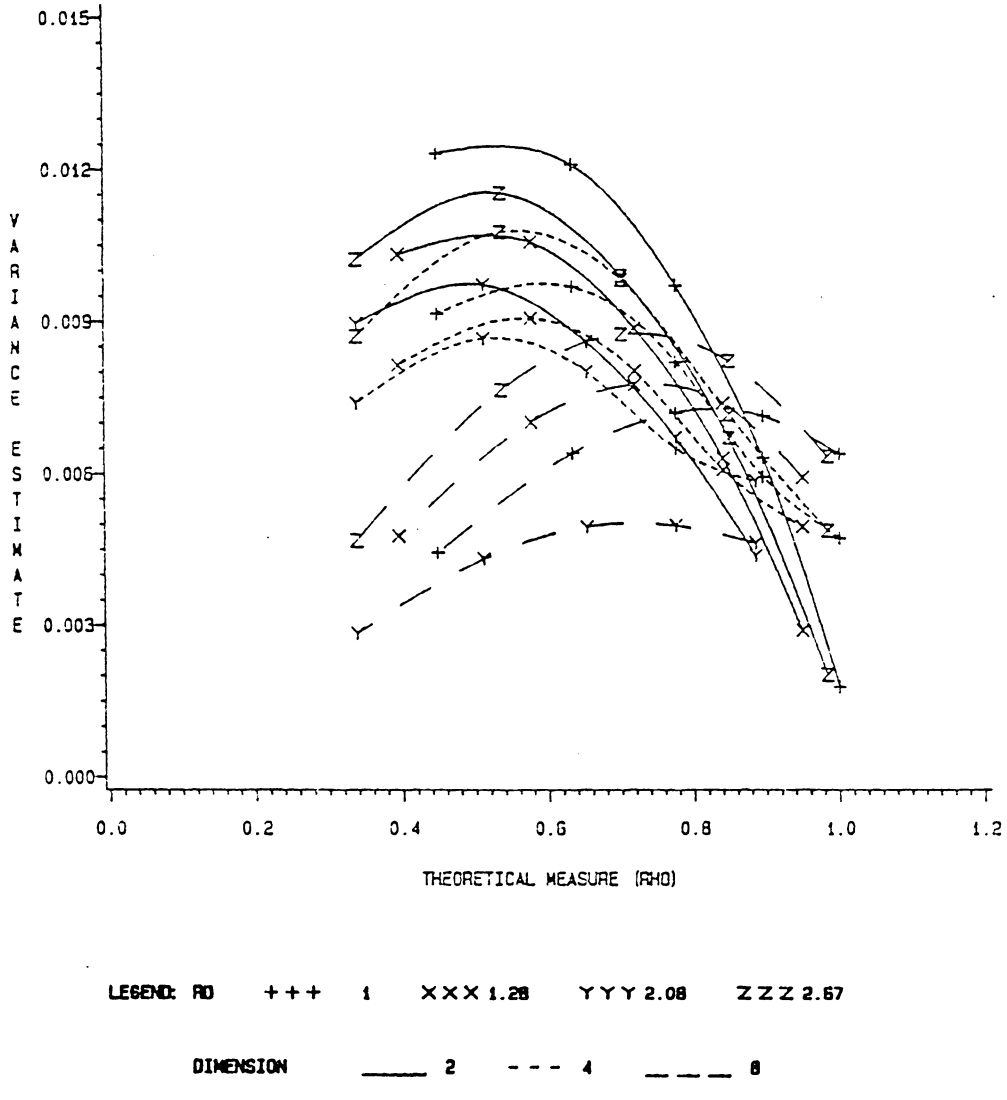

Figure 30. Simulated variance versus Theoretical measure: Simulated variance of $\hat{\rho}^*$ versus ρ^* for different values of mean separation and dimension. Here $N_1 = 20$, $N_2 = 20$.

$\hat{\rho}^*$ varies extensively. When the number of variables increases from 2 to 4 the simulated variance of $\hat{\rho}^*$ decreases a little. When the number of variables increases from 4 to 8 the simulated variance of $\hat{\rho}^*$ comes down except when ρ^* is close to 1. This is a significant result. When there are 8 variables involved in estimating ρ^*, the bias of $\hat{\rho}^*$ is high and the simulated variance of $\hat{\rho}^*$ is low when ρ^* is close to 1. For ρ^* around 0.6, the bias of $\hat{\rho}^*$ is lower and the simulated variance of $\hat{\rho}^*$ is high. Generally, when two populations are far apart, the bias of $\hat{\rho}^*$ is low and the simulated variance of $\hat{\rho}^*$ is small.

In figure 31, the jackknife method is seen to underestimate the simulated variance of $\hat{\rho}^*$ considerably when ρ^* is less than 0.8, and overestimates it otherwise. The bootstrap method (figure 32) underestimates the simulated variance of $\hat{\rho}^*$ severely when ρ^* is less than 0.9 and overestimates it slightly otherwise. When the number of variables is 2, the bootstrap method performs worse than the jackknife method does. When the number of variables is 4 and ρ^* is less than 0.9, the bootstrap method underestimates more than the jackknife method does. If ρ^* is between 0.9 and 1, both methods overestimate the simulated variance of $\hat{\rho}^*$. For dimension of 8, the jackknife performs well when ρ^* is less than 0.8, and overestimates the simulated variance of $\hat{\rho}^*$ more as ρ^* becomes close to 1. The bootstrap method always underestimates the simulated variance of $\hat{\rho}^*$. As the number of variables increases, the higher the index is and the larger the simulated variance estimate will be. The impact of the dimensionality is heavy if the index is close to 1.

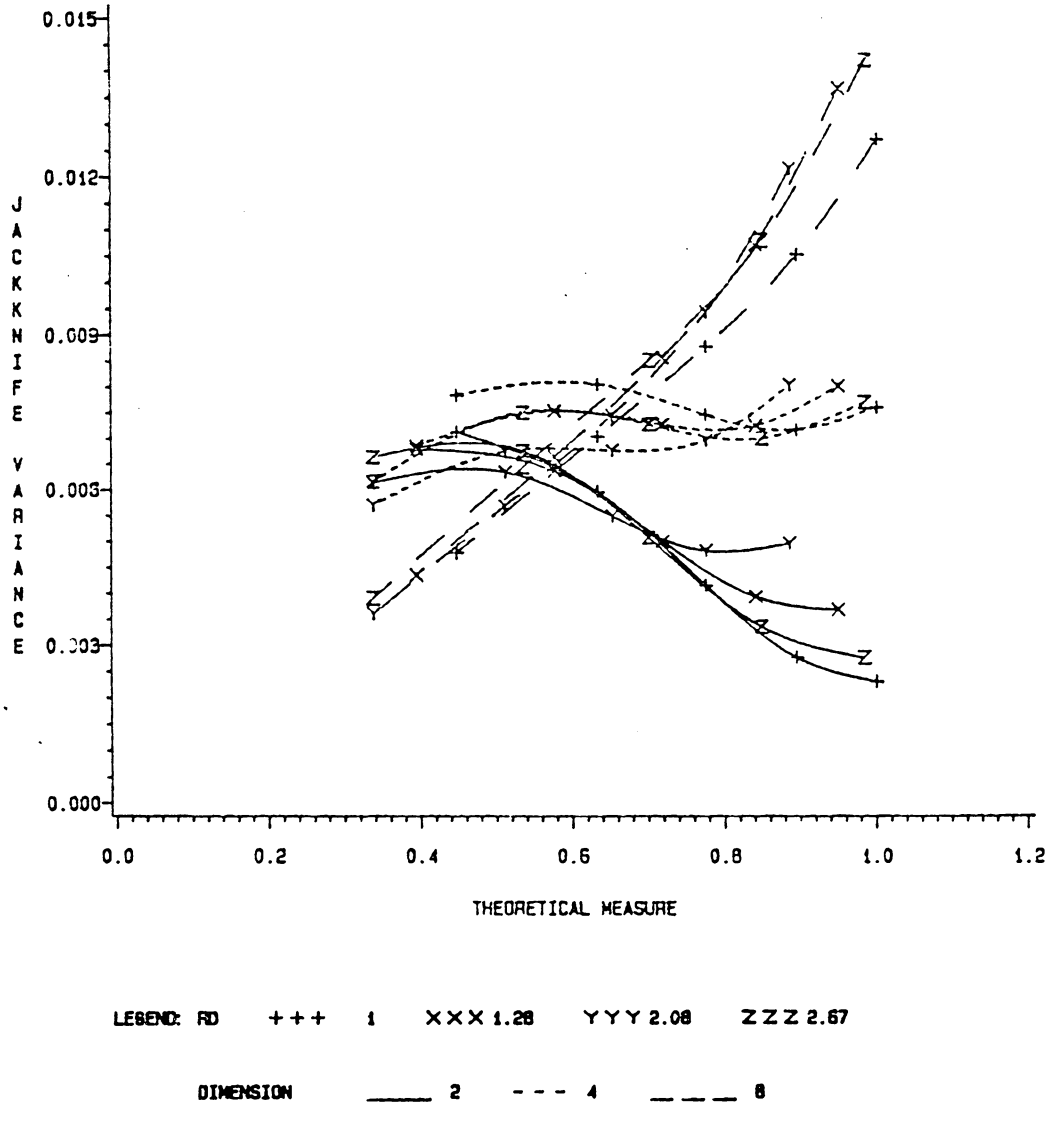

Figure 31. Jackknife variance estimate versus Theoretical measure: Jackknife variance estimate of ρ^* versus ρ^* for different values of mean separation and dimension. Here $N_1 = 20$, $N_2 = 20$

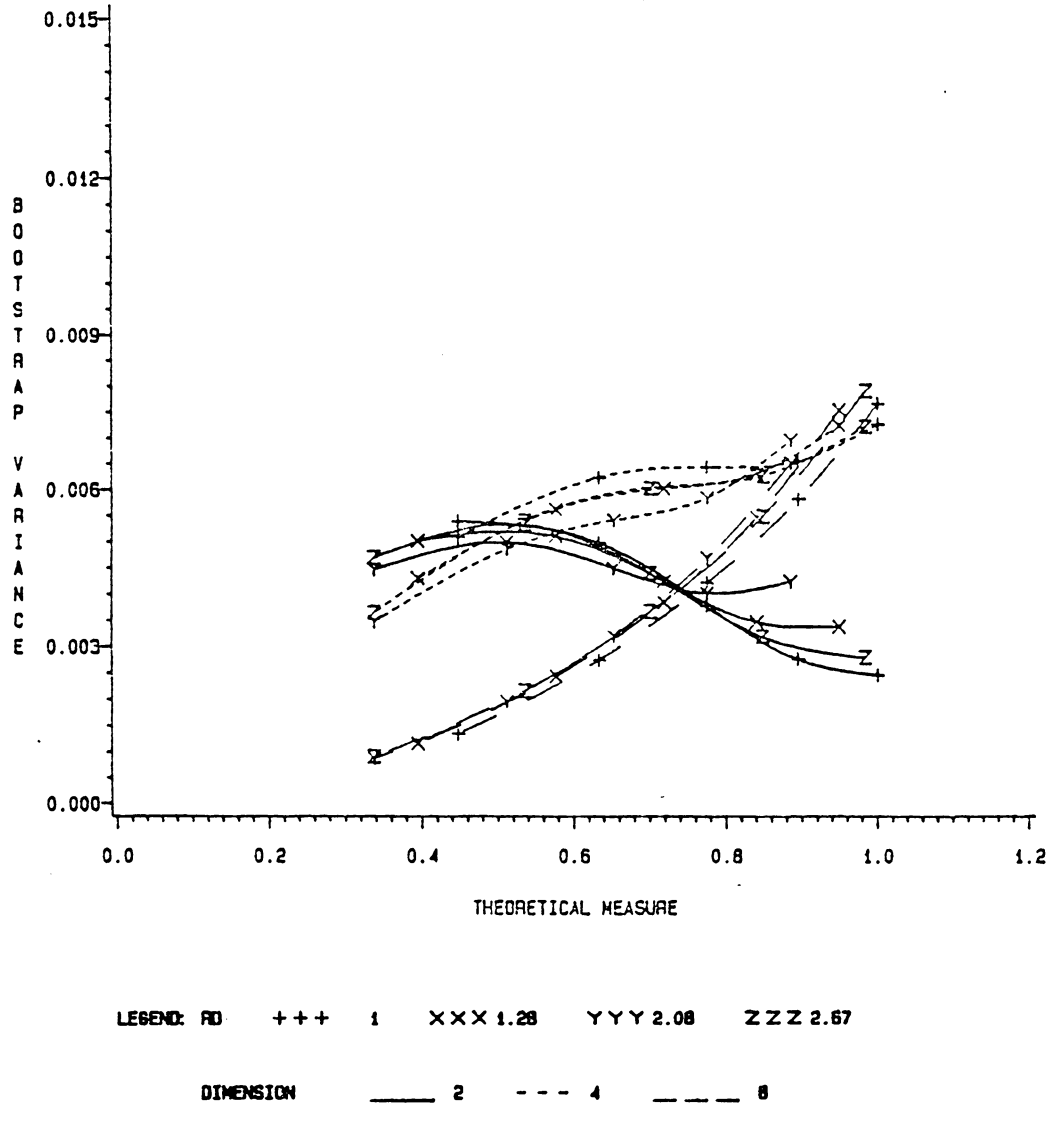

Figure 32. Boqtstrap variance estimate versus Theoretical measure: Bootstrap variance estimate of ρ^* versus ρ^* for different values of mean separation and dimension. Here $N_1 = 20, N_2 = 20$.

