The Annals of Statistics
1996, Vol. 24, No. 3, 1069–1074

REDUCING MULTIDIMENSIONAL TWO-SAMPLE DATA TO ONE-DIMENSIONAL INTERPOINT COMPARISONS[1]

BY JEN-FUE MAA, DENNIS K. PEARL AND ROBERT BARTOSZYŃSKI

Corning Hazelton, Inc. and Ohio State University

The most popular technique for reducing the dimensionality in comparing two multidimensional samples of $\mathbf{X} \sim F$ and $\mathbf{Y} \sim G$ is to analyze distributions of interpoint comparisons based on a univariate function h (e.g. the interpoint distances). We provide a theoretical foundation for this technique, by showing that having both i) the equality of the distributions of within sample comparisons ($h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2)$) and ii) the equality of these with the distribution of between sample comparisons (($h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{X}_3, \mathbf{Y}_3)$) is equivalent to the equality of the multivariate distributions ($F = G$).

1. Introduction. The distribution-free comparison of two high-dimensional samples has attracted substantial interest over the last 20 years. Many authors have reduced this problem to the one-dimensional comparison of interpoint distances. For example, Friedman and Rafsky (1979) proposed a two-sample test based on the minimal spanning tree formed from the interpoint distances; Atkinson (1989) compared a high-dimensional simulated sample with an observed sample, using interpoint distances within samples. Other distance-related tests for high-dimensional data were proposed by Schilling (1986) and Henze (1988).

Our paper will provide a theoretical foundation for this common technique of examining interpoint distances to give a reduction of dimensionality. In fact, the reduction of dimensionality can be based on univariate functions other than distances. In Section 2 we provide the basic theorem, separately for the discrete and continuous cases. In Section 3 we provide a brief description of two simulation-based applications.

2. Main result. In the sequel, we let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ be iid random variables from k-dimensional distributions F and G, respectively. We assume that $\mathbf{X}_i$ and $\mathbf{Y}_j$ are independent for all i, j. Also, we use the symbol $=_{\mathscr{L}}$ to denote the equality of distributions.

THEOREM 1. *Let S_1 and S_2 be two arbitrary countable sets, and let $\mathbf{X}$ and $\mathbf{Y}$ be random variables with values in S_1 and S_2, respectively. If $h(\mathbf{x}, \mathbf{y})$ is any real-valued nonnegative function on $S_1 \times S_2$ such that $h(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$,*

Received June 1993; revised July 1995.
[1]This material is based on work supported by the Cooperative State Research Service, U.S. Department of Agriculture under agreement 91-37302-6745.
AMS 1991 *subject classification*. Primary 62H05.
Key words and phrases. Characterization of distributional equality, multivariate, distances.

then

$$h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_3, \mathbf{Y}_3) \quad \textit{iff } F = G.$$

PROOF. It is trivial that $F = G$ implies $h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_3, \mathbf{Y}_3)$, so one needs only to prove the converse. From $h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_3, \mathbf{Y}_3)$, we have

$$P\big(h(\mathbf{X}_1, \mathbf{X}_2) \le t\big) = P\big(h(\mathbf{Y}_1, \mathbf{Y}_2) \le t\big) = P\big(h(\mathbf{X}_3, \mathbf{Y}_3) \le t\big) \quad \text{for all } t \ge 0.$$

In particular, when $t = 0$,

$$P(\mathbf{X}_1 = \mathbf{X}_2) = P(\mathbf{Y}_1 = \mathbf{Y}_2) = P(\mathbf{X}_3 = \mathbf{Y}_3) \tag{1}$$
$$\text{since } h(\mathbf{x}, \mathbf{y}) = 0 \text{ implies } \mathbf{x} = \mathbf{y}.$$

Let $S_1 = \{\mathbf{x}_i, i = 1, 2, \ldots\}$ and $S_2 = \{\mathbf{y}_i, i = 1, 2, \ldots\}$. Let the probability mass function of $\mathbf{X}_1$ and $\mathbf{Y}_1$ be

$$P(\mathbf{X} = \mathbf{x}_i) = \pi_i, \qquad i = 1, 2, \ldots$$

and

$$P(\mathbf{Y} = \mathbf{y}_i) = \gamma_i, \qquad i = 1, 2, \ldots,$$

with $\sum \pi_i = \sum \gamma_i = 1$. Suppose that $S_1 \cap S_2$ consists of the matched pairs $\mathbf{x}_{[j]} = \mathbf{y}_{[j]}$. Thus, from (1)

