STATISTICAL PROPERTIES OF THE FILE-MERGING METHODOLOGY

by

Thirugnanasambandam Ramalingam
Purdue University

Technical Report #85-21

Department of Statistics
Purdue University

August 1985

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont'd.)

LIST OF TABLES

LIST OF TABLES (Cont'd.)

LIST OF FIGURES

LIST OF FIGURES (Cont'd.)

Figure Page

ABSTRACT

Ramalingam, Thirugnanasambandam, Ph.D., Purdue University, August 1985. Statistical Properties of the File-merging Methodology. Major Professor: Prem K. Goel

Matching is defined as the methodology of merging micro-data files to create larger files of data. Matching is often done to extract statistical information which cannot be obtained from the individual files that are incomplete. Current Federal statistical practice involving multivariate file-merging techniques is typically not based on a formal statistical theory. In view of this situation, a survey on matching is given. All known models for matching are presented under a unified framework, which consists of three situations involving the same or similar individuals.

The properties of a maximum likelihood strategy to match files of data involving the same individuals are derived via ranks and order-statistics from bivariate populations. In addition, the properties of this strategy have been examined with respect to a more reasonable criterion called epsilon-correct matching. Asymptotic results for such situations, including (i) the Poisson approximation for the distribution of the number of correct matches, and (ii) convergence in probability of the average number of epsilon-correct matches have been derived. Small-sample properties, like the monotone behavior of the expected number of matches with respect to the dependence parameters of the underlying models have been proved.

Two matching strategies due to Kadane (1978) and one strategy due to Sims (1978) for merging files of data on similar individuals are discussed. These strategies are evaluated based on a Monte-Carlo study of matching models involving trivariate normal distributions.

CHAPTER 1

INTRODUCTION

One of the most important tools for analyzing economic policies is the micro-analytic model. This technique is used frequently in public decision-making centers. Virtually every Federal Agency uses micro-analytic models for the evaluation of policy proposals.

Direct use of sample observations rather than aggregated data is characteristic of the micro-analytic approach. For this reason, the micro-data that is used as input to the model has a significant bearing on the validity of the results of the model. Furthermore, when all the input-data come from a single sample, the quality of the model depends on, among others, sampling and data-recording procedures. However, if the data from a single source is insufficient or partly aggregated, then typically multiple sources of data are used to provide the necessary input to the model. At the same time, issues such as validity and quality of the results of the model cannot be assessed as easily as when we have a single source of data as input. In such situations, government statisticians have been using a methodology in which multiple sources of data are merged to form a composite data-file. Effective use of the different pieces of data in order to produce sensible but more comprehensive files is a fundamental issue in the file-merging methodology.

Some of the difficulties associated with the merging procedures and techniques for their resolution have been known for quite sometime. Initiated by the Federal Subcommittee on Matching Techniques, there has recently been renewed effort to establish solid theoretical foundation and empirical justification for the file-merging methodology. This research reviews the relevant literature and then presents new statistical properties of some known procedures for merging data-files. We shall now give an example of a typical situation in which merging of two files is carried out.

1.1 A Paradigm

A micro-economic model in heavy use at the Office of Tax Analysis (OTA), Department of the Treasury, is the Federal Personal Income Tax Model. This model is used to assess proposed tax law changes in terms of their effects on the distribution of after-tax income, the efficiency with which the changes will operate in achieving their objectives, etc. The inputs for this model are two sources of micro-data, namely the Statistics of Income File (SOI) and the Current Population Survey (CPS). The SOI file is generated annually by the Internal Revenue Service (IRS) and it consists of personal tax return data. The CPS file is produced monthly by the Bureau of the Census. As we will explain in Section 1.2, such pooling of data from more than one Federal Agency has been severely restricted in recent years by, among others, confidentiality issues such as the privacy of the individuals involved in the aforementioned files of data. For this reason, complete information, especially identifiers such as social security numbers, is typically not released by the IRS and the Census Bureau. The resulting micro-data files are compromises between complete Census files and fully aggregated data-sets. Thus, sufficient detail remains to support micro-analysis of the population, while partial aggregation protects individual privacy and greatly diminishes computational burden.

A typical problem in tax-policy evaluation occurs when no single available data file such as SOI or CPS contains all the information needed for an analysis. For example, consider the variables $\underset{\sim}{W} = (X,Y,Z_1,Z_2)$, where

X = Allowable itemizations and capital gains

Y = Old Age Survivors Disability Insurance (OASDI)

Z_1 = Social security number

Z_2 = Marital status

Suppose that we are interested in estimating a simple correlation $\rho_{X,Y}$ between X and Y or, more generally, the expectation of a known function g, say, of W; that is the integral

$$\gamma = \int g(\underset{\sim}{w})\, dF(\underset{\sim}{w}) \tag{1.1.1}$$

where $F(\underset{\sim}{w})$ is the joint distribution function of the variables in $\underset{\sim}{w}$. Now, the SOI microdata file cannot be used in its original form since it does not include the OASDI benefits (Y). Census files (CPS) with OASDI benefits do not allow a complete analysis of the effect of including this benefit, since it does not contain information on allowable itemizations and capital gains (X). Thus, instead of observing X, Y, Z_1, Z_2 jointly on the same units, we have to get only the following pair of files:

File 1 (SOI): X, Z_1, Z_2

and

File 2 (CPS): Y, Z_1, Z_2

Estimating γ based on the fragmetary data provided by File 1 and File 2 is an important practical problem that has not yet been solved satisfactorily. In an attempt to cope with situations such as the OTA model, Federal Agencies have long been using procedures for matching or merging the two incomplete files so that one can do the usual inference for γ, hoping that the merged-file is a reasonable substitute for the unobserved data on (X, Y, Z_1, Z_2).

The reporting units in CPS are households. In general, the units in a file may refer to other types of legal persons, like corporations, partnerships and fiduciaries. The term "individual" will be used as a generic label in this thesis to refer to the reporting units of the micro-data files.

1.2 A Dichotomy of Matching Problems

Roughly speaking, there are two different categories of matching problem. The first category consists of problems of exact matching in which it is desired to identify pairs of records in the two files that pertain to the same individual. Accurate information on identifiers such as social security number, name, address are assumed to be available when exact-matching the two files. It is clear that all we need to carry out an exact match of two files is, among other tools, an efficient software to sort the individuals by their identifiers. With the help of such software, we can, within reasonable error, link a given individual in File 1 with an individual in File 2 such that these two units possess the same values for the identifiers. The resulting merged file contains data which are more comprehensive than both File 1 and File 2. Also, even after merging, most records will pertain to the same individual, the number of erroneous matches in the enlarged file depending on the particular software used in the process of merging. It is clear that, if accurate identifiers are available for the units in the two files, then no statistical issues are involved in the matching methodology and we shall not discuss this type of problem any more. However, one may refer to, among others, Fellegi and Sunter (1969) and Radner et al. (1980) for work related to the exact-matching methodology. We shall close our discussion of this type of matching problem by noting some of the reasons why exact matching of files is often not possible.

First, over the past several years, there have been significant changes in the laws and regulations pertinent to exact matching of records for statistical and research purposes. New laws, especially the Privacy Act of 1974 and the Tax Reform Act of 1976, have imposed additional restrictions on the matching of records belonging to more than one Federal Agency and on the matching of files of Federal Agencies with those of other organizations. As a result of these laws, some Agencies have limited access to their records for statistical purposes to an even greater extent than seems necessary by statutory requirements.

Second, analyses of microdata often involve data from units that are not available from a single source but are available from several sources. For example, suppose that one is interested in the relationships among two sets of variables, one set consisting of information about health care expenses incurred by individuals and the other set consisting of information about receipt of various types of welfare benefits. Suppose further that no existing data file contains all of the needed variables, but that two samples of a target population, which come from two different surveys, together contain all these variables. If executing a new survey to obtain all the variables from a single sample is not feasible, then one might match the two samples and use the merged file for statistical analyses of variables which are not present in the same sample. Note that the two sample surveys may have information on the <u>same</u> individuals whose identities are either unknown or unreliable. However, in the aforementioned example, it is more appropriate to assume that the two samples contain very few or no individuals in common. In case the two samples are <u>stochastically independent</u>, we shall describe the units in the two samples as <u>similar</u> individuals.

Suppose, then, that exact matching is not feasible in view of the aforementioned reasons. Then the tools that are used in the exact matching methodology are inadequate for the purpose of merging the two files of data. In particular, identifiers are practically useless. However, the probabilistic structure of the populations that generate the data in the two files or other statistical techniques can often be used to combine the two files. Such procedures will be called statistical matching strategies.

In the literature on matching files there is no consensus on rigid definitions of Exact Match and Statistical Match. Indeed, it is traditional to distinguish these two types of problem by verifying whether same (exact) or similar (statistical) individuals are in the two files. Our classification of matching problems is somewhat different from the usual practice in the sense that any procedure for merging files, which may contain the same or similar individuals, will be described as a statistical match if statistical techniques are involved in the process of merging. This convention is in agreement with that of Woodbury (1983), who describes certain matching problems involving the same individuals in two files as "Statistical Record Matching for Files".

1.3 A General Set-up for Statistical Matching

Consider a universe $\mathcal{U}$ of individuals. Let $\underset{\sim}{X}$, $\underset{\sim}{Y}$, $\underset{\sim}{Z}$ denote three groups of random variables and let us assume that we cannot observe the vector $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$ for any unit in $\mathcal{U}$. However, suppose that the following data are available:

(Base) File 1: n_1 individuals, each with information on a function $\underset{\sim}{W}_1^*$, say, of $\underset{\sim}{W}$.

and (Supplementary) File 2: n_2 individuals, each with information on a function, $\underset{\sim}{W}_2^*$, say, of $\underset{\sim}{W}$.

Various matching problems arise depending on what type of data are in $\underset{\sim}{W}_1^*$ and $\underset{\sim}{W}_2^*$. We distinguish only three different situations:

<u>Case I</u>: $\underset{\sim}{W}_1^* = \underset{\sim}{X}$ and $\underset{\sim}{W}_2^* = \underset{\sim}{Y}$; we also assume that the two files contain the <u>same</u> individuals.

<u>Case II</u>: Let $\underset{\sim}{W}_1^* = (\underset{\sim}{X},\underset{\sim}{Z})$, $\underset{\sim}{W}_2^* = (\underset{\sim}{Y},\underset{\sim}{Z})$. As in Case I, we further assume that the two files contain the <u>same</u> individuals.

<u>Case III</u>: Let $W_1^* = (\underset{\sim}{X},\underset{\sim}{Z})$, $W_2^* = (\underset{\sim}{Y},\underset{\sim}{Z})$. Unlike in Cases I and II, we assume that the two files contain <u>similar</u> individuals.

1.4 <u>The Matching Methodology - Some Important Steps</u>

We shall now mention some steps involved in actually creating a statistical match between two given files. First, if the populations represented by the files differ, a "universe adjustment" is carried out to ensure that there is a common universe $\mathcal{U}$ from which the individuals of the two files are sampled. Second, a "units adjustment" might be needed if the units of observation in the two files differ (e.g. persons and tax units). Third, "matching or common variables," $\underset{\sim}{Z}$, are defined and it is assumed that File 1 with n_1 records carries information on $(\underset{\sim}{X},\underset{\sim}{Z})$, whereas File 2 with n_2 records consists of data on $(\underset{\sim}{Y},\underset{\sim}{Z})$. The variables $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are often called non-matching variables. Finally, in the "merging" step, if the records $(\underset{\sim}{X}_i,\underset{\sim}{Z}_i)$, and $(\underset{\sim}{Y}_j,\underset{\sim}{Z}_j)$, respectively from File 1 and File 2, are to be matched, then one completes the i^{th} record in File 1 by substituting $\underset{\sim}{Y}_j$ for the missing value. Thus, we get the synthetic File 1:

$$(\underset{\sim}{X}_i,\underset{\sim}{Y}_j,\underset{\sim}{Z}_i), \quad i = 1,2, \ldots, n_1$$

Clearly, the same methodology can be used to get a synthetic File 2 by finding substitutes for missing $\underset{\sim}{X}$ values of File 2 using $\underset{\sim}{X}$'s from File 1. However, in order to keep our discussion simple, we shall often be concerned with completing only File 1. Although, many different methods have been used in this final step, several basic similarities can be identified. In most matches, certain Z variables are treated as the so-called "cohort" variables. Such variables establish "packets" of the records in each of the two files, with matching permitted only between pairs of cases in the same packet. For example, sex is often a cohort variable so that a male can be matched with another male, and a female with another female. This step about the formation of cells or packets is aimed at diffusing the dissimilarities between units that are being matched. Furthermore, depending on how many of the common variables are used as cohort variables, there may be very little or no within-packet variation with regard to $\underset{\sim}{Z}$. In such situations, File 1 has data on $\underset{\sim}{X}$ and File 2 has data on $\underset{\sim}{Y}$ and we would like to merge the files to get joint information on $\underset{\sim}{X}$ and $\underset{\sim}{Y}$. Note that, in Section 1.3, such a scenario was labeled Case I. The selection of "matching records" within a packet is typically based on a "measure of dissimilarity" by which a "distance" is computed between a given File 1 record and each potential match in the supplementary file. A potential match with the smallest distance is chosen as the match that will provide the missing $\underset{\sim}{Y}$ value to a File 1 record.

1.5 Two Basic Types of Matching Strategies

Suppose that the age of an individual, Z_1, say, is a matching variable. Then, one may define a distance measure d, say, between individuals i in File 1 and j in File 2 by the equation

$$d_{ij} = |Z_{1i} - Z_{2j}| \tag{1.5.1}$$

For fixed $i = 1,2, \ldots, n_1$, one will then match one possible j^* in File 2 with i^{th} record in File 1 if j^* minimizes d_{ij} over j. That is, j^* depends possibly on i and satisfies the restriction

$$d_{ij^*} = \min_{1 \leq j \leq n_2} d_{ij} \tag{1.5.2}$$

If the choice of j^* involves no other restrictions, then the statistical matching strategy is called "Unconstrained Matching". However, there are typically additional restrictions subject to which one must choose the optimal match j^* from File 2. Matching data-files with the restriction that the variance-covariance matrix of data items in each file be identical to the variance-covariance matrix of the same data items in the matched file is an example of a "Constrained Match."

In order to formulate this type of merging mathematically, assume first for simplicity, that both files carry only n records; that is, the common value of n_1 and n_2 is n. Let

$$a_{ij} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ record in File 1 is matched with the } j^{\text{th}} \\ & \text{record in File 2} \qquad 1 \le i,\ j \le n \\ 0 & \text{if the } i^{\text{th}} \text{ record in File 1 is not matched with the} \\ & j^{\text{th}} \text{ record in File 2} \end{cases} \qquad (1.5.3)$$

Then, the following additional conditions will ensure that the aforementioned preservation of moments is achieved by not letting more than one record in File 1 to be matched with the same record in File 2:

$$\sum_{i=1}^{n} a_{ij} = 1, \text{ for } j = 1,2, \ldots, n \qquad (1.5.4)$$

$$\sum_{j=1}^{n} a_{ij} = 1, \text{ for } i = 1,2, \ldots, n \qquad (1.5.5)$$

Now let d_{ij} denote, as in the case of a unconstrained match, a measure of inter-record dissimilarity given by the extent to which the attributes in any one record differ from the same attributes in another record. Then the optimal constrained match minimizes the "objective function"

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}\, a_{ij} \qquad (1.5.6)$$

Subject to the restrictions in (1.5.3) to (1.5.5). Clearly, this extremal problem is the standard linear assignment problem in "Optimization."

A matching situation more typical of problems relating to policy analyses is a constrained merge of two files with variable weights in both files and an unequal number of records in the files. Let α_i be the weight of the i^{th} record in File 1, and let β_j be the weight of the j^{th} record in File 2. If n_1, n_2 are respectively, the number of records in File 1 and File 2, then we minimize the objective function in (1.5.6) subject to the following constraints.

$$\sum_{j=1}^{n_2} a_{ij} = \alpha_i, \; i = 1,2, \ldots, n_1 \qquad (1.5.7)$$

$$\sum_{i=1}^{n_1} a_{ij} = \beta_j, \; j = 1,2, \ldots, n_2 \qquad (1.5.8)$$

$$\sum_{i=1}^{n_1} \alpha_i = \sum_{j=1}^{n_2} \beta_j \qquad (1.5.9)$$

and

$$a_{ij} \geq 0, \; \forall \; i \text{ and } j \qquad (1.5.10)$$

It is clear that an optimal constrained matching strategy when the two files have unequal number of individuals is the solution of a standard transportation problem in which the roles of the "warehouses" and "markets" are respectively played by the records in File 1 and File 2 and the "cost of transportation" is the inter-record distance "d_{ij}". Existing algorithms to solve a linear assignment or transportation problem can be used to complete the final "merge" step, giving us the synthetic sample

$$\underset{\sim}{W}_i^* = (\underset{\sim}{X}_i, \underset{\sim}{Y}_i^*, \underset{\sim}{Z}_i), \quad 1 \leq i \leq n_1, \qquad (1.5.11)$$

where $\underset{\sim}{Y}_i^*$ denotes the value of $\underset{\sim}{Y}$ assigned to the i^{th} record of File 1. The sample in (1.5.11) may now be used to estimate a parameter like γ in (1.1.1).

1.6 Criticisms of Statistical Matching

In Sections 1.4 and 1.5, we described the general form of most matching techniques that have been used by Federal Agencies. Matching records at the "packet" level means basically that the random vectors $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are stochastically independent, given the value of the common variables $\underset{\sim}{Z}$. In the particular case of a multivariate normal distribution for $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$, conditional independence assumption is equivalent to the claim that the partial correlations among $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ variables, controlling on the $\underset{\sim}{Z}$ variables, are all zero. This point was made first by Sims (1972) and repeatedly by others since then. The conditional independence assumption is a strong one for which convincing justifications has generally not been offered. It implies that the relationships between $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ can be totally inferred from $\underset{\sim}{X}$'s relation to $\underset{\sim}{Z}$ and $\underset{\sim}{Y}$'s relationship to $\underset{\sim}{Z}$. Sims (1978) stated that matching the files under such assumptions is unnecessary. He also sketched an alternative statistical procedure that uses the data in the two files to estimate, under conditional independence, a parameter such as γ in (1.1.1). Sims' alternative will be discussed further in Section 3.2.

Fellegi (1978) and many other investigators have expressed great caution about the use of statistical matching because not much is known about the accuracy of the estimates of the joint distribution of $\underset{\sim}{W}$ produced by synthetic files.

Notwithstanding these criticisms of statistical matching, there is no viable alternative statistical procedure that will, in general, provide better estimates of γ than a synthetic file can offer. Given this lack of good alternatives, especially when conditional independence does not hold, the area of statistical matching is wide open and both theoretical and empirical investigations to discover the properties of synthetic data-files are in order.

1.7 Reliability of Synthetic Files

The precision of synthetic-file-based estimators of a given parameter relevant to the population of $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$ is affected by various types of errors that occur while matching two files. To discuss these matching errors, let us first restrict our attention to the cases where the <u>same</u> individuals are in the two files, namely Case I and Case II.

In practice, it is almost inevitable in most matching projects that some matching errors occur, even with the most sophisticated procedure and the most careful execution of matching of the files. These errors fall into two major categories:

(i) Erroneous match (false match) or linking of records that correspond to different individuals.

(ii) Erroneous non-match (false non-match) or failure to link the records that do correspond to the same individual.

The reliability of the results of a statistical matching strategy is often defined (Radner et al., 1980, p. 13) as one of the following coefficients:

(a) the proportion of the correct matches, that is, matches of records on the same individuals.

(b) the proportion of erroneous decisions, that is, false matches and erroneous non-matches.

These reliability coefficients are random variables because, in view of the terminological conventions of Section 1.2, a statistical matching strategy is dependent on the data in the two files. The sampling distribution of the reliability coefficients, either exact or asymptotic (as the sizes of the files grow), are very useful in judging the quality of a given matching procedure.

Now, we will discuss the reliability of a synthetic file in Case III, where the two files contain very few or no overlapping individuals. First, note that the definitions of error in the results of matching, which have been proposed for Case I, are not applicable to Case III because the linkage of records from the two files that pertain to the same unit seldom occurs in Case III. In other words, almost all linkages in Case III are false matches in the sense of the definitions given earlier in this section. In Case III, definitions of error and reliability which are tractable from a theoretical perspective are unavailable at this time. In fact, little theoretical work on the errors present in the synthetic files of Case III has been done. Until now, the evaluation of a given matching strategy in Case III has been done from an empirical point of view. A case in point is the work of Rodgers (1984).

1.8 Thesis Outline

In Section 1.3, three important cases for merging two files of data were distinguished. Of these, Case I and Case II are relevant when the same individuals are represented in the two files. Case III arises when only similar individuals are present in the files. This research is concerned with both theoretical investigations and empirical evaluations of the quality of synthetic files in Case I and Case III. We shall not discuss Case II in this thesis.

In Chapter 2, Case I is discussed at some length. A review of known results for this case is given. New optimality properties of a maximum likelihood matching strategy are established. Some small-sample and large-sample properties of the number of correct matches with regard to this strategy are derived, shedding some light on the reliability of the synthetic file arising from using the maximum likelihood strategy.

Case III is the topic of interest in Chapter 3. The bulk of the discussion in this Chapter is confined to matching two files of data that are sampled from a trivariate normal population. Thus, if (X,Y,Z) is a three-dimensional normal random vector, File 1 has data on (X,Z), while File 2 has data on (Y,Z). Two strategies proposed by Kadane (1978) and one strategy due to Sims (1978) are used to create synthetic files out of simulated data on (X,Z) and (Y,Z). These synthetic files are then evaluated by comparing the estimates of the correlation between X and Y provided by them with the estimates based on unbroken data on (X,Y,Z).

CHAPTER 2

MERGING FILES OF DATA ON SAME INDIVIDUALS

A useful classification of situations involving statistical matching of data-files was discussed in Section 1.3. It may be recalled that in the context of the two files having the same individuals, this classification scheme included two cases. Case I is the scenario where no matching-variables $\underset{\sim}{z}$ are present, while case II is the situation where matching-variables are part of the statistical model. In this chapter, we shall discuss results relevant to case I only.

2.1 A General Model

Let $[\overset{\underset{\sim}{T}}{\underset{\sim}{U}}]$ be a multi-dimensional random vector with C.D.F $H(\underset{\sim}{t},\underset{\sim}{u})$ and P.D.F $h(\underset{\sim}{t},\underset{\sim}{u})$. Let $[\overset{\underset{\sim}{T}_i}{\underset{\sim}{U}_i}]$, $i = 1,2, \ldots, n$ be a random sample of size n from H. We shall assume that these sample values got broken-up into the component vectors T's and U's before the data could be recorded. Thus we do not know which $\underset{\sim}{T}$ and $\underset{\sim}{U}$ values were paired in the original sample and the two files consist of the following data:

File 1 - $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$,

which is an unknown permutation of $\underset{\sim}{T}_1, \ldots, \underset{\sim}{T}_n$, and

File 2 - $\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, \ldots, \underset{\sim}{Y}_n$,

which is an unknown permutation of $U_1 \ldots, U_n$

DeGroot, Feder and Goel (1971) call this a "Broken Random Sample" model for two files.

Two types of statistical decision and inference problems arise from observing a broken random sample. The first type of problem involves trying to pair the $\underset{\sim}{x}$'s with the $\underset{\sim}{y}$'s in the broken data in order to reproduce the pairs in the original unbroken sample. The second type of problem involves making inferences about the values of parameters in the joint distribution $H(\underset{\sim}{t},\underset{\sim}{u})$ of $\underset{\sim}{T}$ and $\underset{\sim}{U}$.

This chapter will be organized into a review of the literature on matching problems in Sections 2.3 to 2.5, followed by a discussion of statistical properties of some matching strategies in Sections 2.6 to 2.9.

2.2 Notations

In this section, we introduce most of the notations that will be used in the present chapter.

(1) $\binom{\underset{\sim}{T}}{\underset{\sim}{U}}$ will denote a multivariate random vector. It is assumed to have an <u>absolutely</u> <u>continuous</u> joint cumulative distribution function (CDF) $H(\underset{\sim}{t},\underset{\sim}{u})$ and joint density $h(\underset{\sim}{t},u)$; the context will make the dimensions of $\underset{\sim}{t}$ and $\underset{\sim}{u}$ clear. In particular, $\binom{T}{U}$ will denote a two-dimensional random vector, with $h(t,u)$ and $H(t,u)$ respectively as the density and CDF of $\binom{T}{U}$. $h_1(\cdot)$ and $h_2(\cdot)$ will respectively denote the marginal densities of T and U and $F(\cdot)$, $G(\cdot)$ will be the respective marginal distribution functions. The symbol $g_{\underset{\sim}{\xi}}(\cdot)$ will be the generic notation for the density

function of the random vector $\underset{\sim}{\xi}$. Without the suffix, $g(\cdot)$ will denote a real-valued function.

(2) Let $\binom{T_i}{U_i}$, $i = 1,2, \ldots, n$ be a random sample from the population of $\binom{T}{U}$. Let $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(T_i \leq x)}$ be the empirical C.D.F based on the variables $T_1, \ldots, T_n$. Similarly, $G_n(x)$ will be the empirical C.D.F based on $U_1, \ldots, U_2$.

Let $R_{1i} = \sum_{\alpha=1}^{n} I_{(T_i \geq T_\alpha)}$ be the rank of T_i among the variables $T_1, \ldots, T_n$, where $i = 1,2, \ldots, n$. Similarly, $R_{21}, \ldots, R_{2n}$ will denote the rank order of the variables $U_1, \ldots, U_n$.

(3) Let $\varphi = (\varphi(1), \ldots, \varphi(n))$ be a permutation of the integers $1,2, \ldots, n$. Φ will stand for the set all such permutations. Also, let $\varphi^* = (1,2, \ldots, n)$.

(4) Let $\varepsilon \geq 0$. $\forall\ i = 1,2, \ldots, n$, define events $A_{ni}\ (\varphi, \varepsilon)$ as follows:

$$A_{ni}(\varphi,\varepsilon) = [|U_{(\varphi(R_{1i}))} - U_i| \leq \varepsilon] \tag{2.2.1}$$

$$\text{Let } A_{ni}(\varepsilon) = A_{ni}(\varphi^*,\varepsilon),\ i = 1,2, \ldots, n, \tag{2.2.2}$$

$$A_{ni} = A_{ni}(\varphi^*,0) \equiv (R_{1i} = R_{2i}),\ i = 1,2, \ldots, n. \tag{2.2.3}$$

$$\text{Let } V_{ni}(\varphi,\varepsilon) = I_{A_{ni}(\varphi,\varepsilon)},\ i = 1,2 \ldots, n. \tag{2.2.4}$$

$$V_{ni}(\varepsilon) = I_{A_{ni}(\varphi^*,\varepsilon)},\ i = 1,2, \ldots, n \tag{2.2.5}$$

$$V_{ni} = I_{A_{ni}},\ i = 1,\ 2, \ldots, n \tag{2.2.6}$$

5) Let $c(x,y)$ be the generic notation for a joint density of two random variables T and U which are marginally uniform. Then,

define the constant λ as $\int_0^1 c(x,x)dx$, which is the density of the random variable T-U evaluated at zero. For any fixed integer d, define

$$\underset{\sim}{S}_n = (S_{n1}, \dots, S_{nd}), \text{ where} \tag{2.2.7}$$

$$S_{nj} = R_{1j} - R_{2j},\ j = 1,2, \dots, n.$$

Note that if

$$\xi_{jk} = I_{(T_j - T_k \geq 0)} - I_{(U_j - U_k \geq 0)},\ \forall\ 1 \leq j \leq d \text{ and } 1 \leq k \leq n \tag{2.2.8}$$

then we get the representation

$$S_{nj} = \sum_{k=1}^{n} \xi_{jk},\ j = 1,2, \dots, d. \tag{2.2.9}$$

Let $\underset{\sim}{\xi}_k = (\xi_{1k}, \dots, \xi_{dk})$ (2.2.10)

Then,

$$\underset{\sim}{S}_n = \sum_{k=1}^{n} \underset{\sim}{\xi}_k \tag{2.2.11}$$

Let $\xi_{1jk} = I_{(T_j - T_k \geq \varepsilon)} - I_{(U_j - U_k \geq 0)},\ 1 \leq j,\ k \leq n$

$$\xi_{2jk} = I_{(U_j - U_k \geq 0)} - I_{(T_j - T_k \geq -\varepsilon)}\ 1 \leq j,\ k \leq n \tag{2.2.12}$$

Let L = T-U and $L_j = T_j - U_j$, where j = 1,2, ... Let Λ_d be the sigma-field $\sigma(\underset{\sim}{W}_1, \dots, \underset{\sim}{W}_d)$ generated by the vectors $\underset{\sim}{W}_i = (\overset{T_i}{U_i})$, 1,2, ..., d. Let $\underset{\sim}{\Psi}_{\underset{\sim}{\eta}}(\underset{\sim}{\theta})$ be the generic notation for the characteristic function of a random vector $\underset{\sim}{\eta}$, $\underset{\sim}{\theta}$ being a vector of dummy variables whose dimension is the same as that of $\underset{\sim}{\eta}$.

Let $\xi_{jk}(\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_d)$ be the variable ξ_{jk}, when $\underset{\sim}{W}_i$ takes the value $\underset{\sim}{w}_i$, $i = 1,2, \ldots, d$.

Let $\underset{\sim}{\xi}_k\ (\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_d)$ and $\underset{\sim}{S}_n = \sum_{k=1}^{n} \underset{\sim}{\xi}_k(\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_d)$ be respectively $\underset{\sim}{\xi}_k$ and $\underset{\sim}{S}_n$ when $\underset{\sim}{W}_i = \underset{\sim}{w}_i$, $i = 1,2, \ldots , d$.

Let $\Psi_d = \Psi_d(\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_d)$ be the negative logarithm of the modulus of the characteristic function of $\underset{\sim}{\xi}_{d+1\cdot}\ (\underset{\sim}{W}_1, \ldots, \underset{\sim}{W}_d)$

2.3 Data-based Matching Strategies

Pairing the observations in the two data-files that were described in Section 2.1 should be distinguished from the problem of matching two equivalent decks of n distinct cards, which is discussed in elementary textbooks such as Feller (1968). One version of card-matching is as follows. Consider a "target pack" of n cards laid out in a row and a "matching pack" of the same number of cards laid out randomly one by one beside the target pack. In this random arrangement of cards, n pairs of cards are formed. A match or coincidence is said to have occurred in a pair if the two cards in the pair are identical. Because the two decks are merged purely by chance and without using any type of observations or other information about the cards, one may describe such problems as no-data matching problems. An excellent survey of various versions of card-matching schemes is found in Barton (1958).

Suppose that N denotes the number of pairs in the aforementioned matching problem which have like cards or matches. The derivation of the probability distribution of N dates back to Montmort (1708). The following is a summary of some of the well-known properties of N (Feller 1968):

Proposition 2.3.1: If $P_{[m]}$ is the probability of having exactly m matches, then

(i) $$P_{[m]} = \frac{1}{m!}[1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \ldots \pm \frac{1}{(n-m)!}] , m = 0,2, \ldots, n-1$$

and
$$P_{[n]} = \frac{1}{n!}$$

(ii) Noting that $\frac{e^{-1}}{m!}$ is the probability that a Poisson random variable with mean 1 takes the value m, we have the following approximation for large n:

$$P_{[m]} \approx \frac{e^{-1}}{m!}$$

(iii) For d = 1,2, ..., n, the dth factorial moment of N, namely $E(N^{(d)})$, is 1.

As one might expect, for certain broken random sample models, it pays to match two files of data using optimal strategies based on such data. Several authors starting with DeGroot, Feder and Goel (1971) have proposed and studied matching strategies based on broken data. In Section 2.9, it will be shown that, for certain matching strategies based on independent variables T and U the distributional properties of the number of correct matches are the same as those mentioned in Proposition 2.3.1. In other words, as far as statistical properties of N are concerned, matching files of data on independent random variables is only as good as no-data matching in which we randomly assign units in one file to the units in the other file.

2.4 Repairing a Broken Random Sample

2.4.1 The Basic Matching Problems

Let us consider matching the broken random sample x_1, x_2, ..., x_n, y_1, ..., y_n by pairing x_i with $y_{\varphi(i)}$, for i = 1,2, ..., n where $\varphi = (\varphi(1), \ldots, \varphi(n))$ is a permutation of 1,2, ..., n. As we seek a φ from Φ that will provide reasonably good pairings of the x's with the y's, we need to clarify the fundamental role of φ in the statistical model described in Section 2.1. If we treat φ as an unknown parameter of the model, then the likelihood of the data will include φ. For instance, if T and U are jointly bivariate normal with means μ_1, μ_2, variances σ_1^2, σ_2^2 and correlation coefficient ρ, then the log-likelihood function of φ, ρ μ_1, μ_2, σ_1^2, σ_2^2, given the broken random sample, is

$$L(\varphi,\rho,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2|x_1, \ldots, x_n, y_1, \ldots, y_n)$$

$$= -\frac{n}{2} \log(1 - \rho^2) - \frac{n}{2} \log \sigma_1^2 - \frac{n}{2} \log \sigma_2^2$$

$$- \frac{1}{2(1-\rho^2)} [\sum_{i=1}^{n} (x_i - \mu_1)^2/\sigma_1^2 + \sum_{i=1}^{n} (y_i - \mu_2)^2/\sigma_2^2$$

$$- 2 \rho \sum_{i=1}^{n} (x_i - \mu_1)(y_{\varphi(i)} - \mu_2)/\sigma_1\sigma_2] \qquad (2.4.1)$$

A constant term not involving the parameters has been omitted in (2.4.1). In subsection 2.4.2, we shall seek φ's that maximize the likelihood such as this. On the other hand, some statisticians would regard φ as some sort of missing data and not as a parameter of the underlying model. The problem of pairing the two files will not arise in such situations. However, one may still want to do statistical inference for other parameters of the model based on the broken random sample. Such issues are not pursued in this thesis and one may refer to DeGroot and Goel (1980) for an approach to estimating the correlation coefficient ρ while treating φ as missing data in the bivariate normal model.

2.4.2 The Maximum Likelihood Solution to the Matching Problem

We start with a bivariate model used in DeGroot et al. (1971) which assumes that the parent probability density function of $\binom{T}{U}$ is

$$h(t,u) = \alpha(t)\ \beta(u)\ \exp[\gamma(t)\ \delta(u)] \tag{2.4.2}$$

where α, β, γ, δ are known but otherwise arbitrary real-valued functions of the indicated variables. Suppose now that $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are the observations in a broken random sample from a completely specified density of the form (2.4.2). If x_i was paired with $y_{\varphi(i)}$ for $i = 1,2, \ldots, n$, in the original unbroken sample, then the joint density of the broken sample would be

$$\prod_{i=1}^{n} h[x_i, y_{\varphi(i)}] = [\prod_{i=1}^{n} \alpha(x_i)][\prod_{i=1}^{n} \beta(y_i)]\exp[\sum_{i=1}^{n} \gamma(x_i)\ \delta(y_{\varphi(i)})] \tag{2.4.3}$$

Thus the maximum likelihood estimate of the unknown permutation φ is the permutation for which $\sum_{i=1}^{n} \gamma(x_i)\ \delta(y_{\varphi(i)})$ is maximum. Without loss of generality, we shall assume that the x_i's and y_j's have been reindexed so that $\gamma(x_1) \leq \cdots \leq \gamma(x_n)$ and $\delta(y_1) \leq \cdots \leq \delta(y_n)$.

Since $\binom{T}{U}$ is assumed to have an absolutely continuous distribution, with probability one, there are no ties among $\gamma(x_i)$'s or $\alpha(y_j)$'s. DeGroot et al. (1971) shows that the maximum likelihood solution is to pair x_i with y_i, for $i = 1, \ldots, n$. In other words, the maximum likelihood pairing (M.L.P) is $\varphi^* = (1, \ldots, n)$.

In particular, if the density in 2.4.2 is that of a bivariate normal random vector with correlation ρ, then M.L.P,can be described knowing only the sign of ρ. If $\rho > 0$, the M.L.P. is to order the observed values so that $x_1 < \ldots < x_n$ and $y_1 < \ldots < y_n$ and then to pair x_i with y_i, for $i = 1,2, \ldots, n$. If $\rho < 0$, the solution is to pair x_i and $y_{(n+1-i)}$, for $i = 1,2, \ldots, n$. If $\rho = 0$, all pairings, or permutations, are equally likely.

Chew (1973) derived the maximum likelihood solution to the (bivariate) matching problem for a larger class of densities $h(t,u)$ with a monotone likelihood ratio. That is, for any values t_1, t_2, u_1 and u_2 such that $t_1 < t_2$ and $u_1 < u_2$,

$$h(t_1,u_1)\ h(t_2,u_2) \geq h(t_1,u_2)\ h(t_2,\ u_1) \qquad (2.4.4)$$

As before, we shall assume that the values $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ in a broken random sample are from a density $h(t,u)$ satisfying (2.4.4). Without loss, relabel the x's and y's so that $x_1 < \ldots < x_n$ and $y_1 < \ldots < y_n$. Then permutation $\varphi^* = (1, \ldots, n)$ is again the M.L.P.

2.4.3 Some Bayesian Matching Strategies

DeGroot et al. (1971) studied the matching problem from a Bayesian point of view as well. They proposed three optimality criteria, subject to which one may choose the matching strategy φ. Before we state these criteria, we need some notation and definitions.

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be the values of a broken random sample from a given parent distribution with density h(t,u). If x_i is paired with $y_{\varphi(i)}$, $i = 1,2, \ldots, n$, then the likelihood function of the unknown permutation φ is given by the equation

$$L(\varphi) = \prod_{i=1}^{n} h(t_i, u_{\varphi(i)}), \tag{2.4.5}$$

Assume that the prior probability of each permutation is $\frac{1}{n!}$. Then the posterior probability that φ provides a completely correct set of n matches is

$$p(\varphi) = L(\varphi) / \sum_{\psi\in\Phi} L(\psi) \tag{2.4.6}$$

For $j = 1,2, \ldots, n$, let

$$\Phi(j) = \{\varphi\in\Phi: \varphi(1) = j\} \tag{2.4.7}$$

be the set of $(n - 1)!$ permutations which specify that x_1 is to be paired with y_j. Using the definitions in (2.4.6) and (2.4.7), we get the posterior probability that the pairing of x_1 and y_j yields a correct match to be

$$p_j = \sum_{\varphi\in\Phi(j)} p(\varphi), \; 1 \le j \le n \tag{2.4.8}$$

For any two permutations φ and ψ in Φ, let

$$K(\varphi,\psi) = \# \{i: \varphi(i) = \psi(i)\}$$

That is, $K(\varphi,\psi)$ is the number of correct matches when the observations in the broken random sample are paired according to φ and the vectors in the original sample were actually paired according to ψ. It then follows that for any permutation $\varphi \in \Phi$, the quantity

$$M(\varphi) = \sum_{\psi \in \Phi} K(\varphi,\psi)\ p(\psi) \tag{2.4.9}$$

is the posterior expected number of correct matches when φ is used to repair the data in the broken random sample.

Finally, let $\Phi_{1,n}$ be the set of all permutations φ such that $y_{\varphi(1)} = y_1$ and $y_{\varphi(n)} = y_n$.

DeGroot, Feder and Goel (1971) have proposed three optimality criteria, subject to which one may choose the matching strategy φ:

(i) maximize the probability, $p(\varphi)$, of a completely correct set of n matches,

(ii) maximize the probability, p_j, of correctly matching x_1 by choosing an optimal j from $\{1,2, \ldots, n\}$ and

(iii) maximize the expected number, $M(\varphi)$, of correct matches in the repaired sample.

Assuming that the bivariate density of T and U was given by $h(t,u) = a(t)b(u)\ e^{tu}$, $(t,u) \in R^2$, the following results, among others, were established by DeGroot et al. (1971):

(a) The M.L.P φ^* maximizes the probability of correct pairing of all n observations.

(b) The probability of pairing $x_1(x_n)$ correctly is maximized by pairing $x_1(x_n)$ with $y_1(y_n)$.

(c) The class of permutations $\Phi_{1,n}$ is complete; that is, given any permutation $\varphi \notin \Phi_{1,n}$, there exists a $\psi \in \Phi_{1,n}$ which is as good as φ in the sense that $M(\psi) \geq M(\varphi)$.

(d) Sufficient conditions in terms of the data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ for the M.L.P φ^* to maximize $M(\varphi)$ were also given.

