Geological Survey of Canada: OPEN FILE REPORT 965, 1984

Q-MODE FACTOR ANALYSIS OF GRAIN SIZE DISTRIBUTIONS

by

James P.M. Syvitski[1]

[1] Geological Survey of Canada, Bedford Institute of Oceanography,
P.O. Box 1006, Dartmouth, N.S., Canada

ABSTRACT

Q-mode factor analysis is investigated for its appropriateness as an analytical tool in evaluating grain size data. Through use of mathematical analysis and simulated data, relationships between the character of size frequency distributions and results of Q-mode factor analysis are provided. Q-mode factor analysis can be applied to the interpretation of depositional processes and environments only to the extent that: (a) the central portions of size-frequency distributions reflect these processes and environments; and (b) the data set is confined to samples representing fairly simple physical processes and environments. The use of Q-mode factor analysis for analyzing the interrelationships in large data sets is therefore of limited value.

INTRODUCTION

> Factor analysis is often considered a part of statistical methodology. But in view of the fervor shown by its advocates as well as by its detractors, it is suggested that factor analysis might better be classified as a religion.
>
> - Wallis, 1968

Analysis of the size frequency distribution (sfd) of sediments and sedimentary rocks remains an important tool for the determination of sedimentary processes and recognition of depositional environments despite the varying degrees of success achieved by its application. A variety of methods, ranging from simple bi-variate plots of 'statistical' parameters derived from sfd, or the dissection of sfd, to complex analytical methods, such as factor analysis, have been used by researchers.

Despite the warnings and criticism of the method (Ehrenberg, 1962; Matalas and Reiher, 1967; Glaister and Nelson, 1974; Temple, 1978) Q-mode factor analysis has provided useful interpretative information (Klovan, 1966; Yorath, 1967; Beall, 1970; Allen et al., 1971; Clague, 1976; Dal Cin, 1976; Chambers and Upchurch, 1979). Its application, like most of the other techniques involving size frequency data, has been based on empirical evidence; the results seem to make 'sense.' Exactly how the method achieves these results from size frequency data has not been studied in detail and that is the purpose of the present paper.

By means of mathematical analysis and simulation, we attempt to determine relationships between characteristics of sfd and the results of Q-mode factor analysis. Some of the fundamental questions raised are: (1) how are differences in means of sfd resolved? (2) how are differences in sorting resolved? and (3) what patterns emerge from the mixing together of end-member sfd having distinct mean and sorting (among other

statistical) values? In short, the approach is to determine what Q-mode factor analysis will produce given some known size frequency distributions to start with.

The general principles of Q-mode factor analysis has been adequately described in the literature (see Armstrong, 1967; Rummel, 1967; Joreskog et al., 1976). We will here discuss those aspects of special importance to its use in the analysis of grain-size data.

The Data

Each sediment sample is considered to be totally described in terms of its size frequency distribution by the weight percent of sediment contained in each of p size classes. Geometrically, each sample can be visualized as a vector whose coordinates on p mutually orthogonal axes are the weight percentages. Because the sum of the weight percentages totals 100 for all samples, the ends of the vectors are constrained to be on a p-1 dimensional hyperplane in p dimensional space.

Similarity

Of the variety of measures which can be used to mathematically describe the degree of similarity between pairs of samples, the cosine theta index of Imbrie and Purdy (1964) provides several advantages. The index between sample i and sample j, assuming p size classes, is computed from:

$$\cos\theta_{ij} = \left(\sum_{k=1}^{p} X_{ik} \cdot X_{jk}\right) \left(\sum_{k=1}^{p} X_{ik}^{2} \sum_{k=1}^{p} X_{jk}^{2}\right)^{-0.5} \qquad (1)$$

where X_{ik} is the amount of sediment in the k^{th} class for sample i,

likewise for sample j.

As is evident from Equation 1, $\cos\theta$ simply measures the cosine of the angle separating any two sample vectors in p space. Two aspects of this index are worth emphasizing: (a) the denominator of the equation, in effect, normalizes the vectors so that they are each of unit length; and (b) the angular separation is determined solely by the proportions of results of sediment in each size class and therefore absolute amounts of sediment are ignored. A $\cos\theta$ value of 1.0 signifies perfect similarity; a value of 0.0 signifies perfect dissimilarity.

In practice, all pairs of samples are substituted into Equation 1 and an N by N matrix of $\cos\theta$ values is computed (where N is the total number of samples present in the study). This matrix contains all the information concerning the mutual similarities (and dissimilarities) between all of the samples and is the starting point of the Q-mode analysis itself.

Q-mode Analysis

Although mathematically complex, the Q-mode procedure as outlined by Miesch (1976) relies on the rather simple notion of principal components. The matrix of similarity coefficients is decomposed into two matrices. One, the matrix of factor loadings, provides the coordinates of the samples in a space of reduced dimensionality. The other, the matrix of factor scores, provides the position of the axes used to determine the reduced space in terms of their coordinates on the original p variables which define the data space.

Miesch (1976) has developed a method which permits the use of the most divergent samples as reference axes and gives their composition in terms of the p weight percent classes used to describe the samples

originally. For grain size analyses, then, the Miesch method attempts to: (1) find the most different types of samples, based on their total sfd; (2) express all other samples as mixtures of these end-member distributions; and (3) describe the composition of the end-members in terms of their total grain-size characteristics.

ANALYTICAL APPROACH

(i) On Understanding $\cos\theta$

Because the $\cos\theta$ matrix of similarities contains all the information on which Q-mode method works, it is important to establish exactly how the $\cos\theta$ index describes similarity between grain-size distributions. We first assume that grain-size distributions are inherently log-normal in character; if grain-size intervals are measured in ϕ units we need consider Gaussian distributions only.

Our first case is to compute $\cos\theta$ between two perfect Gaussian distributions whose properties are solely determined by μ_1, μ_2, σ_1, σ_2, the respective means and standard deviations. The equation for a Gaussian distribution is:

$$y_i = (2\pi\sigma)^{-0.5}\, e^{-0.5\left(\frac{x_i - \mu}{\sigma}\right)^2} \qquad (2)$$

where (in terms of grain size distributions): y_i is the weight of sediment in the i^{th} ϕ interval and less, x_i is the θ value, μ is the mean value, and σ is the standard deviation in ϕ units. The $\cos\theta$ index can be considered as a correlation function between two such distributions.

Denoting the first distribution as $f_1(x)$ and the second as $f_2(x)$ it

is seen that:

$$\cos\theta_{12} = \left[\int_{-\infty}^{\infty} f_1(x)\, f_2(x)\, dx\right] \left[\int_{-\infty}^{\infty} [f_1(x)]^2\, dx\right]^{-0.5} \left[\int_{-\infty}^{\infty} [f_2(x)]^2\, dx\right]^{-0.5} \quad (3)$$

and upon integration we obtain

$$\cos\theta_{12} = \left[\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\right]^{0.5} e^{-\left[\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]} \quad (4)$$

It is thus apparent that $\cos\theta$ is a complex, nonlinear function of the two basic parameters of the Gaussian distributions being compared.

For the special case where the two grain size distributions (samples) have the same standard deviation (sorting), the relationship becomes:

$$\cos\theta_{12} = e^{-\left(\frac{\mu_1 - \mu_2}{2\sigma}\right)^2} \quad (5)$$

Curves generated from this equation (Fig. 1)show that for a given standard deviation, $\cos\theta$ decreases as differences in mean value increase. Also evident is the fact that a pair of poorly sorted samples are designated as being more similar than a pair of well sorted samples having the same difference in mean values. For very well sorted samples, small differences in mean value drastically reduce $\cos\theta$.

For the special case where the distributions have zero mean difference, the relationship becomes:

$$\cos\theta_{12} = \left(2\sigma_1\sigma_2/(\sigma_1^2+\sigma_2^2)\right)^{0.5} \quad (6)$$

Figure 2, generated from this equation, shows that $\cos\theta$ is less sensitive to contrasts in sorting between two poorly sorted distributions than it is to two well sorted distributions.