$$\begin{aligned} P(\mathbf{X}_1 = \mathbf{X}_2) &= \sum_{\mathbf{x}_j \in S_1} \pi_j^2 = P(\mathbf{Y}_1 = \mathbf{Y}_2) = \sum_{\mathbf{y}_j \in S_2} \gamma_j^2 = P(\mathbf{X}_3 = \mathbf{Y}_3) \\ &= \sum_{\mathbf{x}_{[j]} \in S_1 \cap S_2} \pi_{[j]} \gamma_{[j]}, \end{aligned} \tag{2}$$

and it follows that $S_1 \cap S_2 \neq \varnothing$ since $P(\mathbf{X}_1 = \mathbf{X}_2) > 0$ from the assumption of discreteness. Omitting the terms from outside $S_1 \cap S_2$ and using the Cauchy inequality, one has

$$\sum_{\mathbf{x}_j \in S_1} \pi_j^2 \sum_{\mathbf{y}_j \in S_2} \gamma_j^2 \ge \sum_{\mathbf{x}_{[j]} \in S_1 \cap S_2} \pi_{[j]}^2 \sum_{\mathbf{y}_{[j]} \in S_1 \cap S_2} \gamma_{[j]}^2 \ge \left\{ \sum_{\mathbf{x}_{[j]} \in S_1 \cap S_2} \pi_{[j]} \gamma_{[j]} \right\}^2. \tag{3}$$

Since (2) implies that the first and last terms of (3) are equal, we obtain

$$\sum_{\mathbf{x}_j \in S_1} \pi_j^2 \sum_{\mathbf{y}_j \in S_2} \gamma_j^2 = \sum_{\mathbf{x}_{[j]} \in S_1 \cap S_2} \pi_{[j]}^2 \sum_{\mathbf{y}_{[j]} \in S_1 \cap S_2} \gamma_{[j]}^2 = \left\{ \sum_{\mathbf{x}_{[j]} \in S_1 \cap S_2} \pi_{[j]} \gamma_{[j]} \right\}^2.$$

Since the π's and γ's are nonnegative, only $\pi_{[j]}, \gamma_{[j]}$ for $\mathbf{x}_{[j]} \in S_1 \cap S_2$ can be positive. In addition, equality in the Cauchy formula means $\pi_{[j]} = c\gamma_{[j]}$ for $\mathbf{x}_{[j]} \in S_1 \cap S_2$, and c must be equal to 1 since $\sum_{\mathbf{x}_{[j]} \in S_1 \cap S_2} \pi_{[j]} = 1$ and $\sum_{\mathbf{y}_{[j]} \in S_1 \cap S_2} \gamma_{[j]} = 1$. □

The proof above relies on the fact that the probability of two iid discrete random variables assuming the same value is positive. Therefore, it cannot be applied to the continuous case. However, the theorem is also proved in the continuous case with some restriction on the density function. We start from the following lemma.

LEMMA 1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *have densities* f *and* g *with* $\int f^2(\mathbf{x})\,d\mathbf{x} < \infty$ *and* $\int g^2(\mathbf{y})\,d\mathbf{y} < \infty$, *respectively. Assume that* $\mathbf{0}$ *is a Lebesgue point of the function* $u(\mathbf{y}) = \int_{\mathbb{R}^k} g(\mathbf{x}+\mathbf{y})f(\mathbf{x})\,d\mathbf{x}$. *Let* $h\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ *be a nonnegative continuous function such that* $h(\mathbf{x},\mathbf{y}) = 0$ *iff* $\mathbf{x} = \mathbf{y}$, *and* $h(a\mathbf{x}+\mathbf{b}, a\mathbf{y}+\mathbf{b}) = |a|h(\mathbf{x},\mathbf{y})$, $\forall\, a \in \mathbb{R}$, $\forall\, \mathbf{b} \in \mathbb{R}^k$. *Then*

$$\lim_{t \downarrow 0} \frac{P(h(\mathbf{X},\mathbf{Y}) < t)}{t^k} = \alpha \int f(\mathbf{y})g(\mathbf{y})\,d\mathbf{y},$$

where $\alpha = \int_{\{h(\mathbf{x},\mathbf{0})<1\}} d\mathbf{x}$.