6.2.5.The ratio of the generalized variances and mean separation: For the sample sizes $N_1 = 20$, $N_2 = 20$, we compare the different ratios of generalized variances to the simulated variance of $\hat{\rho}^*$ in figure 33. In the bivariate case, the simulated variance of $\hat{\rho}^*$ increases if ρ^* is away from 1 and close to 0. Figure 33 indicates that when the ratio of the generalized variances increases from 1 to 9 the simulated variance of $\hat{\rho}^*$ approaches 0.006. The simulated variance of $\hat{\rho}^*$ stabilizes gradually when the ratio of the generalized variances gets larger. This means that when the variance-covariance matrices are different, the measure is smaller, and the simulated variance of $\hat{\rho}^*$ is consistent. Whenever the ratio of the generalized variances is small, the mean separation affects the simulated variance of $\hat{\rho}^*$. This impact diminishes as the ratio of the generalized variances increases.

In figures 34 and 35, for various ratios of generalized variances, when there is no mean separation, the bootstrap method estimates the variance quite well and the jackknife method overestimates the variance. Both methods underestimate the variance when sample means are different, and the effect of mean separation is diminished gradually as the ratio of generalized variances increases. The bootstrap method again underestimates more than the jackknife method when there exists mean separation.

6.3 COVERAGE OF CONFIDENCE INTERVALS

In previous sections, the jackknife estimate and the bootstrap estimate of ρ^* are discussed, also the jackknife variance estimate and the bootstrap variance es-

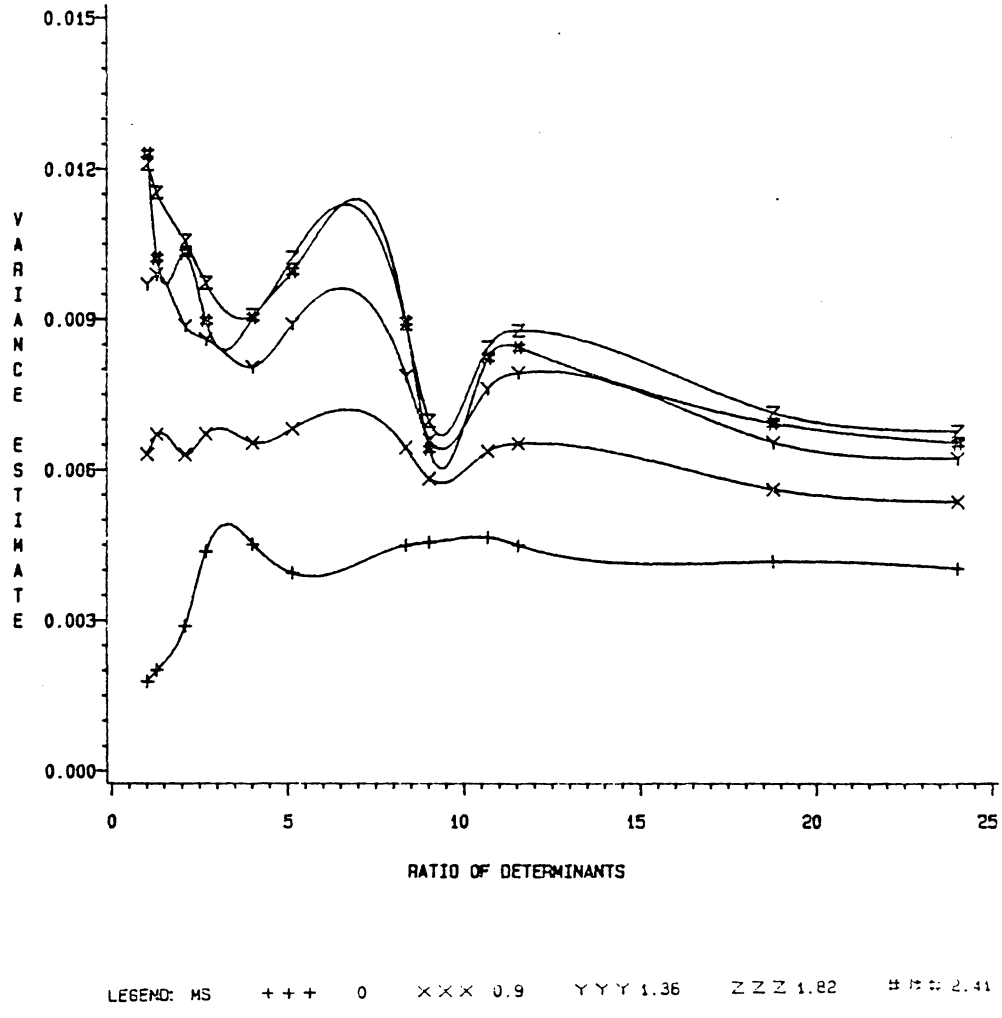

Figure 33. Simulated variance versus RD: Simulated variance of $\hat{\rho}^*$ versus the ratio of the generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

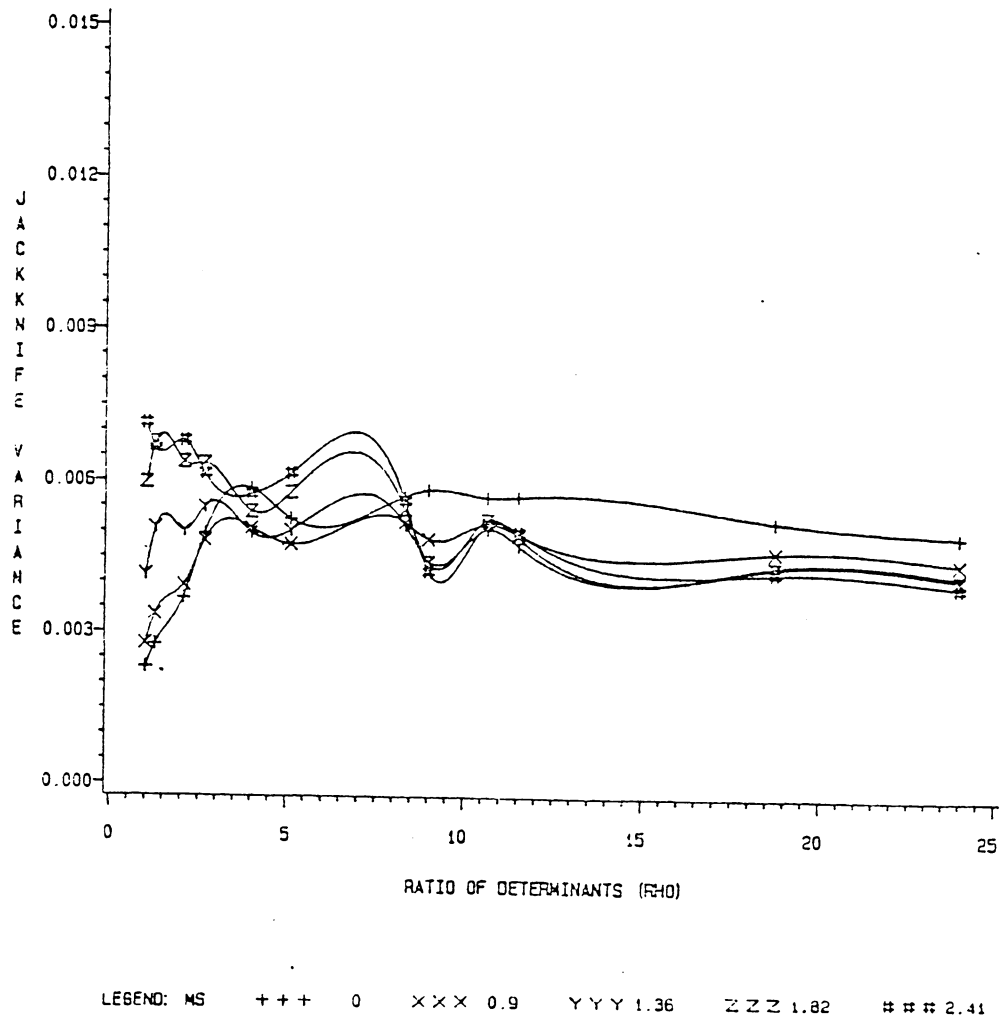

Figure 34. Jackknife variance estimate versus RD: Jackknife variance estimate of $\hat{\rho}^*$ versus the ratio of the generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

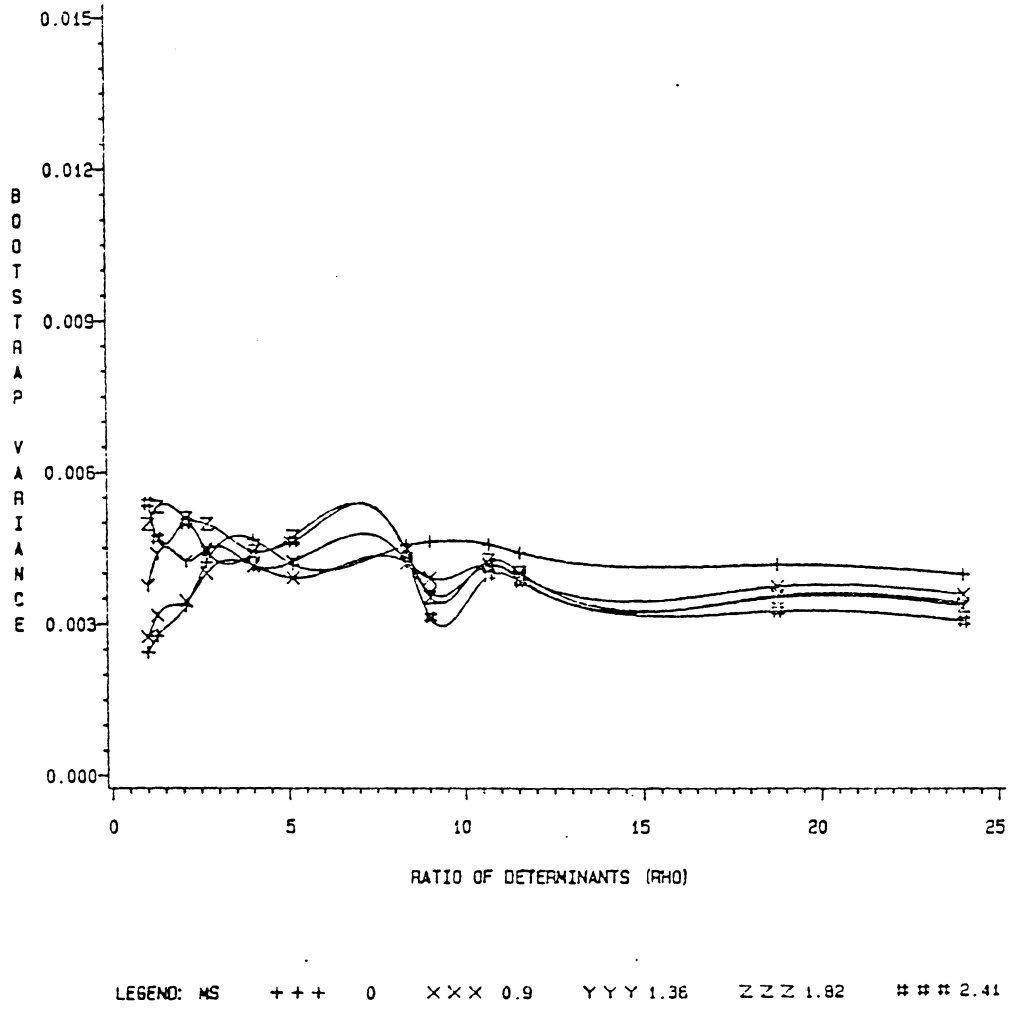

Figure 35. Bootstrap variance estimate versus RD: Bootstrap variance estimate of $\hat{\rho}^*$ versus the ratio of the generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

timate of $\hat{\rho}^*$ are compared. The 95% symmetric confidence interval can be formed through the discussion in chapter 5. Here, we are interested in finding which method attains the higher coverage in estimating the measure ρ^*. In the simulation we repeat each procedure 100 times, so the closer the percentage is to 95 the better the method is. This has a connection with the reliability of the estimation procedures.

It will be clear from the discussion below that in some cases, the methods will perform reasonably well. The coverage of the bootstrap method tends to be lower than that of the jackknife. When the number of extraneous variables is high neither method gives good results.

