Mathematical Geology, Vol. 30, No. 3, 1998

A Bootstrap Test Using Maximum Likelihood Ratio Statistics to Check the Similarity of Two 3-Dimensionally Oriented Data Samples[1]

Sojen Joy[2] and Snigdhansu Chatterjee[3]

Comparing three-dimensionally oriented datasets is a problem encountered in various branches of earth science. A simple statistical tool for the comparison of two 3-dimensionally oriented datasets using the bootstrap method in line with the usual nonparametric permutation test is described here. This bootstrap test involves the estimation of maximum likelihood ratio statistic for properly constructed joint frequency tables of the datasets to be compared. This test does not use asymptotic result and will work well even for small sample sizes. Also this test does not make any specific distributional assumptions.

KEY WORDS: polar coordinates, joint frequency table, nonparametric permutation test, contingency table.

INTRODUCTION

Comparison of oriented data, be it planar or linear, is a common problem encountered in the earth science. A structural geologist might be interested in comparing the orientation of fold axes (or some other structural element) measured from two separate locations. Someone else might be interested in comparing the quartz C-axis fabric diagram prepared for two thin sections of the same specimen. A sedimentologist might want to compare bedding orientations from two locations. More examples of 3-dimensional data are given in Watson (1983). Visual inspection of the contoured equal area lower (or upper) hemi-

[1]Received 24 January 1997; revised 10 May 1997.

[2]Geological Studies Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India. e-mail-res056@isical.ernet.in

[3]Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India. e-mail-res9514@isical.ernet.in

0882-8121/98/0400-0275$15.00/1

sphere projection (Ghosh, 1993; Ramsay, 1967; Turner and Weiss, 1963) is a qualitative approach to the comparison problem. But as with any other qualitative method, interpretation can be never unique and will vary from person to person.

Available two-dimensional tests (Mardia, 1972) such as Kuiper's Kolmogrov type $V_{n1,n2}$ test (Kuiper, 1960), or Watson's two sample U^2 test (Watson, 1962), are not suitable for 3-dimensional data. Available two-sample test of concentration parameters and mean directions of 3-dimensional data assumes a Fisher distribution (point concentration) or a Bingham distribution (orthorhombic symmetry) (Mardia, 1972; Cheeney, 1983), to check whether the two samples are taken from the same population. This assumption is not always satisfied in many of the natural situations encountered. Two sample nonparametric tests proposed by Wellner (1979) and Jupp (1987) are rotationally invariant tests but are too difficult, if not impossible, to apply in practical situations.

Here we present a simple statistical tool for testing the similarity of two 3-dimensional datasets, following a similar approach to the test proposed by Dudley, Perkins, and Gine (1975). However, our methodology is very general, and can be applied to a wide range of test statistics. Only linear data is considered; planar data can be uniquely defined by its pole, which is linear.

THE DATA

Each piece of data, being a vector, has three parameters when represented as a spherical polar coordinate (r, θ, ϕ). But the radius parameter "r" is taken as unity (data points plotted on a unit sphere). θ in the geological case is the plunge direction, with a range from 0 to 2π, and ϕ is the plunge amount of any linear element, which will vary from 0 to $\pi/2$. The value of ϕ can vary from $-\pi/2$ to $\pi/2$ in some cases (e.g., in crystallography).

THE METHOD

Consider a joint frequency distribution table with a properly constructed θ and ϕ classes θ_is and ϕ_is. Any data point with a θ value falling in a particular θ class, say θ_i, and a ϕ class, say ϕ_j, will have a unique position within the two-way joint frequency distribution table. After considering all data points and completing the frequency distribution, each grid in the table, say γ_{ij}, will represent the number of data points with a θ value in the θ_i class and ϕ value in the ϕ_j class (Fig. 1). For axial data, points with a ϕ value of "zero" have to be entered in two cells corresponding to the two end points of the horizontal axis, and therefore the total number of data considered in the table might not be the original total.

↓ θ class φ class →	ϕ_1	ϕ_2	: : :	ϕ_j	: : :	ϕ_n
θ_1						
θ_2						
: : :						
θ_i				γ_{ij}		
: : :						
θ_m						

Figure 1. Frequency distribution table for 3-dimensional data. Each of the θ_is and ϕ_js represents a range of θ and ϕ values. γ_{ij} represents the total number of data points with θ values in the θ_i class and ϕ values in the ϕ_j class.