The results in Chew (1973) and Goel (1975) are extensions of (a) through to (d) to an arbitrary bivariate density $h(t,u)$ possessing the monotone likelihood ratio. The "completeness" property in (c) implies that the permutation φ^E maximizing $M(\varphi)$ satisfies $\varphi^E(1) = 1$ and $\varphi^E(n) = n$, for $n = 2, 3$, $\varphi^* \equiv \varphi^E$. DeGroot et al. (1971) show that for $n > 3$, φ^E is not necessarily equal to the M.L.P φ^* by means of a counter-example.

2.4.4 Matching Problems for Multivariate Normal Distributions

In our review so far, we have discussed optimal matching strategies only in the case of bivariate data, one variable for each of the two files. However, multivariate data are often available in both files. Suppose then that we have a model where $\begin{pmatrix}\underset{\sim}{T}\\ \underset{\sim}{U}\end{pmatrix}$ has a (p+q)-dimensional normal distribution with known variance-covariance matrix Σ. Let us write Σ and its inverse in the following partitioned form:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \text{ and } \Sigma^{-1} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix},$$

where both Σ_{12} and Ω_{12} have dimension p x q.

As before, we shall let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ and $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n$ denote the values in a broken random sample from this distribution, where each $\underset{\sim}{x}_i$ is a vector of dimension p x 1 and each $\underset{\sim}{y}_j$ vector has the dimension q x 1. The results to be presented here were originally described by DeGroot and Goel (1976).

The likelihood function L, as a function of the unknown permutation φ, can be written in the form

$$L(\varphi) = \exp[-\Sigma\ \underset{\sim}{x}_i'\ \Omega_{12}\ \underset{\sim}{y}_{\varphi(i)}], \tag{2.4.10}$$

since the other factors in the joint density of the sample do not depend on φ. If we again assume that the prior probability of each permutation φ is $\frac{1}{n!}$, then the posterior probability that φ provides a completely correct set of n matches is given by (2.4.6). Thus, maximizing $p(\varphi)$ is equivalent to maximizing $L(\varphi)$, or equivalently minimizing

$$Q(\varphi) = \sum_{i=1}^{n} \underset{\sim}{x}_i\ \Omega_{12}\ \underset{\sim}{y}_{\varphi(i)} \tag{2.4.11}$$

There is no simple way, in general, to describe the maximum likelihood solution.

However, if rank $(\Sigma_{12}) = 1$, then rank $(\Omega_{12}) = 1$ and Ω_{12} can be represented in the form $\Omega_{12} = \underset{\sim}{a}'\underset{\sim}{b}$, where $\underset{\sim}{a}$ and $\underset{\sim}{b}$ are vectors of dimensions p x 1 and q x 1. If we let $\gamma(\underset{\sim}{x}_i) = \underset{\sim}{a}'\underset{\sim}{x}_i$ and $\delta(\underset{\sim}{y}_i) = \underset{\sim}{b}'\underset{\sim}{y}_i$ for i = 1,2, ..., n, the φ^* will be the permutation that minimizes

$$Q(\varphi) = \sum_{i=1}^{n} \gamma(\underset{\sim}{x}_i)\ \delta(\underset{\sim}{y}_{\varphi(i)}) \tag{2.4.12}$$

Now, minimizing (2.4.12) is achieved by arranging $\gamma(x_i)$'s from smallest to largest, arranging $\delta(y_j)$'s in the reverse order from the largest to smallest and then pairing the corresponding elements in the two sequences.

Suppose next that rank $(\Omega_{12}) \geq 2$. Without loss of generality, we shall assume that $p \leq q$ and let $v_j = \Omega_{12} y_j$, for $j = 1, 2, \ldots, n$. Then, both $\underset{\sim}{x}_i$ and $\underset{\sim}{v}_j$ are p-dimensional vectors, and the maximum likelihood solution φ^* will be the permutation that minimizes

$$Q(\varphi) = \sum_{i=1}^{n} \underset{\sim}{x}_i'\ \underset{\sim}{v}_{\varphi(i)}$$

Let D denote the n x n matrix $((d_{ij}))$ whose elements are $d_{ij} = \underset{\sim}{x}_i' \underset{\sim}{v}_j$. Then minimizing (2.4.14) is equivalent to minimizing

$$Q(\varphi) = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}\ a_{ij}$$

subject to the constraints

$$\sum_{i=1}^{n} a_{ij} = 1, \text{ for } j = 1, 2, \ldots, n,$$

$$\sum_{j=1}^{n} a_{ij} = 1, \text{ for } i = 1, 2, \ldots, n,$$

$$a_{ij} = 0 \text{ or } 1,$$

which is a standard assignment problem with cost matrix D. Although, there is no simple form for the solution of an arbitrary assignment problem of this type, efficient algorithms are available for finding numerical solutions.

The permutation φ^E that maximizes the expected number of correct matches is very difficult to calculate when p and n are moderately large. No efficient algorithms are known. A Monte Carlo study was reported by DeGroot and Goel (1976) in which they compare φ^E and φ^* for p = 2 and 50 different covariance matrices Σ with the sample size n = 3, 4 and 5. In all cases, the proportion of samples for which φ^E and φ^* were identical was between 0.925 and 0.995. Thus, it is not unreasonable to use φ^* even when the goal is to maximize the expected number of correct matches.

DeGroot and Goel (1976) studied two other simple matching strategies which provide good approximations to the M.L.P φ^* or to the rule φ^E. We shall not discuss them here. In the rest of this chapter, we shall discuss matching problems only in the bivariate case.

2.5 <u>Reliability of Matching Strategies for Bivariate Data</u>

Consider a random sample of size n, $\binom{T_1}{U_1}, \ldots, \binom{T_n}{U_n}$, from a bivariate population with density h(t,u). If the pairings in this sample are lost before the entire data was recorded, we still can observe the marginal order-statistics. In fact, if $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ is the broken random sample corresponding to the unobserved sample on $\binom{T}{U}$, then clearly the order-statistics

$x_{(1)} < \cdots < x_{(n)}$ of the x's are exactly the same as the order-statistics $T_{(1)} < \cdots < T_{(n)}$ of the T's. Similarly, the order-statistics $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ are the same as $U_{(1)} < \cdots < U_{(n)}$. The repairing of the x's and y's was introduced in Section 2.4. Thus for each permutation φ in Φ, there is a matching strategy and the typical merged file consists of the pairs

$$\binom{x_{(i)}}{y_{(\varphi(i))}} , \; i = 1,2, \ldots, n. \qquad (2.5.1)$$

Some optimal matching strategies were discussed in Section 2.4. Here, we are concerned with the quality of the file in (2.5.1).

Ideally, we would like to choose a φ for which the file in (2.5.1) recovers all the $\binom{T}{U}$ pairs that we did not observe. It is therefore natural to look at the random variable N(φ), the number of correct matches due to φ or, equivalently, the number of unobserved sample points which have been recovered in (2.5.1). It should be pointed out that M(φ), which was defined in Section 2.4.3, is different from E[N(φ)] because the former quantity is a posterior expected value given a particular broken random sample and, in the latter, the expectation is taken over all possible samples.

Situations often arise where it is not crucial that, after the two files are matched, the matched pairs are exactly the same as the pairs of the original data. For example, when contingency tables are contemplated for grouped data on continuous variables T and U, we may, in the absence of the knowledge of the pairings, would like to reconstruct the pairs but would not worry too much as long as the

U-value in any matched pair came within a pre-fixed tolerance ε (a non-negative number) of the true U-value that we would get in the ideal match of recovering all the original pairs. This type of "approximate matching" was first introduced by Yahav (1982) who defined ε-correct matching as follows:

Definition 2.5.1 (Yahav): A pair in the merged file (2.5.1), $\begin{pmatrix} X_{(i)} \\ Y_{\varphi(i)} \end{pmatrix}$, say, is ε-correct if $|U_{(\varphi(i))} - U_{[i]}| \leq \varepsilon$, where $\varepsilon > 0$ and $U_{[i]}$ is the concomitant of $X_{(i)}$; that is, the true U-value that was paired with $X_{(i)}$ in the original sample.

The number of ε-correct matches, $N(\varphi,\varepsilon)$, in the merged file (2.5.1) is given by

$$N(\varphi,\varepsilon) = \sum_{i=1}^{n} I_{[|U_{(\varphi(i))} - U_{[i]}| \leq \varepsilon]} \tag{2.5.2}$$

Note that as $\varepsilon \downarrow 0$, $N(\varphi;\varepsilon)$ converges (almost surely) to $N(\varphi;0)$, which is a count of the exact (0-correct) matches. Hence $N(\varphi)$, the number of correct matches due to φ can be obtained from $N(\varphi;\varepsilon)$ by formally letting $\varepsilon = 0$.

In the light of the definition of reliability of a merged file, given in Section 1.7, the counts $N(\varphi)$ and $N(\varphi,\varepsilon)$ are useful indices whose statistical properties reflect the reliability of the merged file resulting from φ. We shall study these performance characteristics in the following sections.

2.6 An Optimality Property of the Maximum Likelihood Pairing φ^*

The known results about the optimality of the maximum likelihood pairing $\varphi^* = (1, \ldots, n)$ with respect to some Bayesian criteria were reviewed in Section 2.4. Here, we shall propose a new criterion and establish that φ^* is optimal with respect to that criterion.

Consider the random variable $N(\varphi)$, the number of correct matches which result when a permutation φ in Φ is used to merge the broken random sample from a bivariate population. In this section, we shall show that φ^* maximizes $E(N(\varphi))$, the expected number of correct matches, provided that the parent density $h(t,u)$ exhibits certain dependence structures.

We begin with quoting a very useful result on the exchangeability of random variables from Randles and Wolfe (1979).

Lemma 2.6.1: If $\underset{\sim}{\xi} \overset{d}{=} \underset{\sim}{\eta}$ and $\underset{\sim}{K}(\cdot)$ is a measurable function (possibly vector valued) defined on the common support of these random vectors, then

$$K(\underset{\sim}{\xi}) \overset{d}{=} K(\underset{\sim}{\eta})$$

We now establish a representation for $N(\varphi,\varepsilon)$ as a sum of exchangeable Bernoulli random variables, which will be useful for extending results of Yahav (1982).

Theorem 2.6.1: Let $N(\varphi,\varepsilon)$ and $V_{ni}(\varphi,\varepsilon)$ be as defined by (2.5.2) and (2.2.4) respectively. Then

$$\forall\ \varphi \text{ in } \Phi,\ N(\varphi,\varepsilon) = \sum_{i=1}^{n} V_{ni}\ (\varphi,\varepsilon), \tag{2.6.1}$$

where the summands are exchangeable random variables.

Proof: The order-statistic $U_{(\varphi(i))}$ and the concomitant $U_{[i]}$ of $T_{(i)}$, used in (2.5.2) can be written in terms of ranks of T's and U's as follows:

$$U_{(\varphi(i))} = \sum_{\alpha=1}^{n} U_\alpha\ I_{(R_{2\alpha} = \varphi(i))} \tag{2.6.2}$$

$$U_{[i]} = \sum_{\alpha=1}^{n} U_\alpha\ I_{(R_{1\alpha}=i)} \tag{2.6.3}$$

Note that $N(\varphi,\varepsilon)$ is simply a count of how many pairs in the merged-file due to φ, namely,

$$\binom{T_{(i)}}{U_{(\varphi(i))}}\ ,\ i = 1,2,\ \ldots,\ n \tag{2.6.4}$$

satisfy

$$|U_{(\varphi(i))} - U_{[i]}| \le \varepsilon \tag{2.6.5}$$

If (2.6.5) holds for some i, then $\exists$ a j such that

$$|U_{(\varphi(i))} - U_j| \le \varepsilon$$

In view of the continuity of (T_i,U_i), this correspondence is one-to-one. Therefore, the count $N(\varphi,\varepsilon)$ must be the same as the count given by

$$N(\varphi,\varepsilon) = \sum_{i=1}^{n} I_{(|U_{(\varphi(R_{1i}))} - U_i| \le \varepsilon)} \tag{2.6.6}$$

Hence, (2.6.1) holds by virtue of the definition (2.2.4) of V_{ni}.

Towards showing the exchangeability of the V_{ni}'s, note that the original sample in (2.6.5) are independent and identically distributed vectors. Hence, using the equal-in-distribution notation, we get

$$(\underset{\sim}{W}_{\alpha 1}, \ldots, \underset{\sim}{W}_{\alpha n}) \overset{d}{=} (\underset{\sim}{W}_1, \ldots, \underset{\sim}{W}_n) \tag{2.6.7}$$

where $(\alpha_1, \ldots, \alpha_n)$ is an arbitrary permutation of $(1,2, \ldots, n)$. Define a function $\underset{\sim}{f} = (f_1, \ldots, f_n)$ from R^{2n} to R^n by the equations

$$f_j = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} I_{(b_j-b_i \ge \varepsilon)} \le \varphi(\sum_{i=1}^{n} I_{(a_j-a_i \ge 0)}) \le \sum_{i=1}^{n} I_{(b_j-b_i \ge -\varepsilon)} \\ 0 \text{ if otherwise} \end{cases}$$

$$j = 1,2, \ldots, n, \tag{2.6.8}$$

where φ is the matching strategy we started with and $(a_1,b_1, \ldots, a_n,b_n)$ is an arbitrary point in R^{2n}.

It follows from (2.6.7) and Lemma 2.6.1 that

$$\underset{\sim}{f}(\underset{\sim}{W}_{\alpha 1}, \ldots, \underset{\sim}{W}_{\alpha n}) \overset{d}{=} \underset{\sim}{f}(\underset{\sim}{W}_1, \ldots, \underset{\sim}{W}_n) \tag{2.6.9}$$

Fix j as an integer in $\{1,2, \ldots, n\}$. Then, using (2.6.8) we see that $f_j(\underset{\sim}{W}_{\alpha 1}, \ldots, \underset{\sim}{W}_{\alpha n})$ is the indicator function of the event

$$\sum_{i=1}^{n} I_{(U_{\alpha j}-U_i \ge \varepsilon)} \le \varphi(\sum_{i=1}^{n} I_{(T_{\alpha j}-T_i \ge 0)}) \le \sum_{i=1}^{n} I_{(U_{\alpha j}-U_i \ge -\varepsilon)} ,$$

or, equivalently, in terms of the ranks $R_{11}, \ldots, R_{1n}$ of the T's and the empirical C.D.F $G_n(\cdot)$ of the U's,

$$G_n(U_{\alpha_j} - \varepsilon) \leq \varphi(R_{1\alpha_j})/n \leq G_n(U_{\alpha_j} + \varepsilon)$$

Observing that $G_n^{-1}(k/n) = U_{(k)}$, $k = 1,2, \ldots, n$, we find $f_j(\underset{\sim}{W}_{\alpha 1}, \ldots, \underset{\sim}{W}_{\alpha_n})$ is 1 iff $|U_{(\varphi(R_{1\alpha_j}))} - U_{\alpha_j}| \leq \varepsilon$. By the same token, $f_j(\underset{\sim}{W}_1, \ldots, \underset{\sim}{W}_n)$ is the indicator of the event $|U_{(\varphi(R_{1j}))} - U_j| \leq \varepsilon$. So that $f_j(\underset{\sim}{W}_1, \ldots, \underset{\sim}{W}_n) = V_{nj}(\varphi,\varepsilon)$. From these facts and (2.6.9) it follows that

$$\begin{aligned}&(V_{n\alpha_1}(\varphi,\varepsilon), \ldots, V_{n\alpha_n}(\varphi,\varepsilon))\\ &\overset{d}{=} (V_{n1}(\varphi,\varepsilon), \ldots, V_{nn}(\varphi,\varepsilon)) \qquad (2.6.10)\end{aligned}$$

Because $(\alpha_1, \ldots, \alpha_n)$ is an arbitrary permutation of $1,2, \ldots, n$, we conclude from (2.6.10) that the summands in (2.6.6) are exchangeable random variables.

<u>Corollary 2.6.1</u>: The number of correct matches resulting from the matching strategy φ has the representation

$$N(\varphi) = \sum_{i=1}^{n} I_{(R_{2i}=\varphi(R_{1i}))} \qquad (2.6.11)$$

<u>Proof</u>: Set $\varepsilon = 0$ in Theorem 2.6.1. □

We will need the following special dependence structures for the population density h(t,u). (see Shaked 1979).

<u>Definition (2.6.1)</u>: Exchangeable random variables T,U are said to

be positive dependent by mixture (PDM) iff the joint distribution of T,U is that of $g(\underset{\sim}{\xi}_0,\xi_1)$ and $g(\underset{\sim}{\xi}_0,\xi_2)$, where ξ_1 and ξ_2 are i.i.d random variables, $\underset{\sim}{\xi}_0$ is a random vector which is independent of ξ_1 and ξ_2 and g is a Borel measurable function.

Definition (2.6.2): Exchangeable random variables T,U are said to be positive dependent by expansion (PDE) iff the joint distribution of T and U admits the following series expansion:

$$dH(t,u) = [1 + \sum a_i \eta_i(t)\eta_i(u)]\ dF(t)dF(u) \qquad (2.6.12)$$

where $F(\cdot)$ is the marginal CDF of T or U, a_i's are nonnegative real numbers, and $\{\eta_i\}$ is a set of functions satisfying

$$\int_{-\infty}^{\infty} \eta_i(x)\ dF(x) = 0,\ i = 1,2,\ \ldots, \qquad (2.6.13)$$

According to the Definitions 2.6.1 and 2.6.2, the dependence concepts will apply only to pairs of exchangeable random variables. It may also be noted that for most of the known expansions of PDE distributions, the set of functions $\{\eta_k(\cdot)\}$ satisfies, in addition to (2.6.13), the orthogonality conditions

$$\int_{-\infty}^{\infty} \eta_k(x)\eta_\ell(x)\ dF(x) = \delta_{k\ell}\ , \qquad (2.6.14)$$

where k, ℓ = 1,2, ..., and $\delta_{k\ell}$ is the kronecker delta.

We now give two examples to illustrate these concepts of dependence.

Example 2.6.1: Let ξ_0, ξ_1, ξ_2 be i.i.d standard normal random variables. Let ρ be any constant in the interval [0,1]. Define new random variables

$$T = \sqrt{1-\rho} \cdot \xi_1 + \sqrt{\rho}\ \xi_0$$

$$U = \sqrt{1-\rho} \cdot \xi_2 + \sqrt{\rho}\ \xi_0$$

Then, it is easy to verify that T,U are jointly normal and that the definition (2.6.1) can be applied to T and U with the above choice of ξ_0, ξ_1 and ξ_2. Hence, the standard bivariate normal distribution with nonnegative correlation has the PDM property.

Also, Mardia (1970, p. 48) gives the following series expansion for the bivariate-normal density

$$h(t,u) = [1 + \sum_{k=1}^{\infty} \rho^k \eta_k(t)\eta_k(u)]\ f(t)\ f(u), \tag{2.6.15}$$

where f(t) is the density of the univariate standard normal random variable and $\{\eta_k(\cdot)\}$ is a set of orthonormal Hermite polynonomials. Thus, if $\rho \geq 0$, bivariate normal distributions possess the PDE property as well.

Example 2.6.2: A class of bivariate densities due to Farlie-Gumbel-Morgenstern is given by the formula

$$h(t,u) = 1 + \alpha(1 - 2t)(1 - 2u), \text{ where } 0 < t,\ u < 1$$

$$-1 \leq \alpha \leq 1 \tag{2.6.16}$$

It is easy to check that T and U are PDE for $\alpha \geq 0$ in (2.6.16). Note that the expansion 2.6.16 has only a finite number of terms, unlike the expansion for the bivariate normal distribution.

We now prove that the PDM/PDE structures are inherited by a pair of new variables obtained from a given sample by computing the <u>same</u> function of the marginals. These results are generalizations of theorems in Shaked (1979), which were proved only for n=2. However, mathematical induction does not help to show the results for an arbitrary n.

<u>Theorem 2.6.2</u>: Let $\binom{T_i}{U_i}$, $i = 1,2, \ldots, n$ be a random sample from a PDM parent with density $h(t,u)$. Then, for any measurable function $g:R^n \to R$, the random variables $g(T_1,T_2, \ldots, T_n)$ and $g(U_1,U_2, \ldots, U_n)$ are jointly PDM.

<u>Proof</u>: By hypothesis, the vectors $\binom{T_i}{U_i}$ are i.i.d, Furthermore, since PDM property is defined only for exchangeable pairs of random variables, we have

$$(T_i,U_i) \overset{d}{=} (U_i,T_i), \; i = 1,2, \ldots, n \tag{2.6.17}$$

Equation (2.6.17) together with the independence of T,U pairs yields

$$(T_1, \ldots, T_n, U_1, \ldots, U_n) \overset{d}{=} (U_1, U_2, \ldots, U_n , T_1, \ldots, T_n) \tag{2.6.18}$$

Consider the function $\underset{\sim}{K}:R^{2n} \to R^n$ defined by the equation

$$\underset{\sim}{K}(a_1, \ldots, a_n; b_1, \ldots, b_n) = (g(a_1, \ldots, a_n), g(b_1, \ldots, b_n))$$

where $(a_1, \ldots, a_n, b_1, \ldots, b_n)$ is any point in R^{2n}. Applying the function $\underset{\sim}{K}$ to both sides of (2.6.18) and invoking Lemma 2.6.1 we get

$$(g(T_1, \ldots, T_n), g(U_1, \ldots, U_n)) \overset{d}{=} (g(U_1, \ldots, U_n), g(T_1, \ldots, T_n)) \tag{2.6.19}$$

Hence, $(g(\underset{\sim}{T}), g(\underset{\sim}{U}))$ is an exchangeable-pair of random variables.

The PDM property of (T_i, U_i), $i = 1,2, \ldots, n$ further implies that there exist n i.i.d. vectors $(\underset{\sim}{\xi}_{0i}, \xi_{1i}, \xi_{2i})$, $i = 1,2, \ldots, n$ and a measurable function f such that

(i) For each j, ξ_{1j}, ξ_{2j} are i.i.d univariate random variables and the vector $\underset{\sim}{\xi}_{0j}$ is independent of ξ_{1j} and ξ_{2j}.

(ii) For each j,

$$T_j = f(\xi_{1j}, \underset{\sim}{\xi}_{0j}) \text{ and } U = f(\xi_{2j}, \underset{\sim}{\xi}_{0j}) \tag{2.6.20}$$

Introducing the random variables,

$$\xi_1^* = \xi_{11}, \ \xi_2^* = \xi_{21}$$

and

$$\underset{\sim}{\xi}_0^* = (\xi_{12}, \ldots, \xi_{1n}, \xi_{22}, \ldots, \xi_{2n}, \underset{\sim}{\xi}_{01}, \ldots, \underset{\sim}{\xi}_{0n}) \tag{2.6.21}$$

We find that ξ_1^* and ξ_2^* are i.i.d univariate random variables and $\underset{\sim}{\xi}_0^*$ is independent of ξ_1^* and ξ_2^* in view of the assumptions (i) and (ii). Note that (2.6.20) and (2.6.21) imply that

$$g(\underset{\sim}{T}) = g(f(\xi_{11}, \underset{\sim}{\xi}_{01}), \ldots, f(\xi_{1n}, \underset{\sim}{\xi}_{0n}))$$

is a measurable function g^*, say, of ξ_1^* and $\underset{\sim}{\xi}_0^*$. Similarly, $g(\underset{\sim}{U})$ is also the <u>same</u> function g^* of the random variables ξ_2^* and $\underset{\sim}{\xi}_0^*$. Hence, by definition, $g(\underset{\sim}{T})$ and $g(\underset{\sim}{U})$ are PDM. □

The next theorem is similar to Theorem 2.6.2 except the parent distribution has the PDE property.

<u>Theorem 2.6.3</u>: Let $\binom{T_i}{U_i}$, i = 1, ..., n be a random sample from a PDE parent. Then, for any measurable function $g:R^n \to R$, the random variables $g(T_1, \ldots, T_n)$ and $g(U_1, \ldots, U_n)$ are PDE.

<u>Proof</u>: The exchangeability of the joint distribution of $g(\underset{\sim}{T})$ and $g(\underset{\sim}{U})$ has already been proved in Theorem 2.6.2 (see equation 2.6.19). It remains to be shown that, when the joint density of each of the n copies of T,U admits an expansion of the type 2.6.12, the joint density of $g(\underset{\sim}{T})$ and $g(\underset{\sim}{U})$ also admits a similar expansion.

Assume therefore that there exists nonnegative constants $\{a_k\}$ and a set of orthonormal functions $\{\eta_k(\cdot)\}$ such that the joint density of T_i and U_i is of the form.

$$dH(t_i,u_i) = dF(t_i)dF(u_i)[1 + \sum_{k=1}^{\infty} a_k\eta_k(t_i)\eta_k(u_i)] \ , \qquad (2.6.22)$$

where i = 1,2, ..., n.

For any real x, define the measurable set in R^n

$$A(x) = \{(x_1, \ldots, x_n): g(x_1, \ldots, x_n) \le x\} \ .$$

Then, the distribution function Q, say, of $(g(\underset{\sim}{T}),g(\underset{\sim}{U}))$ is

$$Q(x,y) = \int \ldots \int_{\underset{\sim}{t}\in A(x)} \int \ldots \int_{\underset{\sim}{u}\in A(y)} \prod_{j=1}^{n} dH(t_j,u_j) \qquad (2.6.23)$$

Using the expansions in equation (2.6.22) we get

$$
\begin{aligned}
Q(x,y) = {} & \tilde{Q}(x)\tilde{Q}(y) + \\
& n \sum_{k=1}^{\infty} a_k \chi_k^{(1)}(x)\chi_k^{(1)}(y) + \\
& \binom{n}{2} \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_k a_\ell \chi_{k,\ell}^{(2)}(x)\chi_{k,\ell}^{(2)}(y) \\
& + \dots + \sum_{k_1=1}^{\infty} \dots \sum_{k_n=1}^{\infty} a_{k_1} \dots a_{k_n} \chi_{k_1,\dots,k_n}^{(n)}(x)\chi_{k_1,\dots,n}^{(n)}(y)
\end{aligned}
\tag{2.6.24}
$$

where $\tilde{Q}(x) = \int_{A(x)} \dots \int \prod_{i=1}^{n} dF(t_i)$

$$\chi_k^{(1)}(x) = \int_{A(x)} \dots \int \eta_k(t_1) \prod_{i=1}^{n} dF(t_i)$$

$$\chi_{k,\ell}^{(2)}(x) = \int_{A(x)} \dots \int \eta_k(t_1)\eta_\ell(t_2) \prod_{i=1}^{n} dF(t_i)$$

and

$$\chi_{k_1,\dots,k_n}^{(n)}(x) = \int_{A(x)} \dots \int \prod_{i=1}^{n} \eta_{k_i}(t_i) \prod_{i=1}^{n} dF(t_i) \tag{2.6.25}$$

Note that $\forall\ k_i = 1,2,\ \dots$ and $\forall\ i = 1,2,\ \dots,\ n$ the signed measure induced by $\chi_{k_1,\dots,k_\ell}^{(\ell)}(x)$ is absolutely continuous with respect to $\tilde{Q}$

so that there exists $\psi^{(\ell)}_{k_1,\dots,k_\ell}(x)$ - the Radon-Nikodym derivative - of $\chi^{(\ell)}(x)$ with respect to $\tilde{Q}$ such that

$$\chi^{(\ell)}_{k_1,\dots,k_\ell}(x) = \int_{-\infty}^{x} \psi^{(\ell)}_{k_1,\dots,k_\ell}(t)\, d\tilde{Q}(t) \ . \qquad (2.6.26)$$

Hence, from equations (2.6.24) to (2.6.26) we get

$$\begin{aligned} dQ(x,y) = d\tilde{Q}(x)d\tilde{Q}(y)[1 &+ n \sum_{k=1}^{\infty} a_k\psi_k^{(1)}(x)\psi_k^{(1)}(y) \\ &+ \binom{n}{2} \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} a_{k_1}a_{k_2}\psi^{(2)}_{k_1,k_2}(x)\psi^{(2)}_{k_1,k_2}(y) \\ &+ \dots \\ &+ \sum_{k_1=1}^{\infty} \dots \sum_{k_n=1}^{\infty} a_{k_1} \dots a_{k_n} \psi^{(n)}_{k_1,\dots,k_n}(x)\psi^{(n)}_{k_1,\dots,k_n}(y) \end{aligned}$$

$$(2.6.27)$$

Representation (2.6.27) holds almost everywhere ($\tilde{Q}$ measure) because Radon-Nikodym derivatives are defined up to sets of measure zero. Also, the coefficients in (2.6.27), being products of the nonnegative a_k's, are themselves nonnegative. Hence, to complete the proof we only have to show that the orthogonality conditions (2.6.13) hold for the ψ_k's of the expansion in (2.6.27)

For $\ell = 1,2, \dots, n$, and $1 \le k_1, \dots, k_\ell < \infty$,

we have

$$\int_{-\infty}^{\infty} \psi^{(\ell)}_{k_1,\ \ldots,}(t)_{k_\ell}\, d\tilde{Q}(t)$$

$$= \lim_{x\to+\infty} \chi^{(\ell)}_{k_1,\ \ldots,\ k_\ell}(x)$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{\ell} \eta_{k_i}(t_i) \prod_{i=1}^{n} dF(t_i)$$

$$= \left[\int_{-\infty}^{\infty} \eta_{k_1}(t_1)dF(t_1)\right]\left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=2}^{\ell} \eta_{k_i}(t_i) \prod_{i=2}^{n} dF(t_i)\right]$$

By hypothesis $\{\eta_k(\cdot)\}$ are a set of orthonormal functions on the marginal distribution $F(\cdot)$ of T so that

$$\int_{-\infty}^{\infty} \eta_{k_1}(t_1)\, dF(t_1) = 0 \qquad (2.6.28)$$

Hence, $$\int_{-\infty}^{\infty} \psi^{(\ell)}(t)\, d\tilde{Q}(t) = 0 \qquad (2.6.29)$$

where $\ell = 1,2, \ldots$ □

and this completes the proof.

The following facts about bivariate ranks are easy consequences of Theorems 2.6.2 and 2.6.3.

Corollary 2.6.1: Let $\binom{T_i}{U_i}$ be a random sample from a PDM (PDE) parent. Consider the marginal ranks

$$R_{1i} = \sum_{\alpha=1}^{n} I_{(T_i \geq T_\alpha)}$$

and

$$R_{2i} = \sum_{\alpha=1}^{n} I_{(U_i \geq U_\alpha)}$$

of T_i and U_i respectively, where $i = 1,2, \ldots, n$. The pair $\binom{R_{1i}}{R_{2i}}$ is PDM (PDE), $i = 1,2, \ldots, n$.

Proof: Fix i and define a function $g: R^n \to R$ by the equation

$$g_i(a_1, \ldots, a_n) = \sum_{\alpha=1}^{n} I_{(a_i \geq a_\alpha)}$$

and observe that

$$R_{1i} = g_i(T_1, \ldots, T_n), \; R_{2i} = g_i(U_1, \ldots, U_n)$$

By invoking Theorems 2.6.2 and 2.6.3, the result follows. □

We need one more result before we establish an optimality property of φ^*.

Theorem 2.6.4: Let random vectors $\binom{T_i}{U_i}$, $i = 1,2, \ldots, n$, be PDM/PDE and denote the ranks of T_1, U_1 among T_i's and U_j's by R_{11}, R_{21} respectively. Consider the joint probability mass function

$$\pi_{ij} = P(R_{11} = i, R_{21} = j), \; 1 \leq i, j \leq n$$

of R_{11} and R_{21}. Then, π_{ij}'s satisfy the following inequalities:

$$\forall \; i,j, \; \pi_{ii} + \pi_{jj} \geq 2\pi_{ij} \qquad (2.6.30)$$

Proof: By hypothesis, the parent distribution is PDM or PDE. According to Corollary 2.6.1, R_{11} and R_{21} are also PDM or PDE. Consequently, R_{11} and R_{21} are exchangeable random variables. Hence,

$$\pi_{ij} = \pi_{ji}, \text{ for } 1 \leq i,\ j \leq n \tag{2.6.31}$$

To establish (2.6.30), first consider the case when T and U are PDM. By Theorem 2.6.2, R_{11} and R_{21} are PDM. Hence, there exists a distribution function Q(•) say, such that

$$\pi_{ij} = \int_{-\infty}^{\infty} \pi_{i\bullet}(t)\ \pi_{\bullet j}(t)\ dQ(t),\ 1 \leq i,\ j \leq n \tag{2.6.32}$$

where $\pi_{i\bullet}(t)$ and $\pi_{\bullet j}(t)$ are the conditional mass functions of R_{11} and R_{21}, given a value t from the Q-distribution.

It follows from equation (2.6.32) that

$$\pi_{ii} + \pi_{jj} - 2\pi_{ij}$$

$$= \int_{-\infty}^{\infty} [(\pi_{i\bullet}(t))^2 + (\pi_{\bullet j}(t))^2 - 2\pi_{i\bullet}(t)\pi_{\bullet j}(t)]\ dQ(t)$$

$$= \int_{-\infty}^{\infty} (\pi_{i\bullet}(t) - \pi_{\bullet j}(t))^2\ dQ(t)$$

$$\geq 0\ .$$

We thus obtain (2.6.30) when T,U are PDM. Suppose now that T and U are PDE. Then, by virtue of Corollary 2.6.1, R_{11} and R_{21} would be PDE. R_{11} and R_{21} are ranks that are based on independent random variables, hence, R_{11} and R_{21} are both discrete uniform random variables on 1,2, ..., n (see Randles and Wolfe (1979), p. 38).

As R_{11} and R_{21} have finite supports the series expansion of R_{11} and R_{21} will have a finite number of terms. In fact, Fisher's identity (see Lancaster (1969), p. 90) holds:

$$\pi_{ij} = \frac{1}{n} \cdot \frac{1}{n} (1 + \sum_{k=1}^{n-1} a_k \eta_k(i) \eta_k(j))$$

$$1 \leq i, j \leq n \qquad (2.6.33)$$

where $\{a_k\}$ are nonnegative constants and $\{\eta_k(\cdot)\}$ are orthogonal functions on 1,2, ..., n. The representation (2.6.33) leads to the following reasoning:

For $1 \leq i, j \leq n$,

$$\begin{aligned}\pi_{ii} + \pi_{jj} - 2\pi_{ij} &= \frac{1}{n^2} [1 + \sum_{k=1}^{n-1} a_k(\eta_k(i))^2 + 1 + \\ &\qquad \sum_{k=1}^{n-1} a_k(\eta_k(j))^2 - 2 - 2 \sum_{k=1}^{n-1} a_k \eta_k(i) \eta_k(j)] \\ &= \frac{1}{n^2} \sum_{k=1}^{n-1} a_k[\eta_k(i) - \eta_k(j)]^2] \\ &\geq 0 \qquad (2.6.34)\end{aligned}$$

Hence, we obtain the inequalities in (2.6.30). An optimality of property φ^* can now be established:

<u>Theorem 2.6.5</u>: Let $\binom{T_i}{U_i}$, i=1,2, ..., n be as in Theorem 2.6.4. Then, $\forall\ \varphi \in \phi$,

$$E(N(\varphi)) \leq E(N(\varphi^*)) \qquad (2.6.35)$$

Proof: In Corollary 2.6.1, $N(\varphi)$ was written as a sum of exchangeable indicator random variables. Hence, using equation 2.6.11, we get

$$E(N(\varphi)) = nP(R_{21} = \varphi(R_{11})) \tag{2.6.36}$$

$$= n \sum_{k=1}^{n} P(R_{21} = \varphi(k),\ R_{11} = k)$$

$$= n \sum_{k=1}^{n} \pi_{k,\varphi(k)}\ ,$$

where π is the joint mass function of R_{11}, R_{21}. Invoking the inequalities on π_{ij} in (2.6.30) we obtain

$$E(N(\varphi)) \leq n \sum_{k=1}^{n} \frac{1}{2} (\pi_{k,k} + \pi_{\varphi(k),\varphi(k)})/2$$

$$= n[\frac{1}{2} \sum_{k=1}^{n} \pi_{k,k} + \frac{1}{2} \sum_{k=1}^{n} \pi_{\varphi(k),\varphi(k)}]$$

$$= n \sum_{i=1}^{n} \pi_{i,i}$$

$$= n\ P(R_{21} = R_{11})$$

$$= E(N(\varphi^*))$$

Which establishes the desired result. □

To interpret Theorem 2.6.5, we first recall from subsection 2.4.2 that $\varphi^* = (1,2,\ \ldots,\ n)$ is M.L.P if the parent density has the monotone likelihood ratio (MLR) property. As demonstrated by Shaked (1979), there is no general relationship between PDM/PDE concepts of positive dependence and the MLR property. We can therefore state the optimality of φ^* in Theorem 2.6.5 as below:

Let T,U have a joint density that has MLR property. In addition, let T and U be either PDM or PDE random variables. Let $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ be a broken random sample from the T-U population. Then the M.L.P φ^* is an optimal strategy to match the x's with the y's in the sense of maximizing the expected number of correct matches.

2.7 Monotonicity of $E(N(\varphi^*))$ with Respect to Dependence Parameters

Repairing of broken random samples based on the available data in two files was discussed in Section 2.4. It was observed that data-based optimal matching strategies exist when data come from populations having certain types of positive dependent structures. It is therefore reasonable to expect an optimal matching strategy to perform better when there is some kind of positive dependence in the population than when the data in the two files are stochastically independent. Our objective in this section is to present a precise account of such intuitive results with regard to the maximum likelihood pairing φ^*. To this end, we will draw upon the results of Section 2.6. We begin with a definition from Shaked (1979):

Definition 2.7.1: Let J be a subset of R. A kernel K defined on JxJ is said to be conditionally positive definite (c.p.d) on JxJ iff

(i) $K(x,y) = K(y,x)$, $\forall\ x,y \in J$; that is K is a symmetric kernel.

(ii) Let m be any positive integer. For arbitrary real numbers $a_1, \ldots, a_m$ and for every choice of distinct numbers $x_1, \ldots, x_m$ from J, it holds that

$$\sum_{i=1}^{m} \sum_{j=1}^{m} K(x_i,x_j)\ a_i a_j \geq 0 \text{ whenever } \sum_{i=1}^{m} a_i = 0 \qquad (2.7.1)$$

It is pertinent to note that this definition is related to the well-known concept of a positive definite kernel, which is used in, among others, the theory of characteristic functions. The nonnegativity of the quadratic form $\sum_{i=1}^{m} \sum_{j=1}^{m} K(x_i,x_j)\ a_i a_j$ without requiring the condition $\sum_{i=1}^{m} a_i = 0$ in (2.7.1) is a standard way of defining positive definite kernels (Widder, 1941, p. 271). We shall now give an example of a c.p.d kernel which will be used in the sequel.

Example 2.7.1: Let $J = \{1,2, \ldots, n\}$, where n is a fixed positive integer. To verify that the kernel $K(x,y) = I_{(x=y)}$ is conditionally positive definite on JxJ, let m be a positive integer. For arbitrary real numbers $a_1, \ldots, a_m$ and for every choice of distinct integers $i_1, \ldots, i_m$ from J, we have

$$\sum_{\alpha=1}^{m} \sum_{\beta=1}^{m} K(i_\alpha,i_\beta)\ a_\alpha a_\beta$$

$$= \sum_{\alpha,\beta:i_\alpha=i_\beta} \sum a_\alpha a_\beta$$

$$= \sum_{\alpha=1}^{m} a_\alpha^2$$

$$\geq 0 \qquad (2.7.2),$$

where we have used the fact that, in view of the integers $i_1, \ldots, i_m$ being distinct, $i_\alpha = i_\beta$ iff $\alpha=\beta$.

Note that we did not have to impose the condition $\sum_{i=1}^{m} a_i = 0$ to arrive at (2.7.2). Also, the function $I_{(x=y)}$ is clearly symmetric in x and y. Hence, it follows from (2.7.2) that K(x,y) is positive definite and, consequently, is also c.p.d.

We will need the following lemma.

Lemma 2.7.1 (Shaked, 1979): Let T and U be PDM or PDE random variables with joint distribution function H(t,u). Letting F(•) stand for the common marginal distribution of T and U, define $H_o(t,u) = F(t) \cdot F(u)$, the distribution function of T and U in the case of independence of the variables. Then we have the ordering

$$E_H(K(T,U)) \geq E_{H_o}(K(T,U)) \qquad (2.7.3)$$

iff K(.,.) is a c.p.d kernel, provided the expectations exist.