6.3.1.The ratio of the generalized variances and different dimensions

For the sample sizes $N_1 = 20$, $N_2 = 20$, we compare the coverage of the 95% jackknife and the 95% bootstrap confidence intervals for different ratios of the generalized variances (1.0, 1.28, 2.08 and 2.67) and different numbers of variables (2, 4 and 8). As indicated in figures 36 and 37, for the same ratio of the generalized variances there are several levels of the coverage. The level of the coverage depends on the number of the variables involved in the model. When there are 2 or 4 variables in the model, the coverage is between 80 and 95. It is reliable to use either method to adjust the bias of $\hat{\rho}^*$ and to obtain the variance of $\hat{\rho}^*$. When the number of variables increases to 8, the coverage is down near 60's when using the jackknife method, and it is down to the 20's in using the bootstrap method. This indicates that if there are a few variables involved, 20 samples from

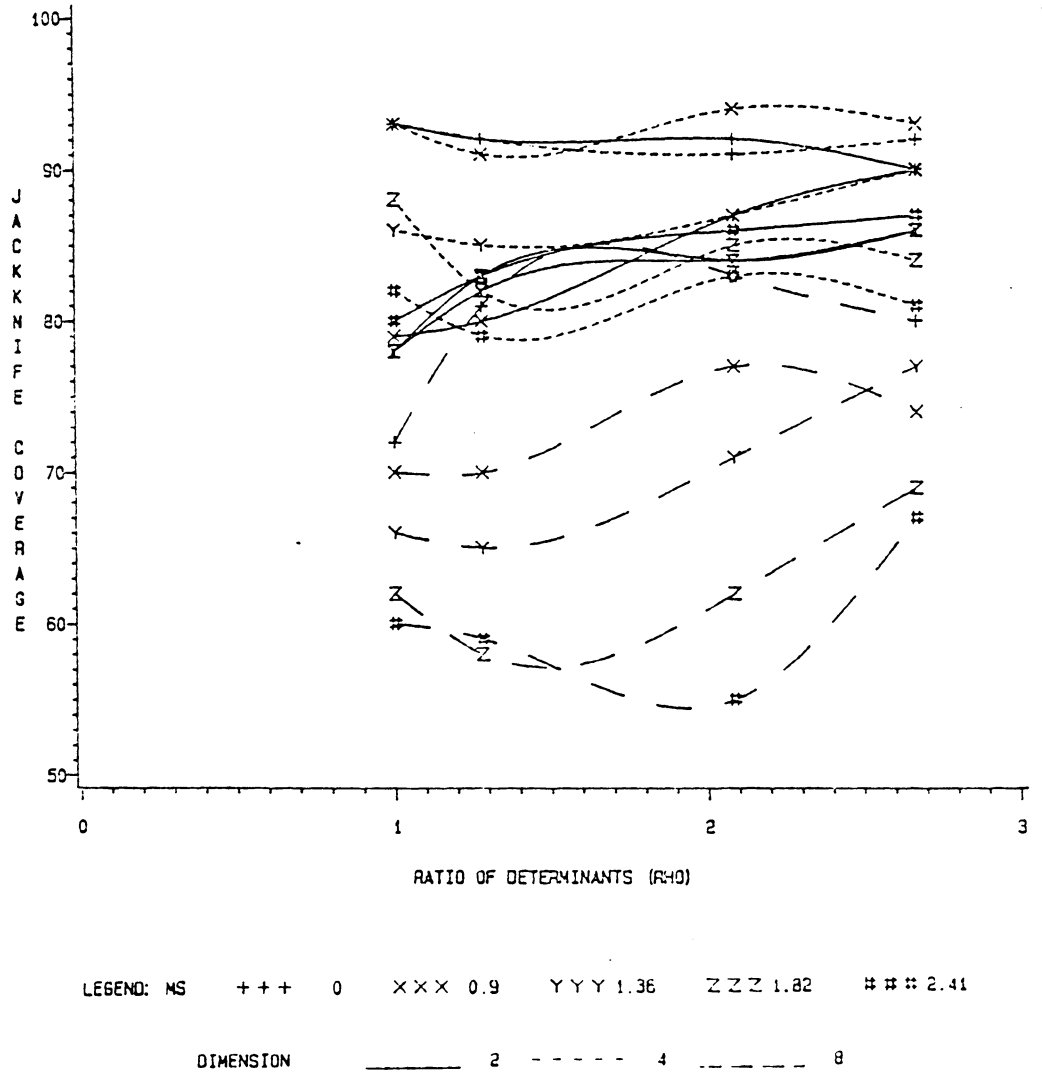

Figure 36. Jackknife coverage versus RD: Coverage of the 95% jackknife CI versus the ratio of the generalized variance for different values of mean separation and dimension. Here $N_1 = 20$, $N_2 = 20$.

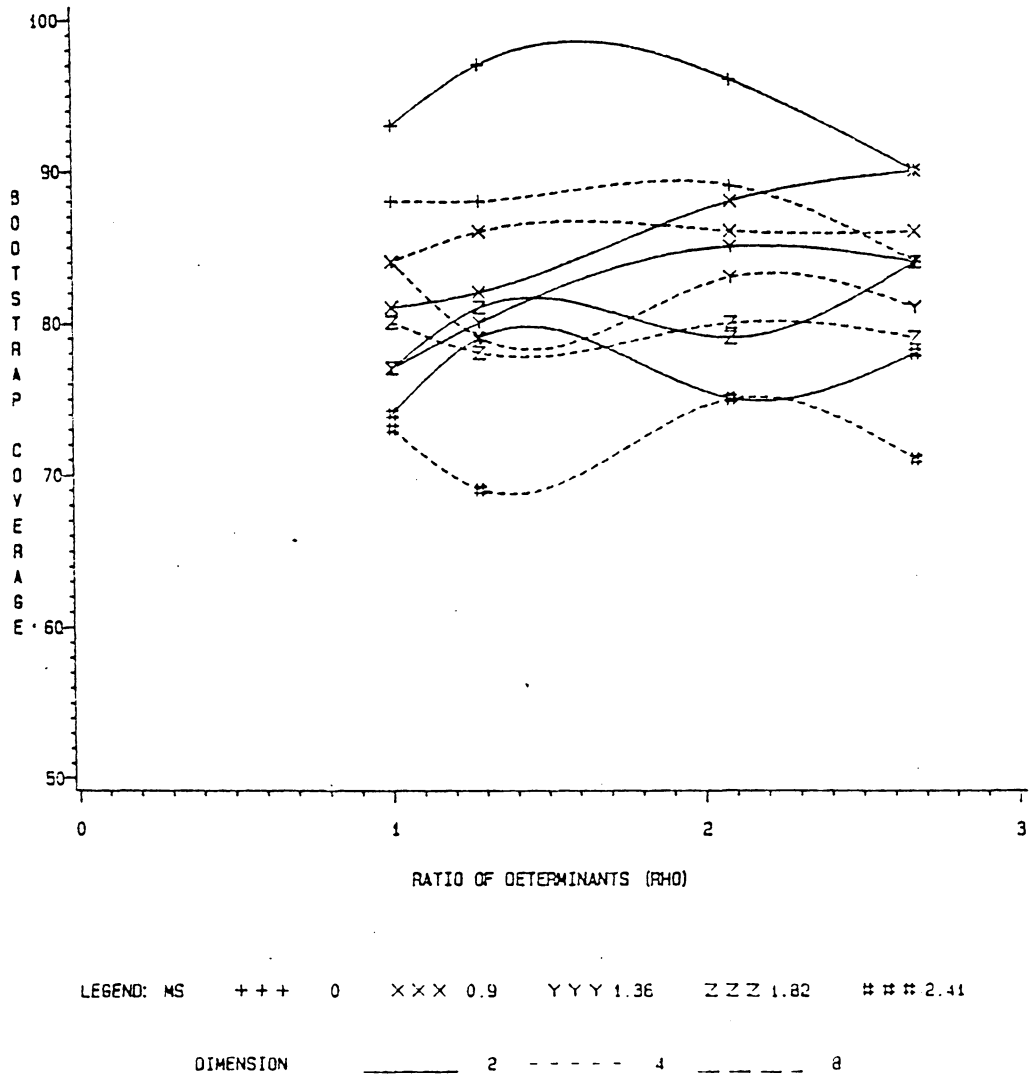

Figure 37. Bootstrap coverage versus RD: Coverage of the 95% bootstrap CI versus the ratio of the generalized variance for different values of mean separation and dimension. Here $N_1 = 20$, $N_2 = 20$.

each population is sufficient to get a reliable sample estimate. When the ratio of the sample sizes versus the dimension is below 5, neither of the methods can give a reliable estimate of ρ^*. There is no significant difference in the reliability of the methods for the different ratios of generalized variances.

6.3.2.The ratio of the generalized variances and unequal samples: We compare the coverage of the 95% jackknife and the 95% bootstrap confidence intervals to the theoretical measure ρ^* for the sample sizes 10,10; 10,20 and 10,30, and the different ratio of the generalized variances (1.0, 1.28, 2.08 and 2.67) when the number of variables is 2 (the bivariate case). In figures 38 and 39, the coverage is smaller as the theoretical measure ρ^* decreases. The ratio of the generalized variances is seen to have no bearing on the coverage. When the sample size increases from 10 to 30 in one population, the coverage stays almost the same. However, the higher the measure ρ^* is, the more reliable the jackknife estimate and the bootstrap estimate of ρ^* are. The reliability of the estimate using either method decreases as the two populations become separated. The bootstrap method is less reliable than the jackknife method.

6.3.3.The theoretical measure ρ^* and different variances

For the sample sizes $N_1 = 20, N_2 = 20$, we compare the coverage of 95% jackknife and the 95% bootstrap confidence intervals to the theoretical measure ρ^* for different ratios of the generalized variances (1.0, 1.28, 2.08 and 2.67) and different variances of X (σ_{x1} = 1, 3 and 5) for the bivariate case in figures 40

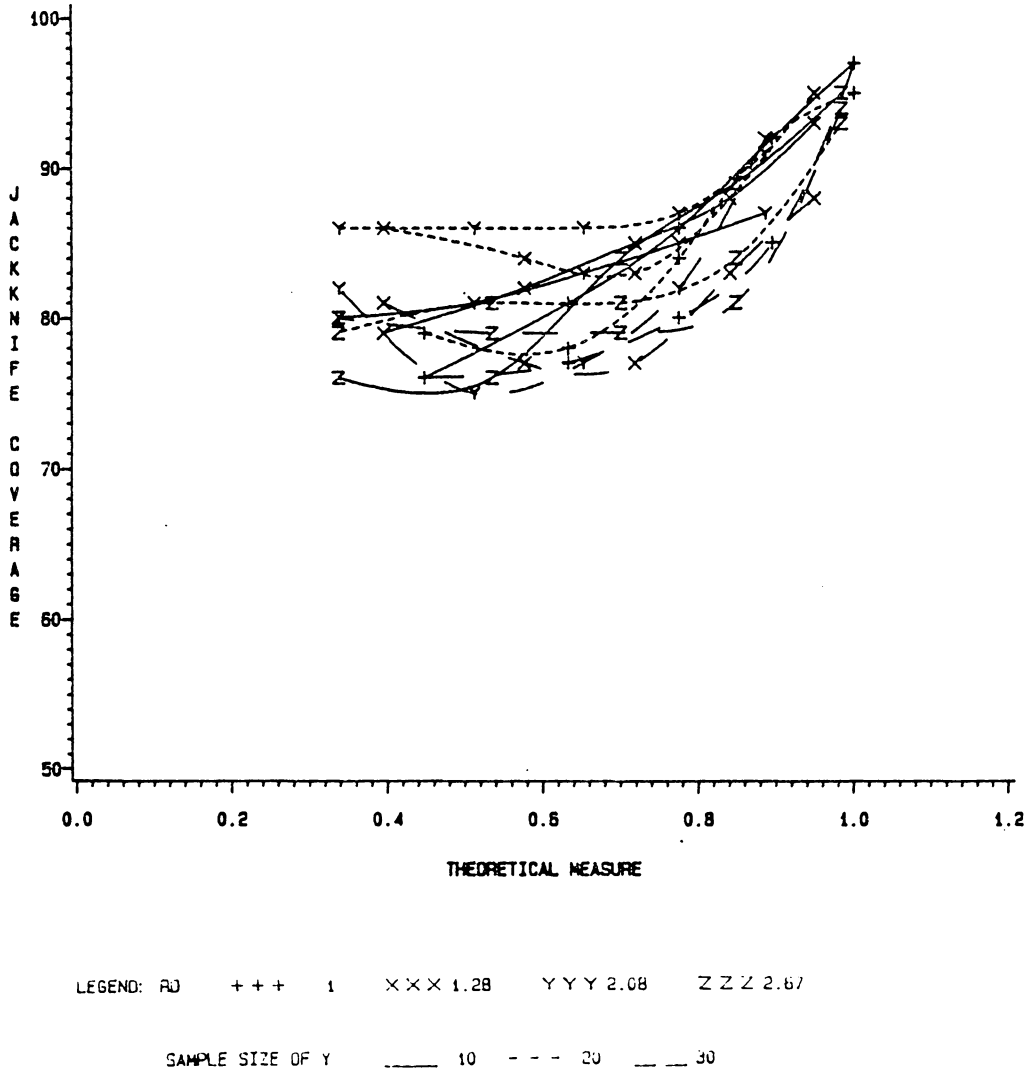

Figure 38. Jackknife coverage versus Theoretical measure: The coverage of 95% jackknife CI versus ρ^* for different sample sizes of Y, with $N_2 = 10, 20, 30$ and various ratios of the generalized variances. Here $N_1 = 10$ and the dimension = 2.