PROOF. We have

$$\begin{aligned} P(h(\mathbf{X},\mathbf{Y}) < t) &= \int_{\{\mathbf{y}:\, h(\mathbf{x},\mathbf{y})<t\}} \left[\int_{\mathbb{R}^k} g(\mathbf{y})f(\mathbf{x})\,d\mathbf{x}\right] d\mathbf{y} \\ &= \int_{\{\mathbf{y}:\, h(\mathbf{0},\mathbf{y})<t\}} u(\mathbf{y})\,d\mathbf{y} \end{aligned}$$

since $h(\mathbf{x},\mathbf{x}+\mathbf{y}) = h(\mathbf{0},\mathbf{y})$. Now, the function $u(\mathbf{y})$ is locally integrable in $\mathbb{R}^k$ since $f, g \in L^2$. Also, our assumptions about h imply that as $t \downarrow 0$ the sets $\{\mathbf{y}: h(\mathbf{0},\mathbf{y}) < t\}$ shrink regularly to $\mathbf{0}$. Thus [see, e.g., Theorem 7.16 of Wheeden and Zygmund (1977)], we obtain

$$\lim_{t \downarrow 0} \frac{\int_{\{\mathbf{y}:\, h(\mathbf{0},\mathbf{y})<t\}} |u(\mathbf{y}) - u(\mathbf{0})|\,d\mathbf{y}}{\int_{\{\mathbf{y}:\, h(\mathbf{0},\mathbf{y})<t\}} d\mathbf{y}} = 0.$$

Next,

$$\int_{\{\mathbf{y}:\, h(\mathbf{0},\mathbf{y})<t\}} d\mathbf{y} = \alpha t^k$$

(note that $\alpha < \infty$). Therefore,

$$\lim_{t \downarrow 0} \frac{\int_{\{\mathbf{y}:\, h(\mathbf{0},\mathbf{y})<t\}} u(\mathbf{y})\,d\mathbf{y}}{t^k} = \alpha u(\mathbf{0}) = \alpha \int f(\mathbf{y})g(\mathbf{y})\,d\mathbf{y}. \qquad \square$$

REMARK 1. The assumption of Lemma 1 regarding $u(\mathbf{0})$ will hold if, for example, g is bounded or continuous. Further, since the role of f and g can be reversed in the proof, it is enough to have such a condition hold for either density.

THEOREM 2. *Let* $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ *be iid* k*-dimensional random variables with density* f *and cdf* F *and let* $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ *be iid* k*-dimensional random variables with density* g *and cdf* G, *and suppose that the* $\mathbf{X}$'s *and* $\mathbf{Y}$'s *are independent. If the densities* f *and* g *and the function* h *all satisfy the conditions of Lemma* 1, *then*

$$h(\mathbf{X}_1,\mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1,\mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_3,\mathbf{Y}_3) \quad \textit{iff } F = G. \tag{4}$$

PROOF. It is trivial that $F = G$ implies $h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_3, \mathbf{Y}_3)$, so one needs only to prove the converse. From

$$h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_3, \mathbf{Y}_3),$$

one has

$$P\big(h(\mathbf{X}_1, \mathbf{X}_2) < t\big) = P\big(h(\mathbf{Y}_1, \mathbf{Y}_2) < t\big) = P\big(h(\mathbf{X}_3, \mathbf{Y}_3) < t\big) \quad \text{for any } t \geq 0.$$

Therefore,

$$\lim_{t \downarrow 0} \frac{P\big(h(\mathbf{X}_1, \mathbf{X}_2) < t\big)}{t^k} = \lim_{t \downarrow 0} \frac{P\big(h(\mathbf{Y}_1, \mathbf{Y}_2) < t\big)}{t^k} = \lim_{t \downarrow 0} \frac{P\big(h(\mathbf{X}_3, \mathbf{Y}_3) < t\big)}{t^k}.$$

From Lemma 1 with $f = g$ for the first two lines, we have

$$\lim_{t \downarrow 0} \frac{P\big(h(\mathbf{X}_1, \mathbf{X}_2) < t\big)}{t^k} = \alpha \int f^2(\mathbf{x})\, d\mathbf{x},$$

$$\lim_{t \downarrow 0} \frac{P\big(h(\mathbf{Y}_1, \mathbf{Y}_2) < t\big)}{t^k} = \alpha \int g^2(\mathbf{x})\, d\mathbf{x},$$

$$\lim_{t \downarrow 0} \frac{P\big(h(\mathbf{X}_3, \mathbf{Y}_3) < t\big)}{t^k} = \alpha \int f(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x}.$$

Therefore, since $0 < \alpha < \infty$,

$$\int f^2(\mathbf{x})\, d\mathbf{x} = \int g^2(\mathbf{x})\, d\mathbf{x} = \int f(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x}. \tag{5}$$

From the Schwarz inequality, one has

$$\left[\int f^2(\mathbf{x})\, d\mathbf{x}\right]\left[\int g^2(\mathbf{x})\, d\mathbf{x}\right] \geq \left(\int f(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x}\right)^2. \tag{6}$$

But (5) shows that we have equality in (6). Thus, since f and g are both density functions, they must be identical a.e. □

REMARK 2. Combining Theorems 1 and 2 shows that our main result is true for a wider class of situations. For example, we can allow for mixtures of continuous and discrete distributions, with any function h satisfying the conditions of Theorem 2. Also, in Theorem 2 we can widen the class of h's by allowing continuous monotone functions of h's satisfying the conditions of the theorem. In fact, we believe (4) is true for all distributions F and G and every h which is a function of the Euclidean metric.