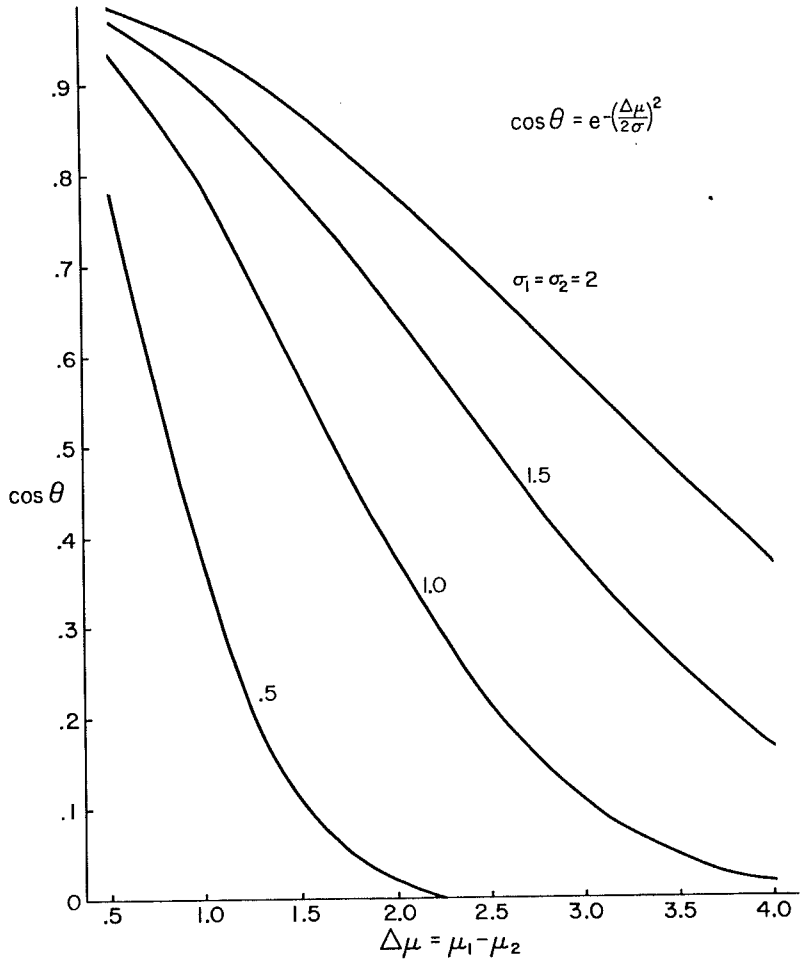


Fig. 1. Relationship between COS θ and difference in mean value ($\Delta\mu$) between two Guassian distributions having the same standard deviation, σ.

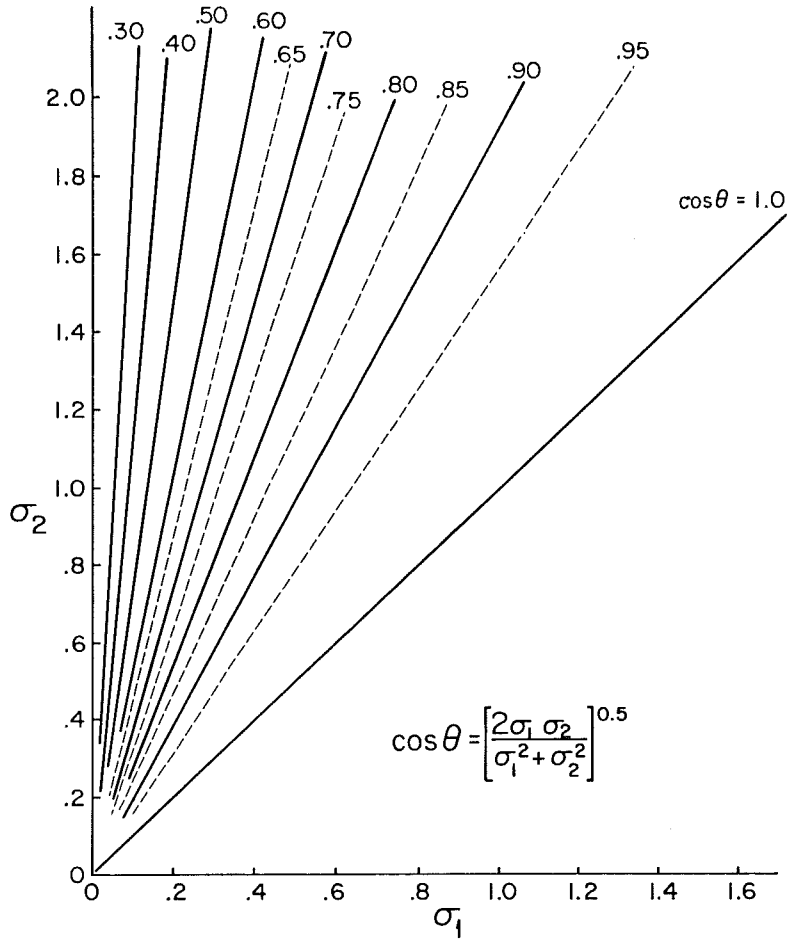


Fig. 2. Relationahip between COS θ and standard deviation of two Guassian distributions with zero mean difference.

A set of ideal, log-normally distributed, grain-size samples was mathematically generated for our second case, with means ranging from 0 $\emptyset$ to 4 $\emptyset$ in 0.5 $\emptyset$ increments and standard deviations from 0.5 $\emptyset$ to 2 $\emptyset$ in increments of 0.5 standard deviations. There are thus 36 samples in the set. The distributions were 'sieved' into 20 intervals between -8 $\emptyset$ and 11 $\emptyset$. Even with this degree of separation considerable deviation from an ideal distribution has been introduced, particularly for those samples whose means fall on a class boundary. Nevertheless, $\cos\theta$ values computed from the simulated discrete distributions differ only in the third decimal from the theoretical $\cos\theta$ values derived from Equation 2.

(ii) On Understanding the Q-mode Factor Analytical Solution

In our first case we compare two perfect Gaussian distributions predetermined to be different by altering the mean and/or standard deviations. Twenty nearly identical samples of each distribution were generated by the Cornish-Fisher expansion equation (outlined in detail by Swan et al., 1978; Kendall and Stuart, 1969, pp. 165-166). Therefore, the factor analysis worked on 40 samples but only two distributions.

The unrotated factor matrix results in two factors accounting for all variance in the data set (Fig. 3). The second factor separates the distributions by the loading sign. When standard deviation is held constant (i.e., at 1 $\emptyset$ in Fig. 3) and the difference between means ($\Delta\mu$) is decreased, the variance (eigenvalue) accounted by the unrotated factor 1 increases with that of factor 2 decreasing proportionally. At $\Delta\mu < 0.15$, factor 1 accounts for nearly 100% of the data variance (Fig. 3). Similarly, when the mean is held constant and the difference between standard deviations ($\Delta\sigma$) is decreased, the variance accounted by factor 1 increases

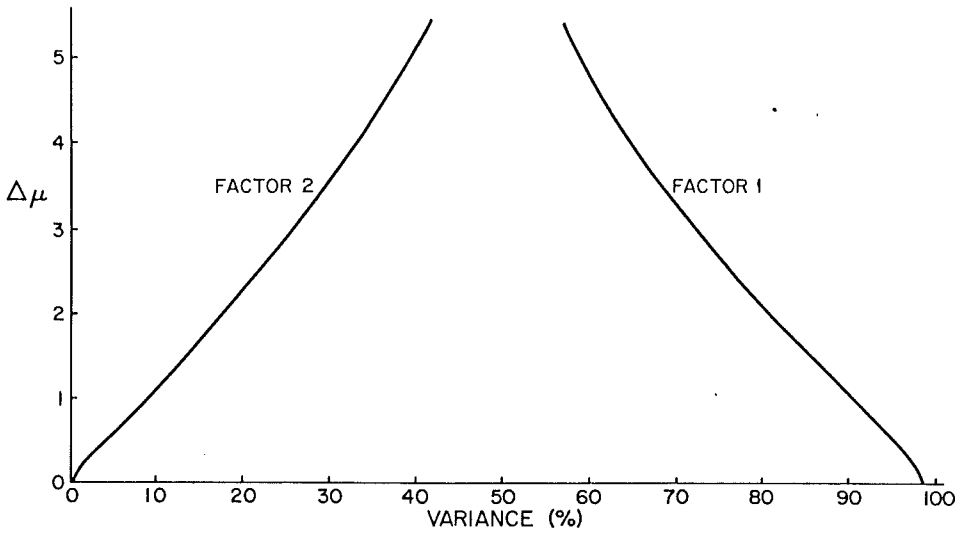


Fig. 3. Relationship between the difference in mean ($\Delta\mu$) between two Gaussian distributions and the percentage of total variance contributed by each eigenvalue of the two unrotated factors.

in the unrotated factor matrix. In both cases, when the factors are rotated, the rotated factor matrix separates the distributions by the loading values (Table 1). When $\Delta\sigma$ or $\Delta\mu$ decreases these loading values approach one another and become indistiguishable at $\Delta\sigma = 0.25$ or $\Delta\mu = 0.5$, respectively.