Theorem 2.7.1: Let the joint density of T,U have MLR property (2.4.4). Let H_o,H be as in Lemma 2.7.1. If $N = N(\varphi^*)$ is the number of correct matches due to the M.L.P φ^*, then

$$E_H(N) \geq 1. \qquad (2.7.4)$$

Proof: It follows from the general representation of $N(\varphi)$ in equation (2.6.11) that

$$E_H(N) = n\ P_H(R_{11} = R_{21}) = n\ E_H((K(R_{11},R_{21})) \tag{2.7.5}$$

where $K(x,y) = I_{(x=y)}$. Now, recall from example 2.7.1 that $K(x,y)$ is c.p.d. on the domain JxJ, where $J = \{1,2, \ldots, n\}$ is the common support of $\underset{\sim}{R}_{11}$ and R_{21}. It was established in Theorems 2.6.2 and 2.6.3 that R_{11} and R_{21} are PDM (PDE) according as T and U are PDM (PDE). Invoking Lemma 2.7.1, we therefore obtain

$$E_H(K(R_{11},R_{21})) \geq E_{H_o}(K(R_{11},R_{21})) \tag{2.7.6}$$

Under H_o, R_{11} and R_{21} are independent. Also, these ranks are marginally discrete uniform random variables on 1,2, ..., n. Hence, we get

$$\begin{aligned} E_{H_o}(K(R_{11},R_{21})) &= P_{H_o}(R_{11} = R_{21}) \\ &= \sum_{k=1}^{n} P(R_{11} = k)\ P(R_{21} = k) \\ &= \sum_{k=1}^{n} \frac{1}{n^2} \\ &= 1/n\ . \end{aligned} \tag{2.7.7}$$

Equations (2.7.5) to (2.7.7) imply the desired inequality:

$$E_H(N) \geq n \cdot \frac{1}{n} = 1\ . \qquad \square$$

We conclude from (2.7.4) that φ^* provides, on the average, more correct matches when the data in the two files come from certain positively dependent populations than when they are independent. In particular, this fact holds for the bivariate normal distribution with positive correlation as well as for Morgenstern distributions in Equation (2.6.14), where the dependence parameter $\alpha \geq 0$. In the light of Theorem 2.7.1, it is natural to conjecture that $E_H(N)$, as a functional of the distribution function H, is order-preserving with regard to certain partial orderings of the space of all continuous bivariate distributions which have fixed marginals (those of T and U) and exhibit positive dependence. Although no proof of this conjecture is available at this time, we offer further evidence in support of this conjecture in the next two theorems.

Theorem 2.7.2: Suppose that a broken random sample comes from the family of densities given by the equation

$$h(t,u) = 1 + \alpha\,(1-2t)(1-2u),\ 0 < t,\ u < 1 \text{ and } 0 \leq \alpha < 1 \quad (2.7.8)$$

Then, $E_\alpha(N)$ is monotone increasing in α.

Proof: Note that in (2.7.8), $\alpha = 0$ means T and U are independent and we might say that the farther α is from 0 the more the positive dependence between T and U. For this family, the marginal distributions of T and U are uniform on [0,1].

It follows from equation (2.6.27) and Corollary 2.6.1 that the joint probability function of the ranks R_{11} and R_{21} can be canonically expanded as follows:

$$\pi_{ij} = P(R_{11} = i,\ R_{21} = j)$$

$$= \frac{1}{n^2} [1 + \sum_{k=1}^{n} \binom{n}{k} \alpha^k \eta_k(i)\eta_k(j)] \qquad (2.7.9)$$

where $i,j = 1,2, \ldots, n$ and $\{\eta_k(\cdot)\}_1^n$ is a set of functions satisfying the orthogonality conditions in (2.6.13). Using the expression (2.7.9) for π_{ij} we get

$$E_\alpha(N) = n\ P(R_{11} = R_{21})$$

$$= n \sum_{i=1}^{n} \pi_{ii}$$

$$= n \cdot \frac{1}{n^2} [n + \sum_{i=1}^{n} \sum_{k=1}^{n} \binom{n}{k}\alpha^k(\eta_k(i))^2]$$

$$= 1 + \frac{1}{n} \sum_{k=1}^{n} \binom{n}{k} b_k \alpha^k \qquad (2.7.10),$$

where, after change of the order of summations on i and k, we have used nonnegative constants b_k given by the equation

$$b_k = \sum_{i=1}^{n} (\eta_k(i))^2, \quad k = 1,2, \ldots, n$$

It follows from (2.7.10) that $E_\alpha(N)$ is a polynomial in α and hence it increases with α, as α goes from 0 to 1. □

<u>Theorem 2.7.3</u>: Suppose that a broken random sample comes from the bivariate normal distributions given by (2.6.15), where we assume

that the correlation parameter ρ is nonnegative. Then $E_\rho(N)$ is increasing in ρ.

Proof: It follows from equation (2.6.27) and Corollary 2.6.2 that

$$\pi_{ij} = P(R_{11} = i,\ R_{21} = j)$$

$$= \frac{1}{n^2} [1 + n \sum_{k=1}^{\infty} \rho\, \psi_k^{(1)} (i)\, \psi_k^{(1)} (j)$$

$$+ \binom{n}{2} \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} \rho^2\, \psi^{(2)}_{k_1,k_2} (i)\, \psi^{(2)}_{k_1,k_2} (j)$$

$$+ \cdots$$

$$+ \sum_{k_1=1}^{\infty} \cdots \sum_{k_n=1}^{\infty} \rho^n\, \psi^{(n)}_{k_1,\ldots,\ k_n} (i)\, \psi^{(n)}_{k_1,\ldots,\ k_n} (j)],$$

(2.7.11)

where, for fixed $\ell = 1,2,\ \ldots,\ \{\psi^{(\ell)}_{k_1,\ldots,k_n}\}$ is a set of orthogonal functions on $\{1,2,\ \ldots,\ n\}$. Using the expression (2.7.11) for π_{ii}, we obtain

$$E_\rho(N) = nP(R_{11}=R_{21})$$

$$= n \sum_{i=1}^{n} \pi_{ii}$$

$$= \frac{1}{n} [n + n\cdot\rho \sum_{k=1}^{\infty} \sum_{i=1}^{n} (\psi_k^{(1)}(i))^2$$

$$+ \binom{n}{2} \rho^2 \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} \sum_{i=1}^{n} (\psi_{k_1,k_2}^{(2)}(i))^2$$

$$+ \cdots$$

$$+ \rho^n \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} \cdots \sum_{k_n=1}^{\infty} \sum_{i=1}^{n} (\psi_{k_1,\ldots,k_n}^{(n)}(i))^2],$$

(2.7.12)

where the order of summations over i and $k_1, \ldots, k_n$ have been reversed because the terms in the expansion (2.7.11) are all non-negative. We conclude from (2.7.12) that $E_\rho(N)$ is a polynomial in ρ and hence it increases with ρ as ρ goes from 0 to 1. □

As we close this section, we shall state a result due to Chew (1973) which somewhat resembles, though conceptually different from, the inequality $E_H(N) \geq 1$ in (2.7.4). Recall the notation $M(\varphi)$ in (2.4.9), which denotes the posterior expected number of correct matches due to the strategy φ. Arguing that $M(\varphi) = 1$ when φ is randomly chosen from Φ, he proved the following result:

Theorem 2.7.3: (Chew, 1973): Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be a broken random sample from a bivariate distribution possessing monotone likelihood ratio. If $x_1 < \ldots < x_n$ and $y_1 < \ldots < y_n$, then the posterior expected number of correct pairings using the M.L.P φ^* is at least unity, that is

$$M(\varphi^*) \geq 1 \tag{2.7.13}$$

It should be noted that the inequality (2.7.13) was derived from a Bayesian perspective, whereas in our inequality (2.7.4) the expectation is over all possible samples. Finally note that while our comparison is between dependent and independent populations for the M.L.P., Chew's inequality compares M.L.P with random pairing.

2.8 Some Properties of $N(\varphi^*,\varepsilon)$

The maximum likelihood pairing, φ^*, was introduced in subsection 2.4.2 and some of its small-sample properties were studied in Section 2.7. Specifically, the behavior of $E(N(\varphi^*))$ was discussed while holding the sample-size n constant and changing only the degree of dependence in the population. We shall now fix the parameters describing dependence in the population of $\binom{T}{U}$ and allow n to tend to infinity in order to study the behavior of $N(\varphi^*,\varepsilon)$. Later, in this section, we shall present the results of a Monte Carlo study about $N(\varphi^*,\varepsilon)$ in which we vary the dependence parameters even as n takes different values.

In this section, the notations of Section 2.2 will be used freely. Recall that $N(\varphi^*)$ and $N(\varphi^*,\varepsilon)$ have the shorter notations N and $N(\varepsilon)$ respectively. We start with a review of Yahav (1982)'s results concerning $E(N(\varepsilon))$.

Assuming that the distribution of T and U is such that the conditional distribution of U given that T = t is (univariate) normal with mean t and variance 1, Yahav (1982) derived the limiting value of $\mu_n(\epsilon) = E(N(\epsilon)/n)$, as $n \to \infty$, by using the representation (2.5.2) in which the summands are functions of the order-statistics of $U_1, \ldots, U_n$ and the concomitants of the order-statistics of $T_1, \ldots, T_n$. His proof relied on an approximation theorem (Bickel and Yahav, 1977) about the order-statistics for the above model. Furthermore, he reported the findings of a Monte-Carlo study for a particular case of his model, namely, T and U are bivariate normal with correlation ρ.

First, we discuss the large-sample behavior of $N(\epsilon)/n$ in case of samples from an arbitrary population. The properties of its expected value are available as a consequence. Second, we indicate how Yahav's simulation study of the small-sample properties of $\mu_n(\epsilon)$ can be improved upon. We shall then present the results of our own Monte-Carlo study of $\mu_n(\epsilon)$ when n is small.

Theorem 2.8.1: For broken random samples from an absolutely continuous distribution, $\frac{N(\epsilon)}{n} \overset{pr}{\to} \mu(\epsilon)$, as $n \to \infty$, (2.8.2)

where $\mu(\epsilon) = P(F(T-\epsilon) \leq G(U) \leq F(T+\epsilon))$.

Proof: Let $L_n = \frac{N(\epsilon)}{n}$. Recall the representation (2.6.6) for $N(\epsilon)$ as a sum of exchangeable indicators:

$$N(\epsilon) = \sum_{i=1}^{n} I_{A_{ni}}(\epsilon) \qquad (2.8.3)$$

It follows that

$$E(L_n) = nP(A_{n1}(\varepsilon))/n = P(A_{n1}(\varepsilon)) \ . \qquad (2.8.4)$$

Note that,

$$E(L_n^2) = n^{-2}[E(N(\varepsilon))^{(2)} + E(N(\varepsilon))], \qquad (2.8.5)$$

where $E(N(\varepsilon))^{(2)}$ is the second factorial moment of $N(\varepsilon)$. Using the exchangeable representation (2.8.3) again, we get

$$E(L_n^2) = n^{-2}[n^{(2)}P(A_{n1}(\varepsilon)A_{n2}(\varepsilon)) + nP(A_{n1}(\varepsilon))] \ .$$

Let $\eta_{1\alpha} = \sum_{i=1}^{n} \xi_{1\alpha i}$

$$\eta_{2\alpha} = \sum_{i=1}^{n} \xi_{2\alpha i}, \ \alpha = 1,2, \ldots, n, \qquad (2.8.6)$$

where the sequences $\{\xi_{1\alpha i}\}$ and $\{\xi_{2\alpha i}\}$ are defined in (2.2.12)

Using (2.8.6), we get

$$A_{n1}(\varepsilon) = (\eta_{11}/n \le 0, \ \eta_{21}/n \le 0) \qquad (2.8.7)$$

and

$$A_{n1}(\varepsilon)A_{n2}(\varepsilon) = \bigcap_{i=1}^{2} \bigcap_{j=1}^{2} (\eta_{ij}/n \le 0) \qquad (2.8.8)$$

Note that, given $\underset{\sim}{W}_1 = \binom{T_1}{U_1}$, the infinite sequence

$\xi_{112}, \xi_{113}, \cdots$ ad inf.

is exchangeable. Hence, by the Strong Law of Large Numbers (SLLN) for exchangeable random variables (see Chow and Teicher, p. 223),

$$\eta_{11}/n \overset{a \cdot s}{\rightarrow} E(\xi_{112}|\underset{\sim}{W}_1) \text{ as } n\to\infty \tag{2.8.9}$$

where the conditional expectation is equal to $F(t_1-\varepsilon) - G(u_1)$. It follows from (2.8.9) that

$$\eta_{11}/n \overset{a \cdot s}{\rightarrow} F(T_1-\varepsilon) - G(U_1) \tag{2.8.10}$$

We can show by similar arguments that

$$\eta_{1\alpha}/n \overset{a \cdot s}{\rightarrow} F(T_\alpha-\varepsilon) - G(U_\alpha) \tag{2.8.11}$$

$$\eta_{2\alpha}/n \overset{a \cdot s}{\rightarrow} G(U_\alpha) - F(T_\alpha+\varepsilon) \tag{2.8.12}$$

where $\alpha = 1,2$.

Using the fact (see Serfling, 1980 p. 52) that a sequence of vectors converges almost surely to a given vector iff the component-wise sequences converge almost surely to the appropriate components of the limit, we get from (2.8.11) and (2.8.12)

$$\begin{pmatrix} n_{11}/n \\ n_{21}/n \\ n_{12}/n \\ n_{22}/n \end{pmatrix} \xrightarrow{a\cdot s} \begin{pmatrix} F(T_1-\varepsilon) - G(U_1) \\ G(U_1) - F(T_1+\varepsilon) \\ F(T_2-\varepsilon) - G(U_2) \\ G(U_2) - F(T_2+\varepsilon) \end{pmatrix} \qquad (2.8.13)$$

It follows from (2.8.7), (2.8.8), (2.8.13) and the independence of $\binom{T_1}{U_1}$ and $\binom{T_2}{U_2}$ that

$$P(A_{n1}(\varepsilon)) \to \mu(\varepsilon) \qquad (2.8.14)$$

and

$$P(A_{n1}(\varepsilon)A_{n2}(\varepsilon)) \to \mu^2(\varepsilon) \qquad (2.8.15)$$

Using (2.8.4), (2.8.5), (2.8.14), (2.8.15) it is easy to verify that, as $n\to\infty$,

$$E(L_n) \to \mu(\varepsilon)$$

and (2.8.16)

$$\mathrm{var}(L_n) \to 0$$

It is well known that (2.8.16) implies the convergence in probability in (2.8.2). □

The following corollary generalizes Yahav (1982)'s result concerning $\mu_n(\varepsilon)$, the first moment of $N(\varepsilon)/n$.

Corollary 2.8.1: For $p > 0$,

(i) $\frac{N(\varepsilon)}{n} \overset{L_p}{\to} \mu(\varepsilon)$, as $n\to\infty$. (2.8.17)

(ii) $E(N(\varepsilon)/n)^p \to [\mu(\varepsilon)]^p$, as $n\to\infty$. (2.8.18)

Proof: The number of ε-correct matches can at most be n, the number of pairs in the unobserved bivariate-data. Hence,

$$0 \leq \frac{N(\varepsilon)}{n} \leq 1, \ \forall\ n = 1,2, \ldots$$

In other words, $\{N(\varepsilon)/n\}$ is a uniformly bounded sequence of random variables. It is well known that convergence in probability and L_p-convergence are equivalent for such sequences. Hence, (i) is an easy consequence of Theorem 2.8.1. It follows from (i) and Theorem 4.5.4 of Chung (1974) that the p^{th} moment of $N(\varepsilon)/n$ converges to $[\mu(\varepsilon)]^p$. Hence (ii) also holds. □

Note that no assumption about the conditional distribution of U given T was made either in Theorem 2.8.1 or Corollary 2.8.1.

Yahav generated samples from a bivariate-normal parent with mean vector $\binom{0}{0}$ and covariance matrix

$$\begin{pmatrix} \rho^2/(1-\rho^2) & \rho^2/(1-\rho^2) \\ \rho^2/(1-\rho^2) & 1/(1-\rho^2) \end{pmatrix} \qquad (2.8.19)$$

Note that in (2.8.19) the variances of T and U are functions of the correlation of T and U because Yahav requires that the conditional distribution of U given T = t be normal with mean t and variance 1.

The limiting value of $\mu_n(\epsilon)$ for his particular model was given by the integral

$$\mu(\epsilon) = \int_{-\infty}^{\infty} \{\Phi(x\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}} + \frac{\epsilon}{\rho}) - \Phi(x\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}} - \frac{\epsilon}{\rho})\} \, d\Phi(x) \qquad (2.8.20)$$

He computed $\mu(\epsilon)$ by numerical integration for $\epsilon = 0.01, 0.05, 0.1, 0.3$. He also provided Monte Carlo estimates of $\mu_n(\epsilon)$, for $n = 10$, 20 and 50 using the simulated data on T and U. The following table is a reproduction of some of his results.

Table 2.1 Expected Average Number of ϵ-Correct Matchings, $\epsilon = .01$

(Yahav (1982))

ρ	$\mu_{10}(\epsilon)$	$\mu_{20}(\epsilon)$	$\mu_{50}(\epsilon)$	$\mu(\epsilon)$
.01	.5864	.5326	.52752	.52269
.01	.1984	.1648	.12712	.11522
.10	.1512	.1058	.07600	.05912
.30	.1084	.0686	.03888	.02144
.50	.1020	.0582	.02720	.01382
.70	.0960	.0614	.02616	.01051
.90	.0972	.0540	.02064	.00864
.95	.0976	.0496	.02144	.00829
.99	.0960	.0484	.02128	.00804

It is clear from Table 2.1 that $\mu_n(\epsilon)$ and $\mu(\epsilon)$ are <u>decreasing</u> as ρ ranges from 0.01 to 0.99. However, one expects that an optimal strategy such as φ^* has the property that $\mu_n(\epsilon)$ as well as $\mu(\epsilon)$ are monotone increasing in ρ. The problem here is not with the M.L.P, φ^*, but with Yahav's model in (2.8.19) because, as the correlation changes its value, so do the marginal variances of T and U. To rectify this problem, we assumed a bivariate normal model for T and U in which the means were zero and the covariance matrix was

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{2.8.21}$$

For each combination of four values of n, namely 10, 20, 50 and 100, and twelve values of ρ, namely 0.00, 0.10 (0.10), 0.90, 0.95, 0.99, a sample of size 1000 was generated from the bivariate normal population using the IMSL subroutines. These data were used to obtain Monte-Carlo estimates of $\mu_n(\varepsilon)$, where ε was given the values 0.01, 0.05, 0.1, 0.3, 0.5, 0.75, 1.0. Furthermore, it is easy to show that, for the model in (2.8.21),

$$\mu(\varepsilon) = P(|Z| \leq \varepsilon/\sqrt{2(1-\rho)}), \tag{2.8.22}$$

where Z is a standard normal random variable. It is clear from (2.8.22) that $\mu(\varepsilon)$ is a monotone increasing function of ρ. Using standard-normal CDF tables, $\mu(\varepsilon)$ in (2.8.22) was computed for each combination of the twelve values of ρ and the seven values of ε mentioned above. We have presented the estimated values of $\mu_n(\varepsilon)$ and the limiting value $\mu(\varepsilon)$ in Table 2.2 to Table 2.8.

Table 2.2 Expected Average Number of ε-Correct Matchings, $\varepsilon = 0.01$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.106	0.054	0.025	0.015	0.008
0.10	0.113	0.059	0.028	0.017	0.008
0.20	0.127	0.068	0.031	0.018	0.008
0.30	0.138	0.075	0.034	0.020	0.008
0.40	0.155	0.083	0.038	0.023	0.008
0.50	0.174	0.095	0.044	0.026	0.008
0.60	0.199	0.109	0.051	0.030	0.008
0.70	0.231	0.129	0.061	0.036	0.008
0.80	0.279	0.162	0.077	0.046	0.016
0.90	0.374	0.222	0.109	0.067	0.016
0.95	0.476	0.296	0.151	0.094	0.024
0.99	0.700	0.521	0.299	0.191	0.056

Table 2.3 Expected Average number of ε-Correct Matchings, $\varepsilon = 0.05$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.127	0.076	0.047	0.037	0.032
0.10	0.134	0.082	0.051	0.040	0.032
0.20	0.149	0.093	0.056	0.043	0.032
0.30	0.161	0.099	0.061	0.047	0.032
0.40	0.180	0.109	0.066	0.052	0.040
0.50	0.201	0.124	0.074	0.057	0.040
0.60	0.228	0.141	0.085	0.065	0.048
0.70	0.262	0.166	0.101	0.076	0.048
0.80	0.317	0.205	0.124	0.094	0.064
0.90	0.420	0.280	0.174	0.135	0.088
0.95	0.529	0.368	0.237	0.186	0.127
0.99	0.769	0.631	0.459	0.377	0.274

Table 2.4 Expected Average Number of ε-Correct Matchings, $\varepsilon = 0.1$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.154	0.102	0.075	0.065	0.056
0.10	0.160	0.110	0.080	0.069	0.056
0.20	0.177	0.121	0.087	0.074	0.064
0.30	0.189	0.130	0.093	0.080	0.064
0.40	0.210	0.143	0.101	0.088	0.072
0.50	0.234	0.161	0.112	0.096	0.080
0.60	0.264	0.181	0.127	0.108	0.088
0.70	0.302	0.210	0.149	0.126	0.103
0.80	0.363	0.258	0.182	0.154	0.127
0.90	0.477	0.347	0.254	0.218	0.174
0.95	0.594	0.452	0.342	0.299	0.251
0.99	0.839	0.744	0.630	0.580	0.522

Table 2.5 Expected Average number of ε-Correct Matchings, $\varepsilon = 0.3$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.255	0.208	0.184	0.175	0.166
0.10	0.265	0.223	0.195	0.186	0.174
0.20	0.284	0.237	0.207	0.197	0.190
0.30	0.305	0.253	0.221	0.211	0.197
0.40	0.334	0.275	0.240	0.229	0.213
0.50	0.363	0.304	0.263	0.250	0.236
0.60	0.401	0.336	0.293	0.278	0.266
0.70	0.455	0.382	0.337	0.320	0.303
0.80	0.532	0.457	0.403	0.386	0.362
0.90	0.670	0.593	0.540	0.519	0.497
0.95	0.802	0.733	0.689	0.674	0.658
0.99	0.978	0.968	0.961	0.961	0.966

Table 2.6 Expected Average Number of
ε-Correct Matchings, $\varepsilon = 0.5$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.353	0.311	0.290	0.281	0.274
0.10	0.367	0.330	0.306	0.298	0.289
0.20	0.390	0.348	0.325	0.315	0.311
0.30	0.417	0.371	0.344	0.336	0.326
0.40	0.452	0.400	0.373	0.362	0.354
0.50	0.485	0.437	0.404	0.393	0.383
0.60	0.528	0.478	0.446	0.435	0.425
0.70	0.591	0.536	0.506	0.495	0.484
0.80	0.675	0.628	0.594	0.584	0.570
0.90	0.811	0.773	0.752	0.744	0.737
0.95	0.917	0.896	0.888	0.885	0.886
0.99	0.998	0.999	0.999	0.999	1.000

Table 2.7 Expected Average number of
ε-Correct Matchings, $\varepsilon = 0.75$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.468	0.433	0.416	0.409	0.404
0.10	0.488	0.454	0.437	0.429	0.425
0.20	0.514	0.477	0.461	0.453	0.445
0.30	0.539	0.505	0.487	0.480	0.471
0.40	0.582	0.542	0.522	0.514	0.503
0.50	0.621	0.586	0.560	0.555	0.547
0.60	0.662	0.633	0.613	0.606	0.599
0.70	0.727	0.694	0.679	0.673	0.668
0.80	0.810	0.786	0.772	0.768	0.766
0.90	0.919	0.908	0.906	0.904	0.907
0.95	0.979	0.976	0.978	0.979	0.982
0.99	1.000	1.000	1.000	1.000	1.000

Table 2.8 Expected Average Number of ϵ-Correct Matchings, $\epsilon = 1.0$

ρ	$\mu_{10}(\epsilon)$	$\mu_{20}(\epsilon)$	$\mu_{50}(\epsilon)$	$\mu_{100}(\epsilon)$	$\mu(\epsilon)$
0.00	0.570	0.545	0.531	0.524	0.522
0.10	0.593	0.566	0.555	0.549	0.547
0.20	0.621	0.595	0.581	0.576	0.570
0.30	0.646	0.622	0.611	0.605	0.605
0.40	0.690	0.664	0.650	0.644	0.627
0.50	0.729	0.707	0.691	0.688	0.683
0.60	0.772	0.753	0.744	0.741	0.737
0.70	0.830	0.812	0.807	0.805	0.803
0.80	0.898	0.889	0.887	0.885	0.886
0.90	0.970	0.970	0.972	0.972	0.975
0.95	0.996	0.996	0.997	0.997	0.998
0.99	1.000	1.000	1.000	1.000	1.000

Note that, as expected, $\mu_n(\epsilon)$ is a monotone increasing function of ρ for each fixed ϵ. Furthermore, the quality of the merged file is quite good if we want to recreate contingency tables with intervals of size $.5\sigma$ or more and the correlation ρ is ≥ 0.5.

2.9 Poisson Convergence of $N(\varphi^*)$

Let us revisit, for a moment, the card-matching problem which was discussed in Section 2.3. Some of the distributional properties of the number of correct matches in randomly arranging one pack of cards against another were stated in Proposition 2.3.1. In particular, the well-known approximation of the distribution of the number of correct matches by a Poisson distribution with mean 1 was mentioned. This Poisson approximation may be motivated by the observation that the occurrence of a match tends to be a rare event when the number of cards in the matching problem grows indefinitely. Inspired by this result, it is natural to ask whether Poisson distributions can approximate the distribution of the number of correct matches due to data-based matching strategies. The answer is in the affirmative in the case of the maximum likelihood pairing φ^*. Our aim in this section is to establish the Poisson convergence of $N(\varphi^*)$.

Using the general representation in Corollary 2.6.1 for the number of correct matches, we can write

$$N = N(\varphi^*) = \sum_{i=1}^{n} I_{A_{ni}} \qquad (2.9.1)$$

where $A_{ni} = (R_{1i} = R_{2i})$, $i = 1,2, \ldots, n$ are exchangeable events. It follows that $E(N) = nP(A_{n1})$. Zolutikhina and Latishev (1978) sketched a proof of the fact that the expectation of N converges to a constant as n tends to ∞. Their approach starts with writing $P(A_{ni})$ as the triple integral

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{\theta=0}^{\rho} \exp[(n-1)\ell n(s(x,y,\theta))]d\theta dH(x,y)$$

where $s(x,y,\theta) = p_3(x,y) + 2\sqrt{p_1(x,y)p_2(x,y)} \cdot \cos 2\theta$,

$$p_1(x,y) = F(x) - H(x,y),$$

$$p_2(x,y) = G(y) - H(x,y),$$

and $p_3(x,y) = 1 - p_1(x,y) - p_2(x,y)$, $\forall\ x,y \in R \quad 0 < \theta < \pi$.

Using the well-known method of Laplace (Bleistein and Handlesman 1975), they expanded this integral in powers of $\frac{1}{n}$ and concluded that $P(A_{n1}) \approx \frac{\alpha}{n}$ for large n, where the constant α is given by

$$\alpha = \int_{-\infty}^{\infty} [h(x,G^{-1}F(x))/h_2(G^{-1}F(x))]dx \tag{2.9.2}$$

They concluded that, in large samples, $E(N) \approx \alpha$.

In this section, we shall generalize the result of Zolutikhina and Latishev (1978) by showing that the d^{th} factorial moment of N, $E(N^{(d)})$, converges to $\alpha^d, d \geq 1$, under certain conditions on the distribution of $\binom{T}{U}$. As a consequence, we shall obtain the weak convergence of N to the Poisson distribution with mean α.

We begin with the observation that the ranks $\underset{\sim}{R}_1 = (R_{11}, \ldots, R_{1n})$ and $\underset{\sim}{R}_2 = (R_{21}, \ldots, R_{2n})$ are invariant under increasing functions of T and U respectively. For this reason, N is also invariant under such transformations. Without loss of generality, we therefore replace T and U by F(T) and G(U) respectively, where F(G) is the marginal distribution function of T(U). This so-called probability integral transformation allows us to assume that T and U are marginally uniform random variables and that the parent CDF, H(t,u), is the joint CDF of F(T) and G(U). Furthermore, the integral (2.9.2) simplifies to $\alpha = \int_0^1 h(x,x)dx$. We might recall from Section 2.2 that this simpler version of α was called λ. We shall henceforth use these simplifications and seek to prove that N weakly converges to the Poisson distribution with mean λ.

Following Schweizer and Wolff (1981), the joint CDF of F(T) and G(U) will be called a copula. In general, a copula is denoted by the symbol C(.,.) and the following Frechét bounds apply to any copula:

$$\max(x+y-1,0) \leq C(x,y) \leq \min(x,y),\ \forall\ (x,y) \in [0,1]^2 \qquad (2.9.3)$$

However, for the purpose of deriving the distribution of N, we shall consider only a part of the spectrum (2.9.3) of all possible copulas. To motivate our choice of the copulas, first note that, in this chapter, only absolutely continuous joint densities are allowed for T and U. This means that the extremes min(x+y-1,0) and min(x,y) are ruled out because these copulas correspond to degenerate joint distributions for T and U (Mardia 1970, p. 32). Second, Goel (1975) has observed that $\varphi^* = (1,2, \ldots, n)$ is M.L.P iff the joint density of T and U has the M.L.R property. However, M.L.R property necessarily implies that the distribution function of $\binom{T}{U}$ must be such that $C(x,y) \geq xy$, for all (x,y) in the unit-square (Tong (1980), p. 80). We shall henceforth assume that the joint CDF of T and U will satisfy the inequalities

$$xy \leq C(x,y) < \min(x,y), \ \forall \ (x,y) \in [0,1]^2. \qquad (2.9.4)$$

Note that, in (2.9.4), T and U are independent iff $C(x,y) \equiv xy$. Positive dependence of T and U occurs when $C(x,y) \geq xy$, for all x and y. In the remainder of this section, the joint CDF of T and U will be a copula C in the class (2.9.4) and the corresponding joint density function will be denoted by c(x,y).

Since $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ are some permutations of (1,2, ..., n), we find it convenient to use the notation φ for realizations of $\underset{\sim}{R}_1$ or $\underset{\sim}{R}_2$. The common support of $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ is denoted by Φ, the set of n! permutations of 1,2, ..., n.

We will now formally establish an equivalence between the card matching problem and the M.L.P in the independence case.

<u>Proposition 2.9.1</u>: Let T and U be independent random variables. Then the distribution of $\underset{\sim}{V} = (V_{n1}, \ldots, V_{nn})$ defined in (2.2.6) is the same as that of the vector $\underset{\sim}{\delta} \equiv (\delta_1, \ldots, \delta_n)$ where

$$\delta_{ni} = I_{(R_{1i}=i)}, \ i = 1,2, \ \ldots, \ n \qquad (2.9.5)$$

Furthermore, the random variables $\delta_1, \ldots, \delta_n$ are exchangeable.

<u>Proof</u>: Note that the rank vectors

$$\underset{\sim}{R}_1 = (R_{11}, \ldots, R_{1n}) \text{ and } \underset{\sim}{R}_2 = (R_{21}, \ldots, R_{2n})$$

are independent because T and U are, by hypothesis, independent random variables, and that $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ are discrete uniform on φ.

That is,

$$P(\underset{\sim}{R}_\alpha = \varphi) = \frac{1}{n!},\ \forall\ \varphi \in \Phi \text{ and } \alpha = 1,2. \tag{2.9.6}$$

As V_{ni}'s are indicators of the occurrence of matches, the Bernoulli variables $\delta_{n1}, \ldots, \delta_{nn}$ in (2.9.5) can be looked upon as indicating whether R_{1i} matches with i or not, i = 1,2, ..., n. It is clear that the common support of $\underset{\sim}{V}$ and $\underset{\sim}{\delta}$ is

$$\Delta = \{(a_1, \ldots, a_n):a_i\text{=0 or 1, i=1,2, } \ldots, n,\ \sum_{i=1}^{n} a_i \neq n - 1\} \tag{2.9.7}$$

Note that Δ has 2^n-n sample points.

Let $\underset{\sim}{a} = (a_1, \ldots, a_n)$ be a fixed but otherwise arbitrary point in Δ. Define the events

$$D(\underset{\sim}{a},\varphi) = [\psi \in \Phi:I_{(\psi(i)=\varphi(i))} = a_i,\ i = 1,2, \ldots, n], \tag{2.9.8}$$

where $\varphi \in \Phi$. Then, using the independence of $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ and (2.9.8) we get

$$\begin{aligned} P(\underset{\sim}{V} = \underset{\sim}{a}) &= P(I_{(R_{1i}=R_{2i})} = a_i,\ i = 1,2, \ldots, n) \\ &= E^{\underset{\sim}{R}_2} P(I_{(R_{1i}=\varphi(i))} = a_i,\ i = 1,2, \ldots, n|\underset{\sim}{R}_2 = \varphi) \\ &= E^{\underset{\sim}{R}_2} P(I_{(R_{1i}=\varphi(i))} = a_i,\ i = 1,2, \ldots, n) \\ &= E^{\underset{\sim}{R}_2} P(\underset{\sim}{R}_1 \in D(\underset{\sim}{a},\varphi)) \end{aligned} \tag{2.9.9}$$

We now observe that the components of $\underset{\sim}{a}$ dictate which positions of $\varphi = (\varphi(1), \ldots, \varphi(n))$ must be matched or mismatched by any permutation ψ in order that $\psi \in D(a,\underset{\sim}{\varphi})$. Clearly, the number of ways in which we can permute the integers 1,2, ..., n and produce ψ's that belong to $D(\underset{\sim}{a},\varphi)$ depends only on the fixed vector a and the fact that φ is an arrangement of n distinct integers. Hence the cardinality of $D(\underset{\sim}{a},\varphi)$ does not change as φ ranges over Φ. In particular, $D(\underset{\sim}{a},\varphi)$ and $D(\underset{\sim}{a},\varphi^*)$ have the same number of sample points, where $\varphi^* = (1,2, \ldots, n)$. Using (2.9.6), we therefore obtain

$$P(\underset{\sim}{R}_1 \in D(\underset{\sim}{a},\varphi)) = P(\underset{\sim}{R}_1 \in D(\underset{\sim}{a},\varphi^*)), \ \forall \ \varphi \in \Phi \qquad (2.9.10)$$

The right-hand-side expression in (2.9.10) is a fixed number depending on φ^* and the chosen $\underset{\sim}{a}$. This means that in (2.9.9), we seek the expectation of a degenerate random variable. Hence, we obtain

$$\begin{aligned} P(\underset{\sim}{V} = \underset{\sim}{a}) &= P(\underset{\sim}{R}_1 \in D(\underset{\sim}{a},\varphi^*)) \\ &= P(I_{(R_{1i}=i)} = a_i, \ i = 1,2, \ldots, n) \\ &= P(\underset{\sim}{\delta} = \underset{\sim}{a}) \qquad (2.9.11) \end{aligned}$$

Because $\underset{\sim}{a}$ was arbitrarily chosen from Δ, we finally infer from (2.9.11) that

$$(V_{n1}, \ldots, V_{nn}) \overset{d}{=} (\delta_{n1}, \ldots, \delta_{nn}) \qquad (2.9.12)$$

The exchangeability of $\delta_1, \ldots, \delta_n$ follows from the fact that the distribution of $\underset{\sim}{R}_1$ is uniform over Φ.

It readily follows from Proposition 2.9.1 that, in the independence case,

$$\sum_{i=1}^{n} V_{ni} \overset{d}{=} \sum_{i=1}^{n} \delta_{ni} \tag{2.9.13}$$

In view of (2.9.13), if we let $Z_n = \sum_{i=1}^{n} \delta_{ni}$, then the exact as well as asymptotic distributions of $N(\varphi^*) = \sum_{i=1}^{n} V_{ni}$ can be derived by studying Z_n, which is same as the no. of matches in the card matching problem. As stated in Proposition 2.3.1, the asymptotic distribution of Z_n is Poisson with mean 1. We now present another proof of this well-known result. The novel part of our proof is that we establish certain dependence properties of $\delta_{n1}, \ldots, \delta_{nn}$ and consequently derive the limiting distribution by using only the first two moments of Z_n.

Our program can be stated as below:

(i) Show that δ_{ni}'s have a certain positive dependence structure.

(ii) Invoke a theorem due to Newman (1982) to arrive at the Poisson convergence of N in the independence case.

We start with the definitions of some concepts of dependence of random variables.

Definition 2.9.1 (Lehmann, 1966): x_1 and x_2 are said to be positive quadrant dependent (PQD) iff

$$P(x_1 > x_1, x_2 > x_2) \geq P(x_1 > x_1)\, P(x_2 > x_2), \ \forall\, x_1, x_2 \in \quad . \tag{2.9.14}$$

Definition 2.9.2 (Newman, 1982): $x_1, \ldots, x_n$ are said to be linearly positive quadrant dependent (LPQD) iff for any disjoint subsets A,B of $\{1,2, \ldots, n\}$ and positive constants $a_1, \ldots, a_n$,

$$\sum_{k \in A} a_k x_k \text{ and } \sum_{k \in B} a_k x_k \text{ are PQD.} \tag{2.9.15}$$

Definition 2.9.3 (Esary, Proschan, Walkup, 1967): $x_1, \ldots, x_n$ are said to be associated iff for every choice of functions $f_1(x_1, \ldots, x_n)$ and $f_2(x_1, \ldots, x_n)$, which are monotonic increasing in each argument,

$$\mathrm{cov}(f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n)) \geq 0, \tag{2.9.16}$$

provided $f_1(x_1, \ldots, x_n)$ and $f_2(x_1, \ldots, x_n)$ have finite variance.

It is well-known that association is a stronger property than LPQD property of n random variables $x_1, \ldots, x_n$. We will now establish that $\delta_{n1}, \ldots, \delta_{nn}$ in (2.9.5) possess a weaker version of the LPQD property.

Lemma 2.9.1: For $k = 1,2, \ldots, n-1$,

$$\sum_{i=1}^{k} \delta_{ni} \text{ and } \delta_{nn} \text{ are PQD.} \tag{2.9.17}$$

Proof: Fix $k = 1,2, \ldots, n-1$. Then, using (2.9.14), we see that

$\sum_{i=1}^{k} \delta_{ni}$ and δ_{nn} are PQD if

$$P(\sum_{i=1}^{k} \delta_{ni} > x_1, \delta_{nn} > x_2) \geq P(\sum_{i=1}^{k} \delta_{ni} > x_1)\, P(\delta_{nn} > x_2), \forall\, x_1, x_2 \in R \tag{2.9.18}$$

Because δ_{ni}'s are binary random variables we obtain

$$P(\delta_{nn} > x_2) = \begin{cases} 1 & \text{if} \quad x_2 < 0 \\ \\ 0 & \text{if} \quad x_2 \geq 1 \end{cases} \tag{2.9.19}$$

It is clear from (2.9.19) that (2.9.18) holds for any x_1, provided $x_2 < 0$ or $x_2 \geq 1$. Hence, it suffices to show (2.9.18) for $0 \leq x_2 < 1$. However, if $0 \leq x_2 < 1$, then $(\delta_{nn} > x_2) = (\delta_{nn} = 1)$. It therefore remains to be shown that

$$P(\sum_{i=1}^{k} \delta_{ni} \geq \ell, \delta_{nn} = 1) \geq P(\sum_{i=1}^{k} \delta_{ni} \geq \ell)\, P(\delta_{nn} = 1),$$
$$\forall\, \ell = 0,1, \ldots, k. \tag{2.9.20}$$

By definition of δ_{ni},

$$P(\delta_{ni} = 1) = P(R_{1i} = i) = \frac{1}{n},$$
$$\text{and} \quad P(\delta_{ni} = 0) = 1 - \frac{1}{n}\ . \tag{2.9.21}$$

Writing $P(\sum_{i=1}^{k} \delta_{ni} \geq \ell)$ in the form

$$P(\sum_{i=1}^{n} \delta_{ni} \geq \ell, \delta_{nn} = 0) + P(\sum_{i=1}^{n} \delta_{ni} \geq \ell, \delta_{nn} = 1)$$

and using (2.9.21) we can rewrite (2.9.20) in a more useful form:

$$P(\sum_{i=1}^{k} \delta_{ni} \geq \ell|\delta_{nn} = 0) \leq P(\sum_{i=1}^{k} \delta_{ni} \geq \ell|\delta_{nn} = 1),$$

$$\ell = 0, \ldots, k. \qquad (2.9.22)$$

Note that, in (2.9.22), k is a fixed integer. For a given k, we now fix the value of ℓ and proceed to establish the inequality in (2.9.22) by means of a combinational argument.