THE TEST

Consider the joint frequency distribution for two datasets X and Y (say x_{ij}s and y_{ij}s) over an m by n contingency tables A and B with n_1 and n_2 being the total number of observations for the samples X and Y, respectively. Add tables A and B to get another contingency table Z, where each of its elements (z_{ij}) are obtained as $z_{ij} = x_{ij} + y_{ij}$ with total number of observations being $n_1 + n_2$.

The sample probability of the ijth cell in the merged table is

$$p_{ij} = (x_{ij} + y_{ij})/(n_1 + n_2)$$
$$= (z_{ij})/(n_1 + n_2) \tag{1}$$

Similarly sample probability of the ijth cell in table A is

$$q_{ij} = x_{ij}/n_1 \tag{2}$$

and sample probability of the ijth cell in table B is

$$r_{ij} = y_{ij}/n_2 \tag{3}$$

Now let $\mathbf{x}$ be a random variable for table A. Therefore,

$$\mathbf{x} \sim \mathbf{M}(S_{11}, S_{12}, \ldots S_{mn})$$

where S_{ij}s are the probability that $\mathbf{x}$ is in the ijth cell of table A. Let $\mathbf{y}$ be another random variable for the second table, so

$$\mathbf{y} \sim \mathbf{M}(T_{11}, T_{12}, \ldots T_{mn})$$

where T_{ij}s are the probability that y is in the ijth cell of table B. Here $\mathbf{M}$ (. . .) denotes the multinomial distribution. Then our null hypothesis is

$$H_0\text{: } S_{ij} = T_{ij} \text{ for all } i \text{ and } j$$

and the alternative hypothesis is

$$H_1\text{: not } H_0$$

Assuming H_0, the likelihood function $L1$ of the data is

$$L1 = \frac{(n_1!)(n_2!)(\mathbf{S}_{11}^{x_{11}+y_{11}})(\mathbf{S}_{12}^{x_{12}+y_{12}}) \ldots (\mathbf{S}_{mn}^{x_{mn}+y_{mn}})}{(x_{11}!x_{12}! \ldots x_{mn}!)(y_{11}!y_{12}! \ldots y_{mn}!)} \tag{4}$$

Maximizing Equation (4) will give $\hat{S}_{ij} = p_{ij}$ with $(mn - 1)$ parameters estimated.

Not assuming H_0, the likelihood function $L2$ of the data is

$$L2 = \frac{n_1!\mathbf{S}_{11}^{x_{11}}\mathbf{S}_{12}^{x_{12}} \ldots \mathbf{S}_{mn}^{x_{mn}}}{x_{11}!x_{12}! \ldots x_{mn}!} * \frac{n_2!\mathbf{T}_{11}^{y_{11}}\mathbf{T}_{12}^{y_{12}} \ldots \mathbf{T}_{mn}^{y_{mn}}}{y_{11}!y_{12}! \ldots y_{mn}!} \tag{5}$$

Maximum likelihood solutions of Equation 5 are $\hat{S}_{ij} = q_{ij}$ and $\hat{T}_{ij} = r_{ij}$. Here we have $(mn - 1) + (mn - 1) = 2(mn - 1)$ parameters estimated. Now Δ, the likelihood ratio is given as

$$\Delta = \frac{p_{11}^{x_{11}+y_{11}} p_{12}^{x_{12}+y_{12}} \ldots p_{mn}^{x_{mn}+y_{mn}}}{(q_{11}^{x_{11}} q_{12}^{x_{12}} \ldots q_{mn}^{x_{mn}})(r_{11}^{y_{11}} r_{12}^{y_{12}} \ldots r_{mn}^{y_{mn}})}$$

and the log likelihood ratio is

$$\log \Delta = \sum_{i=1}^{m} \sum_{j=1}^{n} \{(x_{ij} + y_{ij}) \log p_{ij} - x_{ij} \log q_{ij} - y_{ij} \log r_{ij}\}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \{x_{ij}(\log p_{ij} - \log q_{ij}) + y_{ij}(\log p_{ij} - \log r_{ij})\} \tag{6}$$

THE LIKELIHOOD RATIO STATISTICS FOR TESTING H_0 VERSES H_1

One of the most frequently used tools for quantitatively testing hypotheses is the (maximum) likelihood ratio statistic; $\Delta = L1/L2$. The null hypothesis H_0 is rejected if Δ is not large enough (alternatively, if $\log \Delta$ is not large enough, since the natural logarithm is a monotonic function). However this comparison involves use of the distribution of the Δ or $\log \Delta$. The exact distribution is often intractable, as in the case of Equation 6 above. Usually a large sample approximation to the distribution of Δ or $\log \Delta$ is used. It has been shown that $-2 \log \Delta$, for large sample size, has approximately a χ^2 distribution, under H_0, with

$2(mn - 1) - (mn - 1)$ degrees of freedom (the number of parameters estimated not assuming H_0 − number of parameters estimated assuming H_0) (Bickel and Doksum, 1977).