Figure 4a gives the sfd of two distributions having the same mean but different standard deviation and shows how the factor scores mirror the variance along any class interval i.e. $V_k=(\sum_{i=1}^{N} (X_i-\bar{X}_i)^2)N^{-1}$; where V is the variance in the K^{th} class; Xi is the weight of the i^{th} sample in the K^{th} class, $\bar{X}$ is the mean weight of the K^{th} class, and N is the total number of samples. In an attempt to circumvent this situation, the above experiment was repeated on range transformed data (Table 1, Model 2b). Range transformation essentially eliminates the variance along the variables (class intervals). The total variance in the data set accounted for by the rotated factor matrix is considerably lower than that of the nontransformed data (Table 1, Model 2aii and 2b). However, the loadings produce a better separation of the two distributions. Although the rotated factor scores are hard to interpret, the compositional scores mimic the original distributions (Fig. 4b).

The, 36 discrete, log-normal distributions generated above, each with a different mean and/or standard deviation, were analyzed by Q-mode. The percent sums of squares extracted by successive factors for both the principal components and varimax cases (Figure 5) suggest that a three-factor solution would be optimal. Table 2 shows the goodness of fit statistics; with three to five factors, only the middle seven grain-size

Table 1

Model type	Distribution type	No. of samples	No. of real factor	Unrotated Factor Matrix: factor variance		typical loading f_1	typical loading f_2	Rotated Factor Matrix: factor variance		typical loading f_1	typical loading f_2
1. Non-mixed Populations, variable mean											
a) Two populations	i) A, μ_1=3.1, σ_1=1.0	20	2	f_1=68%	A	.82	.55	f_1=49%	A	.97	-.20
	B, μ_2=6.7, σ_2=1.0	20		f_2=31%	B	.82	-.55	f_2=49%	B	.20	-.97
	ii) A, μ_1=3.1, σ_1=1.0	20	2	f_1=98%	A	.99	.10	f_1=49%	A	.62	-.38
	B, μ_2=3.3, σ_2=1.0	20		f_2= 1%	B	.99	-.10	f_2=49%	B	.38	-.62
2. Non-mixed Populations, variable standard deviation											
a) Two populations	i) A, μ_1=1.0, σ_1=1.0	20	2	f_1=95%	A	.97	-.20	f_1=49%	A	.30	.70
	B, μ_2=2.5, σ_2=1.0	20		f_2= 4%	B	.97	.20	f_2=49%	B	.70	.30
	ii) A, μ_1=1.0, σ_1=1.0	20	2	f_1=98%	A	.99	-.11	f_1=50%	A	.61	.78
	B, μ_2=1.5, σ_2=1.0	20		f_2= 1%	B	.99	.11	f_2=50%	B	.78	.61
b) range transformed	A, μ_1=1.0, σ_1=1.0	20	2	f_1=66%	A	.80	-.32	f_1=40%	A	.30	.83
	B, μ_2=1.5, σ_2=1.0	20		f_2=14%	B	.80	.32	f_2=40%	B	.83	.30

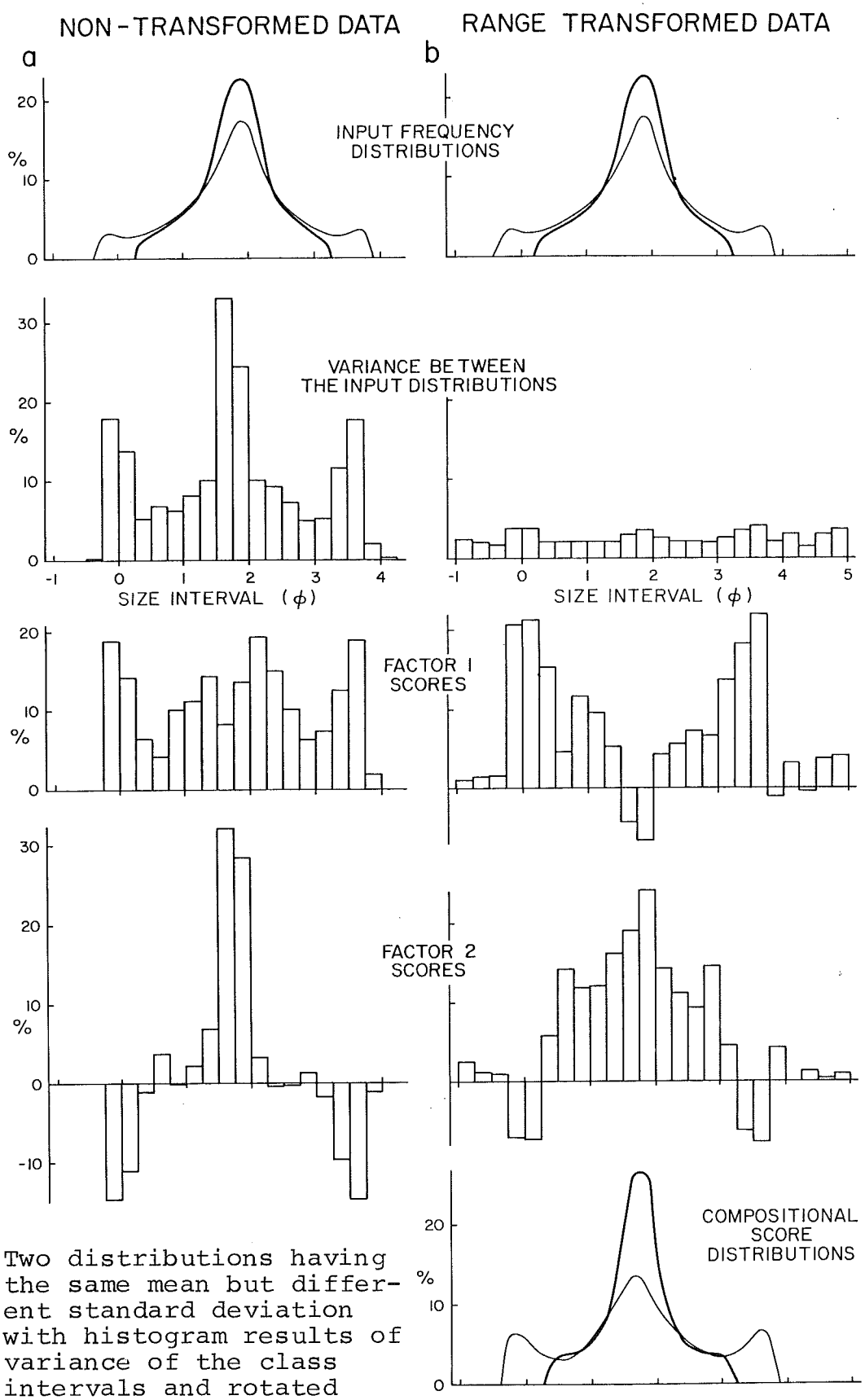


Fig. 4. Two distributions having the same mean but different standard deviation with histogram results of variance of the class intervals and rotated factor score values, for both non-transformed and range transformed data. Compositional scores from the range transformed data indicate the two original distributions.