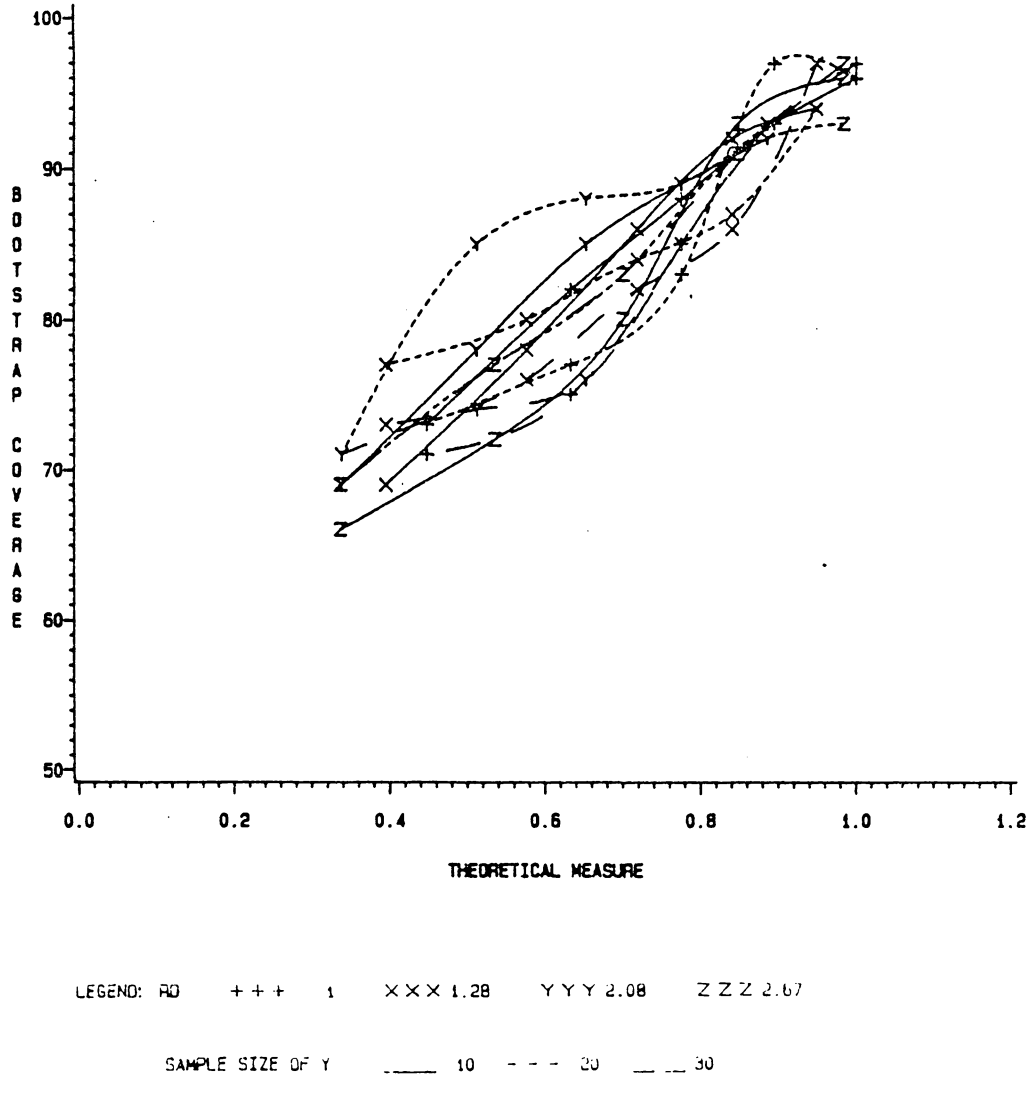

Figure 39. Bootstrap coverage versus Theoretical measure: The coverage of 95% bootstrap CI versus ρ^* for different sample sizes of Y, with $N_2 = 10, 20, 30$ and various ratios of the generalized variances. Here $N_1 = 10$ and the dimension = 2.

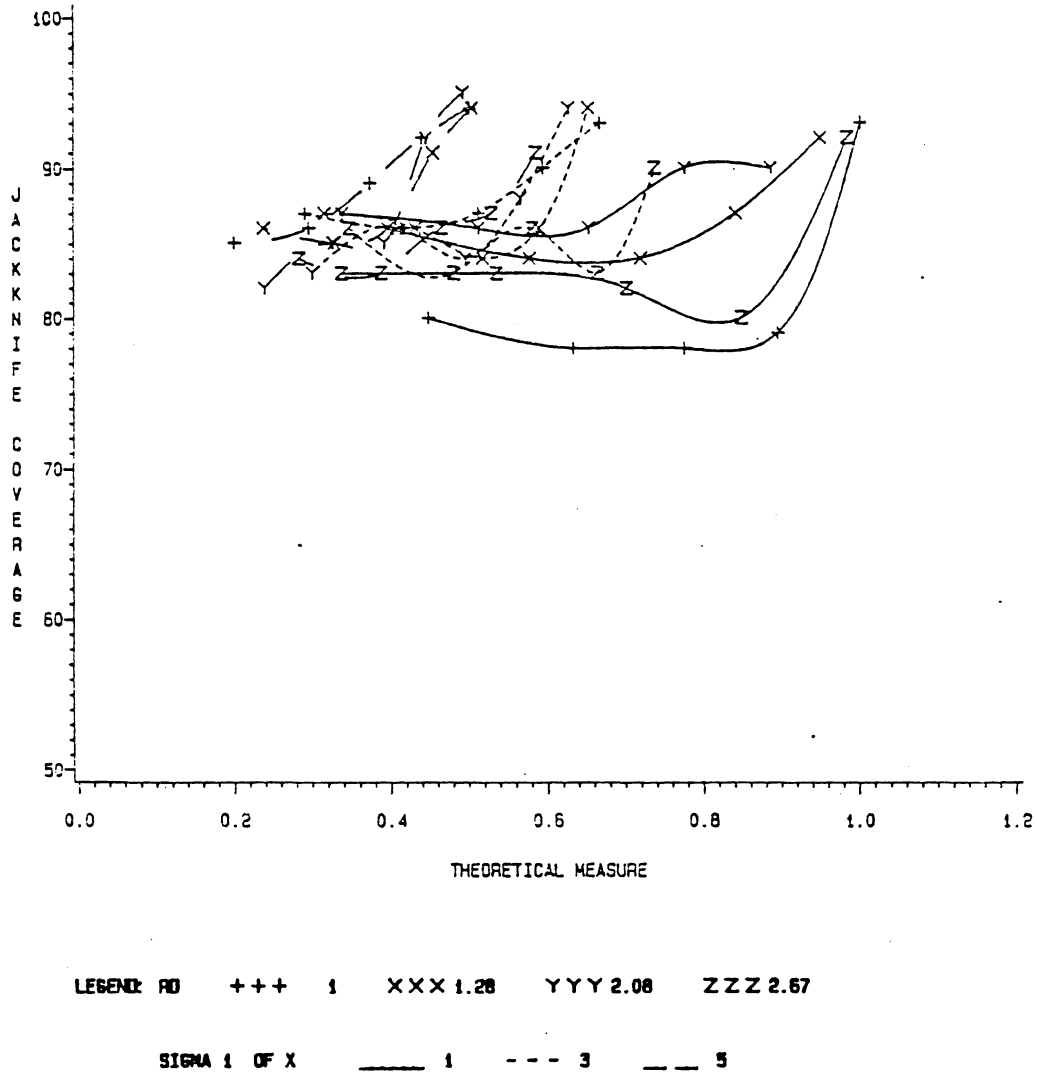

Figure 40. Jackknife coverage versus Theoretical measure: The coverage of 95% jackknife CI versus ρ^* for different variances of X, with σ_{x1} = 1 , 3 , 5 and various ratios of the generalized variances. Here σ_{y1} = 1 and the dimension = 2.

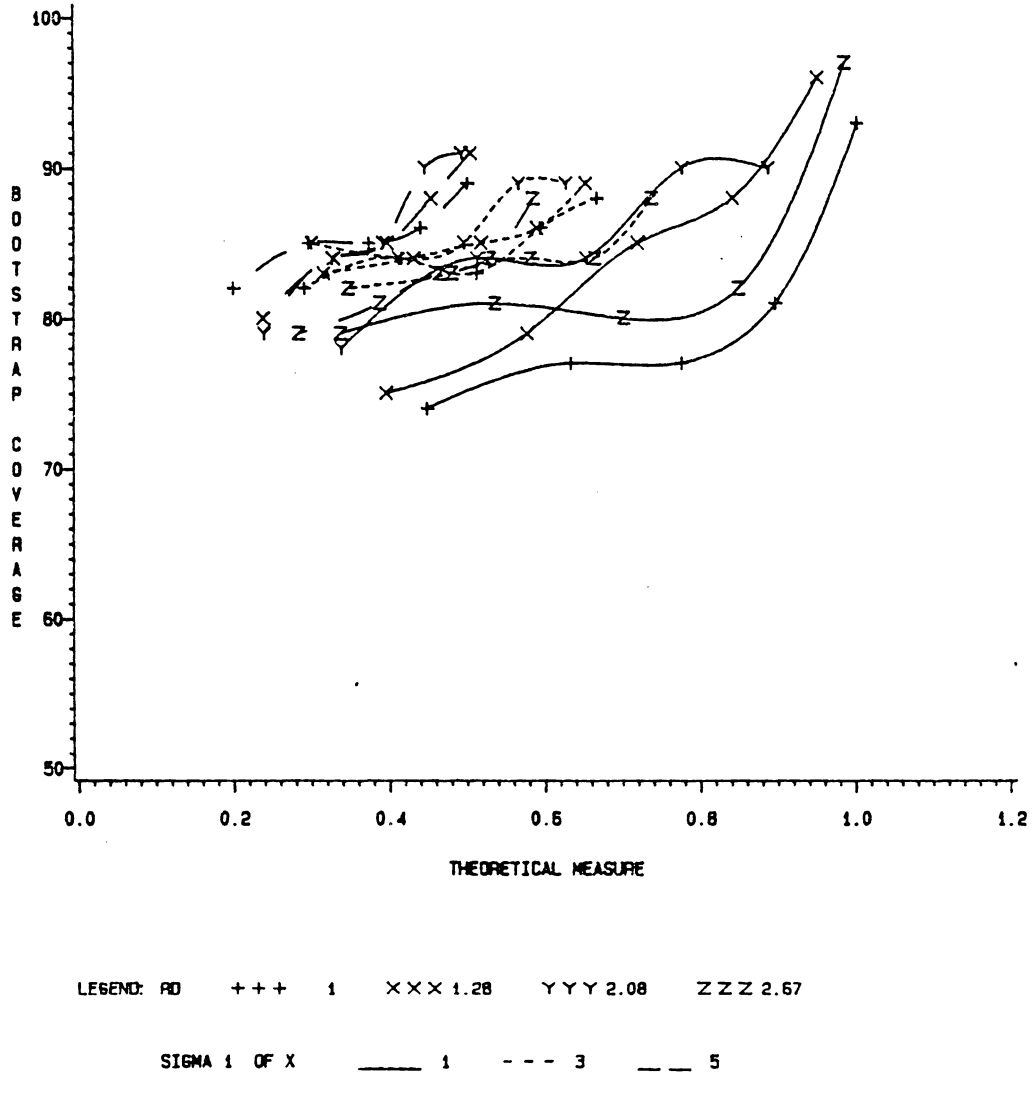

Figure 41. **Bootstrap coverage versus Theoretical measure:** The coverage of 95% bootstrap CI versus ρ^* for different variances of X, with σ_{x1} = 1, 3, 5 and various ratios of the generalized variances. Here σ_{y1} = 1 and the dimension = 2.

and 41. The coverage is between 80 to 95 for different values of the theoretical measure ρ^* . When the ratio of the generalized variances changes from 1.0 to 2.67 the coverage does not alter. When σ_{x1} changes from 1 to 3, there is a significant shift of the theoretical measure ρ^*, but the coverage is kept at the same level. When σ_1 gets larger, there is no significant change in coverage. Using the jackknife method and the bootstrap method result in equally reasonable coverage. The closer the population means are, the better the coverage is.

6.3.4.The theoretical measure ρ^* for various dimensions: For sample sizes $N_1 = 20$, $N_2 = 20$, we compare the effect of the number of variables (2, 4 and 8) on the coverage of the 95% jackknife and the 95% bootstrap confidence intervals of the estimate $\hat{\rho}^*$ in figures 42 and 43. For the same theoretical measure ρ^* there are several levels of the coverage, and the level depends on the number of variables. When the number of variables increases from 2 to 4 the coverage is maintained around 80 using either method. When the number of variables increases from 4 to 8, the coverage decreases to the 60's using the jackknife method, and it drops down in the 20's using the bootstrap method. This is a significant result. When there are more uncorrelated extraneous variables involved in estimating the measure ρ^*, larger samples need to be collected to ensure a reliable estimate. Either method can be used when the number of variables is small (compared to the sample sizes). The impact of the dimensionality is always there if the sample size is small relative to the number of variables.

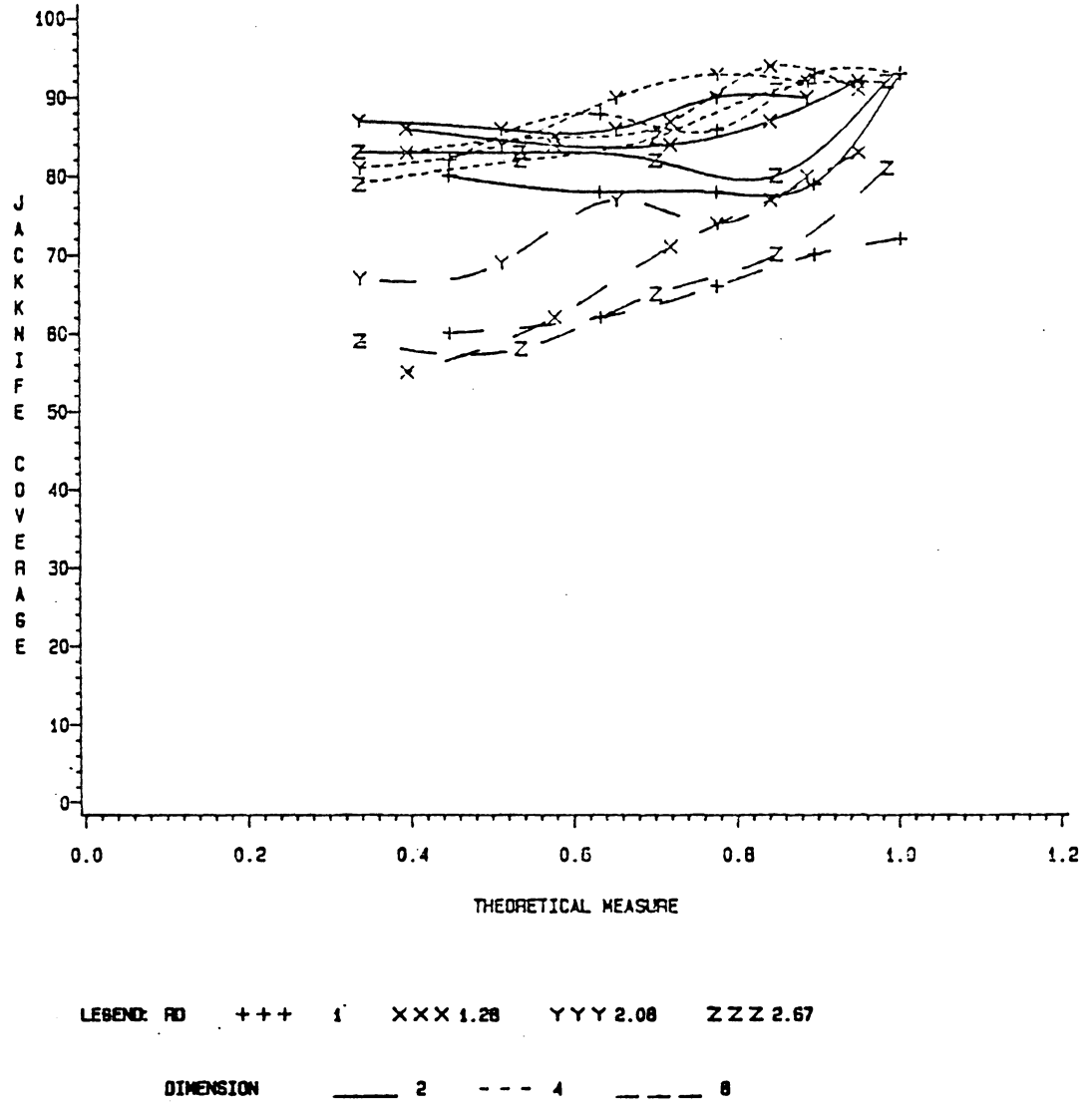

Figure 42. Jackknife coverage versus Theoretical measure: The coverage of 95% jackknife CI versus ρ^* for different dimensions and various ratios of generalized variances. Here $N_1 = 20, N_2 = 20$.