Therefore, the size α likelihood ratio test rejects H_0 if

$$-2 \log \Delta > \chi^2 \, mn - 1, \, 1 - \alpha$$

This test is too stringent as it expects nearly equal cell frequencies for each cell of the grid to accept the null hypothesis, even if H_0 is true. This is difficult to ensure unless there is a large number of observations for each single cell of the grid for both samples. A relaxed test can be constructed by considering only those cells with probability in the combined table greater than a limit value (say by considering cells with the largest 10 probability values). Another relaxation can be made by constructing the grid in such a way to consider the points very near to the periphery of the net in two antipodal cells. The degrees of freedom will also be changed in proportion to the number of cells considered.

THE BOOTSTRAP TEST

The asymptotic test based on χ^2-distribution, although comparatively easy to use, is often very approximate, especially when sample size (n_1 or n_2) is not large enough compared with mn. As an alternative, a bootstrap test (Efron and Tibshirani, 1993) is easy to handle with the help of a computer. The bootstrap method has the following advantages over the test described earlier: (1) It does not use an asymptotic result and will work well even when the sample size is not very large. (2) It does not make specific distributional assumptions, whereas the earlier test assumes a multinomial distribution of the variables with unknown parameters. (3) Bootstrap results are almost always more accurate compared to asymptotic results (Efron and Tibshirani, 1993). The test developed here is similar to the usual nonparametric permutation test (Efron and Tibshirani, 1993).

The joint frequency distribution table constructed in the previous test is used here also. Let $\Omega = \{e_i, i = 1, 2, 3, \ldots, mn\}$ be the sample space where e_i is the vector of length mn (mn-tuple) with 1 at the ith place and zero elsewhere. Let W and U be two random vectors taking values in Ω. Assume $W \sim F$ and $U \sim G$. Then n_1 samples of W are taken, namely, $W_1, W_2, W_3, \ldots, W_{n1}$ and n_2 samples of U, namely, $U_1, U_2, \ldots, U_{n2}$. Each of W_i or U_i is an mn-tuple vector in the sample space Ω.

W_i and U_i can be written as

$$W_i = (W_{i1}, W_{i2}, \ldots, W_{mn}) \qquad i = 1, 2, \ldots, n_1$$

$$U_i = (U_{i1}, U_{i2}, \ldots, U_{mn}) \qquad i = 1, 2, \ldots, n_2$$

Order the entries of table A and table B in a linear manner to present the data in the tables as vectors of length mn. Thus, elements $va_{(k)}$ of vector (A) and $vb_{(k)}$ of vector (B) are obtained by transforming ijth entries of table A and table B as

$$a_{ij} \equiv va_{n(i-1)+j} \quad \text{and} \quad b_{ij} \equiv vb_{n(i-1)+j}$$

For notational convenience, we write vector (A) as a $= (a_1, a_2, \ldots, a_{mn})$ and vector (B) as $b = (b_1, b_2, \ldots b_{mn})$ where

$$a_k = \sum_{i=1}^{n_1} W_{ik}$$

and

$$b_k = \sum_{i=1}^{n_2} U_{ik}$$

Our hypothesis of interest is H_0: $F = G$.

We have presented this scheme for A and B being two-way tables. However, this result could easily be extended to k-way tables of dimensions $(m_1, m_2, \ldots, m_k)$ with any consistent transformation of the table entries into vectors of length $m_1 m_2 \ldots m_k$.

THE TEST

The general algorithm of the test is described here step by step. It should be noted that a small computer program will help to perform the test faster and with ease.

Step 1: Calculate the χ^2 statistic [$-2 \log \Delta$ described in the first test (Eq. 6)] of the datasets and call it T_0.

Step 2: Let a^* and b^* be null vectors of length mn.

Step 3: Fix

$$p = \frac{n_1}{n_1 + n_2}$$

This is the probability of selection from the vector (A).