It is clear that we can express the event ($\delta_{nn} = 0$) or as $\bigcup_{\alpha=1}^{n-1} (R_{1n} = \alpha)$. Hence we can write,

$$(\sum_{i=1}^{k} \delta_{ni} \geq \ell, \delta_{nn} = 0) = \bigcup_{\alpha=1}^{n-1} J_{\alpha} \qquad (2.9.23)$$

where

$$J_{\alpha} = (\sum_{i=1}^{k} \delta_{ni} \geq \ell, R_{1n} = \alpha), \alpha = 1,2, \ldots, n-1 \qquad (2.9.24)$$

Observe that, in (2.9.24), J_{α}'s are mutually disjoint measureable subsets of Φ. Let us now fix $\alpha = 1,2, \ldots, n-1$ as well. Then, any permutation φ in J_{α} satisfies $\varphi(n) = \alpha$ and $(\varphi(1), \ldots, \varphi(n-1))$ is an arrangement of the integers $1,2, \ldots, \alpha-1,\alpha+1, \ldots, n$ producing at least ℓ matches of the type $\varphi(i) = i$ in the positions $i = 1,2, \ldots, k$. On the other hand, any permutation φ in

$(\sum_{i=1}^{k} \delta_{ni} \geq \ell,\ \delta_{nn} = 1)$ satisfies $\varphi(n) = n$ and

$(\varphi(1), \ldots, \varphi(n-1))$ is an arrangement of the integers $1,2, \ldots, n-1$ yielding at least ℓ matches such as $\varphi(i) = i$ in the positions $i = 1,2, \ldots, k$. Because $\alpha \neq n$, it is clear that

$$\#(J_\alpha) \leq \#(\sum_{i=1}^{k} \delta_{ni} \geq \ell,\ \delta_{nn} = 1)\ , \tag{2.9.25}$$

where $\#(A)$ denotes the cardinality of the set A.

Since α, k and ℓ were arbitrary choices, we get from (2.9.23),

$$\#(\sum_{i=1}^{k} \delta_{ni} \geq \ell, \delta_{nn} = 0) \leq (n-1)\ \#(\sum_{i=1}^{k} \delta_{ni} \geq \ell, \delta_{nn} = 1)$$

$$k = 1,2, \ldots, n-1;\ \ell = 0, \ldots, k \tag{2.9.26}$$

Since $\underset{\sim}{R}_1$ is discrete uniform on Φ it follows from (2.9.26) that

$$P(\sum_{i=1}^{k} \delta_{ni} \geq \ell, \delta_{nn} = 0) \leq P(\sum_{i=1}^{k} \delta_{ni} \geq \ell, \delta_{nn} = 1) \cdot (n-1) \tag{2.9.27}$$

Multiplying both sides of the inequality in (2.9.27) by n and using (2.9.21) we establish (2.9.22), which implies that (2.9.20) holds. □

We now state two useful results due to Newman.

<u>Lemma 2.9.2 Newman (1982)</u>: If x_1 and x_2 are PQD, then

$$|E(\exp(irx_1+isx_2)) - E(\exp(irx_1))\ E(\exp(isx_2)|$$

$$\leq |rs|\ \mathrm{cov}(x_1,x_2) \quad \text{for all } r,s \in R \qquad (2.9.28)$$

□

Lemma 2.9.3 Newman (1982): Suppose that $x_1, \ldots, x_n$ are LPQD. Then

$$|\Psi_{x_1,\ldots,x_n}(r_1,\ldots,r_n) - \prod_{j=1}^{n} \Psi_{x_j}(r_j)| \leq \sum_{\substack{k=1 \\ k<\ell}}^{n} \sum_{\ell=1}^{n} |r_k r_\ell|\ \mathrm{cov}(x_k,x_\ell)$$

$$\forall\ r_1, \ldots, r_m \in R\ , \qquad (2.9.29)$$

where Ψ's are given by

$$\Psi_{x_1,\ldots,x_n} = E(\exp(i \sum_{j=1}^{n} r_j x_j))$$

$$\Psi_{x_j} = E(\exp(i\ r_j x_j),\ j = 1,2,\ \ldots,\ n. \qquad □$$

Suppose now that we choose the arguments $r_1, \ldots, r_n$ in (2.9.29) equal to an arbitrary real number r, say. Assume further that $x_1, \ldots, x_n$ are exchangeable random variables so that they have common characteristic function, namely $\Psi_{x_1}(r)$ and that the covariance between any pair of the x_j's is equal to $\mathrm{cov}(x_1,x_2)$. It follows from (2.9.29) that

$$|\Psi_{\Sigma x_i}(r) - \Psi^n_{x_1}(r)| \leq \frac{n(n-1)}{2} |r|^2\ \mathrm{cov}(x_1,x_2) \qquad (2.9.30)$$

This estimate for approximating the characteristic function of $\sum_{i=1}^{n} x_i$

by the product of the marginal characteristic functions of the x's depends on the fact that $x_1, \ldots, x_n$ are LPQD. We now use Lemma 2.9.2 and show that, with regard to the variables $\delta_{n1}, \ldots, \delta_{nn}$, an estimate similar to (2.9.30) can be obtained under the weaker version of the LPQD property which is given by (2.9.17).

Lemma 2.9.4: Let δ_{ni}'s be the Bernoulli variables in (2.9.5) and let $Z_n = \sum_{i=1}^{n} \delta_{ni}$. Then,

$$|\Psi_{Z_n}(r) - \Phi^n_{\delta_{n1}}(r)| \leq \frac{n(n-1)}{2} |r|^2 \operatorname{cov}(\delta_{n1}, \delta_{n2}),$$

$$\forall\ n \geq 2,\ r \in R\ , \tag{2.9.31}$$

Proof: The exchangeability of $\delta_{n1}, \ldots, \delta_{nn}$ was established in Proposition 2.9.1. Hence, we obtain

$$\operatorname{cov}(\delta_{ni}, \delta_{nj}) = \operatorname{cov}(\delta_{n1}, \delta_{n2}),\ \forall\ i \neq j, \tag{2.9.32}$$

$$\Psi_{\delta_{nj}}(r) \equiv \Psi_{\delta_{ni}}(r),\ \forall\ j, \tag{2.9.33}$$

Note also the well-known property that

$$|\Psi_{\delta_{nj}}(r)| \leq 1,\ \forall\ j \text{ and } \forall\ r \tag{2.9.34}$$

From Lemma 2.9.1, we have

$$\sum_{i=1}^{k} \delta_{ni} \text{ and } \delta_{nn} \text{ are PQD, } \forall\ k = 1,2, \ldots, n-1.$$

In view of the exchangeability of $\delta_{n1}, \ldots, \delta_{nn}$, we can restate this property of the δ_{ni}'s as follows:

Let A and B be non-empty <u>disjoint</u> subsets of $\{1,2, \ldots, n\}$ such that B is a <u>singleton</u>. Then

$$\sum_{i \in A} \delta_{ni} \text{ and } \sum_{i \in B} \delta_{ni} \text{ are PQD} \qquad (2.9.35)$$

Fix $n \geq 2$ and consider the following finite sequence of statements:

$$|\Psi_{\sum_{i=1}^{m} \delta_{ni}}(r) - \Psi^{m}_{\delta_{n1}}(r)| \leq \frac{m(m-1)}{2} |r|^2 \operatorname{cov}(\delta_{n1}, \delta_{n2}),$$

$$\forall\ m = 2,3, \ldots, n \qquad (2.9.36)$$

Note that (2.9.31) is obtained from (2.9.36) by letting $m = n$. We shall now establish (2.9.36) by induction on m.

By choosing $A = \{1\}$, $B = \{2\}$ in (2.9.35), we find that δ_{n1} and δ_{n2} are PQD. The Lemma 2.9.2 readily implies that (2.9.36) holds for $m = 2$. Now, let us assume that (2.9.36) holds for $m = 2,3, \ldots, (n-1)$. Splitting $\sum_{i=1}^{n} \delta_{ni}$ as the sum of $\sum_{i=1}^{n-1} \delta_{ni}$ and δ_{nn}, we infer the PQD property of $\sum_{i=1}^{n-1} \delta_{ni}$ and δ_{nn} from (2.9.35). Hence we obtain again from Lemma 2.9.2 and (2.9.32)

$$|\Psi_{\sum_{i=1}^{n} \delta_{ni}}(r) - \Psi_{\sum_{i=1}^{n-1} \delta_{ni}}(r) \cdot \Psi_{\delta_{nn}}(r)|$$

$$\leq |r|^2 \operatorname{cov}(\sum_{i=1}^{n-1} \delta_{ni}, \delta_{nn})$$

$$= |r|\ ^2(n-1) \operatorname{cov}(\delta_{n1}, \delta_{n2}) \qquad (2.9.37)$$

Now, we shall invoke the induction hypothesis that (2.9.36) holds for m = n - 1. Using (2.9.33) to (2.9.37) we finally establish (2.9.36) for m = n as follows:

$$|\Psi_{\sum_{i=1}^{n} \delta_{ni}}(r) - \Psi^{n}_{\delta_{n1}}(r)|$$

$$\leq |\Psi_{\sum_{i=1}^{n} \delta_{ni}}(r) - \Psi_{\sum_{i=1}^{n-1} \delta_{ni}}(r) \cdot \Psi_{\delta_{nn}}(r)|$$

$$+ |\Psi_{\sum_{i=1}^{n-1} \delta_{ni}}(r)\, \Psi_{\delta_{nn}}(r) - \Psi^{n}_{\delta_{n1}}(r)|$$

$$\leq |r|^2 (n-1) \operatorname{cov}(\delta_{n1},\delta_{n2})$$

$$+ |\Psi_{\sum_{i=1}^{n-1} \delta_{ni}}(r) - \Psi^{n-1}_{\delta_{n1}}(r)|$$

$$\leq |r|^2 (n-1) \operatorname{cov}(\delta_{n1},\delta_{n2}) + |r|^2 \cdot \frac{(n-1)(n-2)}{2} \operatorname{cov}(\delta_{n1},\delta_{n2})$$

$$= |r|^2 \operatorname{cov}(\delta_{n1},\delta_{n2})(n-1)[1 + \frac{n-2}{2}]$$

$$= \frac{n(n-1)}{2} |r|^2 \operatorname{cov}(\delta_{n1},\delta_{n2}) \qquad (2.9.38)$$

The proof of (2.9.36) is complete by our inductive argument and (2.9.31) follows from (2.9.38). □

Our preparations so far in this section are adequate for the purpose of establishing the Poisson convergence of N in the independence case.

Theorem 2.9.1: Let T and U be independent random variables. Let the number of correct matches, N, be given by (2.9.1). Then

$$N \to \text{Poisson } (1), \text{ as } n \to \infty \qquad (2.9.39)$$

Proof: We obtain from (2.9.13)

$$N \overset{d}{=} Z_n,$$

where $Z_n = \sum_{i=1}^{n} \delta_{ni}$. Using the exchangeability of δ_{ni}'s, we obtain

$$\text{cov}(\delta_{n1},\delta_{n2}) = P(R_{11}= 1,R_{12}= 2) - [P(R_{11}=1)]^2 \qquad (2.9.40)$$

Since $P(R_{11}=1,R_{12}=2) = 1/n(n-1)$, it follows that

$$n(n-1)\ \text{cov}(\delta_{n1},\delta_{n2}) = \frac{1}{n}\ ,\ \forall\ n \geq 2,$$

and therefore

$$n(n-1)\ \text{cov}(\delta_{n1},\delta_{n2}) = O(1) \text{ as } n\to\infty \qquad (2.9.41)$$

The proof of (2.9.39) consists of showing that the characteristic function of Z_n converges to the characteristic function of the Poisson distribution with mean 1. In other words, we shall show that

$$\Psi_{Z_n}(r) \to \exp(\exp(ir) - 1),\ \forall\ r \in R \quad \text{as } n\to\infty \qquad (2.9.42)$$

To this end, Lemma 2.9.4 gives the following estimate of the difference between the characteristic functions in (2.9.49)

$$|\Psi_{Z_n}(r) - \exp(\exp(ir) - 1)|$$

$$\leq |\Psi_{Z_n}(r) - \Psi^n_{\delta_{n1}}(r)| + |\Psi^n_{\delta_{n1}}(r) - \exp(\exp(ir) - 1)|$$

$$\leq \frac{n(n-1)}{2} |r|^2 \operatorname{cov}(\delta_{n1},\delta_{n2}) + |\Psi^n_{\delta_{n1}}(r) - \exp(\exp(ir) - 1)| \tag{2.9.43}$$

Now, using the distribution of δ_{n1} given by (2.9.21) we get

$$\Psi_{\delta_{n1}}(r) = [1 + \frac{1}{n}(\exp(ir) - 1)] .$$

Clearly,

$$\Psi^n_{\delta_{n1}}(r) \to \exp(\exp(ir) - 1), \forall r \in R, \text{ as } n \to \infty \tag{2.9.44}$$

It readily follows from (2.9.41), (2.9.43) and (2.9.44) that (2.9.42) holds. Hence we obtain

$$Z_n \overset{d}{\to} \text{Poisson } (1) \tag{2.9.45}$$

which is equivalent to (2.9.39). □

We now assume that the broken random sample comes from a population in which T and U are dependent random variables. It should be noted that extensions of some of the techniques used in the proof of the Poisson convergence in the independence case to the dependence case are not available at this time. Specifically, no proof of the counterpart of (2.9.17), namely

$$\sum_{i=1}^{k} V_{ni} \text{ and } V_{nn} \text{ are PQD } \forall\, k = 1,2, \ldots, n-1, \forall\, n \geq 2 \tag{2.9.46}$$

is known. However, direct verification of the association of $V_{n1}, \ldots, V_{nn}$ has been carried out for n=2,3,4 when T and U have the Morgenstern distribution given by (2.6.16). Since association of random variables is a much stronger dependent structure than (2.9.46), it is natural to conjecture that Lemma 2.9.1 holds even when T and U are dependent.

In the absence of a valid proof of Lemma 2.9.1 in the dependence case, we need extra conditions on the distribution of T and U in order to derive the Poisson convergence of N. The following lemma will be useful in deriving the main result of this section.

<u>Lemma 2.9.5</u>: For a fixed d, let $\underset{\sim}{L}_n = \frac{\underset{\sim}{S}_n}{n}$ and $\underset{\sim}{L} = (L_1, \ldots, L_d)'$, $\underset{\sim}{S}_n$ and $\underset{\sim}{L}$ are defined in Section 2.2. Then,

$$\underset{\sim}{L}_n \overset{a.s}{\rightarrow} \underset{\sim}{L}, \text{ as } n \rightarrow \infty \tag{2.9.47}$$

<u>Proof</u>: Fix d $\geq$ 1. It is clear from the definitions of $\underset{\sim}{\xi}_k$ in (2.2.10) and the sigma-field Λ_d in Section 2.2 that the infinite sequence

$$\underset{\sim}{\xi}_{d+1}, \underset{\sim}{\xi}_{d+2}, \ldots, \ldots$$

of d-dimensional vectors are conditionally i.i.d given Λ_d. Hence, using the Strong Law of Large Numbers for exchangeable sequences (Chow and Teicher, p. 223) we get

$$\frac{1}{n-d} \sum_{k=d+1}^{n} \underset{\sim}{\xi}_k \overset{a.s}{\rightarrow} E(\underset{\sim}{\xi}_{d+1} | \Lambda_d) \tag{2.9.48}$$

In order to evaluate the limiting conditional expectation in (2.9.48), note first that, for j = 1,2, ..., d, T_j and U_j are uniform random variables. Now,

$$E(\xi_{jd+1} | T_j = t_j, U_j = u_j)$$

$$= P(t_j - T_{d+1} \geq 0) - P(u_j - U_{d+1} \geq 0)$$

$$= P(T_{d+1} \leq t_j) - P(U_{d+1} \leq u_j)$$

$$= t_j - u_j.$$

$$= L_j. \tag{2.9.49}$$

Therefore, it follows from the definition of $\underset{\sim}{\xi}_{d+1}$ in (2.2.10) and (2.9.49)

$$E(\underset{\sim}{\xi}_{d+1} | \Lambda_d) = (L_1, L_2, \ldots, L_d)'. \tag{2.9.50}$$

Hence, (2.9.48) and (2.9.50) imply that

$$\frac{1}{n-d} \sum_{k=d+1}^{n} \underset{\sim}{\xi}_k \overset{a.s}{\rightarrow} \underset{\sim}{L}, \text{ as } n \rightarrow \infty \tag{2.9.51}$$

Also, d being a fixed integer, we have

$$\frac{1}{n-d} \sum_{k=1}^{d} \underset{\sim}{\xi}_k \overset{a.s}{\to} \underset{\sim}{0}, \text{ as } n \to \infty \qquad (2.9.52)$$

Since,

$$\underset{\sim}{L}_n = \frac{1}{n} \sum_{k=1}^{n} \underset{\sim}{\xi}_k$$

the lemma follows from (2.9.51) and (2.9.52) □

The following sufficient conditions will be used to prove the next theorem.

<u>Assumptions</u>: In the notations of Section 2.2, let

(a) $\lambda < \infty$ (2.9.53)

(b) $\int_{-\infty}^{\infty} |\Psi_L(\theta)| \, d\theta < \infty$ (2.9.54)

and (c) $P(\Psi_d^* \leq t) = O(t^d)$ as $t \to \infty$, $\forall\ d \geq 1$ (2.9.55)

<u>Theorem 2.9.2</u>: If Assumptions (2.9.53) to (2.9.55) hold, then

$$N \overset{d}{\to} \text{Poisson } (\lambda) \text{ as } n \to \infty \qquad (2.9.56)$$

<u>Proof</u>: Proof of (2.9.56) consists in showing that the factorial moments of N converge to those of the Poisson distribution with mean λ, in other words,

$$E(N^{(d)}) \to \lambda^d, \ \forall\ d = 1,2, \ldots, \qquad (2.9.57)$$

By the Fourier inversion theorem,

$$P(\underset{\sim}{S}_n = \underset{\sim}{0}) = (2\pi)^{-d} \int_{-\pi}^{\pi} \dots \int_{-\pi}^{\pi} \Psi_{\underset{\sim}{S}_n}(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}, \qquad (2.9.58)$$

where $\Psi_{\underset{\sim}{S}_n}(\underset{\sim}{\theta})$ is the characteristic function of the d-dimensional random vector $\underset{\sim}{S}_n$ defined in (2.2.7).

The Assumption (2.9.54) ensures that the Fourier inversion theorem can be applied to the continuous random variable L. Noting that $\lambda = \int_0^1 c(x,x)\, dx$ is the value of the density function of L at 0, we get

$$\lambda = g_L(0) = (2\pi)^{-1} \int_{-\infty}^{\infty} \Psi_L(t)\, dt$$

Since $L_j = T_j - U_j$, $j = 1,2, \dots, d$, are i.i.d, with their common density function equal to $g_L(.)$ it follows that

$$\lambda^d = (2\pi)^{-d} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta})\, d\underset{\sim}{\theta} \qquad (2.9.59)$$

Recalling the representation

$$N(\varphi^*) = \sum_{i=1}^{n} I_{A_{ni}}$$

from Corollary 2.6.1, we obtain

$$E(N^{(d)}) = n^{(d)}\, P(A_{n1} A_{n2} \dots A_{nd}),$$

$$= n^{(d)}\, P(\underset{\sim}{S}_n = \underset{\sim}{0}), \qquad (2.9.60)$$

where $n^{(d)} = n(n-1)\ \ldots\ (n-d+1)$.

For fixed d, it is clear that $n^{(d)} \approx n^d$ as $n \to \infty$. It therefore follows from (2.9.60) that, in order to prove (2.9.57), it is sufficient to show that

$$\lim_{n\to\infty} |\Delta(d,n)| = 0, \tag{2.9.61}$$

where $\Delta(d,n) = n^d P(\underset{\sim}{S}_n = \underset{\sim}{0}) - \lambda^d$

From (2.9.58) and (2.9.59), we obtain

$$\Delta(d,n) = n^d(2\pi)^{-d} \int_{-\pi}^{\pi} \ldots \int_{-\pi}^{\pi} \Psi_{\underset{\sim}{S}_n}(\underset{\sim}{u})d\underset{\sim}{u} - (2\pi)^{-d} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta})d\underset{\sim}{\theta} \tag{2.9.62}$$

On making the change of variables $\underset{\sim}{\theta} = (nu_1, \ldots, nu_d)$ in the first term of (2.9.62) and noting that

$\Psi_{\underset{\sim}{S}_n}(\underset{\sim}{\theta}/n) = \Psi_{\underset{\sim}{L}_n}(\underset{\sim}{\theta})$, we get

$$\Delta(d,n) = (2\pi)^{-d} \int_{-n\pi}^{n\pi} \ldots \int_{-n\pi}^{n\pi} \Psi_{\underset{\sim}{L}_n}(\underset{\sim}{\theta})d\underset{\sim}{\theta} - \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta})d\underset{\sim}{\theta} \tag{2.9.63}$$

For positive constants α and β, which will be determined later, define four integrals as follows:

(i) $$J_1 = -\int_{|\underset{\sim}{\theta}|>\alpha} \ldots \int \Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta})\ d\underset{\sim}{\theta} \tag{2.9.64}$$

(ii) $$J_2(n) = \int \underset{|\underset{\sim}{\theta}|\leq\alpha}{\cdots} \int [\Psi_{\underset{\sim}{L}_n}(\underset{\sim}{\theta}) - \Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta})]d\underset{\sim}{\theta} \qquad (2.9.65)$$

(iii) $$J_3(n) = \int \underset{\frac{\alpha}{n}\leq\left|\frac{\underset{\sim}{\theta}}{n}\right|<\beta}{\cdots} \int \Psi_{\underset{\sim}{L}_n}(\underset{\sim}{\theta})d\underset{\sim}{\theta} \qquad (2.9.66)$$

(iv) $$J_4(n) = \int \underset{\beta n\leq|\underset{\sim}{\theta}|\leq\pi n}{\cdots} \int \Psi_{\underset{\sim}{L}_n}(\underset{\sim}{\theta})d\underset{\sim}{\theta} \qquad (2.9.67)$$

It is easy to verify using these integrals and (2.9.62) that

$$\Delta(d,n) = (2\pi)^{-d} \sum_{k=1}^{4} J_k \qquad (2.9.68)$$

For appropriate choices of α and β, we will show that

$$|J_k(n)| \to 0 \text{ as } n \to \infty,\ k = 1,2,3,4,$$

which will imply (2.9.61).

Let $\epsilon > 0$ be a fixed number. Then, assumption (2.9.53) and the expression (2.9.59) imply that $\Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta})$ is absolutely integrable on R^d. Therefore, we can find a large enough α such that

$$|J_1| \leq \int \underset{|\underset{\sim}{\theta}|>\alpha}{\cdots} \int |\Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta})|d\underset{\sim}{\theta}$$

$$< \epsilon/4 \qquad (2.9.69)$$

From Lemma 2.9.5, we have

$$\underset{\sim}{L}_n \overset{a.s}{\to} \underset{\sim}{L},$$

which implies that (cf. Bhattacharya and Ranga Rao, 1976, p.44)

$$\Psi_{\underset{\sim}{L}_n}(\underset{\sim}{\theta}) \to \Psi_{\underset{\sim}{L}}(\underset{\sim}{\theta}) \text{ as } n \to \infty,$$

the convergence being uniform on the compact subset $\{\underset{\sim}{\theta}:\underset{\sim}{\theta}\in R^d \text{and } |\underset{\sim}{\theta}| \leq \alpha\}$

Hence, for the α chosen above, we can find n_1 such that $\forall\ n \geq n_1$,

$$|J_2(n)| < \varepsilon/4 \tag{2.9.70}$$

In order to show that $|J_3(n)| \to 0$, we transform θ to $\underset{\sim}{r} = \underset{\sim}{\theta}/n$ in J_3 and obtain

$$J_3(n) = n^d \underset{\frac{\alpha}{n}<|\underset{\sim}{r}|<B}{\int \dots \int} \Psi_{\underset{\sim}{S}_n}(\underset{\sim}{r})d\underset{\sim}{r} \tag{2.9.71}$$

Note that $\underset{\sim}{S}_n = \sum_{i=1}^{n} \underset{\sim}{\xi}_i$ is a lattice random vector so all its moments exist. Since $\binom{T_i}{U_i}$ are i.i.d, it follows from the definition of $\underset{\sim}{\xi}_i$ in (2.2.10) that

$$E(\underset{\sim}{S}_n) = \underset{\sim}{0} \tag{2.9.72}$$

It was argued in the proof of Lemma 2.9.5 that, for all $n \geq d$, $\underset{\sim}{\xi}_{d+1}, \dots, \underset{\sim}{\xi}_n$ are conditionally i.i.d given Λ_d with mean

$$E(\underset{\sim}{\xi}_j|\Lambda_d) = \underset{\sim}{L},\ \forall\ j = d+1, \dots, n$$

It is easy to verify that the dispersion matrices $D(\underset{\sim}{\xi}_j|\Lambda_d)$, $j = d+1, \dots, n$, are positive definite. Moreover, for $j = 1,2, \dots d$, $\underset{\sim}{\xi}_j$ is degenerate given Λ_d and

$$D(\underset{\sim}{L}) = \sigma^2 \underset{\sim}{I}, \qquad (2.9.73)$$

where $\sigma^2 = \text{var}(T-U)$ and $\underset{\sim}{I}$ is the dxd identity matrix.

The dispersion matrix of $\underset{\sim}{S}_n$ is, for $n > d$,

$$D(\underset{\sim}{S}_n) = D(\sum_{i=1}^{n} \underset{\sim}{\xi}_i)$$

$$= E(D(\sum_{i=1}^{n} \underset{\sim}{\xi}_i | \Lambda)) + D(E(\sum_{i=1}^{n} \underset{\sim}{\xi}_i | \Lambda_d))$$

$$= (n-d)\ ED(\underset{\sim}{\xi}_{d+1} | \Lambda_d) + (n-d)^2 D(\underset{\sim}{L})$$

We finally conclude that

$$D(\underset{\sim}{S}_n) - (n-d)^2 \sigma^2 \underset{\sim}{I} = (n-d)\ ED(\underset{\sim}{\xi}_{d+1} | \Lambda) \qquad (2.9.74)$$

is positive definite.

As the second-order moments of $\underset{\sim}{S}_n$ exist, we expand $\Psi_{\cdot \underset{\sim}{S}_n}(r)$ around $\underset{\sim}{r}=\underset{\sim}{0}$ and using (2.9.72) obtain

$$\log \Psi_{\underset{\sim}{S}n}(\underset{\sim}{r}) = -\frac{1}{2}\, \underset{\sim}{r}' D(\underset{\sim}{S}_n) \underset{\sim}{r} + 0(\|\underset{\sim}{r}\|^2), \text{ as } \|\underset{\sim}{r}\| \to 0 \qquad (2.9.75)$$

In view of (2.9.73), we obtain

$$|\exp(\log \Psi_{\underset{\sim}{S}_n}(\underset{\sim}{r}))| \leq \exp(-\frac{(n-d)^2}{2} \sigma^2 \|\underset{\sim}{r}\|^2 + 0\|r\|^2),$$

$$\text{as } \|r\| \to 0$$

Hence, there exists a constant $\beta > 0$ such that for $n > d$,

$$|\Psi_{\underset{\sim}{S}_n}(\underset{\sim}{r})| \le \exp(-\tfrac{1}{4}(n-d)^2\,\sigma^2\,\|\underset{\sim}{r}\|^2), \qquad \forall\ \|\underset{\sim}{r}\| \le \beta \tag{2.9.76}$$

Now, $\exists\ n_2$ such that $\forall\ \ge n_2$, $\frac{\alpha}{n} < \beta$ so that we obtain using (2.9.72) and (2.9.76)

$$|J_3(n)| \le n^d \underset{\frac{\alpha}{n} < |\underset{\sim}{r}| < \beta}{\int \cdots \int} \exp(-\tfrac{1}{4}(n-d)^2\,\sigma^2\,\|\underset{\sim}{r}\|^2)\, d\underset{\sim}{r}$$

$$\le \underset{|\underset{\sim}{\theta}| > \alpha}{\int \cdots \int} \exp(-\tfrac{1}{4}\sigma^2\,\|\underset{\sim}{r}\|^2)\, d\underset{\sim}{r} \tag{2.9.77}$$

It is clear that we can choose a large enough α in (2.9.77) such that $\forall\ n \ge n_2$,

$$|J_3(n)| < \varepsilon/4. \tag{2.9.78}$$

Finally, to show that $|J_4| \to 0$, we transform $\underset{\sim}{u} = \underset{\sim}{\theta}/n$ in (2.9.67) and obtain

$$|J_4(n)| \le n^d \int_{\beta \le |\underset{\sim}{u}| \le \pi} |\Psi_{\underset{\sim}{S}_n}(\underset{\sim}{u})|\, d\underset{\sim}{u} \tag{2.9.79}$$

In view of the earlier remarks about the conditional distributions of $\underset{\sim}{\xi}_1, \ldots, \underset{\sim}{\xi}_n$ given Λ_d, we obtain for $n \ge d$,

$$|\Psi_{\underset{\sim}{S}_n}(\underset{\sim}{u})| \le E^{\Lambda_d} |\Psi_{\underset{\sim}{\xi}_{d+1}(\underset{\sim}{w}_1,\ldots,\underset{\sim}{w}\ )}(\underset{\sim}{u})|^{\,n-d} \tag{2.9.80}$$

where $\underset{\sim}{\xi}_{d+1} = \underset{\sim}{\xi}_{d+1}(\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_d)$ is the value of $\underset{\sim}{\xi}_{d+1}$ given

$\underset{\sim}{W}_i = (T_i, U_i)$, $i = 1, 2, \ldots, d$. Since the characteristic function $\Psi_{\xi_{d+1}}(\underset{\sim}{u})$ is uniformly continuous on the compact set $\{\underset{\sim}{u}\colon \beta \le |\underset{\sim}{u}| \le \pi\}$ of R^d, it attains its maximum inside this set, say at $\underset{\sim}{u} = \underset{\sim}{u}^*$. Furthermore, $\Psi_{\xi_{d+1}}$ has period 2π so that, for almost all realizations $(\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_d)$,

$$\sup_{\beta \le |u| \le \pi} |\Psi_{\underset{\sim}{\xi}_{d+1}}(\underset{\sim}{u})| < 1 \tag{2.9.81}$$

Letting $\Psi_d^* = -\ln[\Psi_{\xi_{d+1}}(u^*)]$, we get from (2.9.79) and (2.9.80),

$$\begin{aligned} |J_4| &\le n^d\, E^{\Lambda_d}(\exp(-(n-d)\Psi_d^*) \\ &= n^d\, M_{\Psi_d^*}(n-d) \end{aligned} \tag{2.9.82}$$

where

$$M(s) = \int_0^\infty \cdots \int_0^\infty \exp(-s\Psi^*) \prod^d dC(x_j, y_j) \tag{2.9.83}$$

is the moment generating function of Ψ^* with a real positive argument.

Now, using the Abelian Theorem (cf. Widder (1941), p. 181), we obtain

$$\operatorname{Lim\,sup}_{t\to\infty} t^d\, M_{\Psi^*}(t) \le \operatorname{Lim\,sup}_{t\downarrow 0}\left[\frac{P(\Psi_d^* < t)}{t^d}\, \Gamma(d+1)\right] \tag{2.9.84}$$

By Assumption (2.9.55), the right-hand side of (2.9.84) is zero and it follows that

$$n^d M_{\Psi^*_d}(n-d) \to 0, \text{ as } n \to \infty,$$

which implies, in view of (2.9.82),

$$|J_4(n)| \to 0, \text{ as } n \to \infty, \qquad (2.9.85)$$

It follows from (2.9.69), (2.9.70), (2.9.78) and (2.9.85) that

$$\lim_{n\to\infty} |\Delta(d,n)| = 0$$

The convergence of factorial moments in (2.9.57) follows immediately, which in turn implies the Poisson convergence in (2.9.56) □

The validity of Theorem 2.9.2 depends on whether the Assumptions (2.9.53) to (2.9.55) hold or not. We shall now given some examples in order to illustrate the fact that these Assumptions are not vacuous. We start with a discussion of (2.9.53).

For any Copula C(x,y) on $[0,1]^2$, one may define ϕ^2 (possibly an infinite #) by the equation

$$\phi^2 + 1 = \int^1 \Omega^2(x,y)\, dx\, dy, \qquad (2.9.86)$$

where $\Omega(x,y) = dC(x,y)/dxdy$ is the Radon-Nikodym derivative of the Jonit distribution of $\binom{T}{U}$ with respect to the product measure of T and U (i.e., the independent case). C(x,y) is a ϕ^2-bounded distribution (with marginal <u>uniform</u> distribution) if $\phi^2 < +\infty$.

The class of ϕ^2-bounded distributions is large, as is evident from the following general result (see Lancaster 1969, page 95).

<u>Proposition 2.9.3</u>: If H(t,u) is a ϕ^2-bounded bivariate distribu-

tion with marginal distributions F(t) and G(u) then complete sets of orthonormal functions η_{1i}, η_{2i}, $i = 1,2, \ldots$, can be defined on the marginal distributions such that

$$dH(t,u) = [1 + \sum_{i=1}^{\infty} \rho_i \, \eta_{1i}(t) \, \eta_{2i}(u)] \, dF(t) \, dG(u) \qquad (2.9.87)$$

and $\phi^2 = \sum_{i=1}^{\infty} \rho_i^2$ (2.9.88)

It may be recalled from (2.6.12) that, when all $\rho_i \geq 0$ in the above canonical expansion of the joint distribution of $\underset{\sim}{T}$ and $\underset{\sim}{U}$, we say T and U are positive dependent by expansion (PDE). It follows from (2.9.87) that, when a copula C(t,u) is ϕ^2-bounded, λ in (2.9.53) can be evaluated using the orthonormality of $\{\eta_i\}$ as

$$\lambda = \int_0^1 c(x,x)dx$$

$$= 1 + \sum_{i=1}^{\infty} \rho_i \qquad (2.9.89)$$

It follows from (2.9.88) and (2.9.89) that the finiteness of ϕ^2 and λ are related to each other. Specifically, since $\forall \; i \geq 1$, the canonical correlations $\rho_i \leq 1$, we obtain

$$\lambda < \infty \Rightarrow \phi^2 < \infty$$

With regard to the Morgenstern distribution in (2.6.16), we obtain

$$\rho_i = \begin{cases} \alpha & \text{if } i=1 \\ 0 & \text{if } i>1 \end{cases}$$

where $-1\leq\alpha\leq1$. However, we have

$$\lambda = \int_0^1 c(x,x)dx$$

$$= 1 + \frac{\alpha}{3} ,$$

which is finite. Similarly, in the bivariate normal distribution given by (2.6.15),

$$\lambda = \frac{1}{1-\rho} , \; 0 \leq \rho < 1$$

In view of these examples, assumption (2.9.53) is not vacuous.

Bhattacharya and Ranga Rao (1976) (pp. 189-192), gives conditions that are equivalent to the assumption (2.9.54). We cite one here:

Let G_L^{*m} denote the n^{th} convolution of the distribution of L - T - U, where $m \geq 1$. If there exists an integer m such that G_L^{*m} has a bounded (almost everywhere) density, then the modulus of the characteristic function of L is integrable on $(-\infty,\infty)$(that is assumption (2.4.54) is valid) and vice versa.

Another <u>sufficient</u> condition for absolute integrability of $\Psi_L(\theta)$ is due to Bochner and Chandrasekar (1949). If there exists a bounded (almost everywhere density $g_L(t)$ of L = T - U and if its characteristic function $\Psi_L(\theta)$ is (real) and nonnegative, then

$$\int_{-\infty}^{\infty} |\Psi_L(\theta)|\ d\theta < \infty .$$

We illustrate the use of this sufficient (but not a necessary) condition when $\binom{T}{U}$ has the Morgenstern PDF,

$$C(x,y) = 1 + \alpha\ (1 - 2x)(1 - 2y)$$

Clearly, as $|\alpha| \leq 1$, $|x| \leq 1$, $|y| \leq 1$, $\exists$ a positive constant k such that

$$C(x,y) \leq k,\ \forall\ (x,y)\ \epsilon[0,1]^2$$

Note that

$$g_L(t) = \int_{y=0}^{1-t} z(t+y,y)dy,\ \forall\ t > 0$$

By the symmetry of C(x,y) in and, it can be shown that

$g_L(-t) = g_L(t)$, $\forall$ $t > 0$.

Now, using the bound k for C(x,y), and the fact that [-1,1] is the support of L, we get

$$g_L(t) \leq k \int_{0}^{1-t} dy \leq 2k < \infty$$

Hence, it follows that the PDF of L is (almost everywhere) bounded. We now show that $\Psi_L(\theta)$ is real and nonnegative $\forall\ \alpha \geq 0$

$$\Psi_L(\theta) = E(e^{i(T-U)\theta}) = I_1 + \alpha I_2$$

where, $I_1 = \int_0^1 \int_0^1 e^{i(x-y)\theta}\ dxdy$

$$= z_1 \bar{z}_1,$$

with $z_1 = \int_0^1 e^{ix\theta} dx$

$$I_2 = \int_0^1 \int_0^1 e^{i(x-y)\theta} (1-2x)(1-2y)dxdy$$

$$= z_2 \bar{z}_2 ,$$

with $Z_2 = \int_0^1 e^{ix\theta} (1-2x)dx$

Hence, $\Psi_L(\theta) = |Z_1(\theta)|^2 + \alpha|Z_2(\theta)|^2 \geq 0$ if $\alpha \geq 0$.

Invoking Bochner's sufficient condition, we get $\int_{-\infty}^{\infty} |\psi_L(\theta)|d\theta < \infty$, if $\alpha \geq 0$. However, for all α,

$$\int_{-\infty}^{\infty} |\psi_L(\theta)|d\theta = \int_{-\infty}^{\infty} |Z_1(\theta)|^2 d\theta + \alpha \int_{-\infty}^{\infty} |Z_2(\theta)|^2 \tag{2.9.90}$$

so that the two integrals on the right hand side must be finite when $\alpha > 0$. It follows that, even when $\alpha < 0$, $\int_{-\infty}^{\infty} |\psi_L(\theta)|d\theta < \infty$. We conclude that (2.9.54) is valid for any member of the Morgenstern family of densities. It may be remarked, in passing, that, in view of the generality of the conditions of Bhattacharya and Ranga Rao (1976) and Bochner and Chandrasekar (1949). (2.9.54) holds for many distributions of $\binom{T}{U}$.

Lastly, we discuss the validity of (2.9.55). To be specific, when d=1, one can get the bound

$$|\psi_{\xi_2(w)}(\theta)| \le 1 - p_o(1-p_o) + \sin^2(\beta/2)\ \forall\ \beta \le \theta \le \pi,\ \underset{\sim}{w} = \binom{x}{y}$$

where $p_o = p_o(\underset{\sim}{w}) = 1 - x - y + 2C(x,y)$

Therefore,

$$|J_4(n,\beta)| \le \int_0^1 \int_0^1 n\ e^{-(n-1)4\sin^2\beta[P_o(1-P_o)]}\ dxdy.$$

Thus, $J_4 \to 0$ as $n \to \infty$ if we show that $nM_{P_o(1-P_o)}(n\ ) \to 0$ as $n \to \infty$, where $M_\eta(s)$ is the Laplace transform of η. A sufficient condition for this to happen is

$$P(P_o(1-P_o) \le t) = O(t),\ \text{as } t \to 0 \tag{2.9.91}$$

Let $\delta(t)$ and $1-\delta(t)$ be the roots of the equation

$$P_o(1-P_o) = t$$

It suffices to show, as $t \to 0$,

$$P(P_o \le \delta(t)) = O(t)\ \text{and} \tag{2.9.92}$$

$$P(P_o \ge 1 - \delta(t)) = O(t) \tag{2.9.93}$$

If $\binom{T}{U}$ is independent, then the PDF of P_o can be shown to be

$$g_{P_o}(x) = -\ln(|1-2x|)I_{[0,1]}(x)$$

So that (2.9.92) and (2.9.93) are valid when $C(x,y) = C_o$ where $C_o(x,y) = xy$. Also, if $C(x,y) \geq xy$, then $P_o(C) \geq P_o(C_o)$ so that

$$P(P_o(C) \leq \delta(t)) \leq P(P_o(C_o) \leq \delta(t)) \qquad (2.9.94)$$

Thus, using the exact calculations based on the <u>independence case</u>, it follows that

$$\forall\ C \geq xy,\ P(P_o(C) \leq \delta(t)) = O(t)$$

At this time, we are optimistically speculating that, when $\binom{T}{U}$ are dependent, (2.9.93) is also true. We are yet to demonstrate that the assumption (2.9.55) is not vacuous for any $d \geq 1$.

