The Annals of Statistics
1975, Vol. 3, No. 6, 1364-1369

ON RE-PAIRING OBSERVATIONS IN A BROKEN RANDOM SAMPLE

BY PREM K. GOEL

Purdue University

It is assumed that a random sample of size n is drawn from a bivariate distribution $f(t, u)$ which possesses a monotone likelihood ratio (MLR). However, before the sample values are observed, the pairs are 'broken' into components t and u. Therefore, the original sample pairings are unknown, and it is desired to optimally re-pair t- and u-values in order to reconstruct the original bivariate sample. It is observed that for the maximum likelihood pairing (MLP) to be the 'natural' pairing for all t- and u-values, it is necessary that f has MLR. It is shown that if it is desired to maximize the expected number of correct matches, then the class of procedures $\Phi_{1,n}$, which result in pairing the largest t with the largest u and the smallest t with the smallest u, is a *complete class*. A sufficient condition under which the MLP maximizes the expected number of correct matches is also obtained.

1. Introduction and summary. An interesting version of the matching problem was introduced by DeGroot, Feder and Goel [2]. The authors assumed that a random sample of size n is drawn from an infinite bivariate population with pdf $f(t, u)$. However, before the sample values can be observed, each observation vector (t_i, u_i) gets broken into two separate components, namely t_i and u_i, and these observations are available only in the form $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$, where

$$x_1 < x_2 < \cdots < x_n \quad \text{and} \quad y_1 < y_2 < \cdots < y_n \tag{1.1}$$

are the ordered values of $t_1, \cdots, t_n$ and $u_1, \cdots, u_n$ respectively. As a consequence the pairings in the original sample are not known. The observed values are called a *broken random sample* from the given bivariate population. The general problem considered in [2] is to re-pair the observed values $x_1, \cdots, x_n$ with the observed values $y_1, \cdots, y_n$ so as to reproduce most of the vectors (t_i, u_i) from the original sample. Two optimality criteria are suggested: (i) to maximize the probability of correct pairing of all n observations; (ii) to maximize the expected number of correctly matched pairs. Let Φ be the set of all permutations $\phi = (\phi(1), \cdots, \phi(n))$ of the integers $1, 2, \cdots, n$, and let us assume that re-pairing according to the permutation ϕ is to pair x_i with $y_{\phi(i)}$, $i = 1, 2, \cdots, n$. It is desirable to find a permutation ϕ^* which is optimal according to one of the criteria. In [2], it is assumed that the joint distribution of $f(t, u)$ is given by

$$f(t, u) = a(t)b(u) \exp(ut) \quad (u, t) \in R^2 .$$

Received April 1974; revised April 1975.

AMS 1970 *subject classifications*. Primary 62C07; Secondary 62P99.

Key words and phrases. Broken random sample, complete class, matching, monotone likelihood ratio, monotone likelihood pairing.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

The following results, among others, are then established:

(i) The maximum likelihood permutation (MLP) φ^* is given by the 'natural' pairing $\varphi^*(i) = i$ (i.e., pair x_i with y_i) for $i = 1, \cdots, n$.

(ii) The MLP maximizes the probability of correct pairing of all n observations.

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

(iv) Let $\Phi_{1,n}$ denote the class of all permutation ϕ, such that $y_{\phi(1)} = y_1$ and $y_{\phi(n)} = y_n$, and let $M(\phi)$ denote the expected number of correct matches if the permutation ϕ is used to re-pair the observations. (An expression for $M(\phi)$ is given in (5.1) of [2].). Then $\Phi_{1,n}$ is a complete class of permutations; i.e. given any permutation $\phi \notin \Phi_{1,n}$, there exists a $\psi \in \Phi_{1,n}$ which is *as good as* ϕ, i.e., $M(\psi) \geqq M(\phi)$. Sufficient conditions for the MLP to maximize $M(\phi)$ are also given.

Chew [1] extends some of the discussion in [2] to a class of distributions $f(t, u)$ possessing a monotone likelihood ratio (MLR), i.e., for all $t_1 < t_2$ and $u_1 < u_2$, $g(t_1, t_2; u_1, u_2) \geqq 0$ where

$$g(a, b; c, d) = f(a, c)f(b, d) - f(a, d)f(b, c)\,. \tag{1.2}$$

The following results, among others, are established in [1]:

(a) The results in (i) and (ii) above hold.

(b) For a trinomial distribution, the class $\Phi_{1,n}$ is complete. A sufficient condition, similar to the one in [2], for the MLP to maximize $M(\phi)$ is also given.

Various extensions of these results are presented in Section 2 of this paper. In particular, it is proved that for an arbitrary f with MLR, the class $\Phi_{1,n}$ is complete. In Section 3, a set of sufficient conditions for the MLP to maximize $M(\phi)$ is obtained. Thus, the results in this paper extend and complete the results (iv) and (b) obtained in [2] and [1].

