Q

Q-Mode Factor Analysis

Norman MacLeod
Department of Earth Sciences & Engineering, Nanjing University, Nanjing, China

Definition

Q-Mode Factor Analysis – a statistical modelling procedure whose purpose is to find a small set of latent linear variables that estimate the influence of the factors controlling the structure of case similarities/dissimilarities in a multivariate dataset and are optimized to preserve as much of the inter-case similarity or dissimilarity structure as appropriate given an estimate of the number of causal influences.

Introduction

Whereas a linear, multivariate data-analysis procedure whose purpose was to summarize the covariance/correlation structure among a set of variables (*R*-mode) was first introduced by Karl Pearson (1901), the development of linear procedures whose purpose was to summarize the similarity/dissimilarity structure among sets of cases (e.g., objects, samples or events) (*Q*-mode) is of more recent vintage. This type of multivariate data analysis was introduced into the earth-science literature by John Imbrie (1963) based on previous applications in the fields of psychology and biology. Both principal component analysis (PCA) and component-based, *R*-mode factor analysis (FA) can be used to create mathematical ordination spaces into which cases described by the variables can be projected. Cases that plot close to each other in such spaces can be regarded as having similar variable values and cases that plot at some distance from others can be regarded as having different variable values. But in *R*-mode analyses the spaces created by these methods are referenced to the structure of covariance or correlation relations among the variables, not similarity relations among the cases themselves. The *Q*-mode methods of principal coordinates analysis (PCoord) and component-based *Q*-mode factor analysis (QFA) were created to redress this deficiency.

Estimating Similarity

Just as the core of the PCA and FA lies in the manner in which pairwise structural relations among variables are portrayed, the core of QFA lies in the manner in which pairwise structural relations among cases are portrayed. But while in *R*-mode methods this usually comes down to a choice between two quite distinct alternatives – covariance or correlation – in *Q*-mode methods a wide variety of similarity and dissimilarity indices are available. Moreover, across the earth sciences it is often the case that variable values are reported as proportions or percentages, a practice that adds complexity to QFA calculations (Table 1).

The similarity coefficients or indices used in QFA can be classified into several types: binary association indices, distance indices, and proportional similarity indices. Binary association indices, such as the coefficients of Jaccard, Dice, Simpson, and Otsuka (Table 1; see Cheetham and Hazel 2012 for a review), are used exclusively for presence-absence datasets and have been applied most extensively in the context of *Q*-mode cluster analysis. They do, however, produce "similarity" matrices than can be operated on by any number of multivariate data-analysis procedures, including QFA. Owing to the count of either mutual presences, or mutual absences, appearing in the numerators of these ratios, the selection of particular binary association indices will focus the analysis on the structure of similarity associations, dissimilarity associations, or some combination of both (Table 2).

Distance-based indices are the most readily understandable family of *Q*-mode similarity coefficients because they

© Springer Nature Switzerland AG 2021
B. S. Daya Sagar et al. (eds.), *Encyclopedia of Mathematical Geosciences*, Encyclopedia of Earth Sciences Series,
https://doi.org/10.1007/978-3-030-26050-7_259-1

Q-Mode Factor Analysis, Table 1 A selection of six *Q*-mode binary similarity coefficients

$S_{Jaccard} = \frac{C}{N_1+N_2-C}$	$S_{SimpleMatching} = \frac{C+A}{N_t+A}$
$S_{Dice} = \frac{2C}{N_1+N_2}$	$S_{Simpson} = \frac{C}{N_1}$
$S_{Kluczynski2} = \frac{C(N_1+N_2)}{2(N_1N_2)}$	$S_{Otsuka} = \frac{C}{\sqrt[2]{N_1N_2}}$

C Present in both cases, A Absent on both cases, N_1 Total present in case 1, N_2 Total present in case 2, N_t Total present in cases 1 and 2

Q-Mode Factor Analysis, Table 2 A selection of five *Q*-mode distance (dissimilarity) coefficients