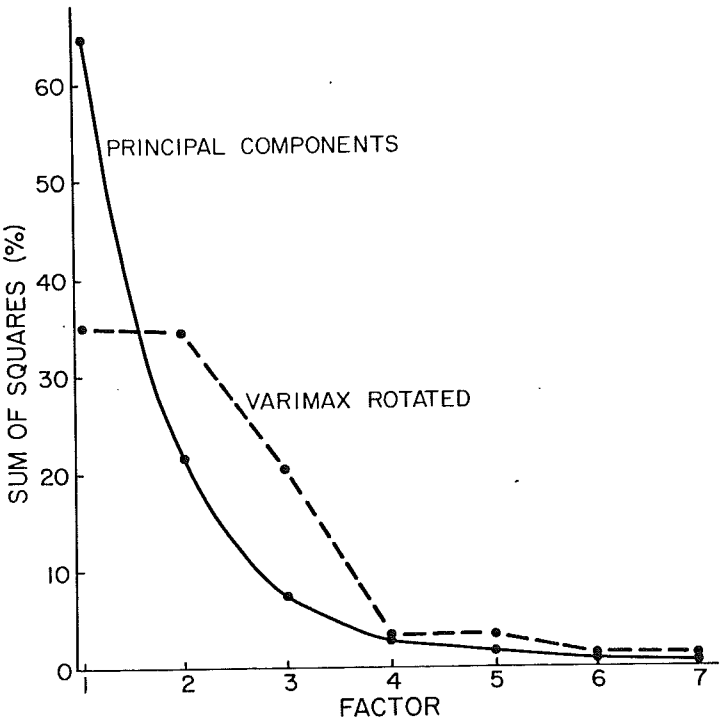


Fig. 5. Percent sums of squares extracted by successive factors for both principal components and varimax solutions. A three factor solution is indicated as being sufficient.

Table 2

COEFFICIENTS OF DETERMINATION

	Number of Factors							
	2	3	4	5	6	7	8	9
var. 1	0.0215	0.0258	0.0388	0.0252	0.1050	0.1494	0.3205	0.4456
var. 2	0.0012	0.0248	0.0521	0.0396	0.2125	0.2967	0.5068	0.6841
var. 3	0.0124	0.0364	0.0658	0.0568	0.2400	0.3345	0.5585	0.7393
var. 4	0.0174	0.0480	0.0848	0.0819	0.2960	0.4107	0.6343	0.8174
var. 5	0.0300	0.0718	0.1221	0.1317	0.3928	0.5358	0.7454	0.9192
var. 6	0.0623	0.1231	0.1964	0.2303	0.5461	0.7274	0.3691	0.9898
var. 7	0.1537	0.2457	0.3536	0.4245	0.7568	0.9487	0.9744	0.9966
var. 8	0.4660	0.6061	0.7781	0.8513	0.9349	0.9761	0.9834	1.0000
var. 9	0.7209	0.8310	0.8779	0.9100	0.9524	0.9995	0.9991	1.0000
var. 10	0.5692	0.6925	0.8755	0.9880	0.9885	0.9989	0.9991	1.0000
var. 11	0.0000	0.9489	0.9489	0.9479	0.9479	0.9997	0.9997	1.0000
var. 12	0.5698	0.6925	0.8573	0.9880	0.9885	0.9889	0.9990	1.0000
var. 13	0.7208	0.8310	0.8781	0.9101	0.9520	0.9995	0.9991	1.0000
var. 14	0.4656	0.6056	0.7777	0.8509	0.9436	0.9762	0.9872	1.0000
var. 15	0.1535	0.2454	0.3531	0.4241	0.7549	0.9486	0.9724	0.9965
var. 16	0.0622	0.1230	0.1962	0.2302	0.5443	0.7277	0.8578	0.9899
var. 17	0.0299	0.0717	0.1219	0.1317	0.3915	0.5389	0.7291	0.9198
var. 18	0.0176	0.0482	0.0850	0.0824	0.2955	0.4118	0.6181	0.8195
var. 19	0.0112	0.0356	0.0650	0.0562	0.2391	0.3350	0.5409	0.7400
var. 20	0.0041	0.0267	0.0533	0.0413	0.2106	0.2958	0.4911	0.6853

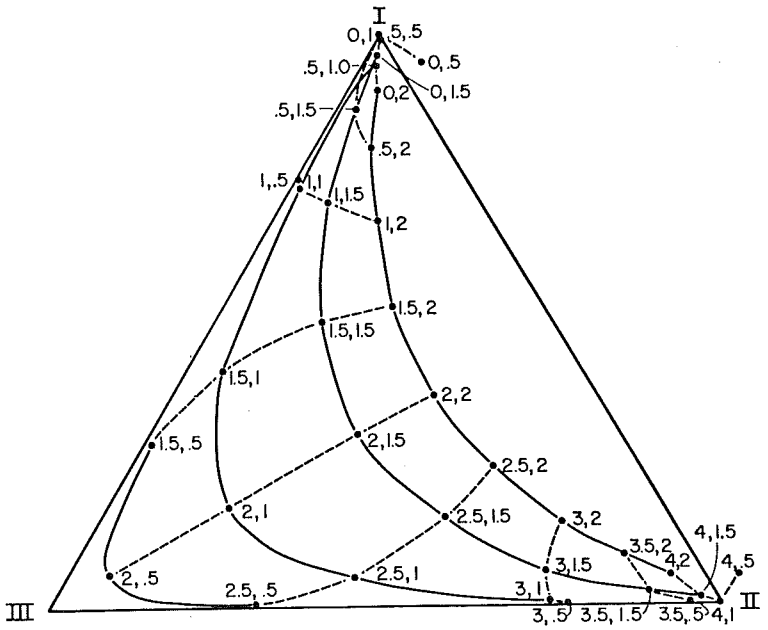
classes show appreciable coefficients of determination (i.e. explained variation divided by the total variation). Only the $\emptyset$ classes between -1 and 5 are actually contributing to the analysis. The low coefficients of the remaining classes reflect the very small variances associated with classes in the tails of the distributions. Communalities for a three-factor solution range from 0.7026 to 0.9990. Thus, even with errorless input data, considerable distortion is present in the factor analysis results.

A plot of the normalized varimax factor components for the three-factor case is shown in Figure 6. The typical 'horse-shoe' pattern caused by a closed data set is evident.

Factor I contains high loadings for samples with means of 0.0 and 0.5; Factor II contains high loadings for samples with means of 3.5 and 4.0; Factor III has a single high loading for the sample with a mean of 2.0 and a standard deviation of 0.5.

The most divergent samples are determined to be (0,0.5), (2,0.5), and (4,0.5) with the first number referring to the mean, the second the standard deviation. A plot using these samples as reference axes is shown on Figure 7. The 'horse-shoe' shape is again evident. Samples with the same standard deviation are arranged along arcs subparallel to the line joining (0,0.5) to (4,0.5). Samples with the same mean are arranged on arcs subparallel to the perpendicular of the line joining (0,0.5) to (4,0.5) which extends to (2,0.5).

Although each sample used in this analysis represents a perfect log-normally distributed sediment, the usual interpretation of this diagram would consider them to be mixtures of three end-member samples (0,0.5), (2,0.5), and (4,0.5) (i.e., Klovan, 1966). Because many samples plot out-



NORMALIZED VARIMAX FACTOR COMPONENTS ON SIMULATED GRAIN-SIZED DISTRIBUTIONS

Fig. 6. Normalized varimax factor components on simulated sfd. Sample points are identified by a two member code; the first describes the mean value, the second, the standard deviation. Heavy lines join samples with equal standard deviations, broken lines join samples with equal means.

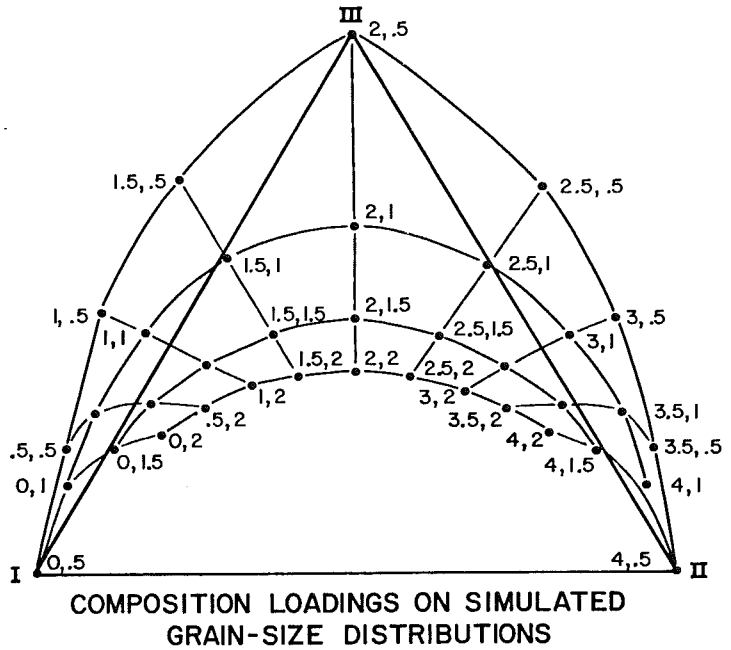


Fig. 7. Composition loadings on simulated sfd. Symbols as in Figure 6.

side the positive triangle defined by these three samples (due to their negative loadings), it is apparent that a simple additive mixing model is not appropriate. Sample (1,0.5), for example, is a mixture of 72.47% of sample (0,0.5), 45.74% of sample (2,0.5), and -18.60% of sample (4,0.5).