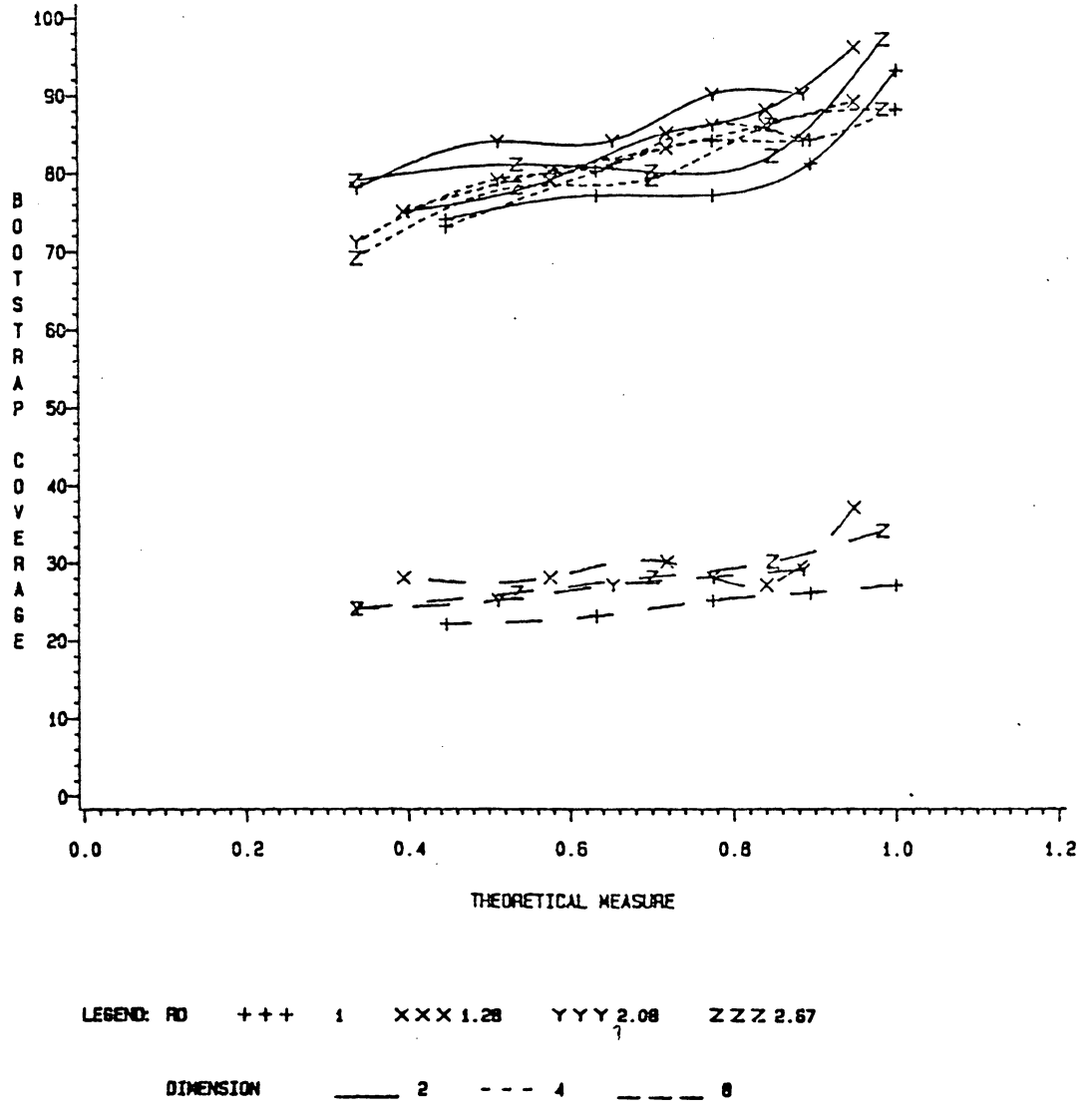

Figure 43. Bootstrap coverage versus Theoretical measure: The coverage of 95% bootstrap CI versus ρ^* for different dimensions and various ratios of generalized variances. Here $N_1 = 20$, $N_2 = 20$.

6.3.5.The ratio of the generalized variances for mean separation

For the sample sizes $N_1 = 20$, $N_2 = 20$, we compare different ratios of generalized variances to the coverage of the 95% jackknife and the 95% bootstrap confidence intervals of the sample estimate $\hat{\rho}^*$ for various mean separations for the bivariate case in figures 44 and 45. When there is no mean separation both methods result in equally good coverage despite various ratios of the generalized variances. When the sample means are different and the ratio of the generalized variances is over 5, the coverage is kept almost at the same level and the differences in means have no impact on the coverage. If the variance-covariance matrices are close to homogeneous, the coverage is bad except when the sample means are the same. Whenever the ratio of the generalized variances is around 1, the mean separation affects slightly the performance of the estimating method.

6.4 LOGARITHMIC TRANSFORMATION OF THE MEASURES

From the previous sections, the variance estimate seems to be a function of the overlap measure. Also by the mathematical forms of the similarity and overlap measures, they are represented as the product of two terms. One term (the exponent) may be regarded as the Mahalanobis generalized distance corresponding to the weighted average of Σ_1 and Σ_2. The second term is a measure of the information contributed by the differences between the covariance matrices Σ_1 and Σ_2. It is reasonable then to look at transformations of the measures to try to stabilize the sample variance of the estimates. There are several transf-

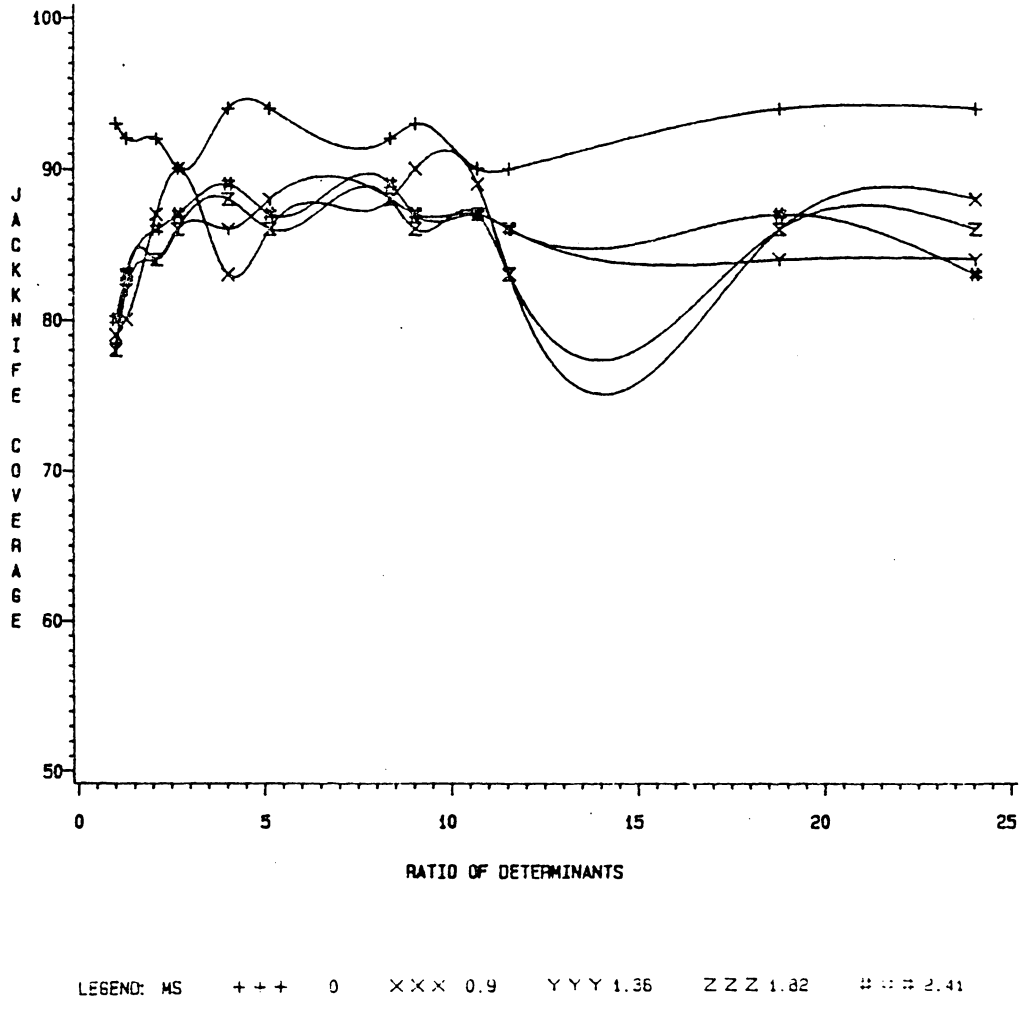

Figure 44. **Jackknife coverage versus RD:** The coverage of 95% jackknife CI versus the ratio of generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

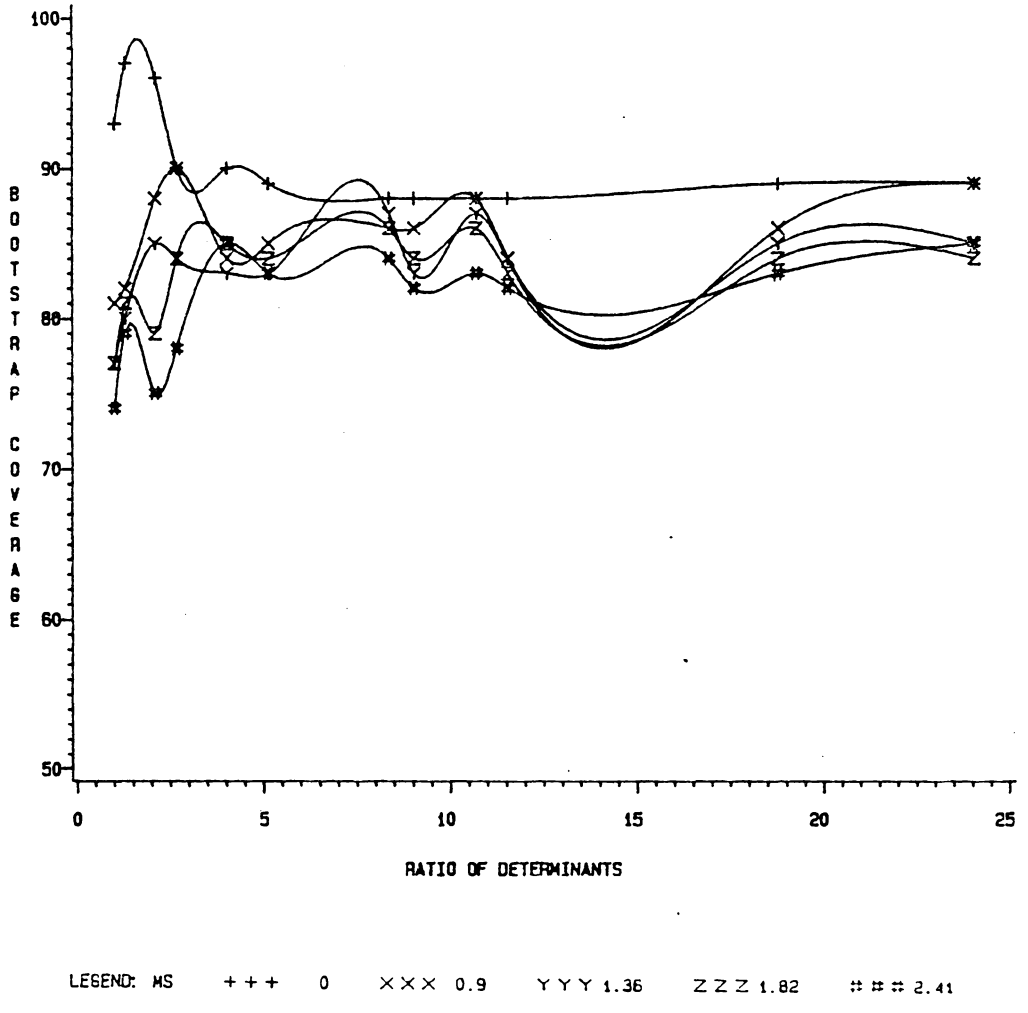

Figure 45. **Bootstrap coverage versus RD:** The coverage of 95% bootstrap CI versus the ratio of generalized variances for different values of mean separation. Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2. For RD, see the beginning of chapter 6.

ormations available, such as: logarithmic, arccosine or arcsine. We will present the logarithmic transformation of the similarity measures in the mathematical forms explicitly, then check the properties of the sampling estimate using the simulation method. The other two transformations did not produce reasonable results.