Euclidean Distance	Squared Euclidean Distance
$d_{i,j} = \sqrt[2]{\sum_{k=1}^{p}\left(x_{i,k} - x_{j,k}\right)^2}$	$d_{i,j}^2 = \sum_{k=1}^{p}\left(x_{i,k} - x_{j,k}\right)^2$
Manhatten Distance	Mahalanobis Distance
$D_{i,j} = \lvert\overrightarrow{x_i} - \overrightarrow{x_j}\rvert$	$D_{i,j} = \left(\overrightarrow{x_i} - \overrightarrow{x_j}\right)^T S^{-1}\left(\overrightarrow{x_i} - \overrightarrow{x_j}\right)$
Gower Distance	
$g_{i,j} = \frac{\sum_{k=1}^{p}\lvert x_{i,k} - x_{j,k}\rvert / range_k}{p}$	

x_i = The i^{th} case. x_j = The j^h case. p = no. of variables, S = covariance/correlation matrix for the p variables

represent case similarity and difference as a spatial construct. Euclidean, squared Euclidean, Manhattan, and Mahalanobis distances (Table 2) are all examples of distance indices that can be used to express the structure of relations between cases by summarizing pairwise differences across all variables. Because the trace of a pairwise, n by n distance matrix will always be occupied by zeros (since the distance between any object, sample or event with itself = 0.0), distance-based indices are often said to record "dissimilarity".

Distance indices are applied typically to sets of interval or ratio data values and in situations where it is desirable to retain the magnitude or range of each variables' values in the analysis. However, if a dataset contains a broad mixture of variable types, or if the data analyst wishes to minimize the influence of different variables' numerical ranges, Gower's distance (Gower 1971) may be used. When using Gower's distance it should be kept in mind that atypical variable values can inflate a variable's range and so have a disproportionate effect on the resulting pairwise similarity estimate. Ideally, outliers should be removed prior to the calculation of Gower distances and the number of different variable types balanced insofar as possible.

Weights may also be employed to control the relative influence of different variables on the overall measure of similarity though, as pointed out by Gower (1971), justification of an appropriate set of weights in the context of particular investigations is usually quite difficult. Here, the temptation to devise a weighting scheme that will either produce a particular ordination or speed calculations will be ever-present. Such temptations should be resisted.

The final type of similarity index often employed in QFA is Imbrie and Purdy's (1962) index of proportional similarity.

$$cos\,\theta_{i,j} = \frac{\sum_{k=1}^{p} x_{i,k}x_{j,k}}{\sqrt[2]{\sum_{k=1}^{p} x_{i,k}^2 \sum_{k=1}^{p} x_{j,k}^2}} \quad (1)$$

In this equation x_i is the i^{th} case; x_j = is the j^h case; and p = is the no. of variables. This is by far the similarity index most commonly employed in QFA and often constitutes the default index in public-domain and commercial statistics and data-analysis software packages. The proportional similarity index is often referred to as the "cosine θ" index because it conceptualizes cases as vectors in the variable space and represents similarity as the angular difference between all pairwise combinations of case vectors, taking no account of these vectors' lengths. Accordingly, the cosine θ index is a bounded index, constrained to vary between 0 and 1. The cosine θ, or proportional similarity, index is also a true similarity index insofar the trace of the resulting matrix is occupied by 1 s. If the variables used to compute the cosine θ index are standardized prior to similarity estimation the resulting values will be identical to those estimated from a calculation of the "correlation" across variables.

The Method

Once a similarity/distance matrix has been obtained the QFA procedure conforms closely to that of FA. The linear model used in QFA derives from Spearman's (1904) original formulation and separates the observed n by n similarity or distance structure among cases, calculated across all variables, into a set of a priori designated generalized factors with a residual, specific factor (e) associated with each of the n cases.

$$\begin{aligned} X_1 &= a_{1,1}F_1 + a_{1,2}F_2 \cdots + a_{1,m}F_m + e_1 \\ X_2 &= a_{2,1}F_1 + a_{2,2}F_2 \cdots + a_{2,m}F_m + e_2 \\ X_3 &= a_{3,1}F_1 + a_{3,2}F_2 \cdots + a_{3,m}F_m + e_3 \\ &\vdots \\ X_j &= a_{j,1}F_1 + b_{j,2}F_2 \cdots + a_{j,m}F_m + e_j \end{aligned} \quad (2)$$

In this expression a_i is a standardized score the i^{th} case, F is the "factor" value for that case across all measured or observed variables, and e_i is the part of the distance or similarity not accounted for by the factor structure. Q-mode factor analysis attempts to find a set of m linear factors (F, where $m < p$) that underly the sample's similarity or distance structure and, hopefully, provide a good estimate of the similarity/distance structure of the larger population from which the sample was drawn.