The compositional scores of the three end-members for the three-factor solution have their mean and standard deviation preserved, but are distorted with numerous negative values of composition. Clearly, a three-factor model is not a satisfactory solution; however, experiments with more factors do not materially improve the situation and there are no good *a priori* reasons for using a higher order factor solution.

(iii) Q-mode factor Analysis of Distribution Mixtures

To clarify the mixing model approach, a series of grain-size distributions were constructed by mixing together various proportions of particular end-members (see Table 3). The first mixture model used two distributions, (4.5,0.5) and (6.5,1.0), as end-members (Table 3, Model 3a). Two factors account for 99% of the variance in the data set. The factor scores of the rotated matrix approximate the original end-member populations A and B (Fig. 8A,B). The squared factor loadings of the rotated factor matrix (times 100 for percentage values) give the approximate mixing ratios (Table 4; the loadings as indicators of mixing ratios are improved upon oblique rotation, see Clarke, 1978). When the experiment was repeated with all six mixtures but no end members, the rotated factor loadings were nearly the same (Table 4).The compositional scores gave nearly the same end-members although none were present in the original data set (Fig. 8A,B).

The second mixture model generated and mixed three distributions (Table 3, model 4; Fig. 9). Three factors account for 100% of the data

Table 3.

Model type	Distribution type	No. of Samples	No. of real factors	Unrotated Factor Matrix: factor variance		typical loading			Rotated Factor Matrix: factor variance		typical loading		
3. Mixed Populations						f_1	f_2				f_1	f_2	
a) Two End Members plus mixtures	A, μ_1=4.5 σ_1=0.5	10	2	f_1=80%	A	.81	-.58		f_1=52%	A	.99	.13	
	B, μ_2=6.5 σ_2=1.0	10	2	f_2=19%	B	.75	.65		f_2=48%	B	.10	.99	
	C = .9A + .1B	10			C	.85	-.51			C	.98	.20	
	D = .7A + .3B	10			D	.95	-.31			D	.91	.39	
	E = .5A + .5B	10			E	.99	-.05			E	.76	.63	
	F = .4A + .6B	10			F	.99	.05			F	.66	.73	
	G = .3A + .7B				G	.96	.24			G	.52	.85	
	H = .1A + .9B	10			H	.99	.52			H	.23	.97	
						f_1	f_2	f_3			f_1	f_2	f_3
b) Three End Members plus mixtures	μ_1=3.5 σ_1=0.5 SK_1=0.5 K_1=5.5 μ_2=4.5 σ_2=0.75 SK_2=0.0 K_2=3.0 μ_3=5.5 σ_3=1.5 $SK_3$0.0 K_3=3.0												
	A,	10	3	f_1=80%	A	.76	-.65	.01	f_1=30%	A	.10	.98	.19
	B,	10		f_2=13%	B	.82	.38	-.42	f_2=30%	B	.36	.17	.92
	C,	10		f_3= 7%	C	.76	.54	.35	f_3=31%	C	.92	.07	.37

Table 3. (cont'd)

Model type	Distribution type	No. of Samples	No. of real factors	Unrotated Factor Matrix				Rotated Factor Matrix			
				factor variance	typical loading			factor variance	typical loading		
	D = .2A + .8B	10		D	.92	.08	-.38	D	.28	.47	.84
	E = .2A + .5B + .3C	10		E	.98	.15	-.14	E	.51	.47	.72
	F = .2A + .2B + .6C	10		F	.98	.15	.08	F	.67	.49	.56
	G = .4A + .2B + .4C	5		G	.99	-.14	.03	G	.49	.72	.49
	H = .6A + .2B + .2C	5		H	.92	-.38	-.02	H	.30	.87	.40
	I = .8A + .2B	5		I	.84	-.53	-.06	I	.15	.93	.32
	J = .5A + .5C	10		J	.93	-.33	.16	J	.45	.84	.29
	K = .2A + .8C	10		K	.94	.18	.29	K	.80	.46	.38

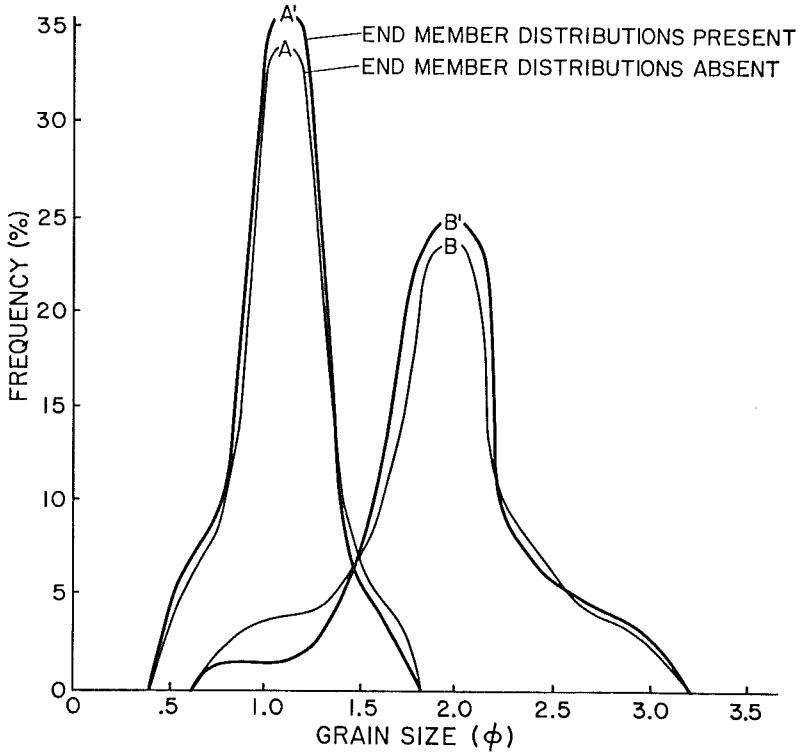


Fig. 8. Compositional score results of the two mixed distributions (Model 3a, Table 3) when the original end members were present in the data set, and absent.

Table 4. Two Population Problem

Population	Generated Mixing Ratio	Factor Loading Ratio (1^2 x 100) End Members Present	Factor Loading Ratio (1^2 x 100) End Members not Present
A	100:0	98:2	-
B	0:100	1:99	-
C	90:10	96:4	94:6
D	70:30	83:15	81:18
E	50:50	58:40	58:42
F	40:60	43:54	41:58
G	30:70	27:72	26:74
H	10:90	5:94	4:94

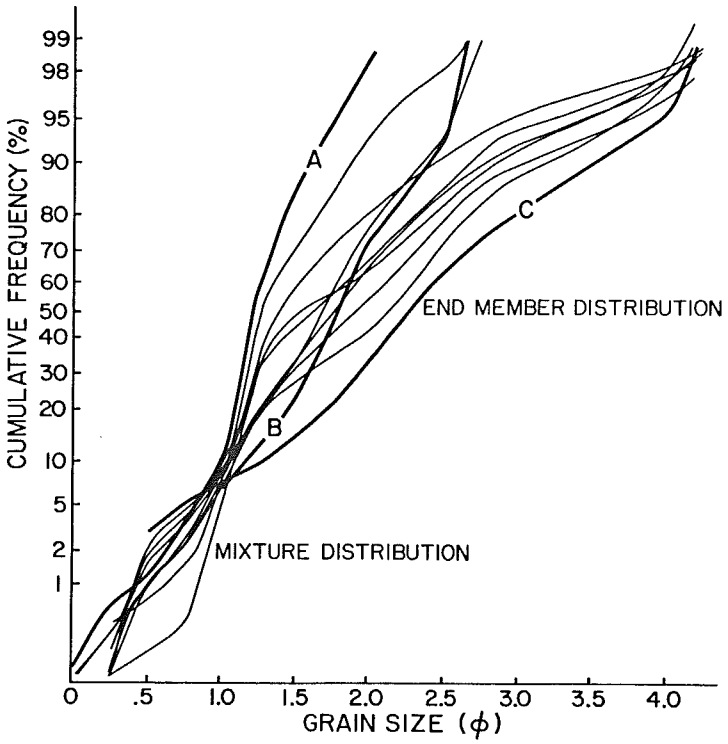


Fig. 9. Log-probability plot of typical distributions from Model 3b, Table 3, the three mixed distributions.

variance. The three rotated factors each account for approximately 33% of the data variance and the varimax scores indicate that each factor corresponds to one of the three end member distributions; again whether or not the original data set included end member data. Mixing ratios calculated from the varimax loadings (Table 5) indicate that an orthogonal rotation is no longer satisfactory for systems as complex as three end member mixtures. Figure 10 indicates part of the problem. The factor model appears only to designate an end member distribution from its central mode. Thus these modal size intervals are over-compensated in the compositonal scores and cause a reduction in score values over the same size intervals in the remaining end members (Figure 10).