After we derived the proof of Theorem 2.9.2, we discussed the Poisson convergence problem with Professor Persi Diaconis, who communicated the problem to Professor Charles Stein. In his Neyman lecture at the IMS Annual (1984) meeting, Professor Stein outlined an alternative proof of the Poisson convergence using his well-known theorem concerning the approximation of probabilities. However, we have not seen any rigorous version of the proof yet.

CHAPTER 3

MERGING FILES OF DATA ON SIMILAR INDIVIDUALS

Problems of statistical matching were discussed in Chapter 2, where we assumed that the two micro-data files being matched consisted of the same individuals. Moreover, the files did not have any common matching variables. In Chapter 1, practical and legal reasons were cited for these assumptions not to hold in certain situations. Suppose, then, we have two files of data that pertain to similar individuals. Allowing for some matching variables to be observed for each unit in the two files, we seek to merge the files so that inference problems relating to the variables not present in the same file can be addressed. This scenario was labeled Case III in Chapter 1. In this chapter,we shall first review the existing literature on Case III, and then briefly discuss some alternatives to matching in certain models in which the non-matching variables are conditionally independent given the values of the matching variables. Finally, we will present the results of a Monte-Carlo study carried out to evaluate certain matching procedures relevant to Case III.

3.1 Kadane's Matching Strategies for Multivariate Normal Models

Distance-based matching strategies were introduced in Section 1.5. The choice of distance measures in the matching methodology can be motivated using a model where the unobserved triplet $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$ has a multivariate normal distribution. The set-up of the two files to be merged is as follows:

File 1 comprises a random sample of size n_1 on $(\underset{\sim}{X},\underset{\sim}{Z})$, while File 2 consists of a random sample of size n_2 on $(\underset{\sim}{Y},\underset{\sim}{Z})$. Furthermore, we expect very few or no records in the two files to correspond to the same individuals. Statistically, this means that, for all practical purposes, the two random samples are themselves independent. For this reason, we shall denote the sample data as follows.

$$\begin{aligned}&\text{(Base) File 1:} && (\underset{\sim}{X}_i,\underset{\sim}{Z}_i),\ i = 1,2,\ \ldots,\ n_1\\ &\text{(Supplementary) File 2:} && (\underset{\sim}{Y}_j,\underset{\sim}{Z}_j),\ j = n_1+1,\ \ldots,\ n_1+n_2\end{aligned} \tag{3.1.1}$$

Once finished, the matching process leads to more comprehensive synthetic files, namely

$$\begin{aligned}&\text{Synthetic File 1:} && (\underset{\sim}{X}_i,\underset{\sim}{Y}_i^*,\underset{\sim}{Z}_i),\ i = 1,\ 2,\ \ldots,\ n_1\\ &\text{Synthetic File 2:} && (\underset{\sim}{X}_j^*,\underset{\sim}{Y}_j,\underset{\sim}{Z}_j),\ j = n_1+1,\ \ldots,\ n_1+n_2\end{aligned} \tag{3.1.2}$$

where, $\underset{\sim}{Y}_i^*$ is an imputed value of $\underset{\sim}{Y}$ that comes from the original File 2 and $\underset{\sim}{X}_j^*$ is an imputed value of $\underset{\sim}{X}$ that is taken from the original File 1 by means of some matching strategy. We shall now review

Kadane (1978)'s development of the matching methodology for a multivariate normal model.

Suppose that $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$ has a multivariate normal distribution with mean vector $(\underset{\sim}{\mu}_x,\underset{\sim}{\mu}_y,\underset{\sim}{\mu}_z)$ and variance-covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{xz} \\ \Sigma_{yx} & \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zx} & \Sigma_{zy} & \Sigma_{zz} \end{bmatrix} \tag{3.1.3}$$

The parameters $\Sigma_{xx},\Sigma_{xz},\Sigma_{yy},\Sigma_{yz},\Sigma_{zz}$ can all be estimated consistently using the marginal information on $(\underset{\sim}{X},\underset{\sim}{Z})$ and $(\underset{\sim}{Y},\underset{\sim}{Z})$ respectively in the two files. However, Σ_{xy} is an unidentified parameter, because the joint likelihood of the data on $(\underset{\sim}{X},\underset{\sim}{Z})$ and $(\underset{\sim}{Y},\underset{\sim}{Z})$ is free of the matrix Σ_{xy}. In fact, in the domain in which Σ_{xy} is such that the matrix $\begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$ is positive semidefinite, nothing is learned from the data about Σ_{xy}, except in a Bayesian framework, where Σ_{xy}, Σ_{xz},Σ_{yz} are, a priori, dependent. Even in this situation, the posterior distribuion of Σ_{xy} is updated only through Σ_{xz} and Σ_{yz}.

Kadane's approach to merging File 1 and File 2 consists of the following steps:

(i) Start with an imputed value of Σ_{xy} via some a priori distribution on the covariance matrix Σ, (ii) Complete Files 1 and 2 by predicting the missing data, $\underset{\sim}{X}$ or $\underset{\sim}{Y}$, using the marginal information in the files, (iii) Match these "completed" files based on a distance measure between records of the two files, (iv) Estimate parameters such as

$$\gamma = \int g(\underset{\sim}{w})\, dF(\underset{\sim}{w}) \ , \qquad (3.1.4)$$

using the synthetic file resulting from Step (iii) and repeating the Steps (ii) through (iv) many times to find the sensitivity of the estimates to the imputed value of Σ_{xy} and finally weight the results using the a priori distribution on Σ.

Some further details of the steps outlined above are as follows:

Suppose that a an imputed value of Σ_{xy} is available. Then we can assume that Σ_{xy} is known and complete the two files by means of conditional expectations. Let $\Sigma_{ab.c}$, for any letters a, b and c, be given by

$$\Sigma_{ab.c} = \Sigma_{ab} - \Sigma_{ac}\, \Sigma_{cc}^{-1}\, \Sigma_{cb}$$

Then the predicted value $\hat{\underset{\sim}{Y}}$, say, of a missing $\underset{\sim}{Y}$ in File 1 is given by

$$\begin{aligned}\hat{\underset{\sim}{Y}} &= E(\underset{\sim}{Y} | \underset{\sim}{X}, \underset{\sim}{Z}) \\ &= \underset{\sim}{\mu}_y + \Sigma_{yx.z}\, \Sigma_{xx.z}^{-1}\, (\underset{\sim}{X} - \underset{\sim}{\mu}_x) + \Sigma_{yz.x}\, \Sigma_{zz.x}^{-1}\, (\underset{\sim}{Z} - \underset{\sim}{\mu}_z), \qquad (3.1.5)\end{aligned}$$

Similarly, the predicted value, $\hat{\underset{\sim}{X}}$, say, of a missing $\underset{\sim}{X}$ in File 2 is given by

$$\begin{aligned}\hat{X} &= E(\underset{\sim}{X} | \underset{\sim}{Y}, \underset{\sim}{Z}) \\ &= \underset{\sim}{\mu}_x + \Sigma_{xy.z}\, \Sigma_{yy.z}^{-1}\, (\underset{\sim}{Y} - \underset{\sim}{\mu}_y) + \Sigma_{xy.y}\, \Sigma_{zz.y}^{-1}\, (\underset{\sim}{Z} - \underset{\sim}{\mu}_z) \qquad (3.1.6)\end{aligned}$$

Using (3.1.3), (3.1.5) and (3.1.6), it is now easy to show that $(\underset{\sim}{X}_i, \hat{\underset{\sim}{Y}}_i, \underset{\sim}{Z}_i)$ is multivariate normal with mean vector $(\underset{\sim}{\mu}_x, \underset{\sim}{\mu}_y, \underset{\sim}{\mu}_z)$ and variance-covariance matrix

$$\Omega_1 = \begin{bmatrix} \Sigma_{xx} & \Lambda_1' & \Sigma_{xz} \\ \Lambda_1 & \Lambda_3 & \Lambda_2' \\ \Sigma_{zx} & \Lambda_2 & \Sigma_{zz} \end{bmatrix} \qquad (3.1.7)$$

where $\Lambda_1 = \Sigma_{yx.z}\, \Sigma^{-1}_{xx.z}\, \Sigma_{xx} + \Sigma_{yz.x}\, \Sigma^{-1}_{zz.x}\, \Sigma_{zx}$

$\Lambda_2 = \Sigma_{zx}\, \Sigma^{-1}_{xx.z}\, \Sigma_{xy.z} + \Sigma_{zz}\, \Sigma^{-1}_{zz.x}\, \Sigma_{zy.x}$

and

$$\begin{aligned} \Lambda_3 = {} & \Sigma_{yx.z}\, \Sigma^{-1}_{xx.z}\, \Sigma_{xx}\, \Sigma^{-1}_{xx.z}\, \Sigma_{xy.z} \\ & + \Sigma_{yz.x}\, \Sigma^{-1}_{zz.x}\, \Sigma_{zz}\, \Sigma^{-1}_{zz.x}\, \Sigma_{zy.x} \\ & + 2\Sigma_{yx.z}\, \Sigma^{-1}_{xx.z}\, \Sigma_{xz}\, \Sigma^{-1}_{zz.x}\, \Sigma_{zy.x} \end{aligned}$$

Also, the vectors $(\hat{\underset{\sim}{X}}_j, \underset{\sim}{Y}_j, \underset{\sim}{Z}_j)$, $j = n_1+1, \ldots, n_1+n_2$, have a common multivariate normal distribution with mean vector $(\underset{\sim}{\mu}_x, \underset{\sim}{\mu}_y, \underset{\sim}{\mu}_z)$ and variance-covariance matrix

$$\Omega_2 = \begin{bmatrix} \Lambda_4 & \Lambda_5' & \Lambda_6' \\ \Lambda_5 & \Sigma_{yy} & \Sigma_{yz} \\ \Lambda_6 & \Sigma_{zy} & \Sigma_{zz} \end{bmatrix} \qquad (3.1.8)$$

where
$$\begin{aligned} \Lambda_4 = {} & \Sigma_{xy.z}\, \Sigma^{-1}_{yy.z}\, \Sigma_{yy}\, \Sigma^{-1}_{yy.z}\, \Sigma_{yx.z} \\ & + \Sigma_{xz.y}\, \Sigma_{zz.y}\, \Sigma_{zz}\, \Sigma^{-1}_{zz.y}\, \Sigma_{zx.y} \\ & + 2\Sigma_{xy.z}\, \Sigma^{-1}_{yy.z}\, \Sigma_{yz}\, \Sigma^{-1}_{zz.y}\, \Sigma_{zx.y} \end{aligned}$$

$$\Lambda_5 = \Sigma_{yy}\, \Sigma^{-1}_{yy.z}\, \Sigma_{yx.z} + \Sigma_{yz}\, \Sigma^{-1}_{zz.y}\, \Sigma_{zx.y}$$

and

$$\Lambda_6 = \Sigma_{zy}\, \Sigma^{-1}_{yy.z}\, \Sigma_{yx.z} + \Sigma_{zz}\, \Sigma^{-1}_{zz.y}\, \Sigma_{zx.y}$$

Note that the distributions given by (3.1.7) and (3.1.8) are singular because the predicted values $\hat{\underset{\sim}{Y}}_i$ and $\hat{\underset{\sim}{X}}_{j+n_1}$ are linear functions of the other components of the random vectors $\hat{\underset{\sim}{T}}_i = (\underset{\sim}{X}_i, \hat{\underset{\sim}{Y}}_i, \underset{\sim}{Z}_i)$ and $\hat{\underset{\sim}{U}}_j = (\hat{\underset{\sim}{X}}_{j+n_1}, \underset{\sim}{Y}_{j+n_1}, \underset{\sim}{Z}_{j+n_1})$ respectively, where $i = 1,2, \ldots, n_1$ and $j = 1,2, \ldots, n_2$. In order to describe Kadane's procedures to match the completed File 1, namely, $\hat{\underset{\sim}{T}}_1, \ldots, \hat{\underset{\sim}{T}}_{n_1}$ with the completed File 2, namely, $\hat{\underset{\sim}{U}}_1, \ldots, \hat{\underset{\sim}{U}}_{n_1}$, let us first assume, for simplicity, that $n_1=n_2=n$. Starting with n records in each file, we will compute the differences

$$\hat{\underset{\sim}{T}}_i - \hat{\underset{\sim}{U}}_j = \begin{bmatrix} \underset{\sim}{X}_i & - & \hat{\underset{\sim}{X}}_{j+n} \\ \hat{\underset{\sim}{Y}}_i & - & \underset{\sim}{Y}_{j+n} \\ \underset{\sim}{Z}_i & - & \underset{\sim}{Z}_{j+n} \end{bmatrix}, \quad 1 \leq i,\ j \leq n \tag{3.1.9}$$

in order to define a measure of dissimilarity between any pair of records, one each from the two completed files. Suppose first that, there exists a vector of constants $\underset{\sim}{\ell} = (\ell_1, \ldots, \ell_n)'$, say, and i and j such that

$$P(\underset{\sim}{\ell}'(\hat{\underset{\sim}{T}}_i - \hat{\underset{\sim}{U}}_j) = 0) = 1. \tag{3.1.10}$$

In view of the independence of the random vectors $\hat{\underset{\sim}{T}}_i$ and $\hat{\underset{\sim}{U}}_j$, it is clear that (3.1.10) cannot hold. Consequently, any of the vectors $\hat{\underset{\sim}{T}}_i - \hat{\underset{\sim}{U}}_j$ is free of any linear relationship among its components. It follows from this fact and (3.1.7) to (3.1.9) that the differences $\hat{\underset{\sim}{T}}_i - \hat{\underset{\sim}{U}}_j$, $1 \leq i, j \leq n$ are identically distributed, each with a nonsingluar multivariate normal distribution with mean $\underset{\sim}{0}$ and variance-covariance matrix $\Omega_1 + \Omega_2$. For any positive definite matrix A, a dissimilarity measure between $\hat{\underset{\sim}{T}}_i$ and $\hat{\underset{\sim}{U}}_j$ can be defined by the quadratic form

$$d_{ij}(A) = (\hat{\underset{\sim}{T}}_i - \hat{\underset{\sim}{U}}_j)'A(\hat{\underset{\sim}{T}}_i - \hat{\underset{\sim}{U}}_j). \qquad (3.1.11)$$

Also, $d_{ij}(A)$ will be referred to as the distance between the i^{th} record of File 1 and the j^{th} record of File 2. Various choices of A in (3.1.11) provide different distance measures.

It may be recalled from Section 1.5 that a constrained matching of the two files is obtained by minimizing

$$C = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} a_{ij} \qquad (3.1.12)$$

subject to the conditions

$$\sum_{j=1}^{n} a_{ij} = 1, \forall\ i = 1,2, \ldots, n \qquad (3.1.13)$$

$$\sum_{i=1}^{n} a_{ij} = 1, \forall\ j = 1,2, \ldots, n \qquad (3.1.14)$$

and

$$a_{ij} = 0 \text{ or } 1, \forall\ i \text{ and } j \qquad (3.1.15)$$

If the d_{ij}'s in (3.1.12) are given by $d_{ij}(A)$'s in (3.1.11) for some choice of A, then we obtain an optimal distance-based constrained match. Note that this type of matching of the two files amounts to solving a linear assignment problem. Sometimes, an optimal matching may be obtained by minimizing (3.1.12) without requiring that the conditions (3.1.13) and (3.1.14) hold. However, as reported in Rodgers (1984), unconstrained optimal matches do not provide good estimates of the distribution $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$. We shall not discuss such "unconstrained matchings."

It is important to note that the aforementioned optimization problem needs to be solved for each realization of the random variables involved. Suppose then that $\underset{\sim}{\hat{T}}_i$ and $\underset{\sim}{\hat{U}}_j$ have been matched in a given problem. Then it might be natural to take $(\underset{\sim}{X}_i,\underset{\sim}{Y}_j,\underset{\sim}{Z}_i)$ and $(\underset{\sim}{X}_i,\underset{\sim}{Y}_j,\underset{\sim}{Z}_j)$ as simulations of the underlying distribution. Now, the parameter γ in (3.1.4) can be estimated using one of the following synthetic samples:

$$\text{Synthetic File 1:} \quad (\underset{\sim}{X}_i,\underset{\sim}{Y}_i^*,\underset{\sim}{Z}_i),\ i = 1,2,\ \ldots,\ n. \tag{3.1.16}$$

$$\text{Synthetic File 2:} \quad (\underset{\sim}{X}_j^*,\underset{\sim}{Y}_j,\underset{\sim}{Z}_j),\ j = n+1,\ \ldots,\ 2n. \tag{3.1.17}$$

where $\underset{\sim}{Y}_i^*$ and $\underset{\sim}{X}_j^*$ are values given by the matching procedure.

Kadane has suggested that matchings based on a fixed A in (3.1.11) and the consequent inferences based on synthetic files such as (3.1.16) or (3.1.17) must be repeated many times and the results must be averaged in some sensible way in order to explore the sensitivity of our findings to the value of Σ_{xy} we started with. We shall not pursue such issues as the actual choice of a prior on Σ and the aforementioned sensitivity studies of inferences based on synthetic data. However, we shall now discuss Kadane's choices of the matrix A, which will be used in our Monte-Carlo Study of Section 3.3.

Kadane has advocated two choices for the matrix A in the definition of distance measure d_{ij}, which is given by (3.1.11):

$$\text{(i)} \quad A = (\Omega_1 + \Omega_2)^{-1}, \tag{3.1.18}$$

where Ω_1 and Ω_2 are the matrices in (3.1.7) and (3.1.8); this A leads to the so-called Mahalanobis distance between the records of the two files, and

$$\text{(ii)} \quad A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Sigma_{zz}^{-1} \end{bmatrix}, \tag{3.1.19}$$

In general, the relative benefits of these two distance measures is an open question, although the empirical studies of Barr et al. (1982) and other investigators reported in Rodgers (1984) indicate that the Mahalanobis distance is worse than the distance provided by (3.1.19) in the sense of distorting the bivariate and multivariate relationships among the variables $\underset{\sim}{X}$, $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$. In view of this, we shall follow Kadane (1978) in calling the measure induced by (3.1.19) the "bias-advoiding distance function." The special case of (3.1.19) when $\underset{\sim}{Z}$ has only one component will be discussed in the next subsection.

3.1.1 Isotonic Matching Strategy

We shall evaluate, in Section 3.3, Kadane's matching strategies in the simple case when the triple $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$ has a trivariate normal distribution. In order to facilitate such evaluations, we now show that, in the special case of a scalar $\underset{\sim}{Z}$, the matching strategy based on (3.1.19) can be implemented without using any algorithm to minimize distances.

Assuming that $\underset{\sim}{Z}$ is scalar and using (3.1.19) in the objective function given by (3.1.12), C is equivalent to

$$C = \sum_{i=1}^{n} \sum_{j=1}^{n} (Z_{1i} - Z_{2j})^2 a_{ij} \qquad (3.1.20)$$

In a constrained match, a_{ij}'s are subject to the conditions (3.1.13) to (3.1.15). Thus, (3.1.20) further simplifies to

$$C = \sum_{i=1}^{n} Z_{1i}^2 + \sum_{j=1}^{n} Z_{2j}^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} Z_{1i}Z_{2j}a_{ij}$$

Hence, the minimization of distances reduces to maximizing

$$C' = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}Z_{1i}Z_{2j} \qquad (3.1.21)$$

subject to the conditions (3.1.13) to (3.1.15) on the a_{ij}'s.

DeGroot and Goel (1976) show that, given the numbers z_{1i}'s and z_{2i}'s, the constrained maximization of C' is equivalent to maximizing $\sum_{i=1}^{n} z_{1i}z_{2\varphi(i)}$ over all permutations φ of the integers 1,2, ..., n. However, this latter extremal problem was encountered in Section 2.4 when we derived the M.L.P φ^* for certain bivariate matching problems. It follows that, with regard to Kadane's distance measure given by (3.1.19), where $\underset{\sim}{Z}$ is scalar, the optimal matching strategy is to order the Z-values in the two files separately and then match the i^{th} largest Z in File 1 with the i^{th} largest Z in File 2. This explicit solution means that, if Kadane's matrix in equation (3.1.19) is used to minimize distances between records of the two files, then the synthetic File 1 is obtained by matching the the $\underset{\sim}{X}$-concomitant of the i^{th} order-statistic among Z's in File 1 with the $\underset{\sim}{Y}$-concomitant of the i^{th} order statistic amont Z's in File 2. We shall refer to this strategy as <u>isotonic matching</u> of the two files because the matching procedure is determined by the order-statistics of the Z's in File 1 and the order-statistics of the Z's in File 2.

3.1.2 <u>Sims' Matching Strategy</u>

In the preceding subsection, it was shown that one of Kadane's matching strategies can be simplified to the point of not using any optimization algorithm in the matching procedure. Such simplification is clearly not possible when the triple $(\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$ has a multidimensional $\underset{\sim}{Z}$. The whole idea of generating very large synthetic data sets by actually minimizing a sum of distances over all potential matches seems computationally profligate. One possible alternative to distance-based strategies, which was suggested by Sims (1978), will now be outlined.

Sims has stressed the importance of exploiting the local sparseness or denseness of the sample data on the matching variables $\underset{\sim}{Z}$. A dense region of the $\underset{\sim}{Z}$-space is one within which we expect that the distributions of $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ given $\underset{\sim}{Z}$ change little. It is, at the same time, a region within which we have many observations. Sims has suggested that, within a dense region, any arbitrary matching procedure will produce results that do not distort the joint distribution of $\underset{\sim}{X}$, $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$. Regions which are <u>not</u> dense have few observations and, within them, statistical matching becomes difficult. Sims felt that in a sparse region, statistical matchings will almost certainly distort the joint distribution of $\underset{\sim}{X}$, $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$. He suggested that, in such a region, we should either not match at all or go beyond matching to more elaborate methods of generating synthetic data. However, Sims did not spell out any specific alternative to matching within sparse $\underset{\sim}{Z}$-regions.

In our Monte-Carlo Study for comparing Kadane's strategies with Sim's, which will be presented in Section 3.3, we created ten bins in the Z-space, namely $(-\infty,-1.00]$, $(-1.00,-0.75]$, $(-0.75,-0.50]$, $(-0.50,-0.25]$, $(-0.25,0.00]$, $(0.00,0.25]$, $(0.25,0.50]$, $(0.50,0.75]$, $(0.75,1.00]$, $(1.00,+\infty)$. The conditional mean of X or Y, given Z did not change much inside the eight bins which were between -1.00

and 1.00. Hence, these latter bins were considered dense bins and the two bins in the left and right tail of the distribution of Z were considered sparse bins. Within each dense bin, we *randomly* matched records of the two files, whereas the isotonic matching strategy of Subsection 3.1.1 was used in the sparse bins.

3.2 Alternatives to Statistical Matching Under Conditional Independence

Several criticisms of the matching methodology were mentioned in Section 1.6. It was observed that the formation of packets on the basis of matching variables $\underset{\sim}{Z}$ and the merging of records within each packet imply that the non-matching variables $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are conditionally independent given the values of $\underset{\sim}{Z}$. Following A. P. Dawid (1979) we shall use the notation $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} \mid \underset{\sim}{Z}$ to denote the conditional independence among the variables $\underset{\sim}{X}$, $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$.

Consider the situation in which we match the fragmentary data provided by the files in (3.1.1). It may be recalled from Section 1.2 that any statistical model for this type of matching should imply that the data in File 1 is stochastically independent of the data in File 2. Clearly, such files of data cannot be used to statistically test the validity of the implicit assumption that $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} \mid \underset{\sim}{Z}$. Furthermore, Sims (1978) has observed that matching itself for the purpose of, among others, estimating γ in (3.1.4) is unnecessary. He pointed out that, when $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} \mid \underset{\sim}{Z}$ holds, one can write

$$dF(\underset{\sim}{w}) = dF^{\underset{\sim\sim}{XZ}}(\underset{\sim}{w})\ dF^{\underset{\sim\sim}{YZ}}(\underset{\sim}{w})/dF^{\underset{\sim}{Z}}(\underset{\sim}{w}), \tag{3.2.1}$$

where $F^{\underset{\sim\sim}{XZ}}(.)$ is the marginal (with regard to $\underset{\sim}{W}$) CDF of $\underset{\sim}{X}$ and $\underset{\sim}{Z}$ and the other terms on the right-hand side of (3.2.1) are analogously defined marginal distribution functions. The two separate samples in (3.1.1) are adequate to estimate all the terms on the right-hand side of (3.2.1) by any of a number of statistical methods. In this section, we will discuss some alternatives to matching. With emphasis on estimating the covariances or correlations between $\underset{\sim}{X}$ and $\underset{\sim}{Y}$, we shall first review a histogram-type alternative which was suggested by Sims (1978).

Suppose that we form a grid in the $\underset{\sim}{W}$ space and estimate the joint density of $\underset{\sim}{W}$ by first counting the number of sample points in each cell of the $\underset{\sim}{z}$ grid. Let i index $\underset{\sim}{X}$-categories, j index $\underset{\sim}{Y}$-categories and k index $\underset{\sim}{Z}$-categories. Let n_{ijk} be the number of sample points in the $(i,j,k)^{th}$ cell and use the dot notation to define counts of sample points with regard to marginal distributions. Thus, we have

$n_{i.k}$ = the number of sample points with $\underset{\sim}{X}$ in the i^{th} category and $\underset{\sim}{Z}$ in the k^{th} category,

$n_{.jk}$ = the number of sample points with $\underset{\sim}{Y}$ in the j^{th} category and $\underset{\sim}{Z}$ in the k^{th} category,

and

$n_{..k}$ = the number of sample points with $\underset{\sim}{Z}$ in the k^{th} category.

Clearly,

$$n_{..k} = \sum_i n_{i.k} = \sum_j n_{.jk}$$

and the data in the two files given by (3.1.1) can be used to compute $n_{i.k}$, $n_{.jk}$ and $n_{..k}$ for all possible values of i, j and k. Thus, $n_{i.k}$ is obtained from File 1, $n_{.jk}$ from File 2 and $n_{..k}$ from the two files together. Finally, for a known function, g(.), say, let $g(\underset{\sim}{w}_{ijk})$ denote the value of g computed at the center, $\underset{\sim}{w}_{ijk}$ of the (i,j,k)th cell of the grid that we started with. Sims has suggested that we could estimate γ in (3.1.4) by the statistic

$$\hat{\gamma} = \sum_{i,j,k} g(\underset{\sim}{W}_{ijk}) \frac{n_{i.k}\, n_{.jk}}{n_{..k}} \tag{3.2.2}$$

With regard to $\hat{\gamma}$ in (3.2.2), theoretical properties such as the asymptotic distribution of $\hat{\gamma}$ (as the sample size tends to ∞) are unknown at the present time. Also, practical problems such as the choice of $\underset{\sim}{W}$-grid and the cells thereof, which would keep the number of terms in the sum (3.2.2) computationally reasonable, have not been studied yet.

Sims (1978) stated that a procedure like the one leading to $\hat{\gamma}$ in (3.2.2), which takes into account the implicit assumption of conditional independence of the matching methodology, had the following advantages over matching to create a synthetic file such as (3.1.16):

(a) the procedure lends itself to computation of standard errors indicating the reliability of computations based on it

(b) the procedure can be connected to the large statistical literature on estimating density functions and multi-dimensional contingency tables, and

(c) it is likely to provide more accurate results than matching.

Given the lack of work on the statistical properties of the alternatives to matching, we can agree with the advantages (a) and (b), but regard (c) as an undemonstrated speculation. We shall not discuss $\hat{\gamma}$ in (3.2.2) any further. Nor shall we elaborate the merits and demerits of alternatives to matching and synthetic-data-based procedures. Nevertheless, in the next subsection, we shall derive the estimators of parameters for conditionally independent normal models without matching the files in (3.1.1).

3.2.1 Maximum Likelihood Estimation in Multivariate Normal Models Using Two Files of Data

Consider the random vectors $\underset{\sim}{X}$, $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$, with respective dimensions p_1, p_2 and p_3. Suppose that $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$ has a nonsingular multivariate normal distribution with unknown mean vector $(\underset{\sim}{\mu}_x,\underset{\sim}{\mu}_y,\underset{\sim}{\mu}_z)$ and unknown variance-covariance matrix Σ, which is partitioned as in (3.1.3). Suppose that the sample data in (3.1.1) is available and that $n_1 \geq p_1+p_3$, $n_2 \geq p_2+p_3$. Note that, in view of the nonsingularity of distribution of $\underset{\sim}{W}$ and the fact that $\underset{\sim}{Z}_1, \ldots, \underset{\sim}{Z}_{n_1+n_2}$ are stochastically independent, the ranks of the matrices $(\underset{\sim}{Z}_1, \ldots, \underset{\sim}{Z}_{n_1})$ and $(\underset{\sim}{Z}_{n_1+1}, \ldots, \underset{\sim}{Z}_{n_2+n_2})$ are equal to P_3 for almost every realization of the Z's.

In this section, we shall find the maximum likelihood estimator of, among others, the covariances among the variables in the vectors $\underset{\sim}{X}$ and $\underset{\sim}{Y}$, without matching the files (3.1.1) but assuming that $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} | \underset{\sim}{Z}$. The maximum likelihood estimation of parameters in multivariate normal models based on various patterns of missing data has been discussed in the literature. See, for example, Eaton and Kariya (1983) Kariya et al. (1983), Anderson (1984) and Srivastava and Khatri (1979). However, the pattern of data given by the set-up (3.1.1) does not seem to have been examined. Note first that, under conditional independence, the density of $\underset{\sim}{w}$ can be written as

$$f_{\underset{\sim}{W}}(\underset{\sim}{w};\underset{\sim}{\theta}) = f_1(\underset{\sim}{z};\underset{\sim}{\theta})f_2(\underset{\sim}{x}|\underset{\sim}{z},\underset{\sim}{\theta})f_3(\underset{\sim}{y}|\underset{\sim}{z},\underset{\sim}{\theta}) \tag{3.2.3}$$

where $$\underset{\sim}{\theta} = (\underset{\sim}{\mu}_x,\underset{\sim}{\mu}_y,\underset{\sim}{\mu}_z,\Sigma_{xx},\Sigma_{xy},\Sigma_{xz},\Sigma_{yy},\Sigma_{zz}) \tag{3.2.4}$$

and $f_{\underset{\sim}{W}}(\underset{\sim}{w})$ is the joint density of $\underset{\sim}{W}$ given by

$$f_{\underset{\sim}{W}}(\underset{\sim}{w}) = (2\pi)^{-(p_1+p_2+p_3)/2}|\Sigma|^{-½}$$

$$\text{x } \mathrm{etr}[-\frac{1}{2}\Sigma^{-1}(\underset{\sim}{w} - \underset{\sim}{\mu})(\underset{\sim}{w} - \underset{\sim}{\mu})'] \text{ ,} \tag{3.2.5}$$

etr being the exponential of the trace of a matrix. Also, $f_1(.)$ is the marginal density functon of $\underset{\sim}{Z}$, $f_2(.)$ and $f_3(.)$ are respectively the conditional densities of $\underset{\sim}{X}$ and $\underset{\sim}{Y}$, given $\underset{\sim}{Z} = \underset{\sim}{z}$. It is well-known (Anderson, 1984, p. 33 and 37) that f_1, f_2 and f_3 also correspond to certain multivariate normal densities like (3.2.5). Using the joint normality of $\underset{\sim}{X}$, $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$, it is easy to verify that (3.2.3) holds iff

$$\Sigma_{xy} = \Sigma_{xz}\,\Sigma_{zz}^{-1}\,\Sigma_{zy} \tag{3.2.6}$$

It follows from (3.2.3) that the likelihood of the observed data in the two files given by (3.1.1) is

$$L(\underset{\sim}{\theta}) = L_1(\underset{\sim}{\theta})L_2(\underset{\sim}{\theta})L_3(\underset{\sim}{\theta}) \ , \tag{3.2.7}$$

where
$$L_1(\underset{\sim}{\theta}) = \prod_{i=1}^{n_1+n_2} f_1(\underset{\sim}{z}_i,\underset{\sim}{\theta}) \tag{3.2.8}$$

$$L_2(\underset{\sim}{\theta}) = \prod_{i=1}^{n_1} f_2(\underset{\sim}{x}_i|\underset{\sim}{z}_i,\underset{\sim}{\theta}) \tag{3.2.9}$$

and

$$L_3(\underset{\sim}{\theta}) = \prod_{i=n_1+1}^{n_1+n_2} f_3(\underset{\sim}{y}_i|\underset{\sim}{z}_i,\underset{\sim}{\theta}) \tag{3.2.10}$$

Taking natural logarithms of both sides of the equation (3.2.7), we obtain

$$\ell(\underset{\sim}{\theta}) = \sum_{\alpha=1}^{3} \ell_\alpha(\underset{\sim}{\theta}) \ , \tag{3.2.11}$$

where $\ell_\alpha(\theta) = \log_e(L_\alpha(\theta))$, $\forall\ \alpha = 1,2,3$

Let $\bar{\underset{\sim}{z}}$ and s_z denote respectively the mean and the matrix of corrected sums of squares and products of the data $\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_{n_1+n_2}$. That is,

$$\bar{\underset{\sim}{z}} = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} \underset{\sim}{z}_i \tag{3.2.12}$$

$$s_z = \sum_{i=1}^{n_1+n_2} (\underset{\sim}{z}_i - \bar{\underset{\sim}{z}})(\underset{\sim}{z}_i - \bar{\underset{\sim}{z}})'$$

Similarly, let $\bar{\underset{\sim}{z}}_1$ $(\bar{z}_2)$ and $s_1(s_2)$ be the mean and the matrix of corrected sums of squares and products of the data $\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_{n_1}$ $(\underset{\sim}{z}_{n_1+1}, \ldots, \underset{\sim}{z}_{n_1+n_2})$. Let, for any lower-case a, b and c, and any vector $\underset{\sim}{z}$,

$$\underset{\sim}{\mu}_{a.b}(\underset{\sim}{z}) = \underset{\sim}{\mu}_a + \Sigma_{ab}\, \Sigma_{bb}^{-1}\, (\underset{\sim}{z} - \underset{\sim}{\mu}_z)$$

$$\Sigma_{ab.c} = \Sigma_{ab} - \Sigma_{ac}\, \Sigma_{cc}^{-1}\, \Sigma_{cb} \qquad (3.2.13)$$

Then using the notations in (3.2.12) and (3.2.13), the equations (3.2.5), (3.2.7) to (3.2.10) and Theorem 2.5.1 of Anderson (1984) (for the expressions defining f_2 and f_3) we obtain

$$\ell_1(\underset{\sim}{\theta}) = -\frac{n_1+n_2}{2} \log|\Sigma_{zz}| + \mathrm{tr}\{-\frac{1}{2}\Sigma_{zz}^{-1}[s_z + (n_1+n_2)(\bar{\underset{\sim}{z}} - \underset{\sim}{\mu}_z)(\bar{\underset{\sim}{z}} - \underset{\sim}{\mu}_z)']\} \qquad (3.2.14)$$

$$\ell_2(\underset{\sim}{\theta}) = -\frac{n_1}{2} \log|\Sigma_{xx.z}| + \mathrm{tr}\{-\frac{1}{2}\Sigma_{xx.z}^{-1}[\sum_{i=1}^{n_1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_{x.z}(\underset{\sim}{z}_i))(\underset{\sim}{x}_i - \underset{\sim}{\mu}_{x.z}(\underset{\sim}{z}_i))]\} \qquad (3.2.15)$$

and

$$\ell_3(\underset{\sim}{\theta}) = -\frac{n_2}{2} \log|\Sigma_{yy.z}|$$

$$+ \operatorname{tr}\{- \frac{1}{2} \Sigma^{-1}_{yy.z}[\sum_{j=n_1+1}^{n_1+n_2} (\underset{\sim}{y}_j - \underset{\sim}{\mu}_{y.z}(\underset{\sim}{z}_j))(\underset{\sim}{y}_j - \underset{\sim}{\mu}_{y.z}(\underset{\sim}{z}_j))']\}$$

(3.2.16)

Note that in (3.2.14) to (3.2.16), certain constant terms have been omitted.

It is clear from (3.2.7) and (3.2.11) that the M.L.E of $\underset{\sim}{\theta}$ is obtained by maximizing $\ell_\alpha(\underset{\sim}{\theta})$ over $\underset{\sim}{\theta}$ for each $\alpha = 1,2,3$ separately. Moreover, this maximization is easier if we reparametrize the distribution of $\underset{\sim}{W}$ by means of

$$\underset{\sim}{\eta} = (\underset{\sim}{\mu}_z, \Sigma_{zz}, \underset{\sim}{\nu}_{xy}, \underset{\sim}{\nu}_{yz}, \Sigma_{xx.z}, \Sigma_{yy.z}, B_{xy}, B_{yz}), \qquad (3.2.17)$$

where, apart from the notations that we have already introduced, we have, for any letters a and b

$$B_{ab} = \Sigma_{ab} \Sigma^{-1}_{bb}$$

and (3.2.18)

$$\underset{\sim}{\nu}_{ab} = \underset{\sim}{\mu}_a - B_{ab} \underset{\sim}{\mu}_b$$

It can be easily shown that there is a one-to-one correspondence between $\underset{\sim}{\theta}$ and $\underset{\sim}{\eta}$. Consequently, if we rewrite $\ell_\alpha(\underset{\sim}{\theta})$'s in terms of $\underset{\sim}{\eta}$, then maximizing $L(\underset{\sim}{\theta})$ over $\underset{\sim}{\theta}$ is equivalent to maximizing $\ell_\alpha(\underset{\sim}{\eta})$ over $\underset{\sim}{\eta}$, for each $\alpha = 1,2,3$. The advantage of the transformation to the $\underset{\sim}{\eta}$-space is that $\ell_\alpha(\eta)$'s are functions of disjoint portions of $\underset{\sim}{\eta}$. In fact, $\ell_1(\underset{\sim}{\eta})$ is the same as $\ell_1(\underset{\sim}{\theta})$, whereas it follows from (3.2.15) to (3.2.16) that

$$\ell_2(\underset{\sim}{\eta}) = -\frac{n_1}{2}\log|\Sigma_{xx.z}|$$

$$+ \mathrm{tr}\{-\frac{1}{2}\Sigma^{-1}_{xx.z}[\sum_{i=1}^{n_1}(\underset{\sim}{x}_i - \underset{\sim}{\nu}_{xz} - B_{xz}\,\underset{\sim}{z}_i)(\underset{\sim}{x}_i - \underset{\sim}{\nu}_{xz} - B_{xz}\,\underset{\sim}{z}_i)']\}$$

(3.2.19)

and

$$\ell_3(\underset{\sim}{\eta}) = -\frac{n_2}{2}\log|\Sigma_{yy.z}|$$
$$+ \mathrm{tr}\{-\frac{1}{2}\Sigma^{-1}_{yy.z}[\sum_{j=n_1+1}^{n_1+n_2}(\underset{\sim}{y}_j - \underset{\sim}{\nu}_{yz} - B_{yz}\,\underset{\sim}{z}_j)(\underset{\sim}{y}_j - \nu_{yz} - B_{yz}\,\underset{\sim}{z}_j)']\}$$

(3.2.20)

In view of Theorem 8.2.1 of Anderson (1984), it can be easily shown using (3.2.14), (3.2.19) and (3.2.20) that M.L.E of $\underset{\sim}{\eta}$ is given by

$$\hat{\underset{\sim}{\mu}}_z = \underset{\sim}{\bar{Z}}$$

$$\hat{\Sigma}_{zz} = \frac{S_z}{n_1+n_2}$$

$$\hat{B}_{xy} = [\sum_{i=1}^{n_1}(\underset{\sim}{X}_i - \underset{\sim}{\bar{X}})(\underset{\sim}{Z}_i - \underset{\sim}{\bar{Z}}_1)']S_1^{-1}$$

$$\hat{\underset{\sim}{\nu}}_{xy} = \underset{\sim}{\bar{X}} - \hat{B}_{xz}\,\underset{\sim}{\bar{Z}}_1$$

$$\hat{B}_{yz} = [\sum_{j=n_1+1}^{n_1+n_2}(\underset{\sim}{Y}_j - \bar{Y})(\underset{\sim}{Z}_j - \underset{\sim}{\bar{Z}}_2)']S_2^{-1} \qquad (3.2.21)$$

$$\hat{\underset{\sim}{\nu}}_{yz} = \bar{\underset{\sim}{Y}} - \bar{B}_{yz} \bar{\underset{\sim}{Z}}_2$$

$$\hat{\Sigma}_{xx.z} = \frac{1}{n_1} \sum_{i=1}^{n_1} (\underset{\sim}{X}_i - \hat{\underset{\sim}{\nu}}_{xz} - \hat{B}_{xz} \underset{\sim}{Z}_i)(\underset{\sim}{X}_i - \hat{\underset{\sim}{\nu}}_{xz} - \hat{B}_{xz} \underset{\sim}{Z}_i)'$$

$$\hat{\Sigma}_{yy.z} = \frac{1}{n_2} \sum_{j=n_1+1}^{n_1+n_2} (\underset{\sim}{Y}_j - \hat{\underset{\sim}{\nu}}_{yz} - \hat{B}_{yz} \underset{\sim}{Z}_j)(\underset{\sim}{Y}_j - \hat{\underset{\sim}{\nu}}_{yz} - \hat{B}_{yz} \underset{\sim}{Z}_j)'$$

Using these estimators and the relationships between $\underset{\sim}{\theta}$ and $\underset{\sim}{\eta}$ we obtain the M.L.E of $\underset{\sim}{\theta}$ by means of the following equations.

$$\hat{\underset{\sim}{\mu}}_x = \hat{\underset{\sim}{\nu}}_{xz} + \hat{B}_{xz} \hat{\underset{\sim}{\mu}}_z$$

$$\hat{\underset{\sim}{\mu}}_y = \hat{\underset{\sim}{\nu}}_{yz} + \hat{B}_{yz} \hat{\underset{\sim}{\mu}}_z$$

$$\hat{\underset{\sim}{\mu}}_z = \bar{Z}$$

$$\hat{\Sigma}_{xx} = \hat{B}_{xz} \hat{\Sigma}_{zz} \hat{B}'_{xz} + \hat{\Sigma}_{xx.z} \qquad (3.2.22)$$

$$\hat{\Sigma}_{xz} = \hat{B}_{xz} \hat{\Sigma}_{zz}$$

$$\hat{\Sigma}_{yy} = \hat{B}_{yz} \hat{\Sigma}_{zz} \hat{B}_{yz} + \hat{\Sigma}_{yy.z}$$

$$\hat{\Sigma}_{yz} = \hat{B}_{yz} \hat{\Sigma}_{zz}$$

and $\hat{\Sigma}_{xy} = \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zy}$

It follows from the above discussion that if we can justify the assumption that $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} \mid \underset{\sim}{Z}$, then we can avoid matching the files in (3.1.1) and estimate, among other parameters, Σ_{xy}, by means of the equations in (3.2.22). Unfortunately, the two data files contain no information regarding the appropriateness of this assumption, and prior information from other sources must be considered. The point here is that, if the matching methodology is based on assumptions like $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} \mid \underset{\sim}{Z}$, then we must look for alternatives to matching whose statistical properties are known. Such alternatives are useful especially because very little is known about the reliatility of synthetic data-files.