Choosing and Extracting the Factors

In a manner analogous to its R-mode counterpart, QFA begins with a principal coordinate analysis of a properly selected basis matrix whose structure reflects those aspects of between-case similarity, distance or association most appropriate for the investigation at hand. The decision of which similarity, distance or association index to use is not a trivial one as it will influence all aspects of the subsequent analysis. In addition, the number of factors the investigator wishes to construct their model around also needs to be specified at the outset of analysis. The guidance for factor number selection offered in the presentation of R-mode FA can also be applied to QFA. As always, there is no substitute for having a detailed understanding of the system being modeled. If there is uncertainty regarding the number of factors controlling the dataset's distance/similarity structure, QFA may proceed, but its results should be interpreted with caution.

Like R-mode FA, QFA scales the loadings of the retained principal coordinates by their eigenvalues, according to the following expression.

$$a_{i,m} = \sqrt[2]{\lambda_m} b_{i,m} \quad (3)$$

Here, λ_m is the eigenvalue associated with the m^{th} principal coordinate with m being set to the number of retained factors and b the eigenvector loading of the i^{th} case on the m^{th} principal coordinate (= factor). This scaling, along with the reduction in the number of factors being considered, represents a basic difference between PCoord and QFA.

Next, the "communality" values of the scaled factors are calculated as the sum of squares of the factor loading values (a) across all factors. This summarizes the proportion of similarity, distance or association provided by each case to each factor. The quantity '1- the sum of communality values for each case, then, expresses the proportion of the case's distance, similarity, or association structure attributable to the error term (e) of the factor model. If the correct number of factors has been chosen, all the summed, squared factor values should exhibit a high communality for each case with a small residual error.

Once the QFA factor equations have been determined, the factor loading values constitute the positions of objects, samples or events projected into the space formed from the eigenvalue-scaled orthogonal factors. These projection scores ($\widehat{X}$) may then be tabulated, plotted, inspected and interpreted in the manner normal for PCoord ordinations. Table 3 and Fig. 1 compare and contrast results of a PCoord and two-factor QFA solution for the trilobite data listed in MacLeod (2006).

Note that, despite the cosine θ basis for this analysis being a square 20 x 20 matrix, only three principal coordinates can be extracted from this small, demonstration dataset. This limitation stems from the fact that the data from which the similarity matrix was calculated included only three variables: body length, glabella length and glabella width. For this simple dataset both the first principal component and first factor would be interpreted as being consistent with generalized size variation owing to their uniformly positive loading coefficients. Similarly, for both analyses the second component/factor would be interpreted as reflecting localized shape variation owing to the contrast between signs among the different taxa.

Figure 1 illustrates the distribution of these trilobite body shapes in the ordination space formed by the first two principal coordinate axes and the two retained factor axes. While the relative positions of the projected data points are also similar, and discontinuities in the gross form distributions are evident in both plots, the scaling of the PCoord and FA axes differ reflecting the additional eigenvalue-based scaling calculation that is part of component-based QFA. Irrespective of this correspondence, an important change has taken place in the ability of the QFA vectors to represent the structure of the original cosine θ similarity matrix (Table 4).