2. Extensions of previous results. The extension of the result (iii) above to the class of distributions with MLR is trivially obtained by replacing the factor $\exp(x_i y_k)$ by $f(x_i, y_k)$ in the proof given in [2]. It can also be proved that if the MLP is the 'natural' pairing for all possible sets of values $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$, then f has MLR.

We shall now show that if f has MLR, then the class $\Phi_{1,n}$ is complete. Let $\phi \in \Phi$ be a fixed permutation such that $x_i < x_j$ *and* $y_{\phi(i)} > y_{\phi(j)}$ for some integers i and j. Let the permutation $\psi \in \Phi$ and the function $A(x, y)$ be defined by

$$\psi(i) = \phi(j)\,, \quad \psi(j) = \phi(i) \quad \text{and} \quad \psi(k) = \phi(k) \quad \text{for all other} \quad k\,, \tag{2.1}$$

and

$$A(x, y) = f(x_i, y)g(x, x_j; y_{\phi(j)}, y_{\phi(i)}) + f(x_j, y)g(x_i, x; y_{\phi(j)}, y_{\phi(i)}) \tag{2.2}$$

where the function g is given by (1.2). Then, by Theorem 1 of [1], a sufficient

condition for $M(\psi) \geqq M(\phi)$ to hold is that $A(x_h, y_q) \geqq 0$ for all $x_h \notin [x_i, x_j]$ and $y_q \notin [y_{\phi(j)}, y_{\phi(i)}]$. We relax this condition and obtain the following stronger results.

THEOREM 1. *Let ϕ be a permutation satisfying $x_i < x_j$* **and** *$y_{\phi(i)} > y_{\phi(j)}$ for some i and j, and let the permutation ψ be given by* (2.1). *If*

$$x_j = x_n \quad \text{or} \quad y_{\phi(j)} = y_1 \quad \text{or} \quad A(x_n, y_1) \geqq 0 \tag{2.3}$$

and

$$x_i = x_1 \quad \text{or} \quad y_{\phi(i)} = y_n \quad \text{or} \quad A(x_1, y_n) \geqq 0\,, \tag{2.4}$$

then the permutation ψ is as good as ϕ, i.e. $M(\psi) \geqq M(\phi)$.

PROOF. First observe that $A(x, y)$ can also be expressed as

$$A(x, y) = f(x, y_{\phi(j)})g(x_i, x_j; y, y_{\phi(i)}) + f(x, y_{\phi(i)})g(x_i, x_j; y_{\phi(j)}, y)\,. \tag{2.5}$$

If we use the assumption that $f(t, u)$ has MLR and let

$$A^*(x, y) = A(x, y)/f(x_j, y)\,, \qquad A^{**}(x, y) = A(x, y)/f(x, y_{\phi(i)})\,, \tag{2.6}$$

then it follows that

Fact 1. $A(x, y) \geqq 0 \Leftrightarrow A^*(x, y) \geqq 0 \Leftrightarrow A^{**}(x, y) \geqq 0$,

Fact 2. $f(x_i, y)/f(x_j, y)$ is a nonincreasing function of y,

and

Fact 3. $f(x, y_{\phi(i)})/f(x, y_{\phi(j)})$ is a nondecreasing function of x.

We shall partition the region $x_h \notin [x_i, x_j]$ and $y_q \notin [y_{\phi(j)}, y_{\phi(i)}]$ into four rectangles, namely

$$\begin{aligned} R_1 &= \{x_1 \leqq x < x_i, y_1 \leqq y < y_{\phi(j)}\}\,, \\ R_2 &= \{x_j < x \leqq x_n, y_{\phi(i)} < y \leqq y_n\}\,, \\ R_3 &= \{x_j < x \leqq x_n, y_1 \leqq y < y_{\phi(j)}\}\,, \qquad \text{and} \\ R_4 &= \{x_1 \leqq x < x_i, y_{\phi(i)} < y \leqq y_n\}\,. \end{aligned}$$

Note. If $x_j = x_n$ or $y_{\phi(j)} = y_1$ then R_3 is a null set; and if $x_i = x_1$ or $y_{\phi(i)} = y_n$, then R_4 is a null set.

Case I. $(x_h, y_q) \in R_1$.