The fact that only the central grain size classes carry much weight in the analysis means valuable information on the character of the tails of the distribution does not contribute significantly in the computation of similarity. To demonstrate this point, all possible mixtures of three end-members (0, 0.5; 2, 0.5; 4, 0.5) in 0.1 incremental values were analyzed by Q-fact. The results are shown on Figure 11. The mixtures are perfectly resolved in terms of their mean and standard deviations. That is, given the compositions of the end members and the mixing proportions, the mean and standard deviation of these mixtures can be perfectly predicted in terms of their position on the factor plot. However, it is not a unique solution. One point on the factor plot can represent samples with quite different grain size characteristics. They will both, however, represent the same proportional mixture of the end members. For instance, a log-normal sample (3.5, 2.0) falls very close to a 1:4:5 mixture on the three end-member factor plot, and a log-normal sample (1.5, 1.0) plots close to a 4:6:0 mixture (Fig. 11), yet the corresponding curves are not very similar

Table 5. Three Population Problem

Population	Generated Mixing Ratio	Factor Loading Ratio (1^2 x 100) End Members Present	Factor Loading Ratio (1^2 x 100) End Members not Present
A	100:0:0	1:96:4	-
B	0:100:0	3:85:13	-
C	0:0:100	0:14:85	-
D	20:80:0	22:52:26	17:73:10
E	20:50:30	22:52:26	16:54:30
F	20:20:60	76:16:9	68:19:13
G	40:20:40	52:24:24	44:27:29
H	60:20:20	76:16:9	68:19:13
I	80:20:0	86:10:2	82:14:05
J	50:0:50	71:8:20	64:11:26
K	20:0:80	21:14:64	15:16:69

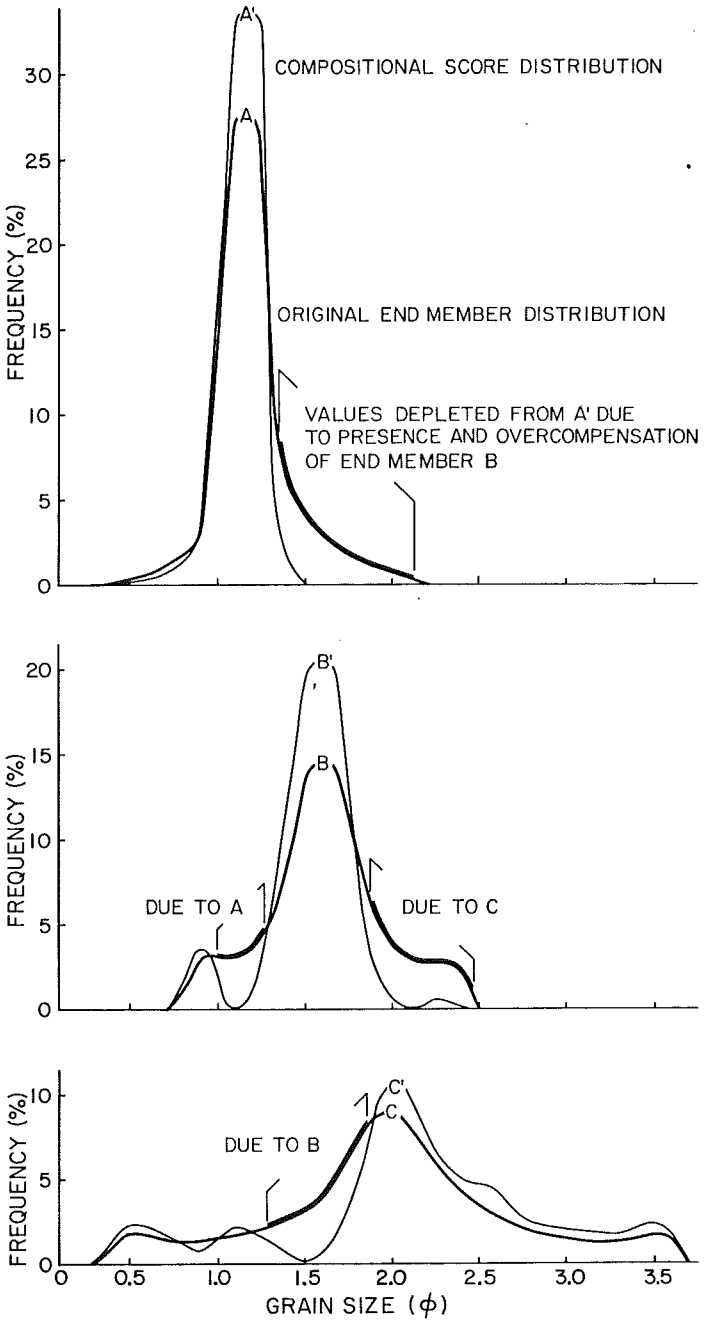


Fig. 10. Frequency plots of each of the end members of Model 3b, Table 3, compared to the compositional scores from the rotated factor matrix. Compositional score A* is affected by B*; B* is affected by A* and C* shows the effect of B*.

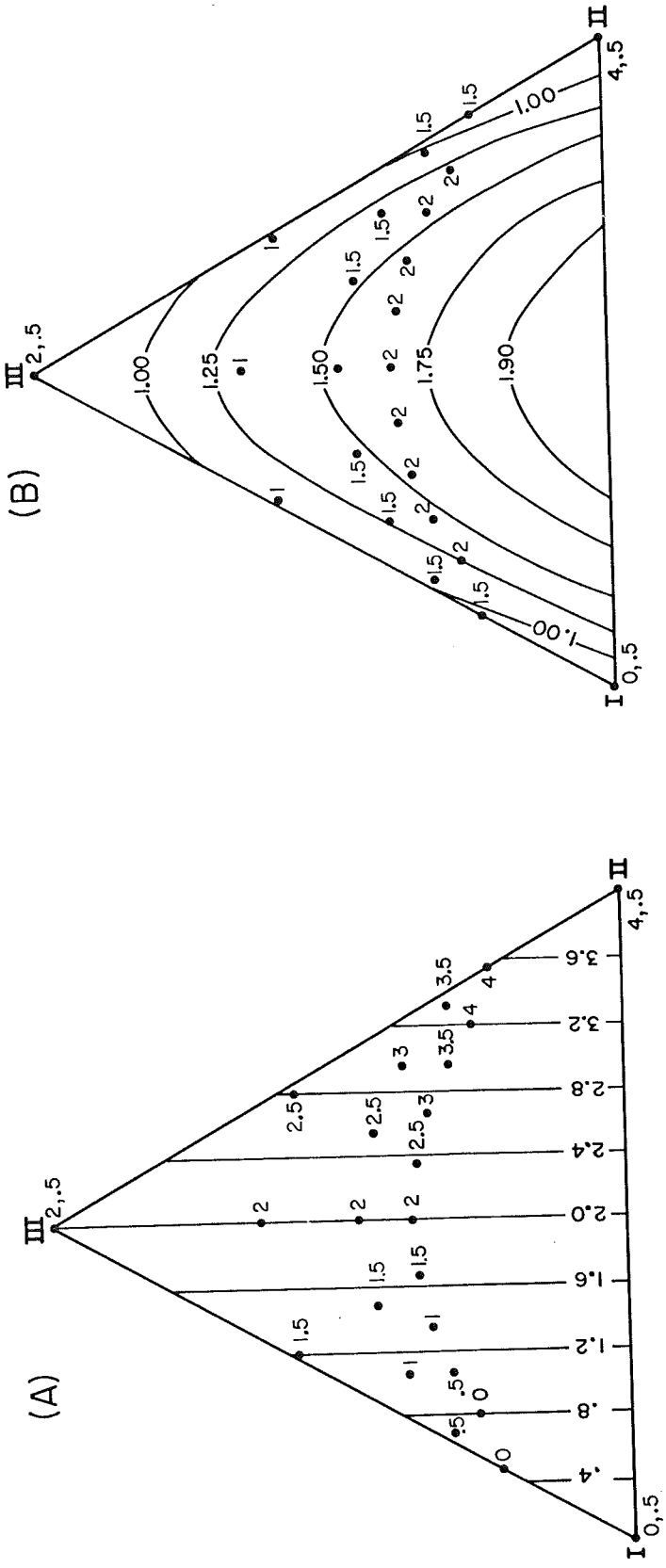


Fig. 11. Plot of composition loadings of all possible mixtures of the three end member distributions 11a. The contours show the values of the mixture standard deviations. The dots show the standard deviations of perfectly log-normal distributions.