For Matusita's measure ρ^* , we get

$$\log(\rho^*) = -\frac{1}{4}[(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)] + \frac{1}{4}\log(|\Sigma_1\Sigma_2|)$$

$$-\frac{1}{2}\log(|\frac{1}{2}(\Sigma_1 + \Sigma_2)|) \qquad [6.1]$$

and for Morisita's measure λ^* , we get

$$\log(\lambda^*) = -\frac{1}{2}[(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)] + \frac{1}{2}\log(|\Sigma_1\Sigma_2|)$$

$$+ \log 2 - \frac{1}{2}\log(|\frac{1}{2}(\Sigma_1 + \Sigma_2)|) - \log(|\Sigma_1|^{1/2} + |\Sigma_2|^{1/2}). \qquad [6.2]$$

For the MacArthur-Levins measure α_{ij}^*, $i,j = 1, 2$

$$\log(\alpha_{ij}^*) = -\frac{1}{2}[(\mu_i - \mu_j)'(\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)]$$

$$+ \log(|\Sigma_i|) - \frac{1}{2}\log(|\frac{1}{2}(\Sigma_i + \Sigma_j)|). \qquad [6.3]$$

Also Pianka's measure α^* becomes

$$\log(\alpha^*) = -\frac{1}{2}[(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)] + \frac{1}{4}\log(|\Sigma_1\Sigma_2|)$$

$$- \frac{1}{2}\log(|\frac{1}{2}(\Sigma_1 + \Sigma_2)|). \qquad [6.4]$$

The above expressions can be treated as the adjustment of the generalized distance due to the differences of variance-covariance matrices.

In figure 46, the bias of the sample estimate of log (ρ^*) is presented, and also the jackknife and bootstrap estimates. Comparing this figure with figures 12, 13 and 14, we note that if ρ^* is 1, $\log(\rho^*)$ is 0, and $\log(\rho^*)$ is bounded above by 0 since $0 < \rho^* < 1$. When ρ^* is close to 1 it is biased heavily, but $\log(\rho^*)$ is less biased when it is close to 0. This transformation changes the scope totally. Note especially the linear nature of the bias. This suggests the possibility of further improvements in bias.

In figure 47, the simulated variance for $\log(\rho^*)$ is presented and also the jackknife and the bootstrap estimates, indicating that both methods underestimate the simulated variance of $\log(\rho^*)$, and that there are only slight differences between the two procedures for this case. The jackknife method tends to underestimate more than the bootstrap method does.

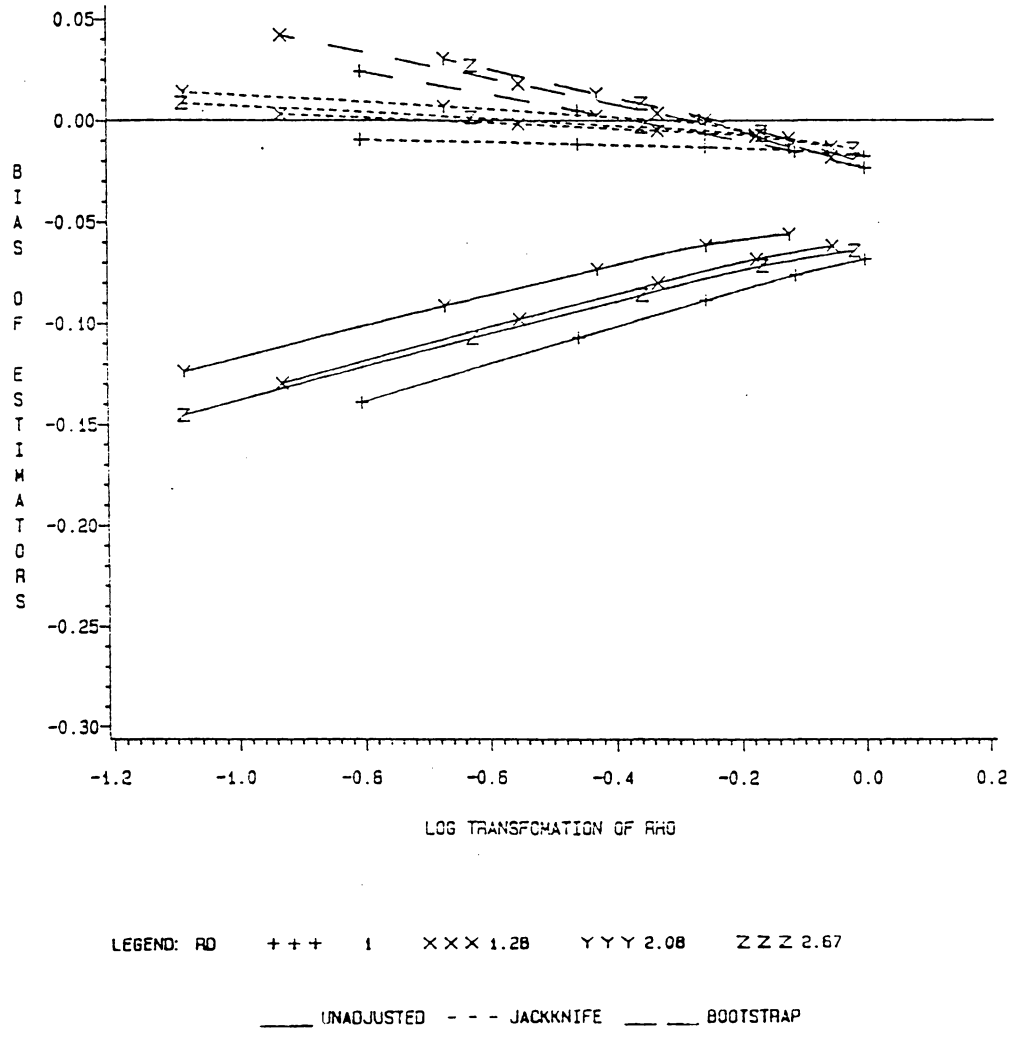

Figure 46. **Bias versus - Log (ρ^*):** Bias of the unadjusted estimate , the jackknife estimate and the bootstrap estimate of - Log (ρ^*) versus - Log (ρ^*). Here $N_1 = 20$, $N_2 = 20$, there is one curve for each of 4 values of RD and the dimension = 2 .

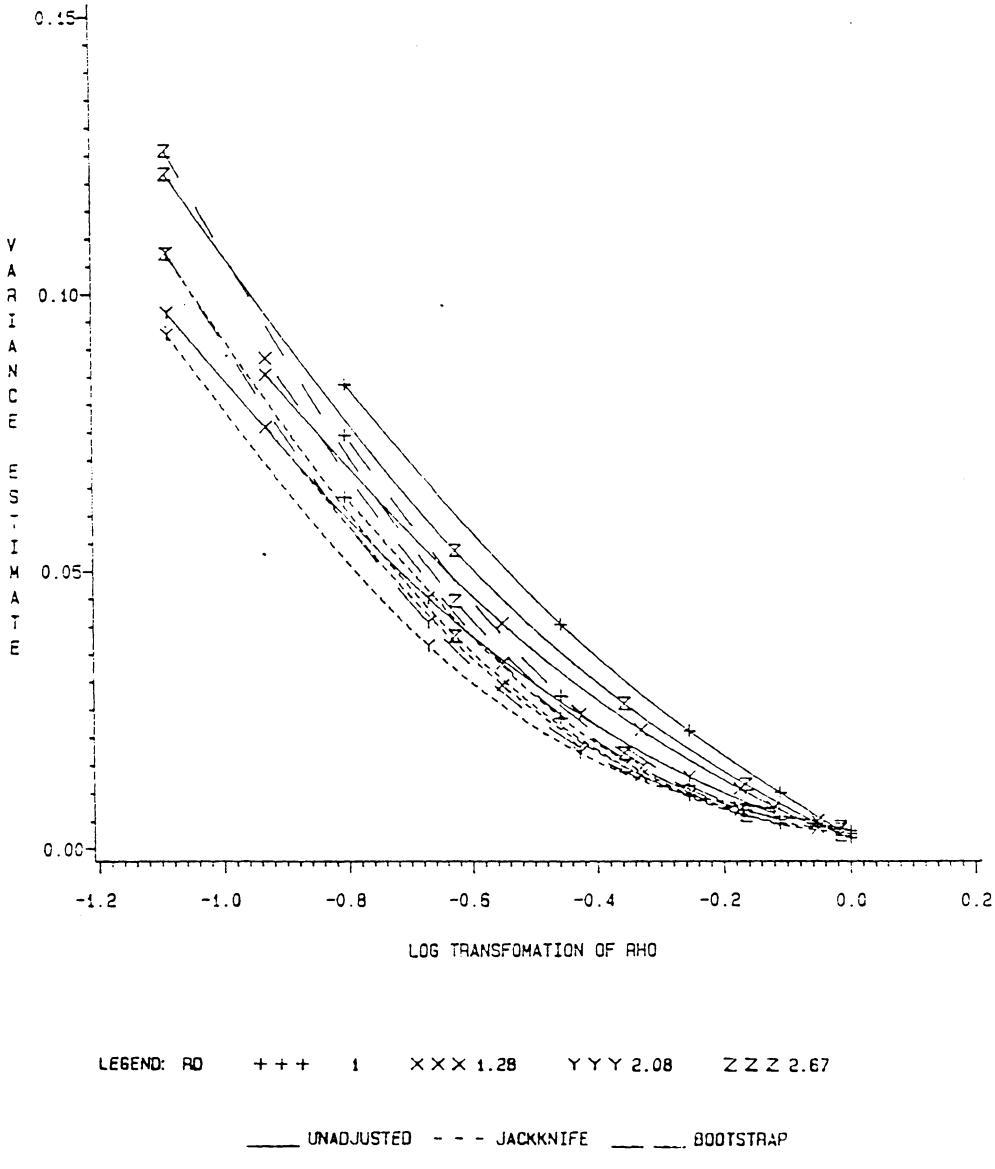

Figure 47. **Simulated variance versus - Log (ρ^*):** The simulated variance , the jackknife variance and the bootstrap variance of - Log (ρ^*) versus - Log (ρ^*). Here $N_1 = 20$, $N_2 = 20$, there are several curves for RD values and the dimension = 2 .

In figure 48, the coverage of the 95% confidence interval for $\log(\rho^*)$ is presented by using the jackknife and the bootstrap methods. It indicates that the coverage is reasonably good and that there are only slight differences between the two procedures for this case. The jackknife has slightly better coverage. Comparing this figure with figures 40 and 41, it is seen that the logarithmic transformation provides some improvement, especially for the bootstrap method.

6.5 COMPARISON OF THE MEASURES

The similarity and the overlap measures discussed in this study are commonly used by ecologists, but are there any discrepancies among them? In this section, we study the sampling properties of estimates using these measures $\rho^*, \lambda^*, \alpha_{12}^*, \alpha_{21}^*, \alpha^*$ and D^2. Generally, the sampling properties of the different measures were quite similar in response pattern with slightly different magnitudes of bias and variability.