REMARK 3. None of the equations in (4) can be dropped. For example, if $X \equiv 0$, $Y = 0$ or 1 with probability $\frac{1}{2}$, then $h(Y_1, Y_2) =_{\mathscr{L}} h(X_1, Y_3)$, but $F \neq G$. However, when $h(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq k} |x_i - y_i|$, then $h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_1, \mathbf{Y}_3)$ does imply $F = G$ under some mild additional restrictions on the characteristic functions of $\mathbf{X}$ and $\mathbf{Y}$.

REMARK 4. The proofs of Theorems 1 and 2 did not require the independence of all three interpoint distances. In particular, we also have

$$h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_1, \mathbf{Y}_1) \quad \text{iff } F = G.$$

This may be useful in applications since it allows **X** to **X**, **Y** to **Y** and **X** to **Y** distances to be computed from the same reference points.

3. Application. The results of this paper were motivated by a multidimensional simulation-based estimation problem [see Pearl, Bartoszyński and Horn (1989) and Maa (1993)]. In this context, we had a number of experimental data points (**X**), each being highly dimensional (observations following the sample path of a stochastic multivariate prey–predator process). Realistic modeling required a large number of parameters, making the likelihood accessible only through extensive computer simulations of sample points (**Y**). Estimation of the parameters was based on "making the simulated **Y**'s as similar to the experimental **X**'s as possible." This was done by optimizing a goodness-of-fit criterion which attempts to align the univariate distributions of interdistances within the **X**-sample, within the **Y**-sample and between the **X**'s and **Y**'s.

Besides the goodness-of-fit type of problem described above, the theorem could be used to disprove conjectures about a possible equality of multidimensional distributions in situations where analytical results are difficult but a computer simulation of the distributions is possible. For example, consider two complicated queueing systems, where for some n, the distribution of the first n departure times is conjectured to be the same for the two systems. Suppose that it is easy to simulate the two systems and let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ represent their random simulated departure times. We may now choose a function h, such as $h(\mathbf{x}, \mathbf{y}) = \Sigma|x_i - y_i|$, and simulate pairs $(\mathbf{X}_1, \mathbf{X}_2)$, $(\mathbf{Y}_1, \mathbf{Y}_2)$ and $(\mathbf{X}_3, \mathbf{Y}_3)$. Then we might use any omnibus test of the equality of three distributions [e.g., David's (1958) three-sample Kolmogorov–Smirnov test] to test the null hypothesis that $h(\mathbf{X}_1, \mathbf{X}_2) =_{\mathscr{L}} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\mathscr{L}} h(\mathbf{X}_3, \mathbf{Y}_3)$. Rejection of this hypothesis is strong evidence against the conjecture. Importantly, the user has complete control over the power of the test through an increase in the number of simulations. If the conjecture is false (and h satisfies the assumptions of the theorem), then the test will lead to a rejection with probability 1 as the simulation size goes to ∞.

REFERENCES

ATKINSON, E. N., BROWN, B. W. and THOMPSON, J. R. (1989). Parallel algorithms for fixed seed simulation based parameter estimation. In *Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface* 259–261.

DAVID, H. T. (1958). A three-sample Kolmogorov–Smirnov test. *Ann. Math. Statist.* **29** 842–851.

FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717.

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783.

MAA, J. (1993). Simulation-based parameter estimation for multivariate distributions. Ph.D. dissertation, Dept. Statistics, Ohio State Univ.

PEARL, D. K., BARTOSZYŃSKI, R. and HORN, D. J. (1989). A stochastic model for simulation of interactions between phytophagous spider mites and their phytoseiid predators. *Exp. Appl. Acarol.* **7** 143–151.

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 779–806.

WHEEDEN, R. L. and ZYGMUND, A. (1977). *Measure and Integral: An Introduction to Real Analysis*. Dekker, New York.

J.-F. MAA
CORNING HAZELTON, INC.
PO BOX 7545
MADISON, WISCONSIN 53707

D. K. PEARL
R. BARTOSZYŃSKI
DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210
E-MAIL: pearl.1@osu.edu
bartoszynski.1@osu.edu