It is important to note that (3.2.6) is a necessary condition even if $\underset{\sim}{W}$ is not normal, provided only that $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} \mid \underset{\sim}{Z}$ holds and that the appropriate moments of the distribution of $\underset{\sim}{W}$ exist. Hence, we can use the estimator $\hat{\Sigma}_{xy}$ in (3.2.22) even for non-normal populations. We now show that $\hat{\Sigma}_{xy}$ is consistent for $\hat{\Sigma}_{xy}$ without assuming that $\underset{\sim}{W}$ has a multi-variate normal distribution.

<u>Theorem 3.2.1</u> Suppose the joint distribution of $\underset{\sim}{W}$ is such that its second-order moments exist and that the dispersion matrix, Σ, of $\underset{\sim}{W}$ is partitioned as in (3.1.3). If $\underset{\sim}{X} \perp\!\!\!\perp \underset{\sim}{Y} \mid \underset{\sim}{Z}$ then $\hat{\Sigma}_{xy}$, given by (3.2.22), is strongly consistent for Σ_{xy}.

<u>Proof</u>: We first note that $\hat{\Sigma}_{xz}$ and $\hat{\Sigma}_{zy}$ are stochastically independent because they are functions of the independent data in File 1 and File 2 respectively. However, $\hat{\Sigma}_{zz}$ involves Z_i's in both files so that the elements of the vector

$$(\hat{\Sigma}_{xz},\hat{\Sigma}_{zz},\hat{\Sigma}_{zy}) \tag{3.2.23}$$

are dependent. The almost sure convergence of the vector in (3.2.23) will follow from the almost sure convergence of $\hat{\Sigma}_{xz},\hat{\Sigma}_{zz},\hat{\Sigma}_{zy}$ individually (cf. Serfling, 1980, p. 52). In view of the similarities of the proofs of the convergence of these matrices, we shall only show that, as $n_\alpha \to \infty$, $\alpha = 1,2$,

$$\hat{\Sigma}_{zz} \overset{a.s}{\to} \Sigma_{zz} \tag{3.2.24}$$

We obtain from (3.2.21),

$$\hat{\Sigma}_{zz} = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} \underset{\sim}{Z}_i \underset{\sim}{Z}_i' - \bar{\underset{\sim}{Z}}\, \bar{\underset{\sim}{Z}}' \tag{3.2.25}$$

Recalling our assumption that the files in (3.1.1) are independent random samples and that the vector $\underset{\sim}{Z}$ has a finite dispersion matrix, it readily follows that the Strong Law of large numbers (cf. Serfling, p. 27) applies to independent sequences $\{\underset{\sim}{Z}_i\}$ and $\{\underset{\sim}{Z}_i\underset{\sim}{Z}_i'\}$. Hence, we obtain, as $n_\alpha \to \infty$

$$\frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} \underset{\sim}{Z}_i \underset{\sim}{Z}_i' \overset{a.s}{\to} E(\underset{\sim}{Z}\, \underset{\sim}{Z}') \tag{3.2.26}$$

and

$$\bar{\underset{\sim}{Z}} \overset{a.s}{\to} E(\underset{\sim}{Z}) \tag{3.2.27}$$

It follows from (3.2.25) to (3.2.27) that

$$\hat{\Sigma}_{zz} \overset{a.s}{\rightarrow} \Sigma_{zz}$$

We conclude from our remarks earlier in this proof that, $n_\alpha \rightarrow \infty$

$$(\hat{\Sigma}_{xz},\hat{\Sigma}_{zz},\hat{\Sigma}_{zy}) \overset{a.s}{\rightarrow} (\Sigma_{xz},\Sigma_{zz},\Sigma_{zy}) \qquad (3.2.28)$$

Let us now observe that

$$\hat{\Sigma}_{xy} = \hat{\Sigma}_{xz}\,\hat{\Sigma}_{zz}^{-1}\,\hat{\Sigma}_{zy}$$

is a continuous function of the random variables in the vector (3.2.23). Hence, the strong consistency of $\hat{\Sigma}_{xy}$ follows from (3.2.28).

□

3.3 An Empirical Evaluation of Certain Matching Strategies

Several distance-based matching strategies for creating synthetic data have been discussed in Section 3.1. Specifically, two strategies due to Kadane (1978) and a strategy which was proposed by Sims (1978) were mentioned. In this section, we shall evaluate these three strategies, individually as well as in relative terms, in the special case where $\underset{\sim}{W} = (\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{Z})$, the unobservable vector, has a trivariate normal distribution. Before we discuss the Monte-Carlo Study of the aforementioned strategies, we shall review some of the earlier simulation studies of statistical matching procedures, which have certain bearing on our study. A more comprehensive review of evaluations of statistical matching procedures can be found in Rodgers (1984).

Barr et al. (1982) used, among others, a statistical model in which a vector $\underset{\sim}{W} = (X,Y,Z_1,Z_2)$ had a four-dimensional normal distribution with zero means, unit variances and various levels of covariances among the four variables. Altogether, these investigators generated 100 pairs of independent files, namely File 1 comprising 200 observations on (X,Z_1,Z_2) and File 2 consisting of 200 observations on Y, Z_1 and Z_2, for each of 12 populations, where the populations differed with respect to the covariances of the variables. Then, for each such pair of files, six statistical matches were performed, namely three constrained matches and three unconstrained matches. In each of these six matches, they used three distance functions for each type of match. The first was a weighted sum of the absolute differences of the two Z variables between records of the two files and the last two were the Mahalanobis-distance and the "bias-avoiding" distance, which were discussed in Section 3.1. A summary of the findings of Barr et al. is as follows.

All three distance measures provided accurate estimates of the variance of the Y variable when the constrained matching procedure was used. They also found that all three unconstrained matching procedures produced Y distributions that had means which were significantly different from the corresponding population values. The estimated covariances of Y with Z_1,Z_2, which were computed only for constrained matches, tended to be underestimated. With respect to the most important question in the context of merging files, namely the estimation of relationships between X and Y variables, it was reported that, if the conditional independence assumption was invalid, all statistical matching procedures provided estimates of the X-Y covariance that were extremely poor. On the other hand, for the cases in which the conditional independence assumption was valid, all six procedures provided estimates of the X-Y covariance that were generally quite accurate. Their simulations also indicated that the Mahalanobis distance measure produced less accurate matching than subjectively weighted distance measures.

As we mentioned earlier, our own Monte-Carlo study was confined to a trivariate normal model. However, our findings were sufficiently interesting to justify their inclusion in this thesis. In fact, some new facts about Kadane's bias-avoiding matching strategy have already been mentioned in Section 3.1. Suppose, then, that $\underset{\sim}{W} = (X,Y,Z)$ is tri-variate normal with zero means and variance-covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{xy} & 1 & \rho_{yz} \\ \rho_{xz} & \rho_{yz} & 1 \end{bmatrix} \tag{3.3.1}$$

Assume further that the following data is available for the purpose of estimating the three unknown correlations in (3.3.1):

$$\text{File 1:} \quad (X_i, Z_i),\ i = 1, 2,\ \ldots,\ n \tag{3.3.2}$$

$$\text{File 2:} \quad (Y_j, Z_j),\ j = n+1,\ \ldots,\ 2n \tag{3.3.3}$$

In view of the discussions in Section 3.2, if the conditional independence assumption $X \perp\!\!\!\perp Y \mid Z$ or, equivalently,

$$\rho_{xy} = \rho_{xz}\, \rho_{yz} \tag{3.3.4}$$

were true, then we can avoid merging the files in (3.3.2) and (3.3.3) because File 1 and File 2 can be used to get the sample correlations $\hat{\rho}_{xz}$ and $\hat{\rho}_{yz}$, which in turn provide the maximum likelihood estimator of ρ_{xy}, namely

$$\hat{\rho}_{xy} = \hat{\rho}_{xz}\, \hat{\rho}_{yz} \tag{3.3.5}$$

We shall say X and Y are conditionally dependent, given Z, iff (3.3.4) does not hold; that is

$$\rho_{xy} \neq \rho_{xz}\, \rho_{yz}$$

For the sake of simplicity, we shall consider hereinafter only the conditional positive dependence case of the model in (3.3.1), namely

$$\rho_{xy} > \rho_{xz}\, \rho_{yz} \tag{3.3.6}$$

The complementary case of conditional negative dependence, namely

$$\rho_{xy} < \rho_{xz}\, \rho_{yz}$$

can, however, be handled by methods similar to ours. We shall also include the case when $X \perp\!\!\!\perp Y \mid Z$ holds mainly for comparing and contrasting our results for the positive dependence case. Finally, we shall evaluate matching strategies only from the point of view of estimating ρ_{xy}, the correlation between variables which are <u>not</u> in the same file, because File 1 and File 2 can respectively be used to estimate the remaining parameters ρ_{xz} and ρ_{yz}.

It is clear that, if the condition $X \perp\!\!\!\perp Y \mid Z$ does not hold, then we should not estimate ρ_{xy} by means of (3.3.5). In such a case, matching the files (3.3.2) and (3.3.3) for estimation purposes is an alternative that we shall study in this section. Thus, if after merging, File 1 becomes the synthetic File 1 namely

$$(X_i, Y_i^*, Z_i),\ i = 1,2,\ \ldots,\ n \tag{3.3.7}$$

where Y_i^* is the value of Y assigned to the i^{th} record in the process of merging, then we shall use the synthetic data (X_i, Y_i^*), $i = 1,2,\ \ldots,\ n$ to estimate ρ_{xy}.

It was mentioned in Section 1.7 that performance characteristics, which can help us assess the reliability of synthetic data generated by independent files in (3.3.2), are not known. Given this paucity, our program for an empirical evaluation of matching strategies is as follows

(i) Starting with a <u>known</u> correlation matrix given by (3.3.1), generate data from the normal population of $\underset{\sim}{W} = (X,Y,Z)$ and create independent files (3.3.2) and (3.3.3). Note that data on (X,Y), which is typically missing in actual matching situations, is available in simulation studies.

(ii) Using any given matching strategy, merge the two files created in Step (i) and compute the "synthetic correlation", denoted by $\hat{\rho}_s$, which is defined to be the sample correlation coefficient based on the (X,Y*) data given by the synthetic file (3.3.7)

(iii) Compare $\hat{\rho}_s$ of Step (ii) with the following sample correlations:

(a) $\hat{\rho}_{m\ell 1}$, the sample correlation coefficient based on the unbroken data (X_i,Y_i), i = 1,2, ..., n which was generated in Step (i). Observe that, if there is no apriori restriction on the model parameters in (3.3.1), then $\hat{\rho}_{m\ell 1}$ is the maximum likelihood estimator of ρ_{xy}.

(b) $\hat{\rho}_{m\ell 2}$, the estimator of ρ_{xy} given by (3.3.5), which is also the maximum likelihood estimator of ρ_{xy} when conditional independence holds.

Because $\hat{\rho}_{m\ell 1}$ and $\hat{\rho}_{m\ell 2}$ are respectively based on one sample on (X,Y) and two independent samples on (X,Z) and (Y,Z), we shall also refer to these as one-sample and two-sample estimates of ρ_{xy}.

Using the aforementioned program, we shall evaluate Kadane's distance-based matching strategies discussed in Section 3.1, namely the isotonic matching strategy and the procedure induced by the Mahalanobis distance, and the method of matching in bins, which, as explained in Subsection 3.1.2, is an adaptation of a strategy due to

Sims (1978). The synthetic correlations resulting from the use of these three strategies will be denoted by $\hat{\rho}_{s1}$, $\hat{\rho}_{s2}$ and $\hat{\rho}_{s3}$ respectively.

Our study has been conducted for three values of n, namely 10, 25 and 50. The values of the population correlation ρ_{xy} which are used, among others, to generate random deviates from the normal population of $\underset{\sim}{W} = (X,Y,Z)$, were chosen from the following categories:

$$\begin{aligned} &\text{Low } \rho_{xy}\text{: } 0.00, 0.25 \\ &\text{Medium } \rho_{xy}\text{: } 0.50, 0.60, 0.65, 0.70 \\ &\text{High } \rho_{xy}\text{: } 0.75\ (0.05)\ 0.95, 0.99 \end{aligned} \tag{3.3.8}$$

Combined with low as well as high values of ρ_{xz} and ρ_{yz}, there were 15 choices of ρ_{xy} from (3.3.8) such that the conditional independence restriction (3.3.5) was satisfied. As remarked earlier, these correlations were chosen mainly to provide a basis such that the estimates of ρ_{xy} resulting from the case of conditional positive dependence can be compared with those resulting from conditional independence. The fifteen values of ρ_{xy} in the conditional independence case were increased in such a way that the positive dependence was achieved. Altogether, nineteen such Σ's were selected.

For n=10, $\underset{\sim}{W}$ was generated 1000 times by using the IMSL subroutines. The calculation of $\hat{\rho}_{S1}$ was based on sorting Z's in the two files, as discussed in Section 3.1.1. Furthermore, $\hat{\rho}_{s2}$ was computed for each realization by solving a linear assignment problem.

The Ford-Fulkerson algorithm (Zionts, 1974) was used for this purpose. The computational cost for solving assignment problems grew quite rapidly with n. Therefore, only 700 independent samples of size n=25 were generated. A comprehensive examination of the results for n=10,25, revealed $\hat{\rho}_{s1}$ and $\hat{\rho}_{s2}$, the correlations corresponding to Kadane's two distance measures, were, for all practical purposes, identical (see Figures 3.1 and 3.2). In view of this and the high computational costs, we compared only two strategies, the isotonic and the method of matching in bins for n=50 (2500 independent samples).

Four summary statistics, namely the mean, the standard deviation, the minimum and the maximum for the simulated data on $\hat{\rho}_{m\ell 1}, \hat{\rho}_{m\ell 2}, \hat{\rho}_{s1}, \hat{\rho}_{s2}, \hat{\rho}_{s3}$ were calculated for 34 Σ's selected for the study. However, we provide these statistics only for a representative collection of 15 Σ's in tables 3.1 to 3.7. For each Σ and for any $\hat{\rho}$, the first entry in the tables is the mean, the second entry (in parentheses) is the standard deviation and the third and the fourth entries are respectively the minimum and the maximum. Also, the General Plotting Package at The Ohio State University was used to plot the following pairs of estimates of ρ_{xy}

(i) $\hat{\rho}_{s1}$ vs. $\hat{\rho}_{s2}$

(ii) $\hat{\rho}_{s1}$ vs. $\hat{\rho}_{s3}$

(iii) $\hat{\rho}_{s1}$ vs. $\hat{\rho}_{m\ell 1}$

(iv) $\hat{\rho}_{s1}$ vs. $\hat{\rho}_{m\ell 2}$

(v) $\hat{\rho}_{s2}$ vs. $\hat{\rho}_{m\ell 1}$

(vi) $\hat{\rho}_{s2}$ vs. $\hat{\rho}_{m\ell 2}$

(vii) $\hat{\rho}_{s3}$ vs. $\hat{\rho}_{m\ell 1}$

(viii) $\hat{\rho}_{s3}$ vs. $\hat{\rho}_{m\ell 2}$

Figures 3.1 to 3.20 provide an illustration of these comparisons.

3.3.1 Conclusions of the Monte Carlo Study

Tables 3.1 to 3.4 clearly show that the two estimates $\hat{\rho}_{s1}$ and $\hat{\rho}_{s2}$, provided by the isotonic matching strategy and the Mahalanobis-distance based strategy, respectively have nearly identical summary statistics. In fact, an examination of all the results showed that, for all values of n and Σ in our study, the estimates $\hat{\rho}_{s1}$ and $\hat{\rho}_{s2}$ were the same for most of the realizations of $\underset{\sim}{W}$. Figures 3.1 and 3.2 provide the empirical evidence of this fact.

Now we shall discuss our results in the case of conditional independence. As noted in Section 3.2, $\hat{\rho}_{m\ell 2}$ is the maximum likelihood estimator of ρ_{xy} under this model, whereas $\hat{\rho}_{m\ell 1}$, the method of moments estimator based on paired-data, is computed for comparison purposes. As expected, $\hat{\rho}_{m\ell 1}$ and $\hat{\rho}_{m\ell 2}$ behave equally well on the average even though the estimated standard error of $\hat{\rho}_{m\ell 1}$ is consistently higher than that of $\hat{\rho}_{m\ell 2}$. Furthermore the ranges of $\rho_{m\ell 1}$

are consistently larger than those of $\hat{\rho}_{m\ell 2}$ (see Tables 3.1, 3.3 and 3.5).

For low correlation and each n, $\hat{\rho}_{s1}$, $\hat{\rho}_{s2}$ and $\hat{\rho}_{s3}$ compare well with the estimates $\hat{\rho}_{m\ell 1}$, or $\hat{\rho}_{m\ell 2}$ as far as the averages are concerned (see Tables 3.1, 3.3 and 3.5). However, the synthetic data estimators have larger variation than $\hat{\rho}_{m\ell 2}$, as shown in Fig. 3.3 - Fig. 3.5. Furthermore, all the synthetic data estimators have variation comparable to that of $\rho_{m\ell 1}$ as shown in Fig. 3.6 - Fig. 3.8.

For medium and high values of ρ_{xy}, all three synthetic estimators exhibit some amount of negative bias with regard to both $\hat{\rho}_{m\ell 1}$ and $\hat{\rho}_{m\ell 2}$. Also, ρ_{s3}, the estimator given by the method of matching in bins, is more negatively biased than $\hat{\rho}_{s1}$ and $\hat{\rho}_{s2}$. Tables 3.1, 3.3 and 3.5, Fig. 3.9 - Fig. 3.14 illustrate these points. Again, $\hat{\rho}_{s3}$ is worse than $\hat{\rho}_{s1}$ and $\hat{\rho}_{s2}$. These patterns among the five estimates exist for any sample size even though the difference between synthetic data estimators and $\hat{\rho}_{m\ell 2}$ tends to decrease as n increases.

Turning to the conditional positive dependence case, we first note that $\hat{\rho}_{m\ell 1}$ is a reasonable estimator of ρ_{xy}, even though it would not be available to the practitioner. On comparing $\hat{\rho}_{m\ell 1}$ with the synthetic data estimators $\hat{\rho}_{s1}$, $\hat{\rho}_{s2}$, and $\hat{\rho}_{s3}$ and $\hat{\rho}_{m\ell 2}$, we find that these estimators perform very badly, in that all of them are consistently underestimates and therefore heavily negatively biased (See Tables 3.2, 3.4, 3.6 and 3.7 and Fig. 3.15).

For each n, and low or medium choices of ρ_{xy}, the synthetic data estimators are comparable to $\hat{\rho}_{m\ell 2}$, whereas for high values of ρ_{xy}, the three synthetic data estimators have a definite negative bias compared with $\hat{\rho}_{m\ell 2}$. Tables 3.2, 3.4, 3.6 and 3.7 and Fig. 3.16 - Fig. 3.19 support this conclusion. Furthermore it is observed that $\hat{\rho}_{s3}$ based on binning is worse than $\hat{\rho}_{s1}$ ($\hat{\rho}_{s2}$) as illustrated by Fig. 3.20. However, the difference between the average $\hat{\rho}_{m\ell 2}$ and $\hat{\rho}_{si}$, i = 1,2,3 tends to decrease as n increases.

Finally it must be pointed out that as the positive dependence increases; ie,$\rho_{xy}-\rho_{xz}\rho_{yz}$ increases, the bias in the three synthetic data estimators and $\rho_{m\ell 2}$ increases. Tables 3.4 and 3.7 illustrate this fact.

Based on these observations, we must conclude that when conditional independence model holds, the synthetic data estimators do not provide any advantage over $\hat{\rho}_{m\ell 2}$, the no-matching estimator. In fact, they are slightly worse than the $\hat{\rho}_{m\ell 2}$. On the other hand, in the case of conditional positive dependence, $\hat{\rho}_{m\ell 2}$ and all the synthetic data estimators perform badly, the performance of synthetic data estimators being slightly worse than that of $\hat{\rho}_{m\ell 2}$. Thus estimators based on matching strategies do not seem to provide any advantage over the estimators based on the assumption of conditional independence and no matching. Thus for estimating ρ_{xy} in Case III models, the extra work involved in matching data files is almost worthless. Further studies are in order for much larger sample sizes to examine if this picture changes at all. We should point out that it is possible that matching may be useful for extracting some other features of the joint distribution and further Monte Carlo studies are warrented to explore this.

Table 3.1 Summary Statistics of Sample Correlations - Files with n=10 Records Conditional Independence Case

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
0.00	0.10	0.00	0.0149	-0.0032	-0.0101	-0.0100	-0.0114
			(0.3384)	(0.1127)	(0.3296)	(0.3297)	(0.3212)
			-0.8170	-0.5844	-0.7575	-0.7575	-0.8506
			0.8472	0.4675	0.8590	0.8590	0.7708
0.92	0.65	0.60	0.5879	0.5794	0.5457	0.5457	0.5105
			(0.2212)	(0.2006)	(0.2337)	(0.2337)	(0.2396)
			-0.6523	-0.4040	-0.6058	-0.6058	-0.6058
			0.9753	0.9431	0.9626	0.9626	0.9681
0.93	0.75	0.70	0.6830	0.6638	0.6150	0.6151	0.5748
			(0.1986)	(0.1728)	(0.2087)	(0.2086)	(0.2230)
			-0.3369	-0.1437	-0.3115	-0.3115	-0.3396
			0.9936	0.9609	0.9576	0.9576	0.9696

Table 3.1 (Cont'd.)

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
			0.7863	0.7775	0.7302	0.7302	0.6874
			(0.1445)	(0.1182)	(0.1522)	(0.1522)	(0.1731)
0.94	0.85	0.80	-0.3432	0.2058	-0.2367	-0.2367	-0.2367
			0.9879	0.9566	0.9799	0.9799	0.9723
			0.8937	0.8901	0.8252	0.8251	0.7789
			(0.0764)	(0.0625)	(0.0994)	(0.0995)	(0.1236)
0.95	0.95	0.90	0.3247	0.3508	0.3821	0.3821	0.1796
			0.9949	0.9814	0.9850	0.9850	0.9725
			0.9448	0.9421	0.8758	0.8760	0.8238
			(0.0419)	(0.0317)	(0.0741)	(0.0741)	(0.1063)
0.97	0.97	0.95	0.5329	0.7364	0.5027	0.5027	0.2123
			0.9973	0.9910	0.9898	0.9898	0.9868

Table 3.2 Summary Statistics of Sample Correlations - Files with n=10 Records Conditional Positive Dependence Case

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
0.00	0.10	0.95	0.9413	-0.0046	-0.0289	-0.0395	-0.0153
			(0.0474)	(0.1142)	(0.3310)	(0.3327)	(0.3269)
			0.5942	-0.5723	-0.8425	-0.8525	-0.8962
			0.9959	0.5302	0.8897	0.8897	0.8181
0.92	0.65	0.88	0.8676	0.5729	0.5276	0.5108	0.4919
			(0.0885)	(0.2021)	(0.2403)	(0.2443)	(0.2483)
			0.2744	-0.5510	-0.6166	-0.6248	-0.6119
			0.9914	0.9407	0.9621	0.9621	0.9621
0.93	0.75	0.92	0.9103	0.6771	0.6310	0.6262	0.5834
			(0.0666)	(0.1617)	(0.2018)	(0.2050)	(0.2085)
			0.4811	-0.2063	-0.3529	-0.3529	-0.2667
			0.9918	0.9448	0.9722	0.9722	0.9892

Table 3.2 (Cont'd.)

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
0.94	0.85	0.96	0.9558	0.7741	0.7188	0.7165	0.6687
			(0.0353)	(0.1153)	(0.1573)	(0.1578)	(0.1781)
			0.6288	0.2202	-0.2325	-0.2325	-0.1806
			0.9960	0.9798	0.9707	0.9707	0.9535
0.95	0.95	0.98	0.9775	0.8871	0.8225	0.8211	0.7770
			(0.0177)	(0.0640)	(0.1036)	(0.1040)	(0.1231)
			0.8491	0.4165	0.2546	0.2546	0.0215
			0.9986	0.9783	0.9922	0.9922	0.9727
0.97	0.97	0.99	0.9888	0.9439	0.8770	0.8774	0.8258
			(0.0088)	(0.0329)	(0.0760)	(0.0755)	(0.1039)
			0.9184	0.6081	0.4432	0.4432	0.3541
			0.9992	0.9919	0.9894	0.9894	0.9857

Table 3.3 Summary Statistics of Sample Correlations - Files with n=25 Records Conditional Independence Case

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
0.00	0.10	0.00	-0.0068	0.0001	-0.0025	-0.0026	-0.0040
			(0.2059)	(0.0479)	(0.2013)	(0.2014)	(0.2008)
			-0.6576	-0.2851	-0.5749	-0.5749	-0.6980
			0.5450	0.2501	0.6196	0.6196	0.5087
0.92	0.65	0.60	0.5915	0.5788	0.5568	0.5564	0.5171
			(0.1336)	(0.1231)	(0.1365)	(0.1365)	(0.1476)
			-0.0576	-0.0890	0.0259	0.0259	-0.0468
			0.8704	0.8189	0.8663	0.8663	0.8096
0.93	0.75	0.70	0.6859	0.6859	0.6620	0.6627	0.6111
			(0.1087)	(0.0935)	(0.1096)	(0.1097)	(0.1216)
			0.2953	0.2697	0.1828	0.1828	0.1642
			0.9022	0.8959	0.8955	0.8955	0.8973

Table 3.3 (Cont'd.)

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
0.94	0.85	0.80	0.7993	0.7934	0.7644	0.7643	0.7129
			(0.0754)	(0.0617)	(0.0789)	(0.0790)	(0.0964)
			0.4274	0.4778	0.4617	0.4617	0.2724
			0.9380	0.9087	0.9139	0.9139	0.9241
0.95	0.95	0.90	0.8967	0.8961	0.8648	0.8643	0.8049
			(0.0416)	(0.0313)	(0.0473)	(0.476)	(0.0676)
			0.7057	0.7592	0.6580	0.6580	0.4614
			0.9753	0.9636	0.9632	0.9632	0.9297
0.97	0.97	0.95	0.9479	0.9473	0.9117	0.9123	0.8485
			(0.0211)	(0.0154)	(0.0327)	(0.0326)	(0.0605)
			0.8446	0.8638	0.7636	0.7636	0.5102
			0.9874	0.9755	0.9735	0.9735	0.9519

Table 3.4 Summary Statistics of Sample Correlations - Files with n=25 Records Conditional Positive Dependence Case

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
0.00	0.10	0.95	0.9475	-0.0019	0.0058	-0.0372	-0.0004
			(0.0222)	(0.0439)	(0.2061)	(0.2038)	(0.1989)
			0.8249	-0.2817	-0.5665	-0.5480	-0.7596
			0.9857	0.1963	0.6964	0.6964	0.5557
0.92	0.65	0.88	0.8758	0.5857	0.5643	0.5149	0.5277
			(0.0503)	(0.1207)	(0.1331)	(0.1436)	(0.1425)
			0.6051	0.1442	0.1621	0.0617	0.0404
			0.9738	0.8344	0.8896	0.8896	0.8512
0.93	0.75	0.92	0.9143	0.6907	0.6627	0.6489	0.6190
			(0.0361)	(0.0851)	(0.1058)	(0.1093)	(0.1125)
			0.6844	0.2967	0.2949	0.2641	0.1829
			0.9774	0.8876	0.8661	0.8642	0.9020

Table 3.4 (Cont'd.)

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
0.94	0.85	0.96	0.9578	0.7931	0.7641	0.7539	0.7127
			(0.0174)	(0.0624)	(0.0832)	(0.0853)	(0.0948)
			0.8756	0.5449	0.3612	0.3647	0.3425
			0.9893	0.9226	0.9181	0.9174	0.9128
0.95	0.95	0.98	0.9792	0.8956	0.8614	0.8543	0.7998
			(0.0096)	(0.0308)	(0.0496)	(0.0516)	(0.0691)
			0.9131	0.7693	0.6315	0.6226	0.5157
			0.9959	0.9661	0.9647	0.9647	0.9413
0.97	0.97	0.99	0.9895	0.9475	0.9123	0.9139	0.8499
			(0.0042)	(0.0158)	(0.0339)	(0.0336)	(0.0584)
			0.9685	0.8769	0.7182	0.7352	0.5685
			0.9972	0.9833	0.9769	0.9849	0.9773

Table 3.5 Summary Statistics of Sample Correlations - Files with n=50 Records Conditional Independence Case

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s3}$
0.00	0.10	0.00	-0.0004	-0.0003	-0.0019	-0.0044
			(0.1436)	(0.0242)	(0.1474)	(0.1445)
			-0.4381	-0.1663	-0.4872	-0.5205
			0.4746	0.1244	0.4398	0.4574
0.92	0.65	0.60	0.5936	0.5952	0.5823	0.5391
			(0.0916)	(0.0794)	(0.0909)	(0.0959)
			0.2530	0.2219	0.2242	0.1098
			0.8377	0.8103	0.7998	0.7873
0.93	0.75	0.70	0.6950	0.6953	0.6807	0.6279
			(0.0756)	(0.0612)	(0.0709)	(0.0815)
			0.2796	0.3696	0.3760	0.2526
			0.8768	0.8426	0.8718	0.8543

Table 3.5 (Cont'd.)

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell1}$	$\hat{\rho}_{m\ell2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s3}$
0.94	0.85	0.80	0.7959	0.7974	0.7797	0.7198
			(0.0528)	(0.0408)	(0.0527)	(0.0645)
			0.5689	0.5664	0.4919	0.4531
			0.9204	0.9082	0.9222	0.8821
0.95	0.95	0.90	0.8982	0.8978	0.8778	0.8110
			(0.0289)	(0.0200)	(0.0306)	(0.0493)
			0.7152	0.7845	0.7331	0.6079
			0.9634	0.9467	0.9595	0.9149
0.97	0.97	0.95	0.9486	0.9490	0.9276	0.8559
			(0.0151)	(0.0103)	(0.0199)	(0.0419)
			0.8549	0.9100	0.8039	0.6529
			0.9808	0.9743	0.9761	0.9576

Table 3.6 Summary Statistics of Sample Correlations - Files with n=50 Records Conditional Positive Dependence Case

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s3}$
0.00	0.10	0.95	0.9491	0.0001	0.0015	0.0025
			(0.0148)	(0.0245)	(0.1475)	(0.1427)
			0.8700	-0.1447	-0.5256	-0.5157
			0.9828	0.1506	0.4727	0.5145
0.92	0.65	0.88	0.8776	0.5934	0.5809	0.5358
			(0.0336)	(0.0817)	(0.0928)	(0.0981)
			0.6908	0.2791	0.1519	0.1593
			0.9576	0.8031	0.8181	0.8338
0.93	0.75	0.92	0.9183	0.6944	0.6771	0.6257
			(0.0225)	(0.0638)	(0.0752)	(0.0834)
			0.8119	0.4028	0.3506	0.2950
			0.9698	0.8628	0.8599	0.8595

Table 3.6 (Cont'd.)

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s3}$
0.94	0.85	0.96	0.9595	0.7967	0.7803	0.7198
			(0.0116)	(0.0415)	(0.0512)	(0.0627)
			0.8793	0.6023	0.5699	0.3595
			0.9853	0.8960	0.9158	0.8824
0.95	0.95	0.98	0.9794	0.8973	0.8776	0.8106
			(0.0061)	(0.0200)	(0.0294)	(0.0468)
			0.9390	0.8096	0.7596	0.6273
			0.9932	0.9506	0.9570	0.9279
0.97	0.97	0.99	0.9898	0.9492	0.9281	0.8555
			(0.0029)	(0.0107)	(0.0200)	(0.0426)
			0.9736	0.8927	0.8181	0.6501
			0.9964	0.9757	0.9713	0.9555

Table 3.7 Summary Statistics of Sample Correlations - Files with n=25 Records Conditional Positive Dependence Case

ρ_{xz}	ρ_{yz}	ρ_{xy}	$\hat{\rho}_{m\ell 1}$	$\hat{\rho}_{m\ell 2}$	$\hat{\rho}_{s1}$	$\hat{\rho}_{s2}$	$\hat{\rho}_{s3}$
			0.4933	0.0008	-0.0027	-0.0063	0.0012
			(0.1574)	(0.0451)	(0.2117)	(0.2105)	(0.2044)
0.00	0.10	0.50	-0.0632	-0.1632	-0.6421	-0.6421	-0.0035
			0.8777	0.1976	0.6186	-0.6186	0.5807
			0.7425	0.5876	0.5655	0.5622	0.5236
			(0.0940)	(0.1108)	(0.1292)	(0.1301)	(0.1430)
0.92	0.65	0.75	0.2986	0.1141	-0.0065	-0.0065	0.0205
			0.9390	0.8326	0.8621	-0.8621	0.8285
			0.7943	0.6919	0.6683	0.6691	0.6249
			(0.0762)	(0.0889)	(0.1109)	(0.1102)	(0.1180)
0.93	0.75	0.80	0.3982	0.3129	0.1844	0.1844	0.2023
			0.9373	0.8978	0.9047	0.9047	0.8853

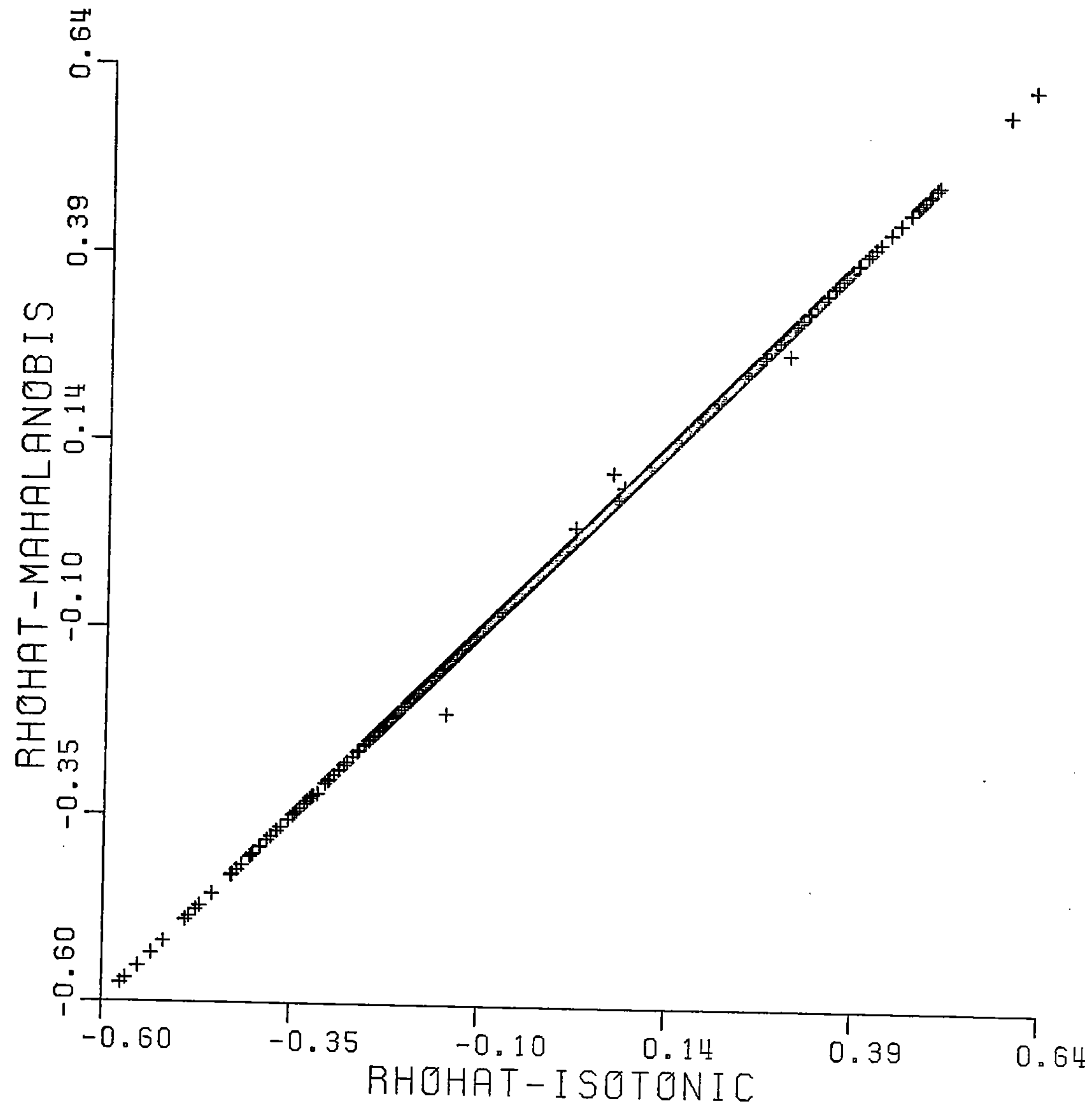

Figure 3.1 Isotonic vs. Mahalanobis.
$\rho_{xz} = 0.00$, $\rho_{yz} = 0.10$, $\rho_{xy} = 0.00$, $n = 10$.