(Fig. 12). The attempt to circumvent the situation through use of range transformation was not successful in that the same inconsistencies emerged as in the untransformed three factor solution.

(iv) Data sets with both mixtures and non-mixed distributions

Data input into Model 3a (Table 3) was used with the addition of two unrelated distributions (Fig. 13, Table 6, Model 4). Four factors resulted, one for every type of distribution excluding those that are mixtures of the two end-member distributions. There are some noteable similarities and differences between Model 3a (Table 3) and Model 4 (Table 6). The end-member factors of the rotated factor matrix account for less variance in Model 4; the factor loadings, however, are nearly identical. Non-mixed distribution I (of Model 4) has significantly higher rotated factor loadings than non-mixed distribution J, the numerical difference being due to distribution I having more easily recognized marker variables (class intervals). Distribution J is nearly hidden within the mixed distributions (Fig. 13).

When additional non-mixed distributions are added to Model 4 factors, they have smaller eigenvalues (or factor variance) and this decreases the chance of getting true compositional scores. Factors that have accountable variance of less than 5% have unrealistic compositional scores.

SUMMARY

1. COSθ is a complex, non linear function of the means and standard deviations of Gaussian grain size distributions (eqn's. 4, 5, 6). COS θ is more affected by differences that occur between well-sorted) samples than those between poorly sorted samples.

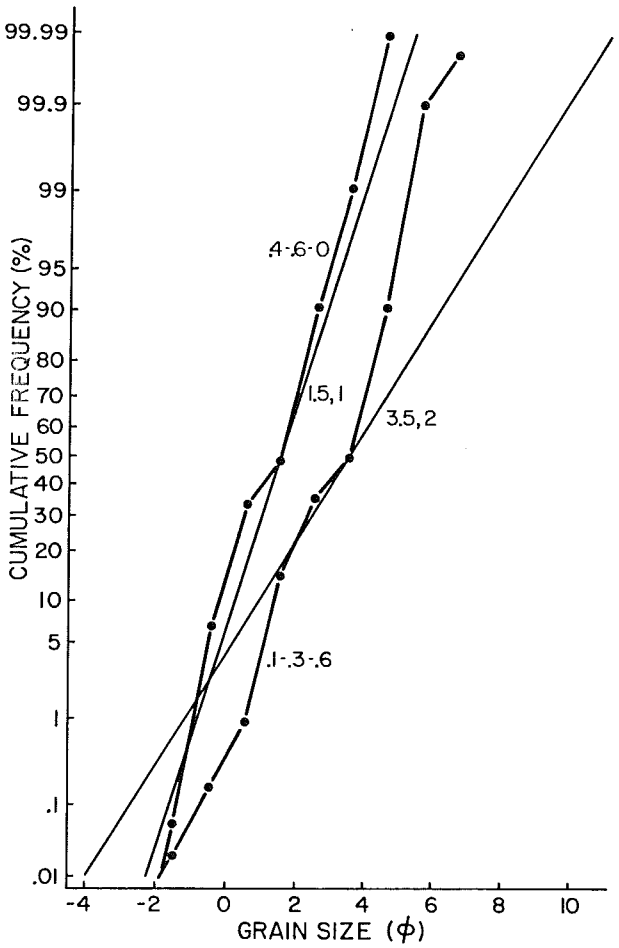


Fig. 12 Comparison of sfd for samples falling close to each other on Figure 11. Curve (1.5, 1) is an ideal log-normal distribution with a mean of 1.5 and standard deviation of 1.0; curve (3.5,2) is interpreted the same way. Curve .4-.6-0 is a mixture of the three proportions 40%, 60% and 0% respectively. Curve 0.1-.3-.6 is interpreted in the same way.

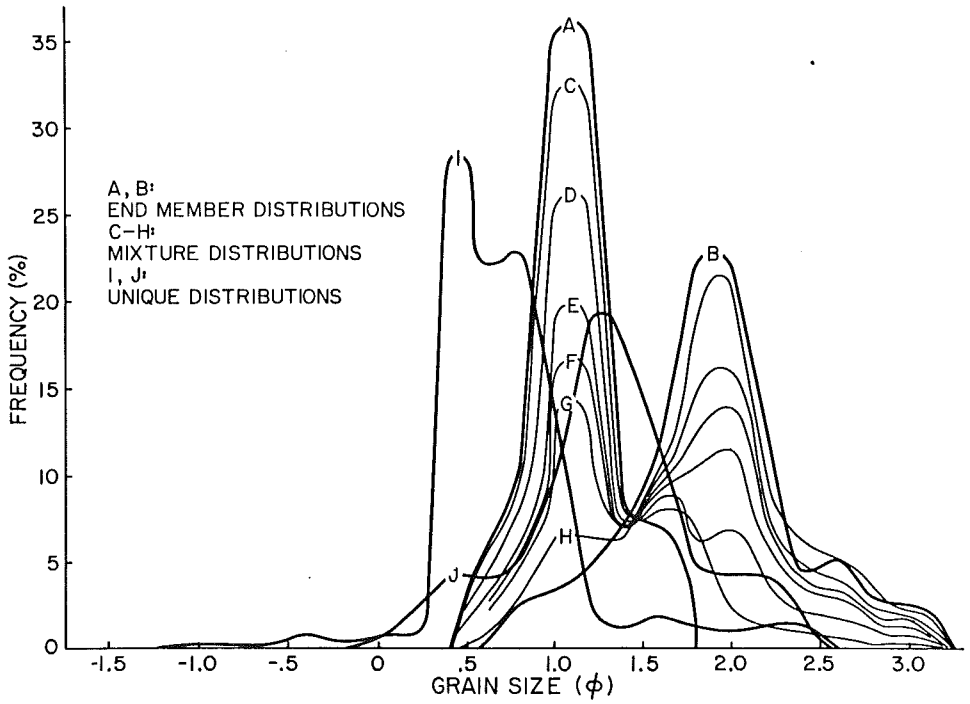


Fig. 13 Frequency plot of Model 4a, Table 6, of mixed with non-mixed distributions.

Table 6.

Model type	Distribution type	No. of Samples	No. of real factors	Unrotated Factor Matrix						Rotated Factor Matrix					
				factor variance		typical loading				factor variance		typical loading			
						f_1	f_2	f_3	f_4			f_1	f_2	f_3	f_4
4. Two End Members plus mixtures plus two non-mixed populations.	A, μ_1= 4.5 σ_1= 0.5	10	4	f_1= 73%	A	.83	-.51	-.19	-.08	f_1= 45%	A	.98	.12	.15	.08
				f_2= 17%	B	.72	.68	.11	-.00	f_2= 39%	B	.09	.99	.04	.08
	B, μ_2= 6.5 σ_2= 1.0	10		f_3= 8%	C	.87	-.45	-.18	-.08	f_3= 11%	C	.96	.20	.14	.09
				f_4= 2%	D	.95	-.26	-.14	-.07	f_4= 4%	D	.90	.40	.14	.09
	C = .9A + .1B	10			E	.99	.00	-.08	-.06		E	.75	.63	.12	.10
	D = .7A + .3B	10			F	.98	.16	-.04	-.05		F	.64	.75	.11	.10
	E = .5A + .5B	10			G	.94	.32	-.00	-.04		G	.50	.85	.09	.10
	F = .4A + .6B	10			H	.79	.60	.08	-.02		H	.21	.97	.05	.08
	G = .3A + .7B	10			I	.41	-.38	.83	-.02		I	.24	.06	.97	.03
	H = .1A + .9B	10			J	.90	-.20	-.05	.38		J	.73	.39	.20	.53
	I, μ = 3.5 σ_3 = 0.75 sk_3= 1.0 k_3= 10.5														
	J, μ_4= 5.0 σ_4= 1.0	10													

2. The Q-mode factor solution can be used to separate distributions with $\Delta\sigma > 0.25$ and for $\Delta\mu > 0.5$. It concentrates on similarities and differences involving only the central portions of the distributions and thus effectively classifies samples on the basis of mean and standard deviation only.