In figures 49 and 50, the variance-covariance matrices are assumed to be homogeneous, (the ratio of generalized variance is 1). The sample sizes are 20 and 20, and the number of variables is 2 (the bivariate case). Note that because the exponent of Matusita's measure is different from other measures the curves do not line up. The generalized distance measure has the smallest bias and Morisita's measure has the most bias. There is monotonic increase in the magnitude of bias. When the two populations are far apart, there is no difference in the bias of estimation. Figure 50 shows that the generalized distance measure,

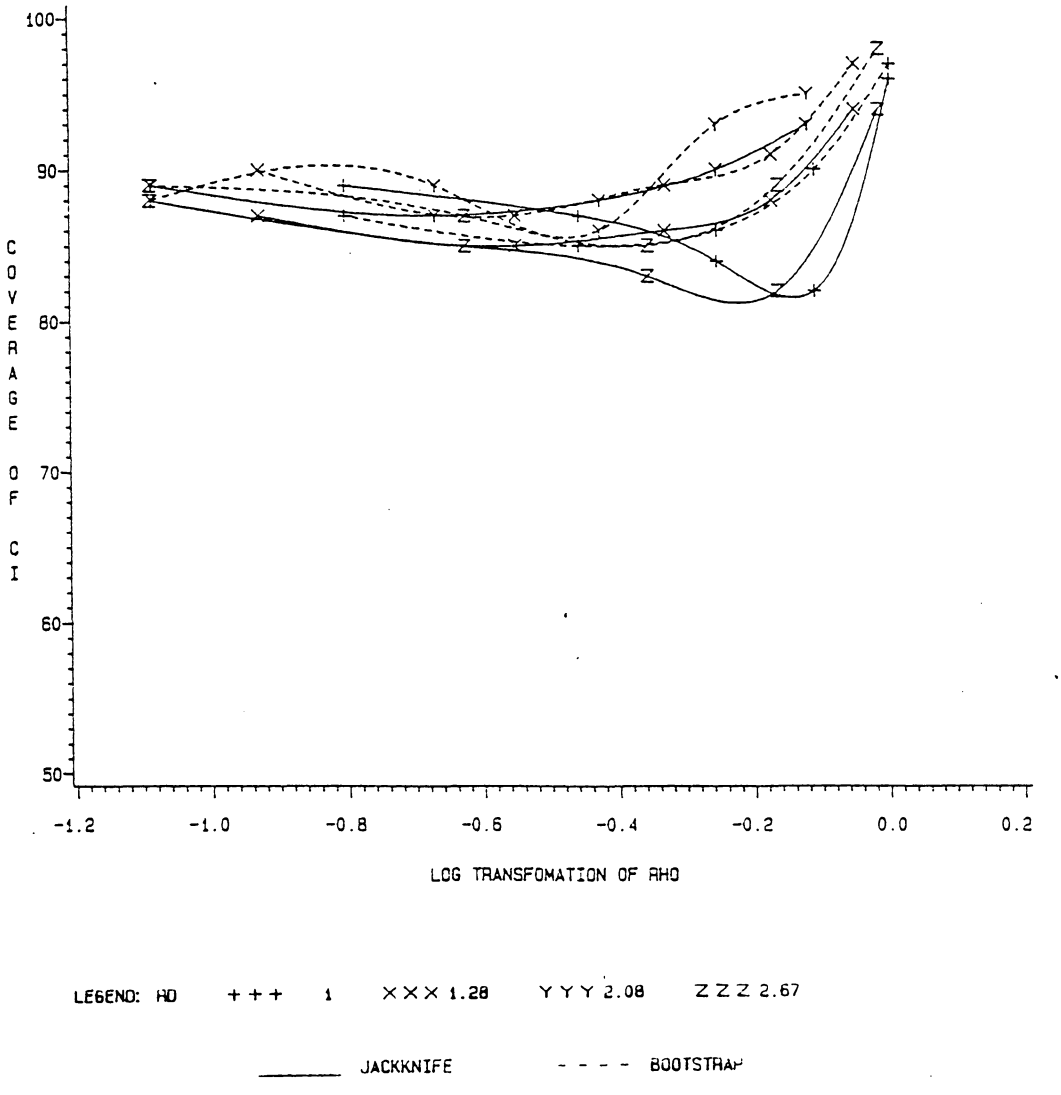

Figure 48. Coverages versus Log (ρ^*): The coverages of the 95% confidence intervals using the jackknife and the bootstrap methods versus log(ρ^*). Here $N_1 = 20$, $N_2 = 20$, the dimension = 2, and there are different curves for different values of RD.

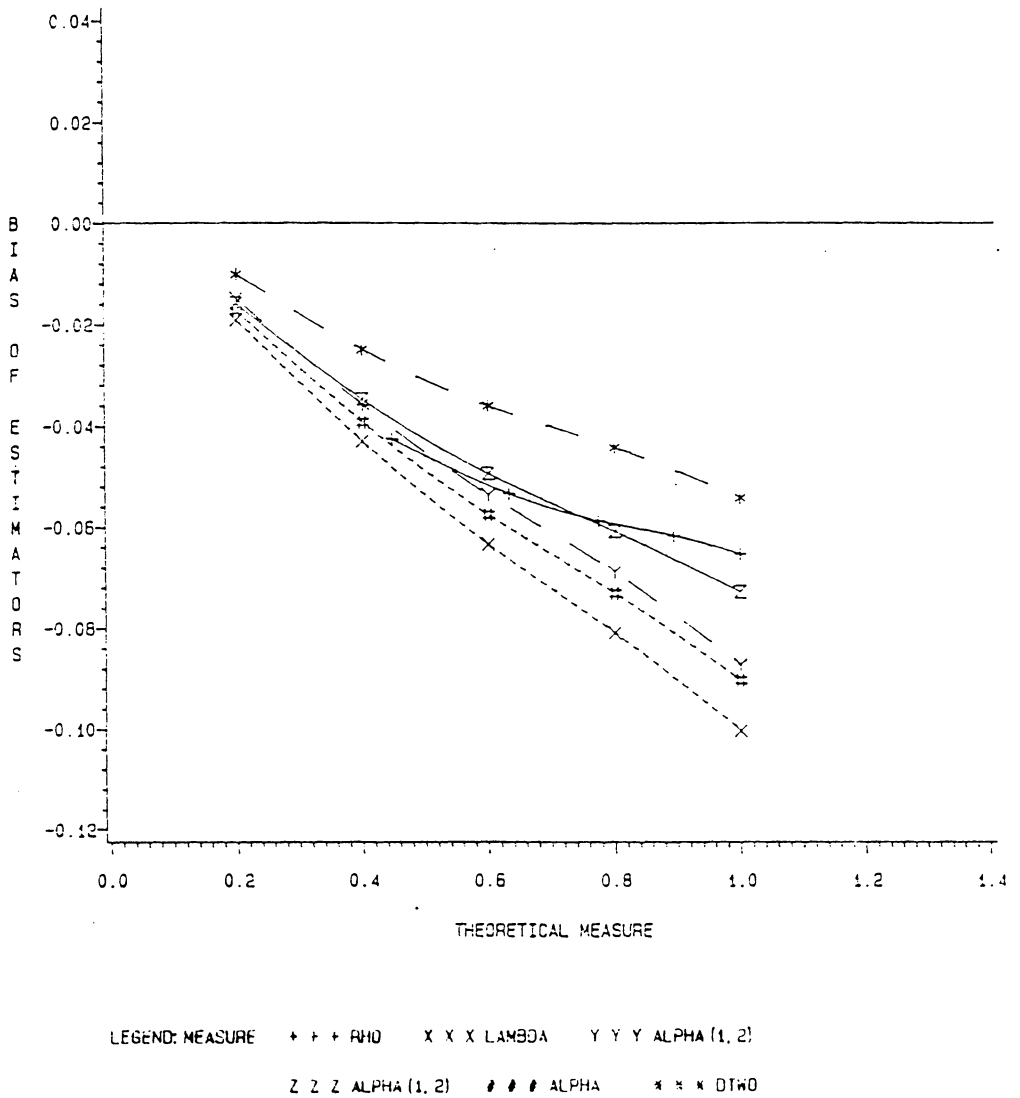

Figure 49. **Bias of estimators versus Theoretical measures (RD = 1):** Bias of the unadjusted estimate of (ρ^*, λ^*, α^*_{12}, α^*_{21}, α^* and D^*). Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2.

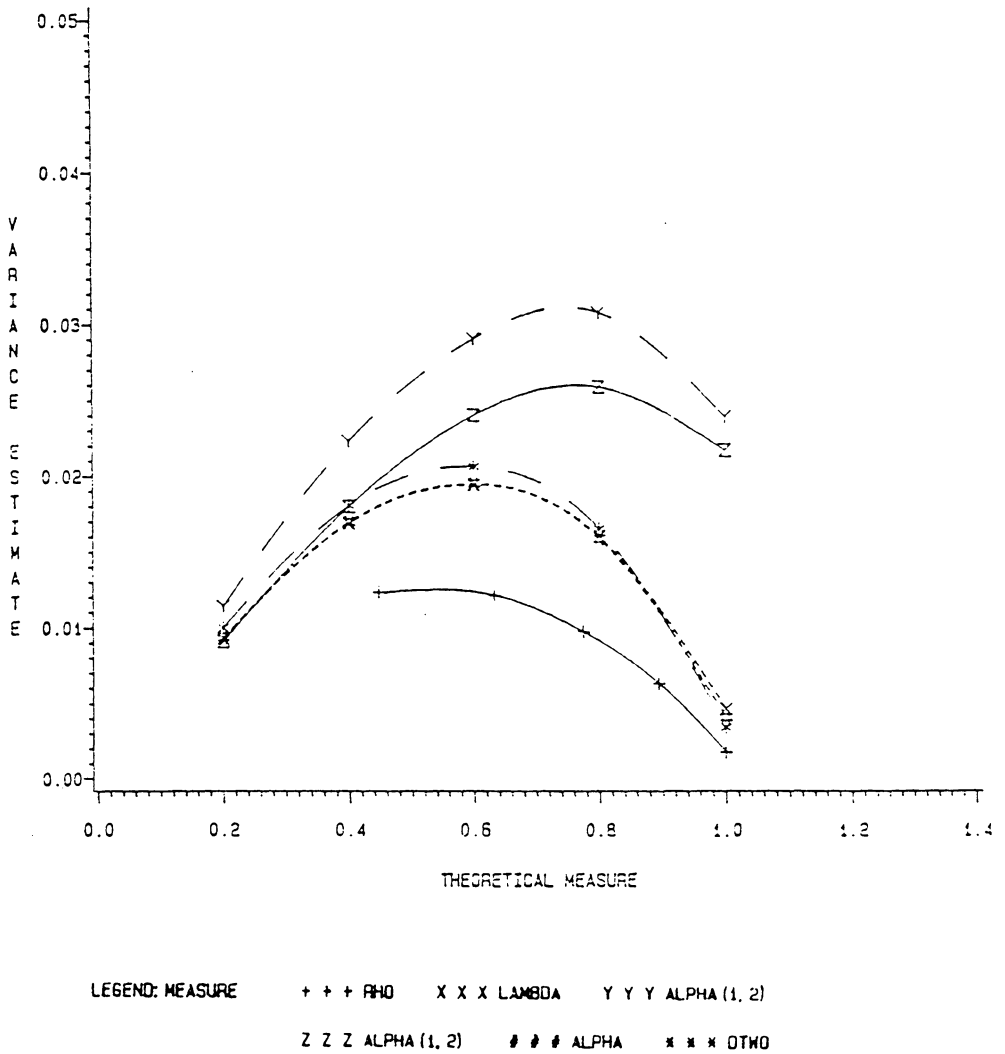

Figure 50. **Variance estimates versus Theoretical measures (RD = 1):** The simulated variances of (ρ^*, λ^*, α^*_{12}, α^*_{21}, α^* and D^2) versus (ρ^*, λ^*, α^*_{12}, α^*_{21}, α^* and D^2). Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2.

Morisita's measure and Pianka's measure have roughly the same variance estimate. The MacArthur-Levins measures ($\alpha^*_{12}, \alpha^*_{21}$) have more variability. The Matusita's measure has the least variation in the sample estimate.

In figures 51 and 52, the bias and variance estimates of these measures are compared. The sample sizes are 20 and 20 for the bivariate normal densities. The ratio of generalized variances is 24. This can be treated as severe heterogeneity. The biases of the estimate have no significant difference among them, but the variance estimates have large discrepancy among them. α^*_{12} has the largest variance estimate due to the asymmetry of the measure. The generalized distance measure has a much larger variance than the other measures. The reason for this phenomenon is that the generalized distance measure does not account for the difference of the variance-covariance matrices. This indicates significant discrepancy among these measures.

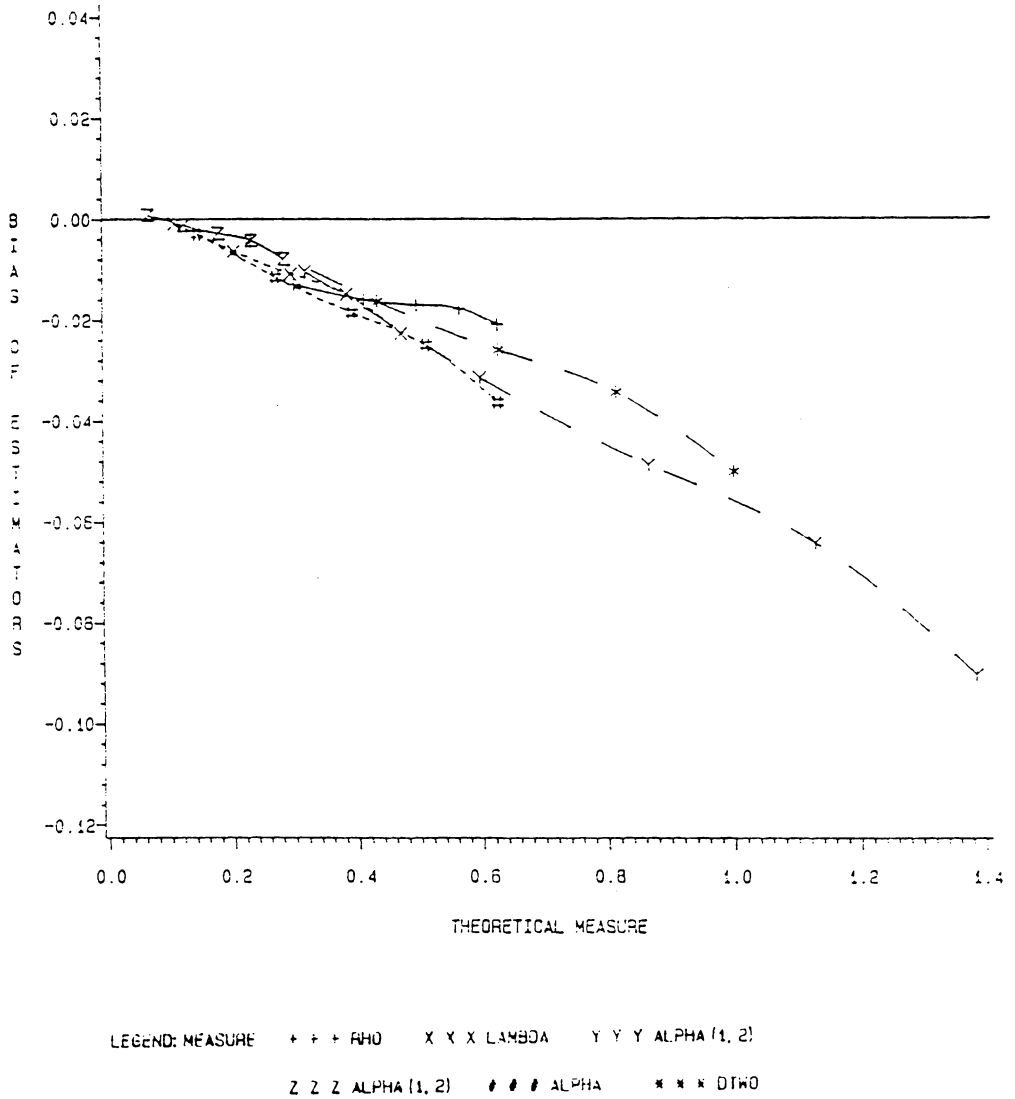

Figure 51. Bias of estimators versus Theoretical measures (RD = 24): Bias of the unadjusted estimate of (ρ^*, λ^*, α^*_{12}, α^*_{21}, α^* and D^*). Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2.