Step 4: let

$$C_1 = \frac{a_1}{n_1}, \; C_2 = \frac{a_1 + a_2}{n_1}, \; C_3 = \frac{a_1 + a_2 + a_3}{n_1}, \text{ and}$$

$$C_{mn} = \frac{a_1 + a_2 + \cdots + a_{mn}}{n_1} = 1$$

such that

$$C_k - C_{k-1} = \frac{a_k}{n_1}$$

similarly construct D_ks such that $D_k - D_{k-1} = b_k/n_2$ for table B.

Step 5: Draw a random sample from Bernoulli(p) and call it X. Hint: select a random number (r) between zero and one, and if (r) is less than p then set X as zero, else assign one to X.

Step 6: Draw a random sample from $U(0, 1)$ and call it Y.

Step 7:

if $X = 0$ and $Y \in [C_{k-1}, C_k)$ then $a_k^* = a_k^* + 1$, otherwise

if $X = 1$ and $Y \in [D_{k-1}, D_k)$ then $a_k^* = a_k^* + 1$

Note that a^* is a mixed type random variable generated by the stochastic law from which the matrix A is drawn with probability $p = n_1/(n_1 + n_2)$ and by the stochastic law from which matrix B is drawn with probability $p = n_2/(n_1 + n_2)$. Under H_0, these two stochastic laws will be the same.

Repeat steps 5, 6, and 7 n_1 times to get the a^* vector and similarly repeat the steps 5, 6, and 7 n_2 times to get the b^* vector.

Step 8: Transform a^* back to matrix form ($m \times n$) to get the pseudo matrix $A^* = ((a_{ij}^*))$ and similarly transform b^* to get $B^* = ((b_{ij}^*))$, where $i = i^*/n$; $j = n$ if $i^* \bmod n = 0$, else $i = [i^*/n] + 1$; $j = i^* \bmod n$. Here $[x]$ denotes the largest integer $\leq x$.

Step 9: Calculate the χ^2 statistic T^* [$-2 \log \Delta$ described for the first test (Eq. 6)] based on tables A^* and B^*

Step 10: Let $I = I(T_0 > T^*)$ where I is an indicator function, i.e., I has a value of 1 if T_0 is greater than T^* and zero if T^* is greater than or equal to T_0.

Step 11: Repeat the above scheme for some larger number (B) of times. Let I_b be the I at the bth iteration ($b = 1, 2, 3, \ldots, B$), If

$$b_0 = \frac{\sum_{b=1}^{B} I_b}{B} > 1 - \alpha \tag{7}$$

then the hypothesis H_0 is rejected at level α.

It should be noted here that b_0 has no probabilistic interpretations (Efron and Tibshirani, 1993).

EXAMPLE

We have taken quartz C-axis measurements for two sections of the same specimen of quartz L-S tectonite from the Singhbhum Shear Zone (Sarkar and

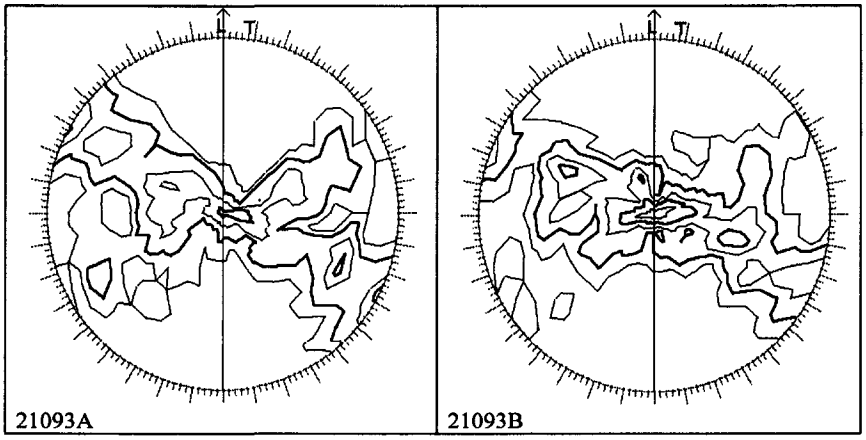

Figure 2. Quartz *c*-axis fabric diagram from mutually perpendicular sections of the same specimen. Fabric from perpendicular to lineation section (21093B) has been rotated for the comparison of the diagrams. *L* is plunge direction of lineation, *T* is top of foliation, Vertical line is the trace of foliation. Contour levels are 0.5, 1.5, 3.0, 4.5, 6.0, 7.5% per 1% area of the hemisphere.