3. Q-mode factor analysis provides a mixing model for the interpretation of grain size distributions whether such a solution is applicable or not.

3a. When not applicable, the composition factor scores reflect only the standard deviation along the class intervals between the different distributions, and not the original sample compositions.

3b. When applicable, the factor scores of the rotated matrix approximate the original end member distributions, even if the original data set only contained mixtures and no end member distributions. If the original data set is complex (i.e., contains mixtures of 3 or more end-members) the compositional scores begin to show marked deviations from the actual end member distributions.

4. Consequently Q-mode factor analysis can be an effective method of dissecting mixtures of sediments. The squared factor loadings of the rotated factor matrix give the approximate mixing ratios, but the accuracy decreases with increasing complexity: i.e., when 3 or more end-members, or extraneous distributions are included in the data set.

5. When a data set contains mostly (>50%) distributions resulting from the mixing of end-members, other extraneous distributions which are not part of the mixing model do not interfere with the end-member solution to the mixture distributions. Other factors, accounting for little variance in the data, appear to be related to each of the

non-mixed distributions. However, when the accountable variance per factor is less than 5%, these factors produce unrealistic compositional scores.

6. There is no unique place on a factor plot of a given size frequency distribution. Grain size distributions only similar in the central portions of their frequency distribution will be interpreted as being the same, even if one is generted as a mixture of end-member distributions and another is a perfect Gaussian distribution.

IMPLICATIONS FOR REAL DATA SETS

The results of Q-mode factor analysis can be applied to the interpretation of depositonal processes and environments of deposition only to the extent that: (1) the central portions of sediment distributions (i.e., as given by mean and standard deviation) reflect these processes and environments; and (2) the data sets are confined to samples that delimit only certain simple physical processes or environment(s).

It is these a priori conditions that constrain the value of Q-mode factor analytical interpretation. In many coastal areas, depositional environments are complex, with energy contributions from fluvial, tidal and wave processes. Many shelf environments are a mixture of sediments deposited from recent processes and those comprised of erosional surfaces of exposed paleoenvironments. Even some single process environments can show no distinguishable trend in the central portion of grain size distributions; i.e., suspended sediments collected with depth from the breaking wave environment (Kennedy et al., 1981).

Yet it has been demonstrated that trends in the central portion of size distributions are prevalent in some sedimentary environments (e.g.,

surficial sediments collected down-inlet in fjords, Syvitski and MacDonald, 1982; many other environments, McLaren, 1981). It does, however, seem counter productive to have to test whether Q-mode factor analysis is applicable to one's data when the normal reason for its use is to simplify the variance in large data sets.

The study of suspended sediments in terms of vertical or horizontal flux may be a realistic environment to apply Q-mode factor analysis. Many studies have already demonstrated the trends in size distributions of suspended particulate matter, both with depth, time or distance from a source (Syvitski and Murray, 1981; Kranck, 1981; Clarke, 1978; Yamamoto, 1976).

In conclusion, Q-mode factor analysis of size frequency data has distinct limitations to real world data. Caution in its use is very much stressed, and we suggest that all indiscriminate use of the technique be avoided.

ACKNOWLEDGMENT

Our study was supported by grants from the Natural Science and Engineering Research Council of Canada and the Geological Survey of Canada. I thank Ed Krebes for his development of equation (2) and Ed Klovan and Bob Dalrymple for their encouragement and discussions.

REFERENCES

Allen, G.P., Castaing, P., and Klingebiel, A., 1971, Preliminary investigation of the surficial sediments in the Cape Breton Canyon (southwest France) and the surrounding continental shelf; Marine Geology, V. 10, p. M27-M32.

Armstrong, J.S., 1967, Deviation of theory by means of factor analysis: Tom Swift and his electic factor analysis machine: The American Statistician, V. 21, p. 17-21.

Beall, A.O. Jr., 1970, Textural differentiation within the fine sand grade: Journal of Geology, V. 78, p. 77-93.

Chambers, R.L., and Upchurch, S.B., 1979, Multivariate analysis of sedimentary environments using grain-size frequency distributions: Mathematical Geology, V. 11, p. 27-43.

Clague, J.J., 1976, Surficial sediments of the Northern Strait of Georgia, British Columbia; Geological Survey of Canada, paper 75-1, part B, p. 151-156.

Clarke, T.L., 1978, An oblique factor analysis solution for the analysis of mixtures: Mathematical Geology, V. 10, p. 225-241.

Dal Cin, R., 1976, The use of factor analysis in determining beach erosion and accretion from grain size data: Marine Geology, V. 20, p. 95-116.

Ehrenberg, A.S.C., 1962, Some questions about factor analysis: The Statistician, V. 12, p. 191-208.

Glaister, R.P. and Nelson, H.W., 1974, Grain-size distributions an aid in facies identification: Bulletin of Canadian Petroleum Geology, V. 22, p. 203-240.

Imbie, J., and Purdy, E.G., 1962, Classification of modern Behemian carbonate sediments: American Association of Petroleum Geologists, Memoir 1, p. 252-272.

Joreskog, K.G., Klovan, J.E., and Reyment, R.A., 1976, Geological Factor Analysis. Elsevier, Amsterdam.

Kendall, M.G., and Stuart, A., 1969, The Advanced Theory of Statistics. Volume 1 - Distribution Theory (3rd edition): Hafner Publishing Co., New York, 439 pp.

Kennedy, S.K., Ehrlich, R., and Kana, T.W., 1981, The non-normal distribution of intermittent suspension sediments below breaking waves: Journal of Sedimentary Petrology, V. 51, p. 1103-1108.

Klovan, J.E., 1966, The use of factor analysis in determining depositional environments from grain-size distributions. Journal of Sedimentary Petrology, V. 36, p. 115-125.

Kranck, K., 1981, Particulate matter grain-size characteristics and flocculation in a partially mixed estuary: Sedimentology, V. 28, p. 107-114.

Matalus, N.C., and Reiher, B.J., 1967, Some comments on the use of factor analyses: Water Resources Research, V. 3, p. 213-223.

McLaren, P., 1981, An interpretation of trends in grain size measures: Journal of Sedimentary Petrology, V. 51, p. 611-624.

Miesch, A.T., 1976, Q-mode factor analysis of geochemical and petrologic data matrices with constant row-sums: U.S. Geological Survey Prof. Paper 574 G.

Rummel, R.J., 1967, Understanding factor analysis: Journal of Conflict Resolution, V. 11, p. 444-480.

Swan, D., Clague, J.J., and Luternauer, J.L., 1978, Grain-size statistics I. Evaluation of the Folk and Ward graphic measures: Journal of Sedimentary Petrology, V. 48, p. 863-878.

Syvitski, J.P.M., and MacDonald, R.D., 1982, Sediment character and provenance in a complex fjord; Howe Sound, British Columbia: Canadian Journal of Earth Sciences, V. 19, p. 1025-1044.

Syvitski, J.P.M., and Murray, J.W., 1981, Particle interaction in fjord suspended sediment: Marine Geology, V. 39, p. 215-242.

Temple, J.T., 1978, The use of factor analysis in Geology: Mathematical Geology, V. 10, p. 379-387.

Wallis, J.R., 1968, Factor analysis in hydrology an agnostic view: Water Resources Research, V. 4, p. 521-527.

Yamamoto, S., 1976, Multivariate correlation analysis of suspended sediment characteristics: Journal of Mathematical Geology, V. 8, p. 57-74.

Yorath, C.J., 1967, Determination of sediment dispersal patterns by statistical and factor analysis northestern Scotian Shelf. Ph.D. Thesis, Queen's University, Ontario Canada.