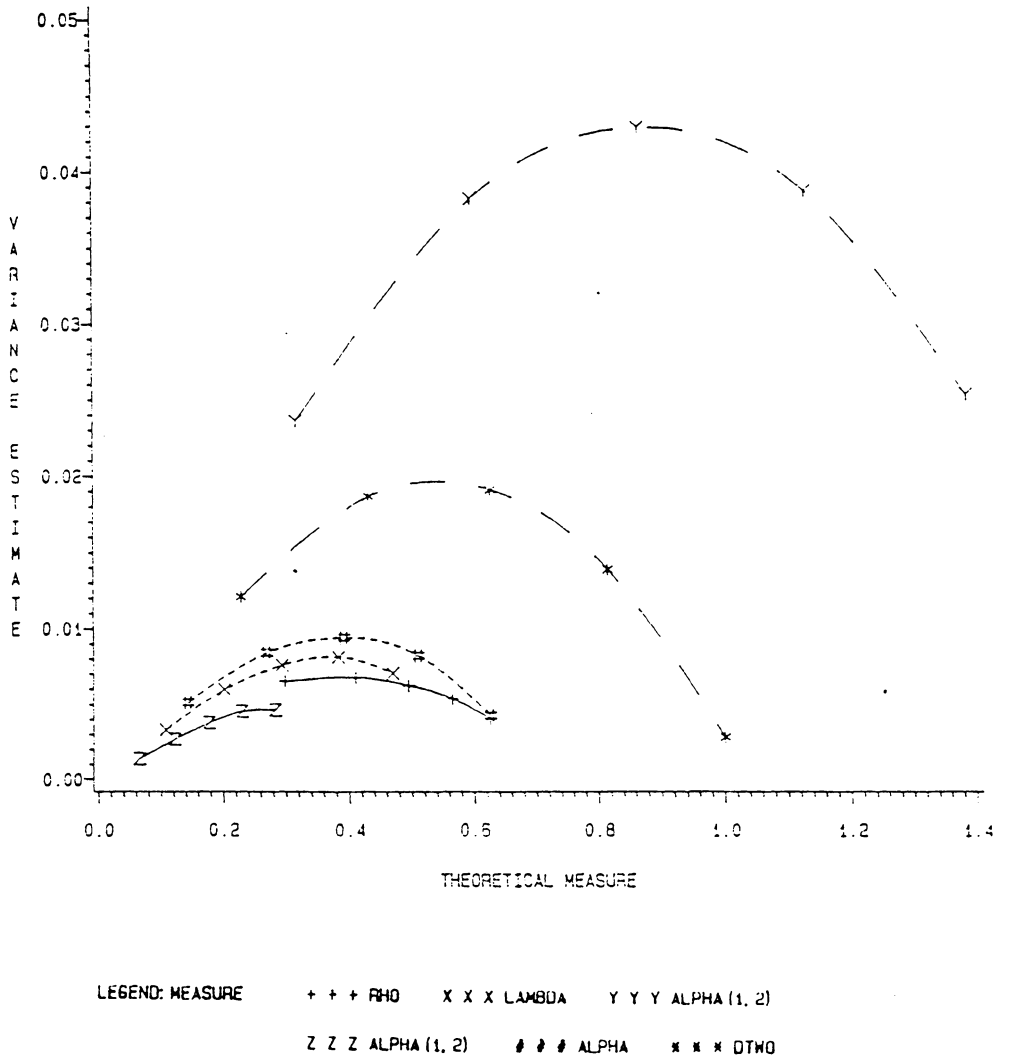

Figure 52. Variance estimates versus Theoretical measures (RD = 24): The simulated variances of estimated measures versus (ρ^*, λ^*, α_{12}^*, α_{21}^*, α^* and D^*). Here $N_1 = 20$, $N_2 = 20$ and the dimension = 2.

Chapter VII

CONCLUSION

In the study of ecological community structure, the multivariate niche model has always been the assumed structural model. This model is closely connected to the multivariate two-sample problem. To measure the degree to which niches of two species overlap is of interest to ecological researchers, and discriminant analysis is the tool used most often to analyze the similarity. In this study, we discuss the most commonly used similarity measures, and develop measures that are less dependent on the assumptions of the usual discriminant analysis. The derivations of the measures assuming normal distributions with heterogeneous variance-covariance matrices are in chapter 3.

In practice, researchers usually collect data with heterogeneous variance-covariance matrices, and the measures developed can give an exact interpretation in the above conditions. The problem of estimating the measures and their precision and accuracy is not a simple problem. Two methods, the jackknife and the

bootstrap, are described for estimating the bias and variance of an estimated measure. The performance of these methods was evaluated using simulation. When the number of variables involved in the model is large, the estimates of these measures may be severely biased, and the bias is consistently negative. By collecting larger samples the bias can be reasonably adjusted. Two potentially important factors affecting results are the disparity in the means and the ratio of the generalized variances. It is shown that when the mean separation is small, these have a moderate effect on the bias, but the effect is limited when the mean separation becomes larger. The usual assumption of homogeneous variance-covariance matrices always gives higher similarity. It is more biased when one takes into account the disparity as the similarity decreases. The variance of the similarity estimates is also related to the estimate and is a quadratic function of the similarity. For Matusita's measure, for example, when the true measure is near 0.6, the variation is the largest and it curves down as the measure approaches 0 or 1. The logarithmic transformation of the similarity is seen to linearize the variance of the similarity estimate.

The jackknife method gives better adjustment of the bias of the estimated measures. Generally, the bootstrap method performs worse than the jackknife method. In some cases, especially when there are many redundant variables neither method gives reliable results.

From the discussion of chapter 6, the number of variables is a crucial factor in estimating the similarity accurately. We think a variable selection procedure is worthy of further investigation. Also it may be useful to apply discriminant

analysis or principal component analysis to select classified variables in the model.

Multivariate normality is assumed in this study. If this condition is loosened, one has to adopt a nonparametric analysis to assess these similarity measures.

Another aspect of multivariate nichemetrics occurs when the data are from a mixture of distributions. The variables that are used may be either continuous or discrete. For each discrete state, a continuous density is assumed and the assessment of the similarity measure in the mixed model is more complicated and involves more computation.

Bibliography

1. Battacharyya, A. (1943) On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin Calcutta Mathematical Society 35: 99-109.

2. Carnes, B. A. and N. A. Slade (1982) Some comments on niche analysis in canonical space. Ecology 63: 888-893.

3. Carter, E. M., C. G. Khatri and M. S. Srivastava (1979) The effect of inequality of variances on the t-test. Sankhya B. 41: 216-225.

4. Chaddha, R. L. and L. F. Marcus (1968) An empirical comparison of distance statistics for populations with unequal covariance matrices, Biometrics 24: 683-694.

5. Chernoff, H. (1973) Some measures for discriminating between normal multivariate distributions with unequal covariance matrices 337-344. Multivariate Analysis III. P. R. Krishnaiah, editor, Academic Press, New York.

6. Dueser, R. D. and H. H. Shugart, Jr. (1978) Microhabitat in forest-floor small-mammal fauna. Ecology 59: 89-98.

7. Dueser, R. D. and H. H. Shugart, Jr. (1979) Niche pattern in a forest-floor small-mammal fauna. Ecology 60: 108-118.

8. Efron, B. (1982) The Jackknife, the Bootstrap and Other Resample Plans. S. I. A. M. Philadephia.

9. Fatti, L. P. and D. M. Hawkins (1986) Variable selection in heteroscedastic discriminant analysis. Journal of American Statistical Association. 81: 494-500.

10. Gilbert, E. S. (1969) The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. Biometrics 25: 505-515

11. Giller, P. S. (1984) Community Structure and the Niche. Chapman and Hall, New York.

12. Good, I. J. and E. P. Smith (1985) The variance and covariance of a generalized index of similarity especially for a generalization of an index of Hellinger and Bhattacharyya. Comm. Statist. - Theor Meth. 14 (12) 3053-3061.

13. Green, R. H. (1971) A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs in central Canada. Ecology 52: 543-546.

14. Green, R. H. (1974) Multivariate niche analysis with temporally varying environmental factors. Ecology 55: 73-83.

15. Green, R. H. (1980) Multivariate approaches in Ecology: The assessment of ecologic similarity. Annals Review of Ecological System. 11: 1-14.

16. Harner, E. J. and R. C. Whitmore (1977) Multivariate measures of niche overlap using discriminant analysis. Theoretical population biology. 12: 21-36.

17. Hellinger, E. (1904) Die orthogonalinvarienten quadratischer Formen von unendlich vielen Variabelen. Gottingen dissertation pp. 84.

18. Hurlbert, S. H. (1978) The measurement of niche overlap and some relatives. Ecology 59: 67-77.

19. Hutchinson, G. E. (1957) Concluding remark. Cold Spring Harbor Symposium in Quantitative Biology 22: 415-427.

20. IMSL (1982) The International Mathematics Subroutine Library. Houston. Texas.

21. Ito, K. and W. J. Schull (1964) On the robustness of the T^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. Biometrika 51: 71-82.

22. Jeffrey, S. H. (1948) Theory of Probability. 2nd edition., Oxford University Press. London.

23. Johnson, R. A. and D. W. Wichern (1982) Applied Multivariate Statistical Methods. Prentice-Hall, Englewood Cliffs, New York.

24. Kullback, S. (1968) Information Theory and Statistics. Dover Publication, New York.

25. MacArthur, R. and R. Levins (1967) The limiting similarity, covergence, and divergence of coexisting species. The American Naturalist. 101:377-385.

26. Mahalanobis, P. C. (1936) On the generalized distance in statistics. Proceeding of National Institute of Science, India 2: 49-55.

27. Marks, S. and O. J. Dunn (1974) Discriminant functions when covariance matrices are unequal. Journal of American Statistical Association 69: 555-559.

28. Matusita, K (1955) Decision rules, based on the distance, for problems of fit, two samples, and estimation. Annals of Mathematical Statistics 26: 631-640.

29. Matusita, K (1964) Distance and decision rules. Annals of Institute of Statistics and Mathematics 16: 305-315.

30. Matusita, K (1966) A distance and related statistics in Multivariate Analysis. pp. 187-200; in Multivariate Analysis I. P. R. Krishnaiah, editor. Academic Press, New York.

31. Matusita, K. (1973) Correlation and affinity in Gaussian cases. pp. 345-349; in Multivariate analysis III. P. R. Krishnaiah, editor. Academic Press, New York.

32. Maurer, B. A. (1982) Statistical inference for MacArthur-Levins niche overlap. Ecology 63: 1712-1719.

33. Miller, R. G. (1974) The jackknife-a review. Biometrika 61: 1-15.

34. Morisita, M. (1959) Measuring of interspecific association and similarity between communities. Mem. Fac. Sci. Kyushu Univ. ser. E. 65-80.

35. Muirhead, R. J. (1982) Aspects of Multivariate Statistical Theory. John Wiley. New York.

36. Pianka, E. R. (1974) Niche overlap and diffuse competition. Proceedings of the National Academy of Science 71: 2141-2145.

37. Porter, J. H. and R. D. Dueser (1982) Niche overlap and competition in an insular small mammal fauna: A test of the niche overlap hypothesis. Oikos 39:228-236.

38. Rao, C. R. (1982) Diversity and dissimilarity coefficients: A unified approach. Theoretical Population Biology 21: 24-43.

39. Ricklefs, R. E. (1980) Ecology. 2nd edition. Nelson, Walton on Thames.

40. Seber, G. A. F. (1984) Multivariate Observations. John Wiley, New York.

41. Smith, E. P. (1982) Niche breadth, resource availability and inference. Ecology 63: 1675-1681.

42. Smith, E. P. (1984) On the asymptotic variance of Socransky's proportional similarity index. Journal of Statistical Computation and Simulation 19: 90-94.

43. Smith, E. P. (1985) Estimating the reliability of diet overlap measures. Environmental Biology of Fishes. 13:125-138.

44. Shugart, H. H. Jr. and B. C. Patten (1982) Niche quantification and the concept of nich pattern. 284-327. Systems Analysis and Simulation in Ecology, II. B. C. Patten, editor. Academic Press, New York.

45. Van Belle, G. and I. Ahmad (1971) Measuring affinity of distributions, pp. 651-668. Reliability and Biometry. F. Proschan and R. J. Serfling, editors. S. I. A. M.

46. Van Horne, B. and R. D. Ford (1982) Niche breadth calculation based on discriminant analysis. Ecology 63: 1172-1174.

47. Williams, B. K. (1983) Some observations on the use of discriminant analysis in ecology. Ecology 64: 1283-1291.

48. Zaret, T. M. and E. P. Smith (1984) On measuring niches and not measuring them. 127-138. Evolutionary Ecology of Neotropical Freshwater Fishes. T. M. Zaret, editor. Dr. W. Junk Publishers, The Hague.

The vita has been removed from the scanned document