Since $g(x_h, x_j; y_{\phi(j)}, y_{\phi(i)}) \geqq 0$ for all $(x_h, y_q) \in R_1$, it follows from Fact 2, (2.2), and (2.6) that

$$A^*(x_h, y_q) \geqq A^*(x_h, v_{\phi(j)})\,. \tag{2.7}$$

However,

$$\begin{aligned} A^*(x_h, y_{\phi(j)}) = \{f(x_j, y_{\phi(i)})/f(x_j, y_{\phi(j)})\}f(x_i, y_{\phi(j)})f(x_h, v_{\phi(j)}) \\ - f(x_i, y_{\phi(i)})f(x_h, y_{\phi(j)})\,. \end{aligned} \tag{2.8}$$

Together, (2.7), (2.8), and Fact 3 imply that

$$A^*(x_h, y_q) \geqq f(x_i, y_{\phi(i)})f(x_h, y_{\phi(j)}) - f(x_i, y_{\phi(i)})f(x_h, y_{\phi(j)}) = 0\,.$$

Therefore, from Fact 1, it follows that the condition $A(x_h, y_q) \geqq 0$ holds automatically for all $(x_h, y_q) \in R_1$.

Case II. $(x_h, y_q) \in R_2$.

Since $g(x_h, x_j; y_{\phi(j)}, y_{\phi(i)}) \leqq 0$ for all $(x_h, y_q) \in R_2$, it follows from Fact 2, (2.2), and (2.6) that

$$A^*(x_h, y_q) \geqq A^*(x_h, y_{\phi(i)}) . \tag{2.9}$$

However,

$$\begin{aligned} A^*(x_h, y_{\phi(i)}) &= f(x_i, y_{\phi(j)})f(x_h, y_{\phi(i)}) \\ &\quad - \{f(x_j, y_{\phi(j)})/f(x_j, y_{\phi(i)})\}f(x_i, y_{\phi(i)})f(x_h, y_{\phi(i)}) . \end{aligned} \tag{2.10}$$

Together, (2.9), (2.10) and Fact 3 imply that

$$A^*(x_h, y_q) \geqq f(x_i, y_{\phi(j)})f(x_h, y_{\phi(i)}) - f(x_i, y_{\phi(j)})f(x_h, y_{\phi(i)}) = 0 .$$

Therefore, from Fact 1, it follows that the condition $A(x_h, y_q) \geqq 0$ holds automatically for all $(x_h, y_q) \in R_2$.

Case III. $(x_h, y_q) \in R_3$.

Since $g(x_h, x_j; y_{\phi(j)}, y_{\phi(i)}) \geqq 0$, for all $(x_h, y_q) \in R_3$, it follows from Fact 2, (2.2), and (2.6) that

$$A^*(x_h, y_q) \geqq A^*(x_h, y_1) . \tag{2.11}$$

However, $(x_h, y_1) \in R_3$ and, therefore, it follows from Fact 1 and (2.11) that the condition $A(x_h, y_1) \geqq 0$ for all $x_h \in (x_j, x_n]$ is equivalent to $A(x_h, y_q) \geqq 0$ for all $(x_h, y_q) \in R_3$. But, $g(x_i, x_j; y_1, y_{\phi(i)}) \geqq 0$. Therefore, Fact 3, (2.5) and (2.6) together imply that

$$A^{**}(x_h, y_1) \geqq A^{**}(x_n, y_1) . \tag{2.12}$$

However, $(x_n, y_1) \in R_3$ and, therefore, if follows from Fact 1 and (2.12) that the condition $A(x_h, y_q) \geqq 0$ holds for all $(x_h, y_q) \in R_3$ iff $A(x_n, y_1) \geqq 0$.

Case IV. $(x_h, y_q) \in R_4$. Using an argument similar to Case III, it can be proved that the condition $A(x_h, y_q) \geqq 0$ holds for all $(x_h, y_q) \in R_4$ iff $A(x_1, y_n) \geqq 0$. The result in Theorem 1 follows from the above discussion and Theorem 1 of [1].

Now the completeness property of the class $\Phi_{1,n}$ of permutations follows from the next theorem.

THEOREM 2. *For each permutation* $\phi \notin \Phi_{1,n}$, $\exists\, \phi^* \in \Phi_{1,n}$ *such that* $M(\phi^*) \geqq M(\phi)$.

PROOF. Let $\phi \notin \Phi_{1,n}$ be given. If $y_{\phi(1)} = y_1$, let $\psi = \varphi$. However, if $y_{\phi(1)} > y_1$ and $y_{\phi(j)} = y_1$ for some j, then choose the ψ defined by (2.2) for $i = 1$. Now, if $x_j = x_1$, then $M(\psi) = M(\varphi)$ and if $x_j > x_1$, then by Theorem 1, $M(\psi) > M(\phi)$. If ψ satisfies $y_{\psi(n)} = y_n$, then let $\psi^* = \psi$. However, if $y_{\psi(n)} < y_n$, and $y_{\psi(i)} = y_n$

for some $i < n$, then obtain the ϕ^*, which is as good as ϕ, by applying Theorem 1. Now $\phi^* \in \Phi_{1,n}$ and $M(\phi^*) \geqq M(\phi)$.