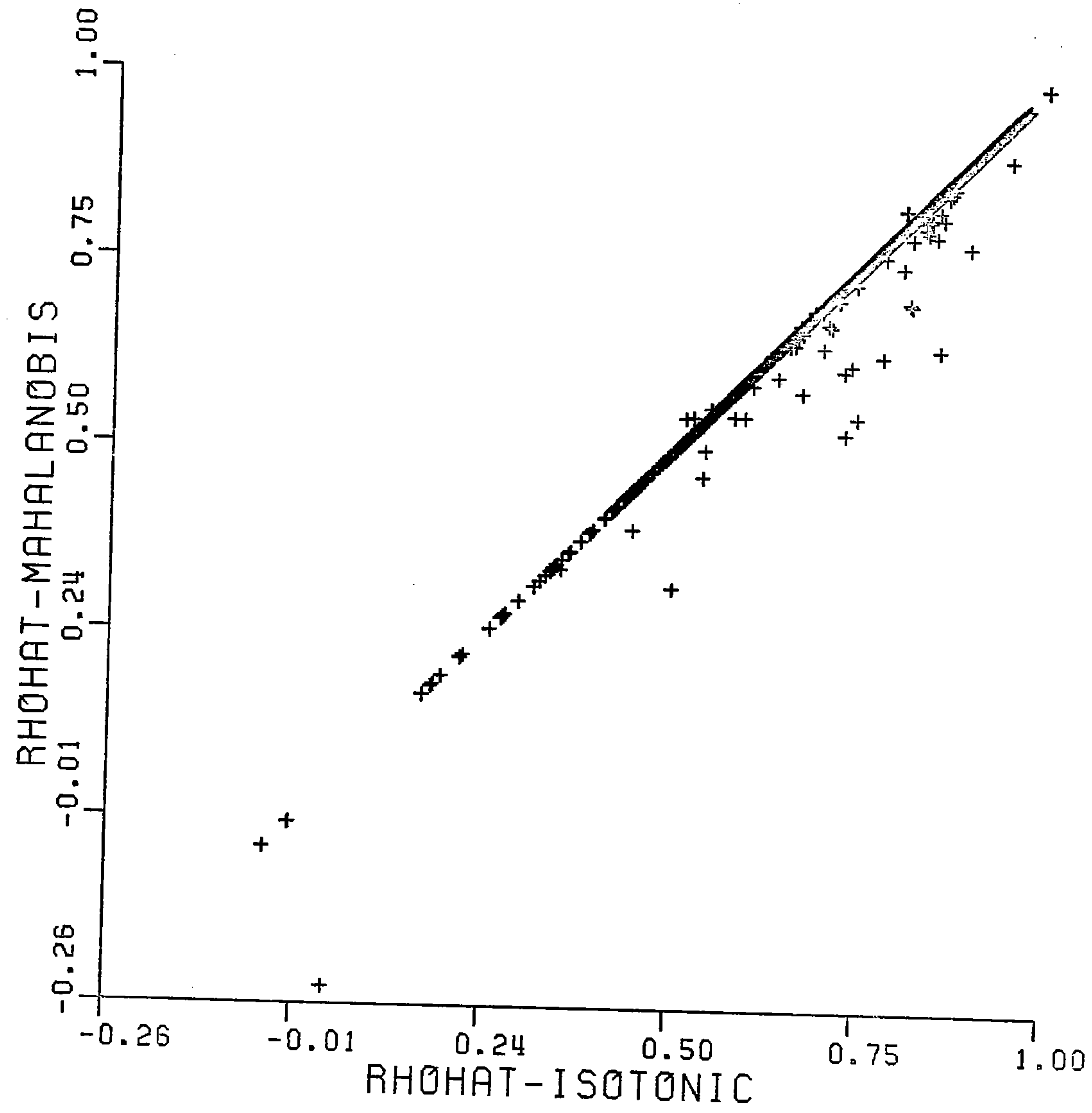

Figure 3.2 Isotonic vs. Mahalanobis.
$\rho_{xz} = 0.94$, $\rho_{yz} = 0.85$, $\rho_{xy} = 0.96$, n = 50.

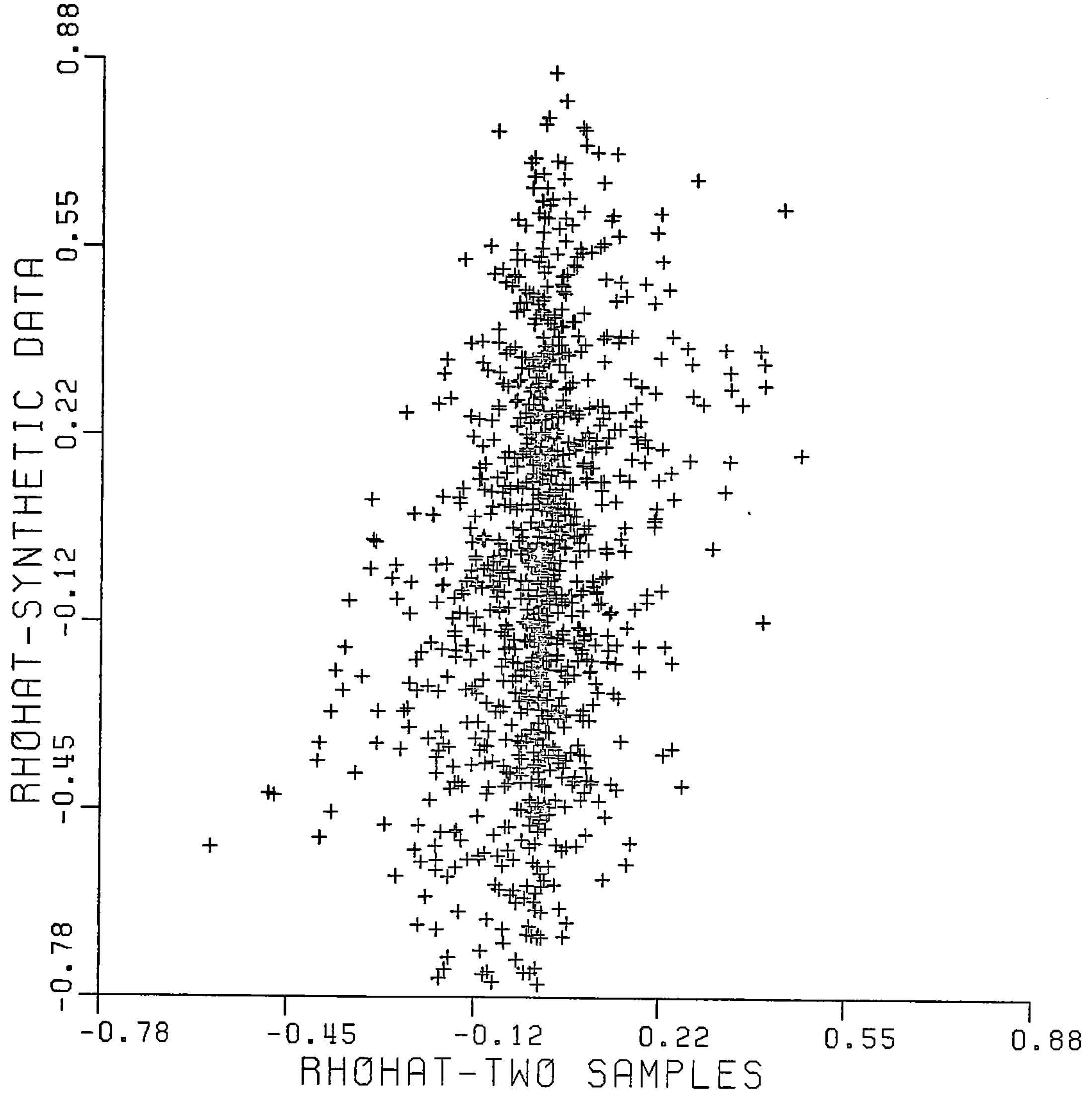

Figure 3.3 **Isotonic vs. Nomatching.**
$\rho_{xz} = 0.00$, $\rho_{yz} = 0.10$, $\rho_{xy} = 0.00$, n = 10.

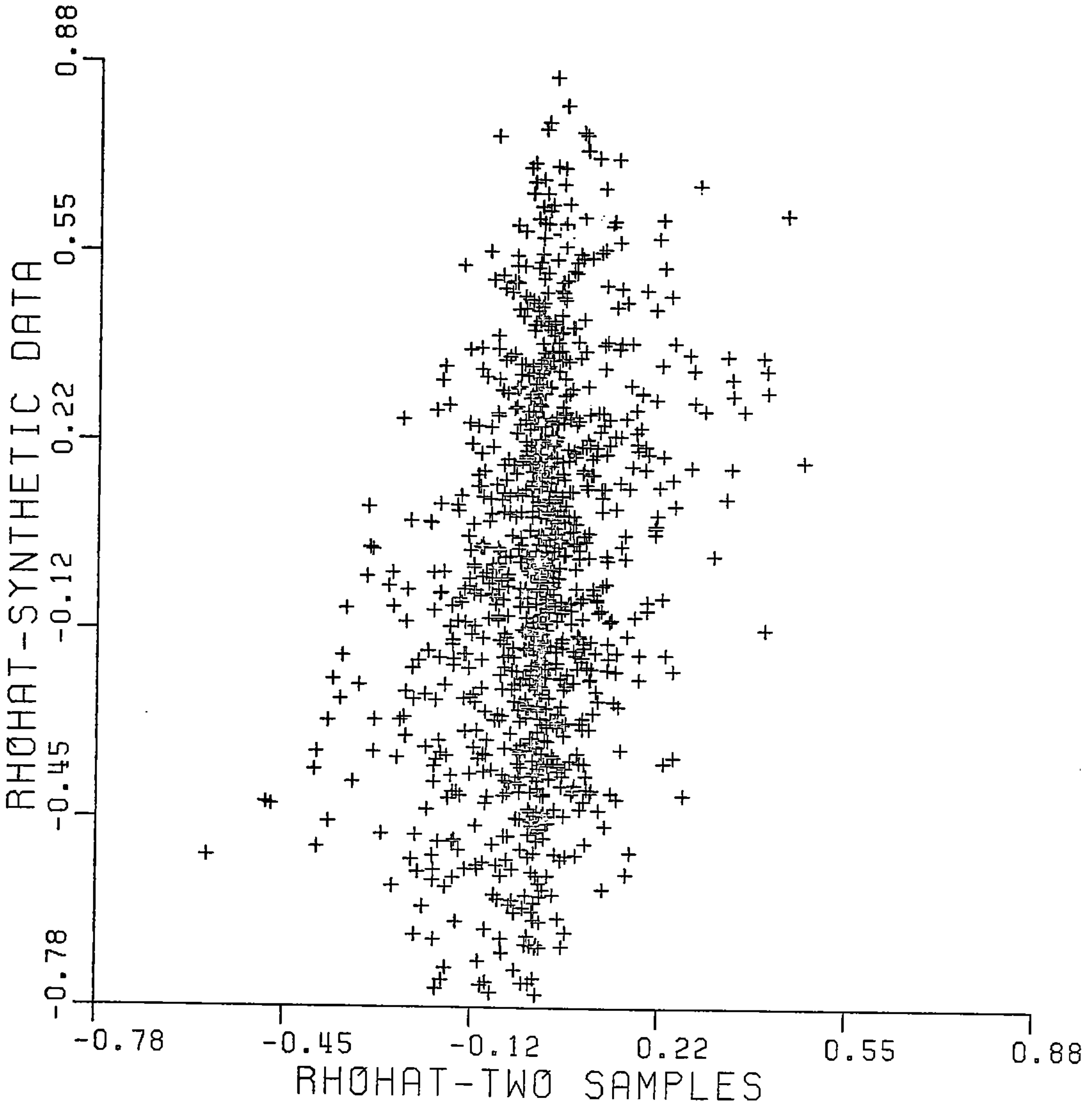

Figure 3.4 Mahalanobis vs. Nomatching.
$\rho_{xz} = 0.00$, $\rho_{yz} = 0.10$, $\rho_{xy} = 0.00$, $n = 10$.

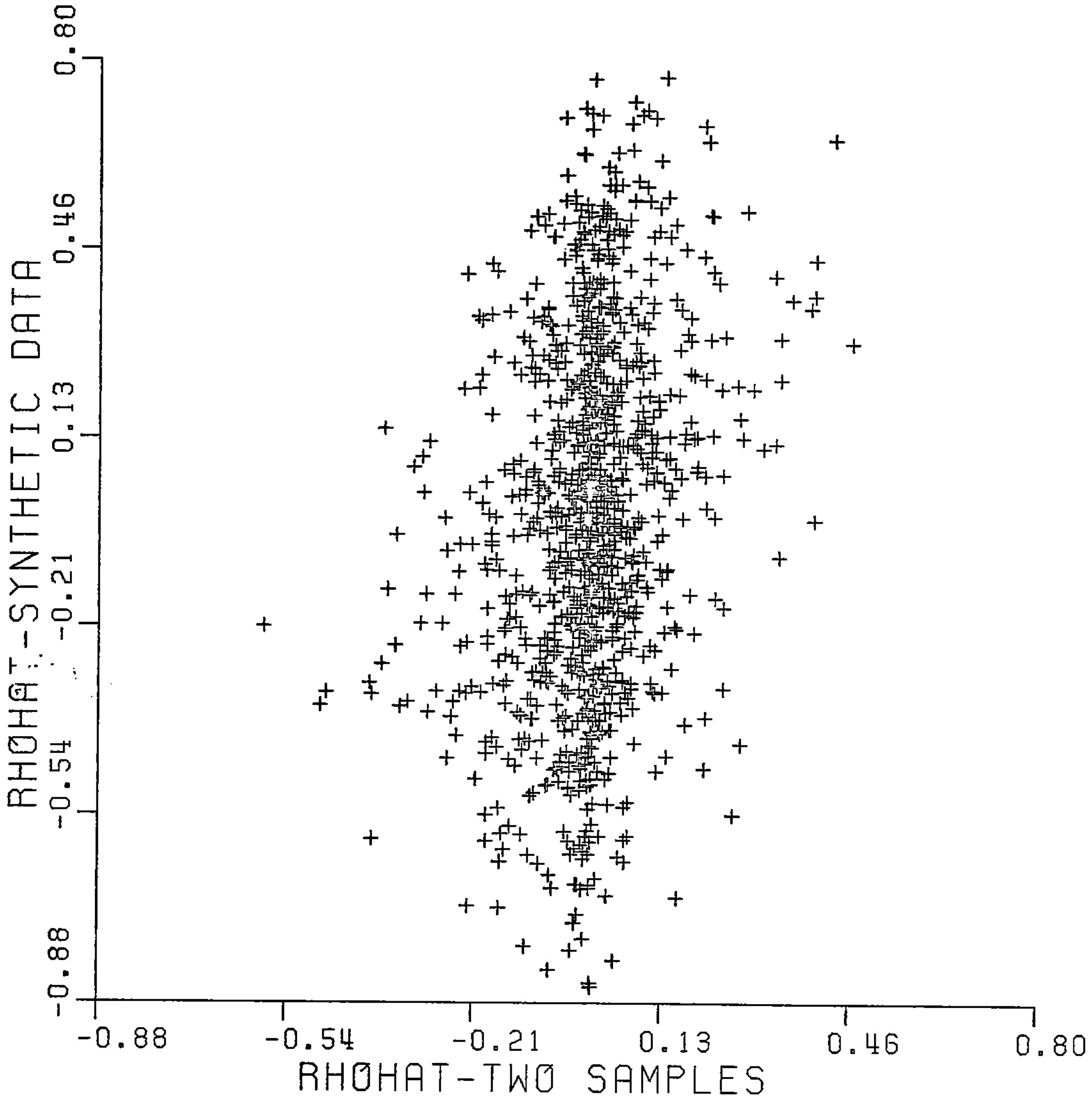

Figure 3.5 Matching in Bins vs. Nomatching.
ρ_{xz} = 0.00, ρ_{yz} = 0.10, ρ_{xy} = 0.00, n = 10.

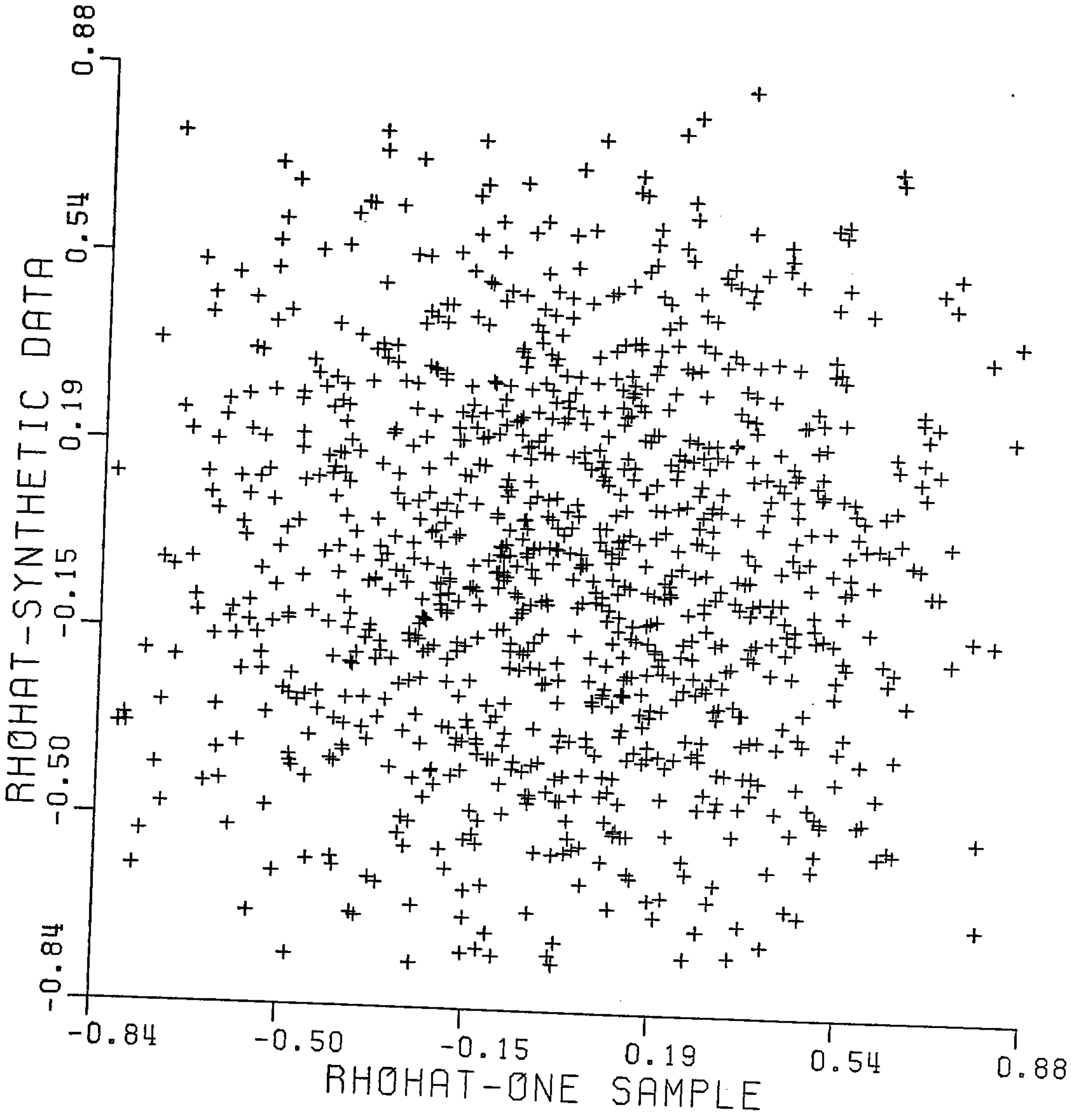

Figure 3.6 Isotonic vs. Nomatching.
ρ_{xz} = 0.00, ρ_{yz} = 0.10, ρ_{xy} = 0.00, n = 10.

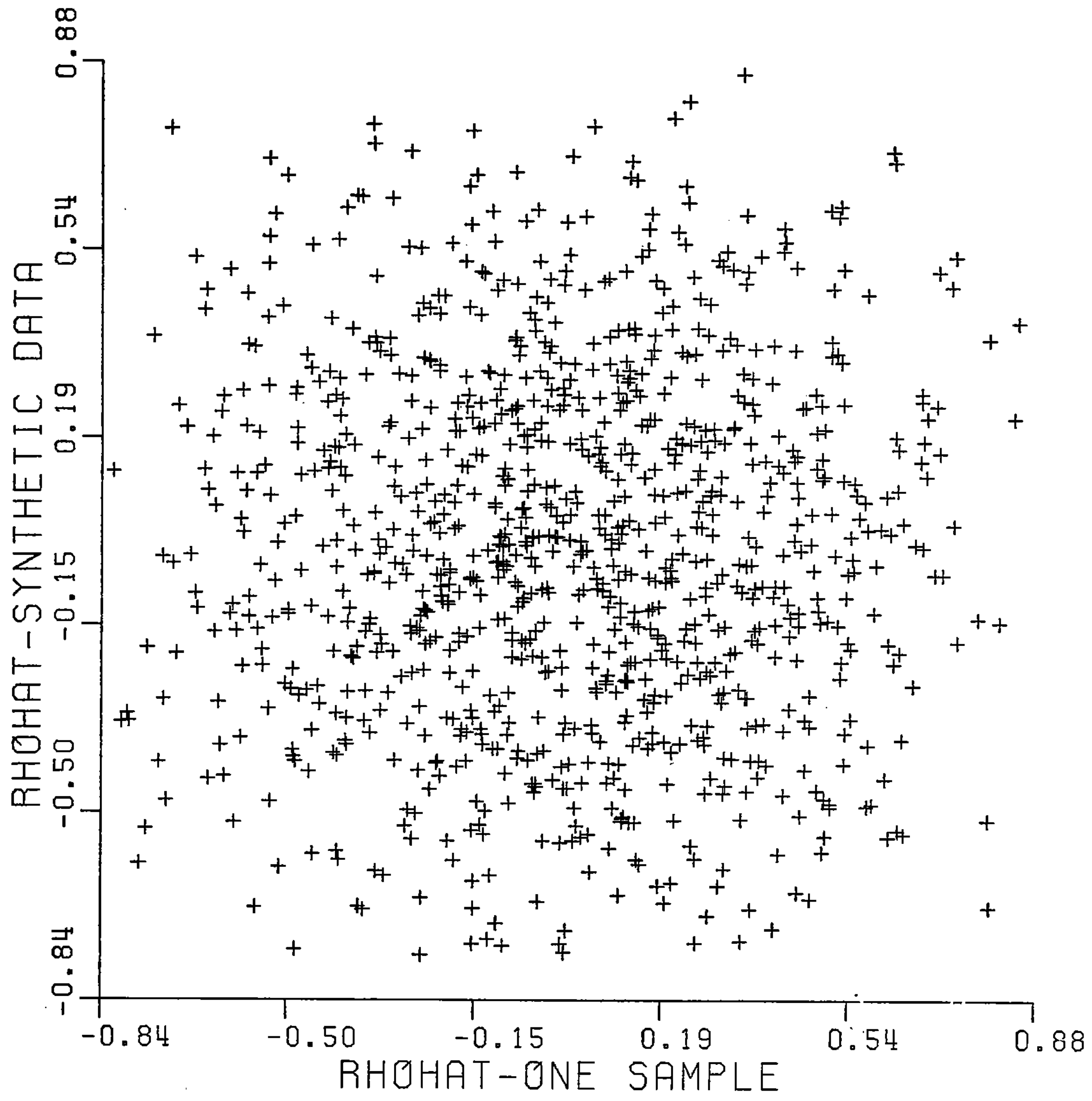

Figure 3.7 Mahalanobis vs. Nomatching.
ρ_{xz} = 0.00, ρ_{yz} = 0.10, ρ_{xy} = 0.00, n = 10.

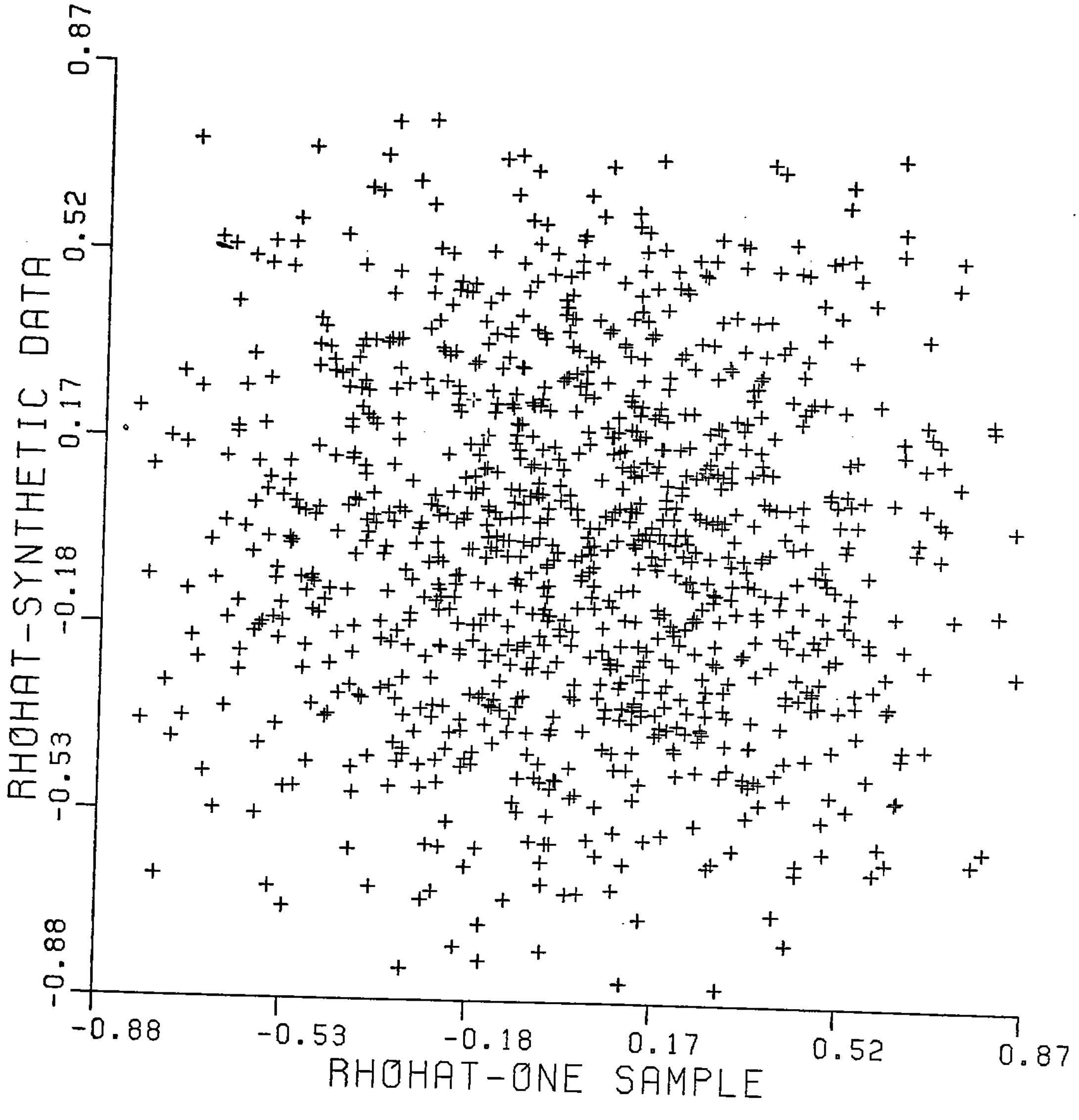

Figure 3.8 Matching in Bins vs. Nomatching.
ρ_{xz} = 0.00, ρ_{yz} = 0.10, ρ_{xy} = 0.00, n = 10.

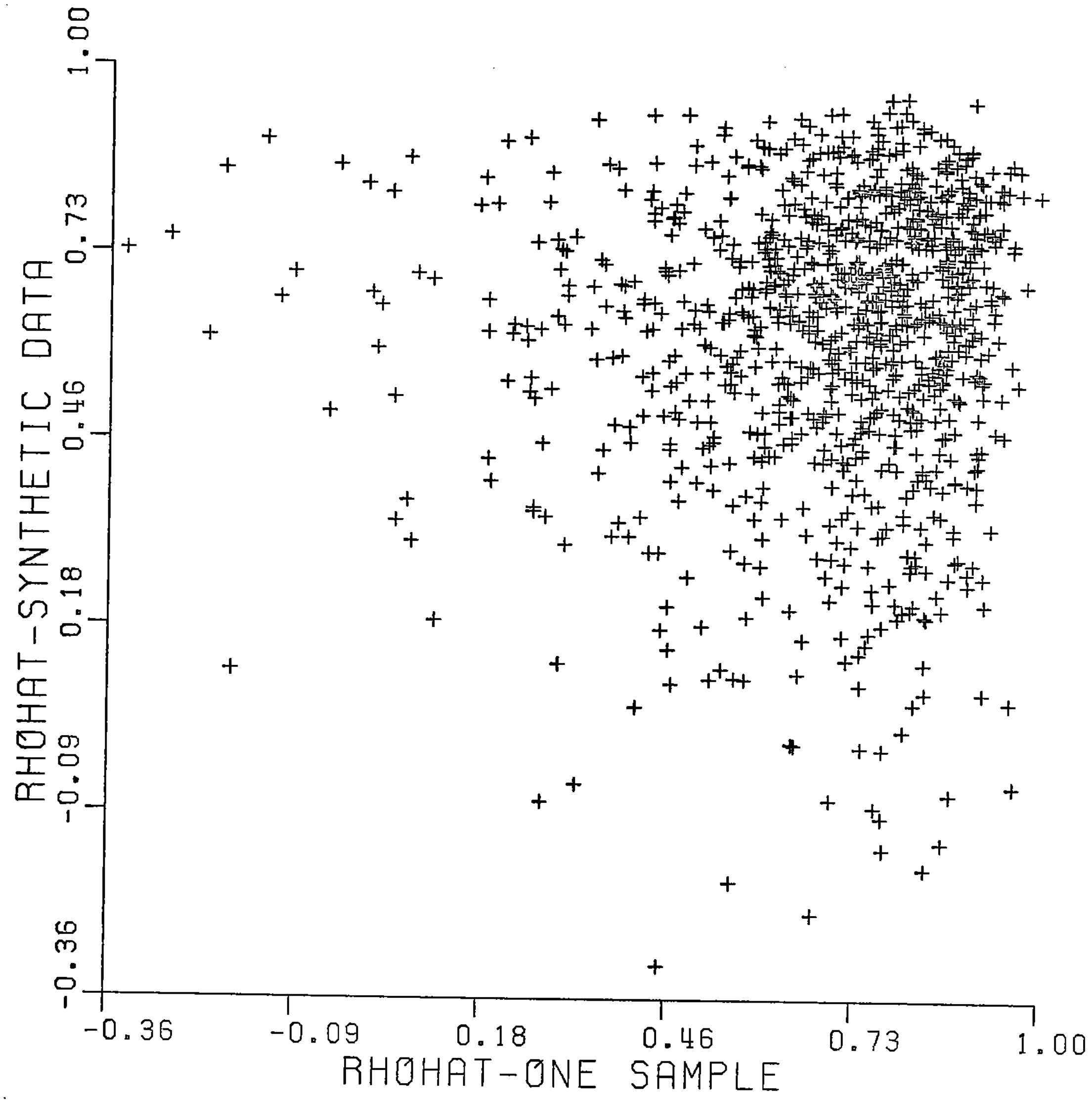

Figure 3.9 **Isotonic vs. Nomatching.**
$\rho_{xz} = 0.93$, $\rho_{yz} = 0.75$, $\rho_{xy} = 0.70$, n = 25.

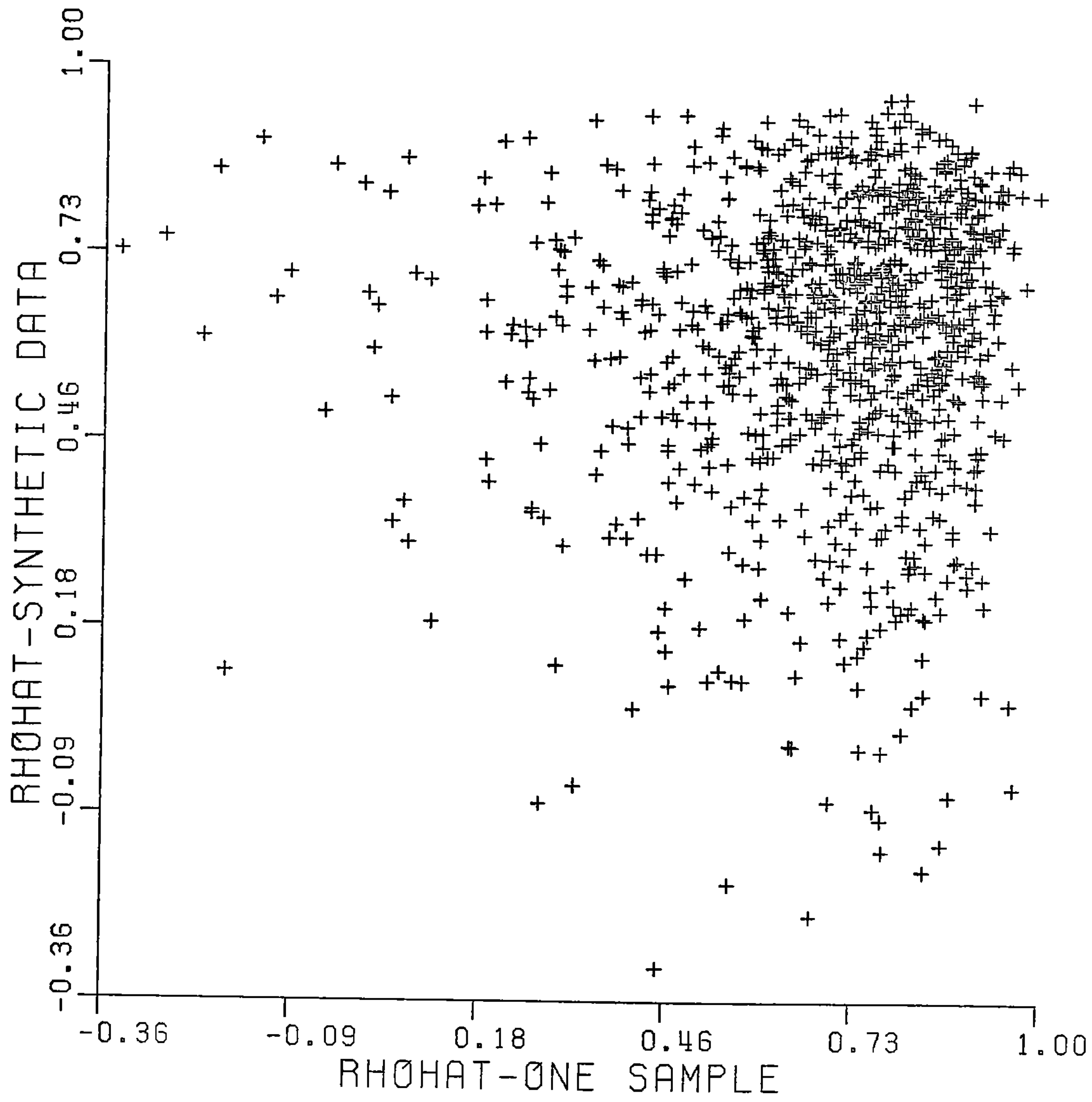

Figure 3.10 Mahalanobis vs. Nomatching.
$\rho_{xz} = 0.93$, $\rho_{yz} = 0.75$, $\rho_{xy} = 0.70$, n = 25.

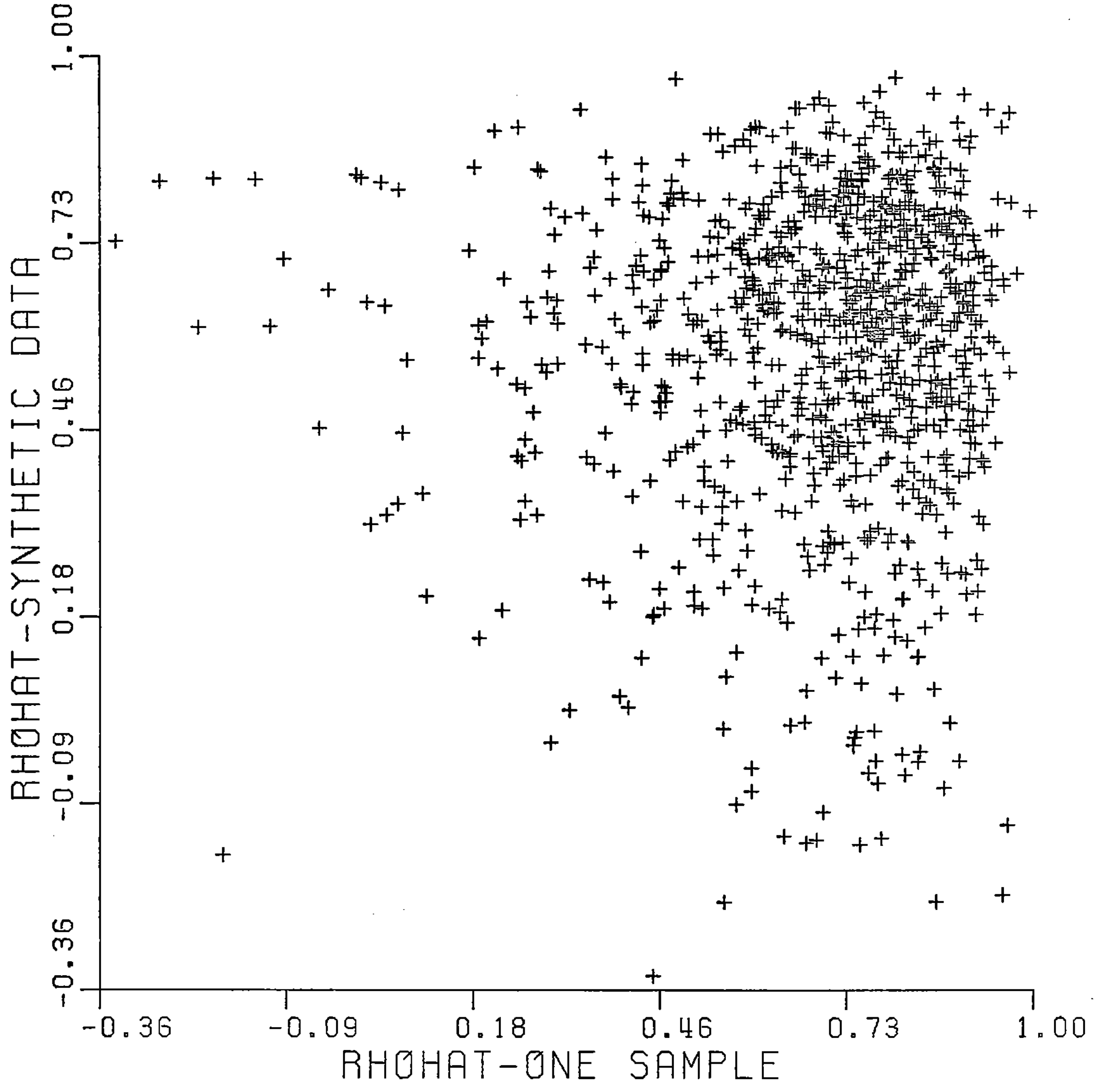

Figure 3.11 **Matching in Bins vs. Nomatching.**
ρ_{xz} = 0.93, ρ_{yz} = 0.75, ρ_{xy} = 0.70, n = 25.