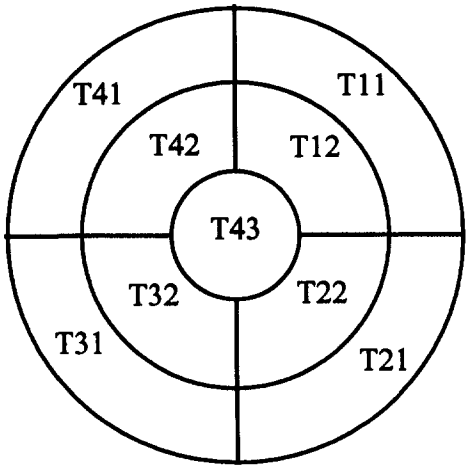

Figure 3. The grid preparation method adopted. Each element of the grid is named as members of a 4 × 4 matrix T (see the joint frequency distribution tables in Fig. 4). The first inner circle is kept at a ϕ value of 20 degrees and the innermost circle is placed at a ϕ value of 60 degrees.

a) Joint frequency table for 21093A			
6	26	0	0
19	30	0	0
7	18	0	0
22	33	53	0
Total data = 214			

b) Joint frequency table for 21093B			
6	23	0	0
18	33	0	0
6	15	0	0
18	22	70	0
Total data = 211			

Figure 4. Frequency distribution tables for specimens 21093A and 21093B constructed using the grid preparation method shown in Figure 3.

Saha, 1962; Joy and Saha, personal communication), Eastern India. One section is prepared perpendicular to the foliation and parallel to the lineation (21093A) and the other section (21093B) is prepared perpendicular to both lineation and foliation. The contoured equal area lower hemisphere projections of the two sections are shown in Figure 2. The contoured diagram of 21093B has been rotated to parallel orientation with 21093A.

For grid preparation, the method adopted was as follows. The total data were split into nine grid cells as shown in Figure 3. Data with (dip) values less than 5 degrees were added to antipodal cells also. To keep a 4×4 matrix form, other elements were set as zeroes (Fig. 4). The estimated T_0 (Eq. 6) is 5.65537. The bootstrap statistic estimated (Eq. 7) is 0.29380 (B was taken as 5000). A bootstrap statistic value of less than 0.95 can be taken as an acceptable limit for not rejecting the null hypothesis (i.e., a test of 5% level). Therefore the estimated bootstrap statistic of 0.29380 is not significant at this bootstrap test level and the null hypothesis (the two sections are similar) cannot be rejected.

REFERENCES

Bickel, P. J., and Doksum, K. A., 1977, Mathematical statistics: Basic ideas and selected topics: Holden-Day, Inc., San Francisco, 493 p.

Cheeney, R. F., 1983, Statistical methods in geology: For field and lab decisions: George Allen & Unwin, London, 169 p.

Dudley, R. M., Perkins, P. C., and Gine, E. M., 1975, Statistical tests for preferred orientation: Jour. Geology, v. 83, no. 6, p. 685–705.

Efron, B., and Tibshirani, R. J., 1993, Introduction to the bootstrap: Chapman & Hall, New York, 436 p.

Ghosh, S. K., 1993, Structural geology: Fundamentals and modern developments: Pergamon Press, Oxford, 598 p.

Jupp, P. E., 1987, A nonparametric correlation coefficient and two sample test for random vectors or directions: Biometrika, v. 74, no. 4, p. 887–890.

Kuiper, N. H., 1960, Tests concerning random points on a circle: Indag. Math., v. 22, no. 1, p. 38–47.

Mardia, K. V. 1972, Statistics of orientation data: Academic Press, New York, 357 p.
Ramsay, J. G., 1967, Folding and fracturing of rocks: McGraw-Hill Book Co., New York, 568 p.
Sarkar, S. N., and Saha, A. K., 1962, A revision of the Precambrian stratigraphy and tectonics of the Singhbhum and adjacent regions: Quart. Jour. Geol. Min. Met. Soc. India, v. 34, no. 2 & 3, p. 97–136.
Turner, F. J., and Weiss, L. E., 1963, Structural analysis of metamorphic tectonites: McGraw-Hill Book Co., New York, 545 p.
Watson, G. S., 1962, Goodness-of-fit tests on circle-II: Biometrika, v. 49, no. 1 & 2, p. 57–63.
Watson, G. S., 1983, Statistics on sphere: John Wiley & Sons, New York, 237 p.
Wellner, J. A., 1979, Permutation tests for directional data: Ann. Statist., v. 7, no. 5, p. 929–943.