COROLLARY 1. *Let $\varphi^{**} \in \Phi$ be a permutation such that $M(\varphi^{**}) = \max_{\phi \in \Phi} M(\phi)$. Then $y_{\varphi^{**}(1)} = y_1$ and $y_{\varphi^{**}(n)} = y_n$.*

COROLLARY 2. *If $n = 3$, then φ^{**} is the* MLP φ^*.

3. Sufficient conditions for MLP to maximize $M(\phi)$. We shall now obtain conditions, similar to (6.8), (6.9) and (6.10) of [2], under which the MLP maximizes $M(\varphi)$ for every $n > 3$.

Let us define

$$\lambda_{rs}(x, y) = \{f(x_r, y_s)f(x, y)/f(x_r, y)f(x, y_s)\}, \tag{3.1}$$

and

$$A_{rs}(a, b; c, d) = \lambda_{rs}(b, d) - \lambda_{rs}(b, c) - \lambda_{rs}(a, d) + \lambda_{rs}(a, c). \tag{3.2}$$

On multiplying $A(x, y)$ by $f(x, y)/\{f(x, y_{\phi(i)})f(x, y_{\phi(j)})f(x_i, y)f(x_j, y)\}$, it follows that the last conditions in (2.3) and (2.4) are equivalent to

$$A_{n1}(x_i, x_j; y_{\phi(j)}, y_{\phi(i)}) \geqq 0, \quad \text{and} \quad A_{1n}(x_i, x_j; y_{\phi(j)}, y_{\phi(i)}) \geqq 0. \tag{3.3}$$

The following lemma is a consequence of Theorem 1 and Theorem 2.

LEMMA 1. *If $A_{n1}(a, b; c, d) \geqq 0$ for all $x_2 \leqq a < b < x_n$ and $y_1 < c < d \leqq y_{n-1}$, and $A_{1n}(a, b; c, d) \geqq 0$ for all $x_1 < a < b \leqq x_{n-1}$ and $y_2 \leqq c < d < y_n$, then the* MLP φ^* *maximizes $M(\phi)$.*

Let us now assume that $f(t, u)$ has second order partial derivatives and observe that $A_{rs}(a, b; c, d)$ is a second mixed difference of the function $\lambda_{rs}(x, y)$. The next theorem is a consequence of the above lemma.

THEOREM 3. *If $(\partial^2/\partial x\, \partial y)\lambda_{n1}(x, y) \geqq 0$ for all x and y such that $x_2 \leqq x < x_n$ and $y_1 < y \leqq y_{n-1}$ and if $(\partial^2/\partial x\, \partial y)\lambda_{1n}(x, y) \geqq 0$ for all x and y such that $x_1 < x \leqq x_{n-1}$ and $y_2 \leqq y < y_n$, then the* MLP φ^* *maximizes $M(\phi)$.*

The proof is similar to that of Theorem 5 of [2] and is omitted.

COROLLARY 3. *If $(\partial^2/\partial x\, \partial y)\lambda_{n1}(x, y)$ and $(\partial^2/\partial x\, \partial y)\lambda_{1n}(x, y)$ are nonnegative for all (x, y) such that $x_1 < x < x_n$ and $y_1 < y < y_n$, then the* MLP φ^* *maximizes $M(\phi)$.*

REMARK. 1. As noted in [2], the sufficient conditions given in Theorem 3 and Corollary 3 are typically more restrictive then is necessary.

2. It is conjectured that there exists no distribution for which the conditions in Theorem 3 hold for all rectangles (x_1, x_n), (y_1, y_n). However, it is obvious from the examples considered in [1] and [2], that these conditions do hold for some data sets.

EXAMPLE. An example of a pdf with MLR, not satisfying the condition in

[2], is the bivariate logistic distribution, Gumbel [3], given by

$$f(t, u) = 2\exp(-t - u)[1 + \exp(-t) + \exp(-u)]^{-3}$$
$$-\infty < t < \infty, \quad -\infty < u < \infty.$$

For this pdf, the sufficient condition, given in Corollary 3, reduces to

$$\frac{(\exp(-x_1) - \exp(-x_n))(\exp(-y_1) - \exp(-y_n))}{(1 + \exp(-x_1) + \exp(-y_1))(1 + \exp(-x_n) + \exp(-y_n))} < \tfrac{1}{3}.$$

Acknowledgment. I wish to thank the referees for useful comments and suggestions which have improved the presentation of the material.

REFERENCES

[1] CHEW, M. C., JR. (1973). On pairing observations from a distribution with monotone likelihood ratio. *Ann. Statist.* **1** 433–445.

[2] DEGROOT, M. H., FEDER, P. I. and GOEL, P. K. (1971). Matchmaking. *Ann. Math. Statist.* **42** 578–593.

[3] GUMBEL, E. J. (1961). Bivariate logistic distribution. *J. Amer. Statist. Assoc.* **56** 335–349.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907