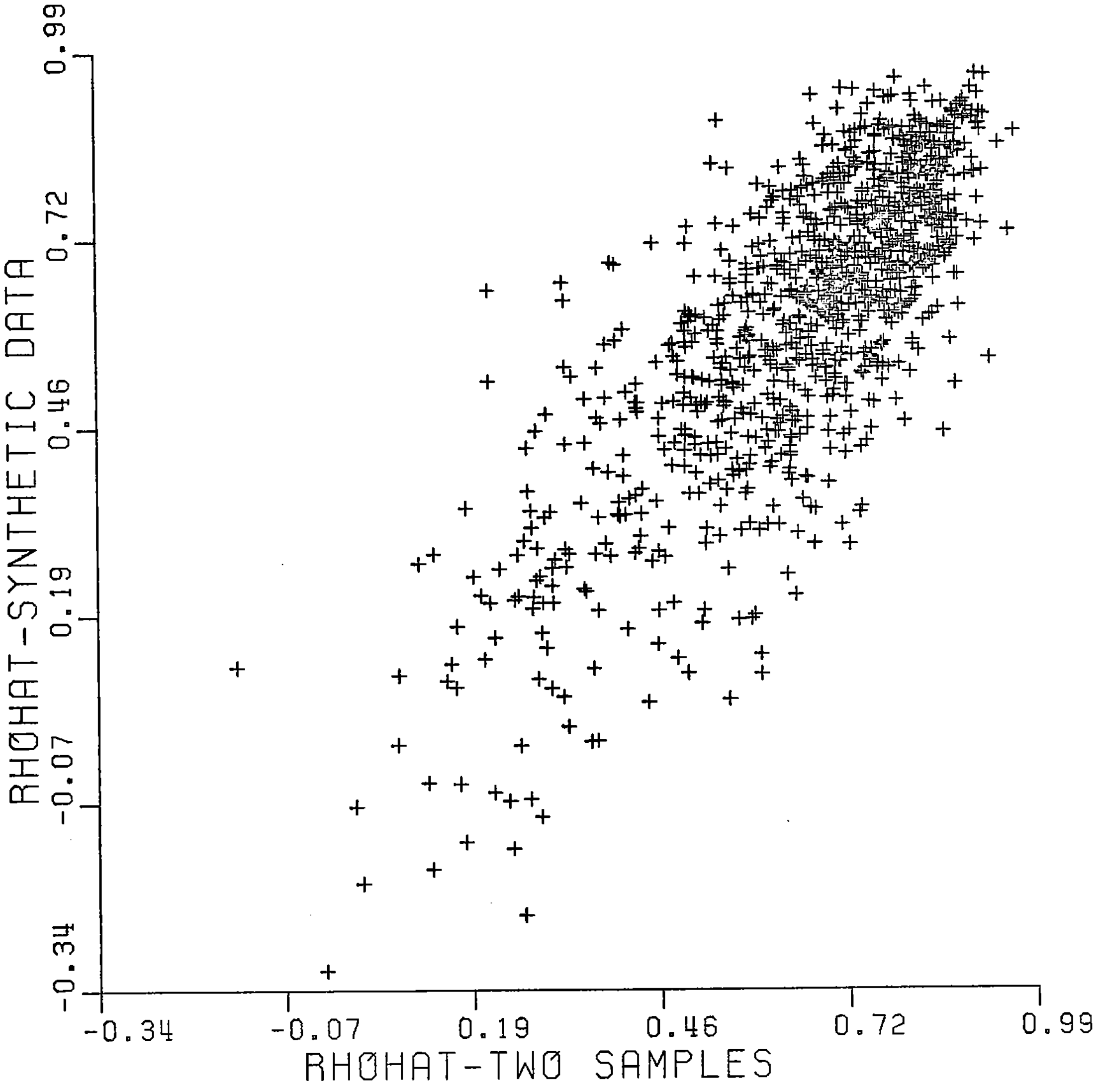

Figure 3.12 Isotonic vs. Nomatching.
ρ_{xz} = 0.93, ρ_{yz} = 0.75, ρ_{xy} = 0.70, n = 25.

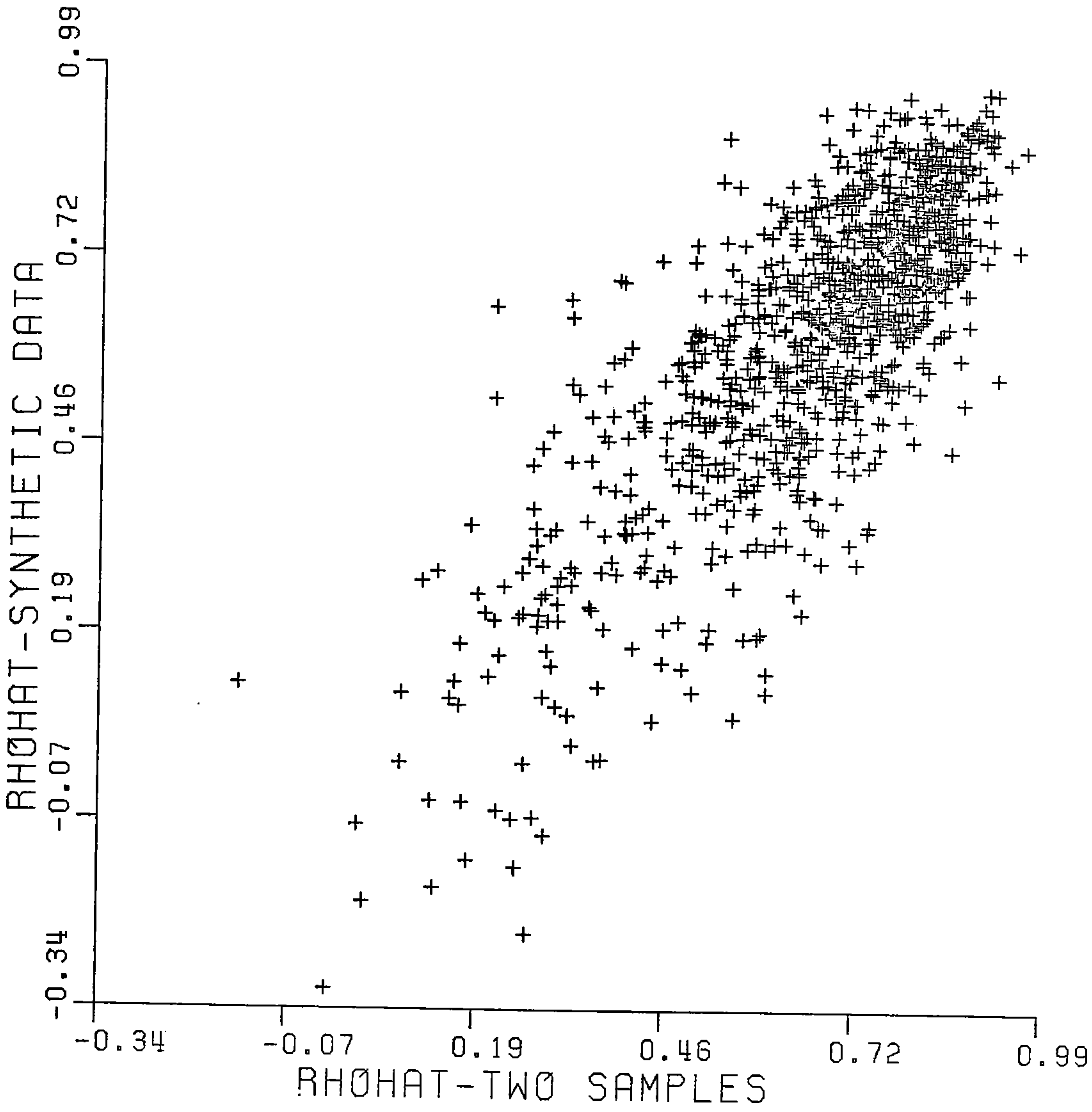

Figure 3.13 Mahalanobis vs. Nomatching.
$\rho_{xz} = 0.93$, $\rho_{yz} = 0.75$, $\rho_{xy} = 0.70$, n = 25.

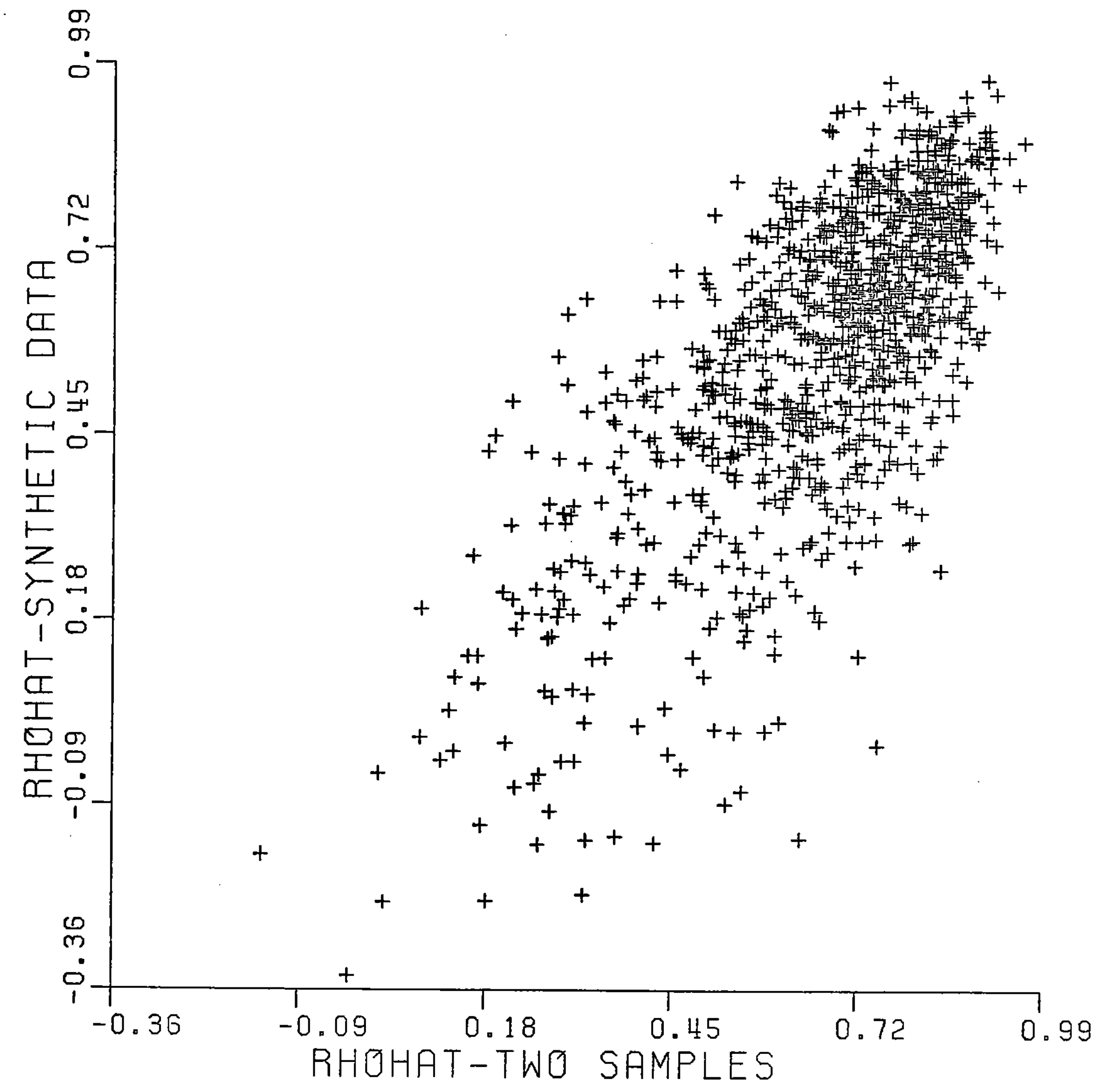

Figure 3.14 Matching in Bins vs. Nomatching.
ρ_{xz} = 0.93, ρ_{yz} = 0.75, ρ_{xy} = 0.70, n = 25.

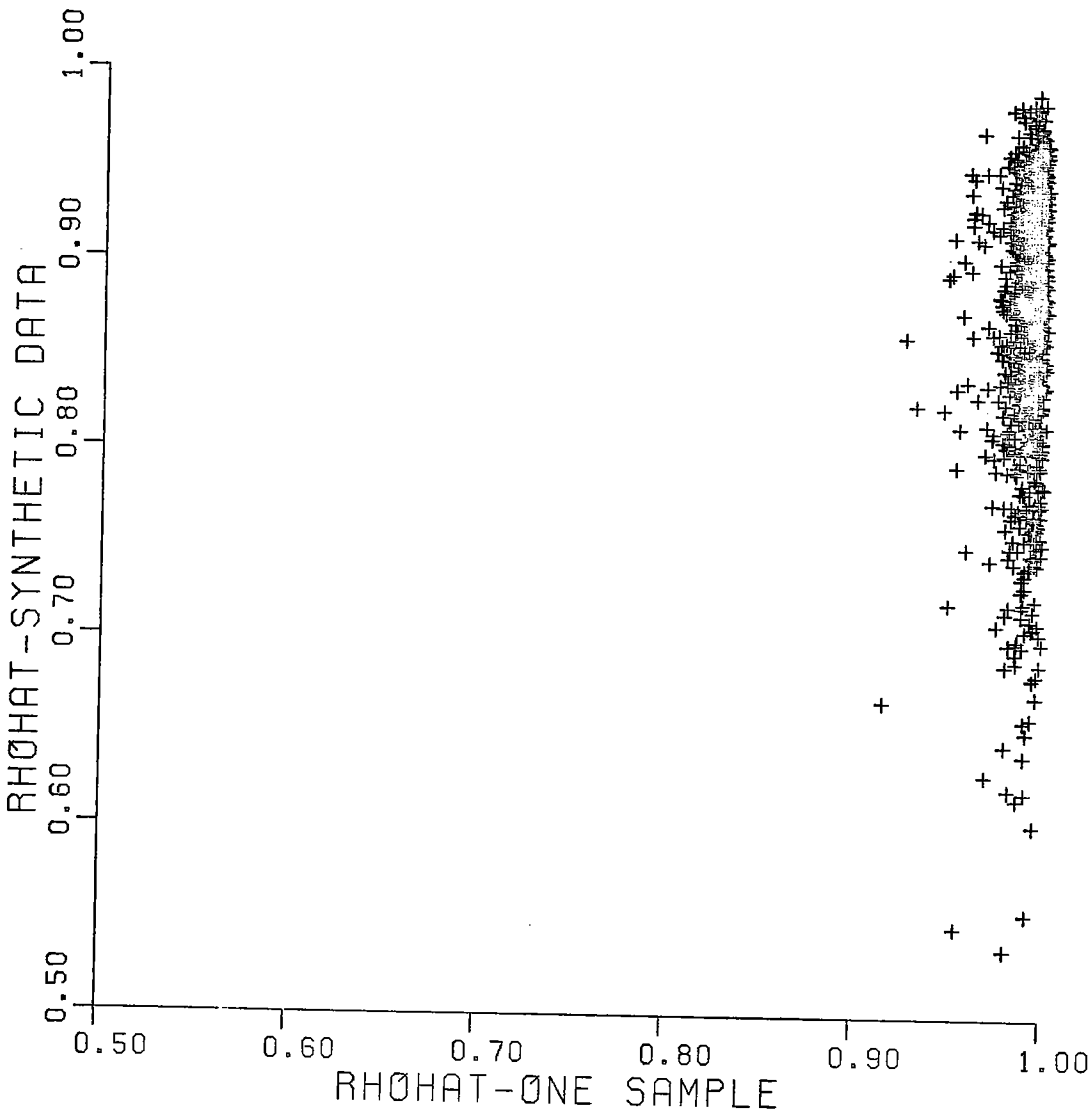

Figure 3.15 **Mahalanobis vs. Nomatching.**
$\rho_{xz} = 0.00$, $\rho_{yz} = 0.10$, $\rho_{xy} = 0.95$, n = 25.

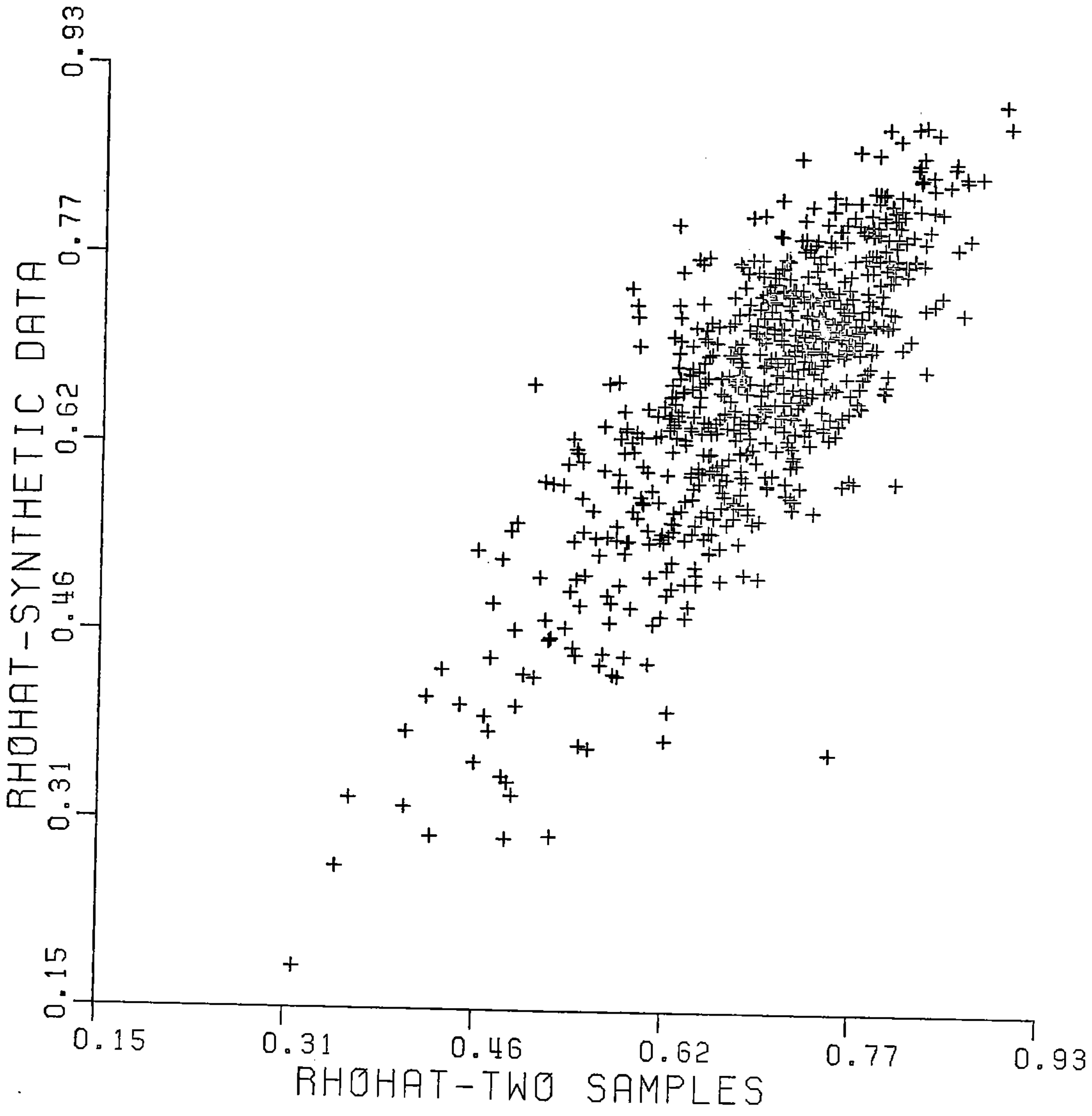

Figure 3.16 Isotonic vs. Nomatching.
ρ_{xz} = 0.93, ρ_{yz} = 0.75, ρ_{xy} = 0.80, n = 25.

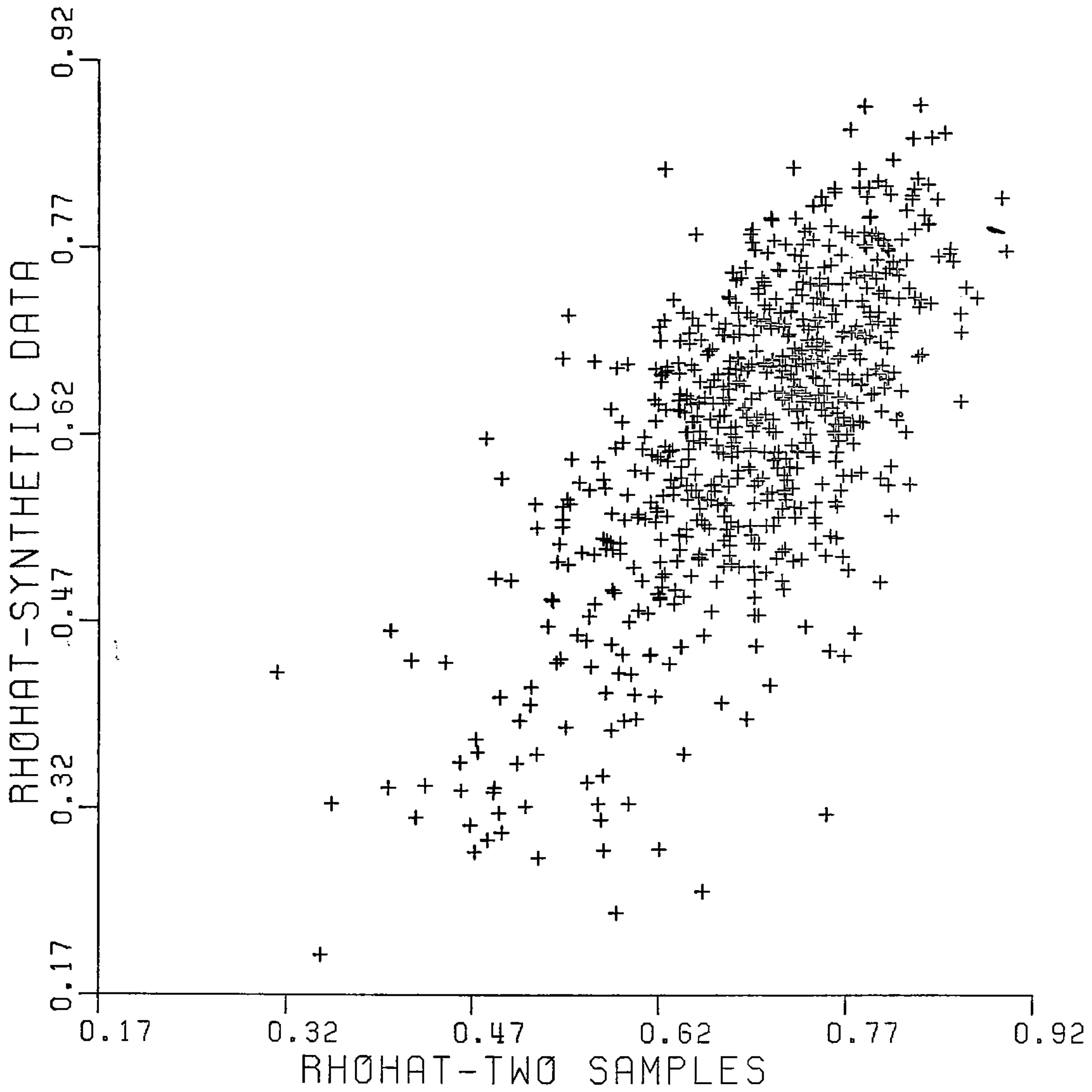

Figure 3.17 Matching in Bins vs. Nomatching.
$\rho_{xz} = 0.93$, $\rho_{yz} = 0.75$, $\rho_{xy} = 0.80$, n = 25.

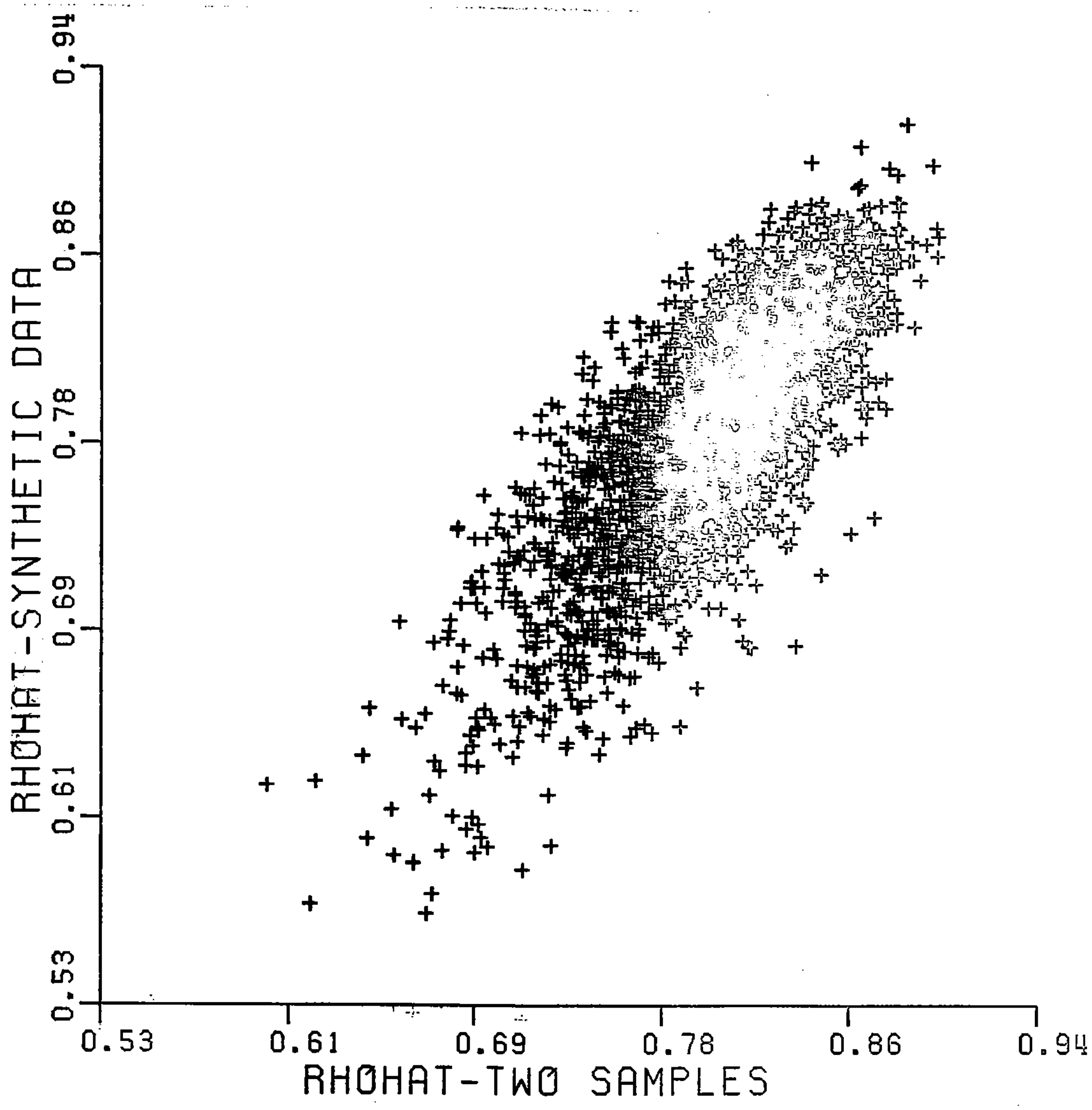

Figure 3.18 Isotonic vs. Nomatching.
ρ_{xz} = 0.94, ρ_{yz} = 0.85, ρ_{xy} = 0.96, n = 50.

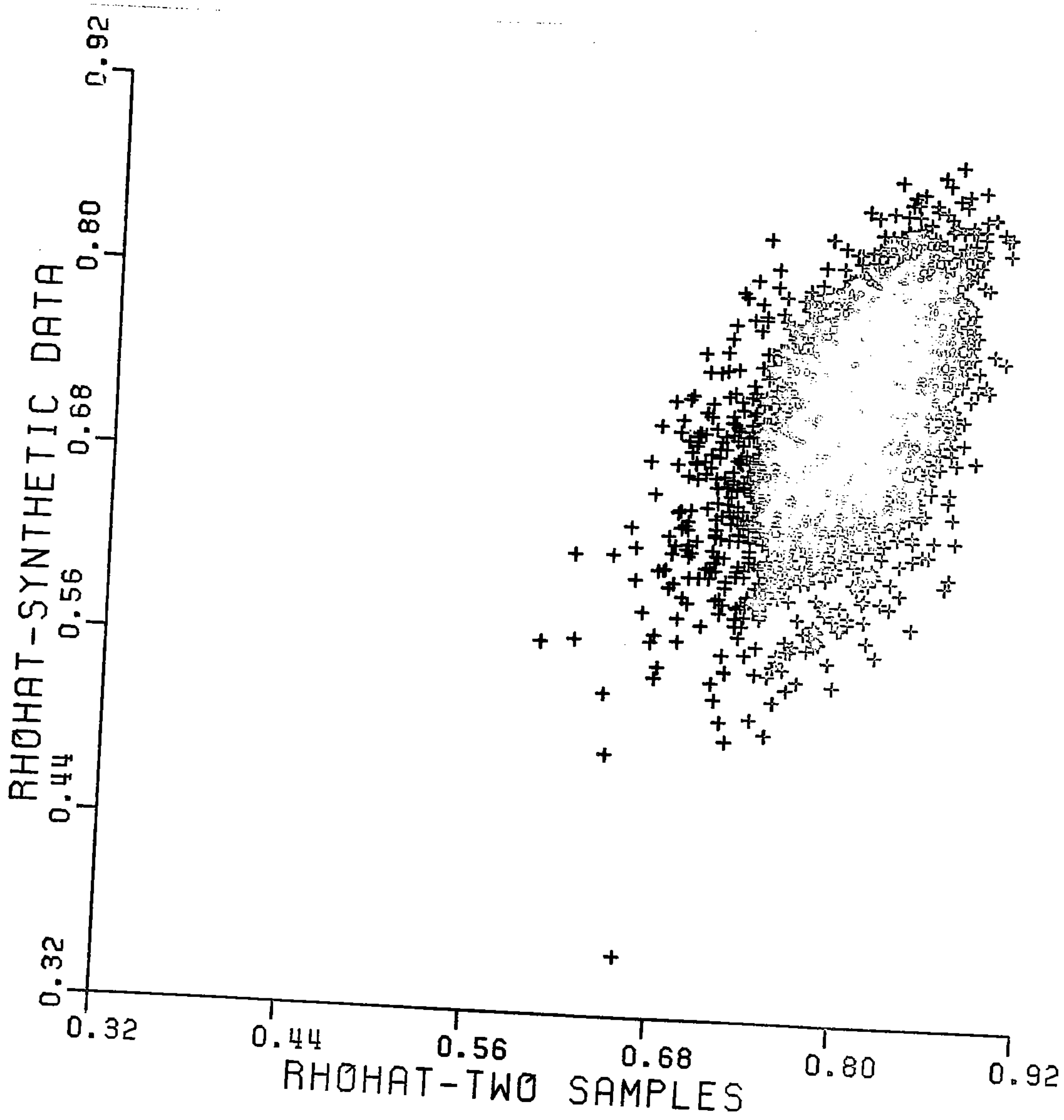

Figure 3.19 Matching in Bins vs. Nomatching.
ρ_{xz} = 0.94, ρ_{yz} = 0.85, ρ_{xy} = 0.96, n = 50.

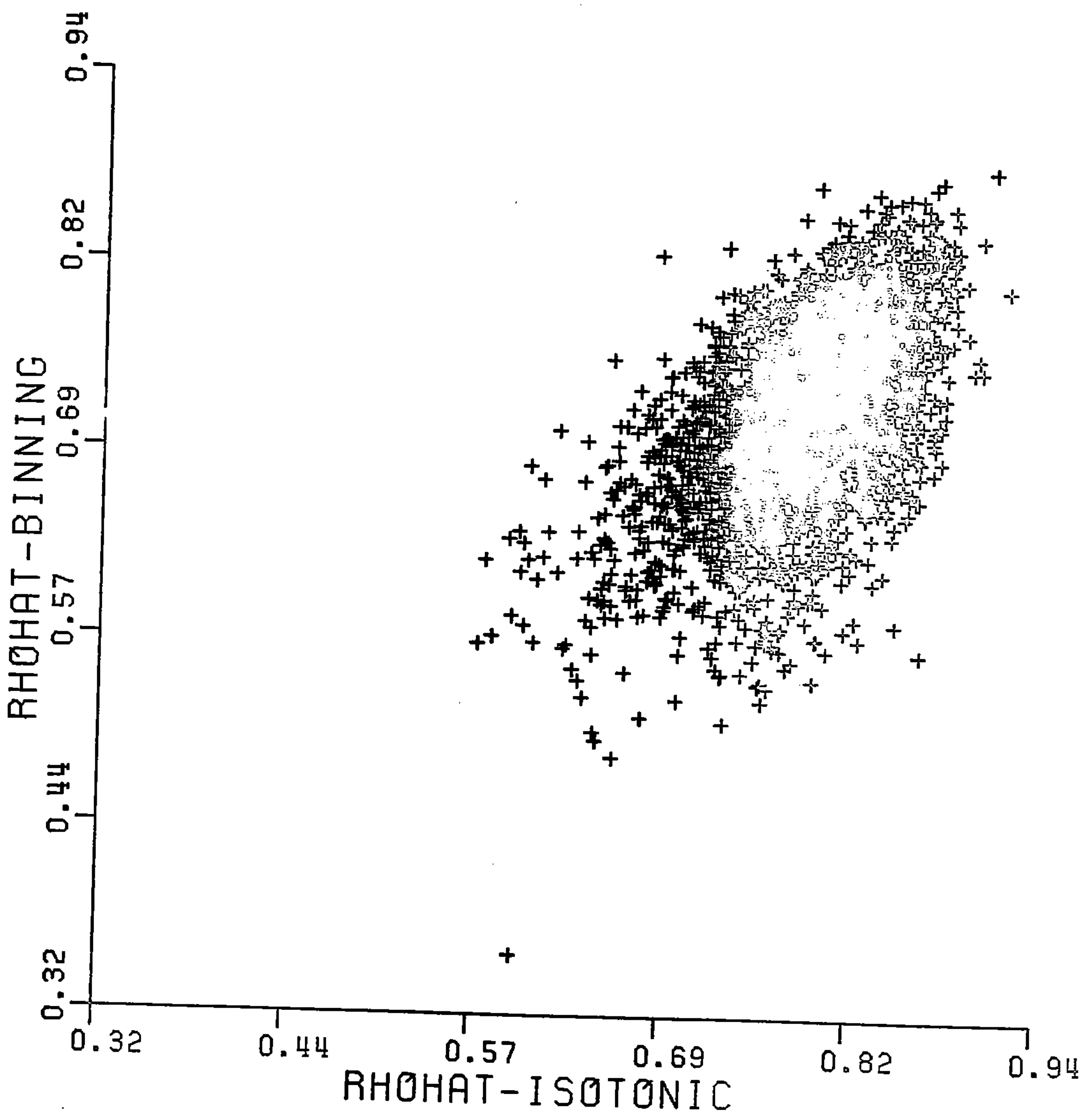

Figure 3.20 Isotonic vs. Binning.
ρ_{xz} = 0.94, ρ_{yz} = 0.85, ρ_{xy} = 0.96, n = 50.

REFERENCES

Anderson, T. W (1984), An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc. New York.

Barr, R. S., Stewart, W. H. and Turner, J. S. (1982), An Empirical Evaluation of Statistical Matching Methodologies, Unpublished Mimeo, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.

Barton, D. E. (1958) The Matching Distributions: Poisson Limiting Forms and Derived Methods of Approximation, Jour. Royal Statist. Soc. Ser-B, 20,73-92.

Bhattacharya, R. N. and Ranga Rao, R. (1976), Normal Approximation and Asymptotic Expansions, John Wiley and Sons, New York.

Bickel, P. J. and Yahav, J. A. (1977), On Selecting a Subset of Good Populations, Statistical Decision Theory and Related Topics II, (Ed.) Gupta, S. S. and Moore, D. S., Academic Press, New York.

Bleistein, N. and Handelsman, R. A. (1975) Asymptotic Expansions of Integrals, Holt, Rinehart, & Winston, New York.

Bochner, S. and Chandrasekar, K. (1949), Fourier Transforms, Annals of Mathematical Studies Series, #19, Princeton University Press.

Chew, M. C., Jr (1973), On Pairing Observations from a Distribution with Monotone Likelihood Ratio, Annals of Statistics 1, 433-445.

Chow, Y. S. and Teicher, H. (1978), Probability Theory: Independence Interchangeability Martingales, Springer-Verlag, New York.

Chung, K. L. (1974), A Course in Probability Theory, 2nd ed. Academic Press, New York.

Dawid, A. P. (1979), Conditional Independence in Statistical Theory, Jour. Royal Statist. Soc. Ser-B,Vol. 41, 1-31.

DeGroot, M. H., Feder, P. I., and Goel, P. K. (1971), Matchmaking, Annals of Mathematical Statistics, 42, 578-593.

DeGroot, M. H., and Goel, P. K. (1976), The Matching Problem for Multivariate Normal Data, Sankhya, Series B 38, 14-29.

DeGroot, M. H., and Goel, P. K. (1980), Estimation of the Correlation Coefficient from a Broken Random Sample, Annals of Statistics, 8, 264-278.

Eaton, M. and Kariya, T. (1983), Multivariate Tests with Incomplete Data, Annals of Statistics 11, 654-665.

Esary, J. D., Proschan, F., and Walkup, D. W. (1967), Association of Random Variables with Applications, Ann. Math. Statist. 38, 1466-1474.

Fellegi, I. P, and Sunter, A. B. (1969), A Theory for Record Linkage, Journal of the American Statistical Association, 64, 1183-1210.

Fellegi, I. P. (1978), 1977 Procedings of the American Statistical Association Social Statistics Section, 762-4.

Feller, W. (1968), An Introduction to Probability Theory and Its Applications, Vol. I, 3rd ed., New York: John Wiley and Sons.

Goel, P. K. (1975), On Re-pairing Observations in a Broken Random Sample, Annals of Statistics, 3, 1364-1369.

Kadane, J. B. (1978), Some Statistical Problems in Merging Data Files, Compendium of Tax Research, Washington, D.C.: Office of Tax Analysis, Dept. of the Treasury, 159-179.

Kariya, T. Krishnaiah, P. R. and Rao, C. R. (1983), Inference on Parameters of Multivariate Normal Populations when Some Data is Missing, Developments in Statistics, vol.4, (Ed.) P. R. Krishnaiah, 137-181.

Lancaster, H. O. (1969) The Chi-squared Distribution, John Wiley and Sons, New York.

Lehmann, E. L. (1966), Some Concepts of Dependence, Ann. Math. Statist., 37, 1137-1153.

Mardia, K. V. (1970), Families of Bivariate Distributions, #27, Griffin's Statistical Monographs and Courses, (Ed.) Alan Stuart, Charles Griffin & Co., London.

Montmort, P. R. de (1708), Essay d'Analyse Sur les Jeux des Hazards, 1st ed. Paris.

Newman, C. M. (1982), Asymptotic Independence and Limit Theorems for Positively and Negatively Dependent Random Variables, Inequalities in Statistics and Probability, (Ed.) Tong, Y. L.

Radner et al. (1980), Report on Exact and Statistical Matching Techniques, Statistical Policy Working Paper 5, Office of Federal Statistical Policy and Standards, U.S. Dept. of Commerce.

Randles, R. H. and Wolfe, D. A. (1979), Introduction to the Theory of Nonparametric Statistics, John Wiley and Sons, New York.

Rodgers, W. L. (1984), An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, 91-102.

Schweizer, B. and Wolff, E. F. (1981), On Nonparametric Measures of Dependence for Random Variables, Annals of Statistics, 9, 879-885.

Serfling, R. J. (1980), Approximation Theorems of Mathematical Statistics, John Wiley & Sons, Inc., New York.

Shaked, M. (1979), Some Concepts of Positive Dependence for Bivariate Interchangeable Distributions, Ann. Inst. Statist. Math. 31, Part A, 67-84.

Sims, C. A. (1972), "Comments" (On Okner 1972), Annals of Economic and Social Measurement, 1, 343-345.

Sims, C. A. (1978), "Comments" (On Kadane 1978), 1978 Compendium of Tax Research, Office of Tax Analysis, Dept. of the Treasury, Washington, D.C.: U.S. Govt. Printing Office, 172-177.

Srivastava, M. S. and Khatri, C. G. (1979), An Introduction to Multivariate Statistics, Elsevier North Holland, New York.

Tong, Y. L. (1980), Probability Inequalities in Multivariate Distributions, Academic Press, New York.

Widder, D. V. (1941), The Laplace Transform, Princeton University Press.

Woodbury, M. A. (1983), Statistical Record Matching for Files, Incomplete Data in Samples Surveys, Vol. 3, (Eds.) Madow, W. G. et al, 173-202.

Yahav, J. A. (1982), On Matchmaking, Statistical Decision Theory and Related Topics III, vol. 2 (Eds.) S. S. Gupta and J. O. Berger), New York: Academic Press, 497-504.

Zionts, S. (1974), Linear and Integer Programming, Prentice-Hall, Englewood Cliffs, N. J.

Zolutuchina, L. A. and Latishev, K. P. (1978), Asymptotic Behavior of the Expected Number of Coincidences of Elements in a Sequence of Bivariate Samples (in Russian), Leningrad Older. Matimal. Inst., Akad. Nauk SSSR, 79, 4-10.