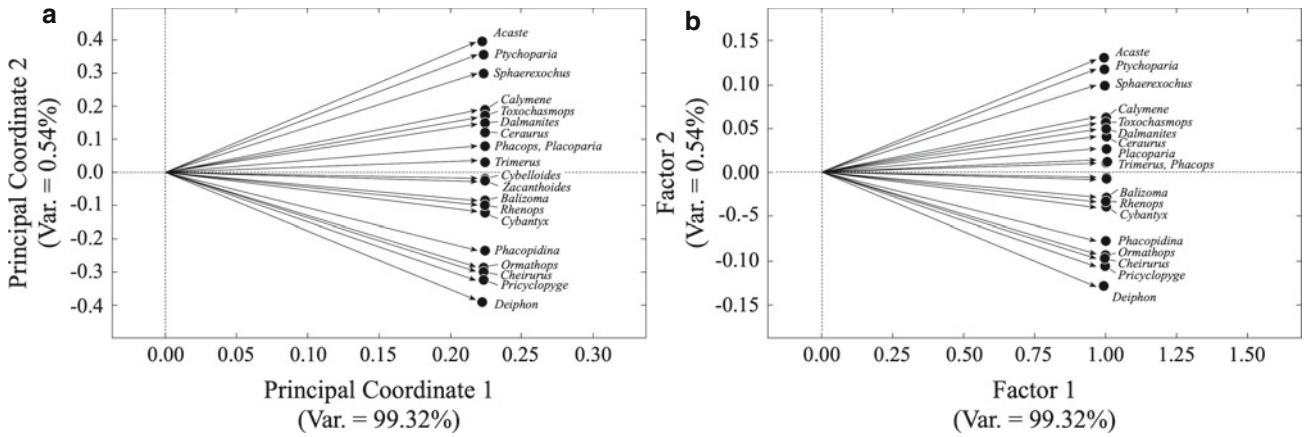
From these results it is evident that the two-factor QFA has been much more successful in reproducing the entire structure of the original cosine θ similarity matrix than principal coordinate analysis. If all 20 principal coordinates had been used in these calculations the trace of the original similarity matrix would be reproduced exactly, but none of the off-diagonal elements. This simple demonstration illustrates how the eigenvalue-scaling operation, that stands at the core of both FA and QFA, render these approaches to data analysis more information-rich than their PCA and PCoord counterparts, especially if a dramatic reduction in causal-factor dimensionality is required. The validity of applying either the FA or

QFA models to a dataset is predicated, however, on the dataset exhibiting a factor-based variation/similarity structure.

Aside from the use of different indices to quantify the structure of pairwise relations among cases, QFA also differs from *R*-mode FA in a number of other, more subtle ways. For example, it is standard practice to mean center each variable prior to analysis in PCA and FA so the ordination space created as a result of those procedures will be centered on the origin of the coordinate system within which the ordination results will be displayed. This operation is typically not performed in QFA (or PCoord) for either the set of variables or the set of cases. In this way, the ordination actually expresses the orientation of eigenvectors all of whose tails emanate from the coordinate system origin, irrespective of whether the shafts or bodies of the eigenvectors are drawn in the *Q*-mode ordination space plots. It would also make little

Q-Mode Factor Analysis, Table 3 Principal coordinate loadings, *Q*-mode factor loadings, and factor communalities for the example trilobite morphometric variables

	Principal Coordinates			Factors		
Parameter/ Variable	1	2	3	1	2	Communality
Eigenvalue	19.865 (99.33%)	0.108 (0.540%)	0.027 (0.134%)	19.865 (99.33%)	0.108 (0.540%)	
Acaste	0.222	0.400	0.090	0.991	0.131	1.000
Balizoma	0.224	−0.084	0.032	1.000	−0.028	1.000
Calymene	0.224	0.192	0.009	0.998	0.063	1.000
Ceraurus	0.224	0.124	−0.022	0.999	0.041	1.000
Cheirurus	0.223	−0.294	0.326	0.994	−0.097	0.997
Cybantyx	0.224	−0.119	−0.137	0.999	−0.039	0.999
Cybeloides	0.224	−0.016	0.075	1.000	−0.005	1.000
Dalmanites	0.224	0.152	−0.277	0.998	0.050	0.998
Deiphon	0.222	−0.388	0.344	0.990	−0.128	0.997
Ormathops	0.223	−0.283	−0.209	0.995	−0.093	0.999
Phacopidina	0.224	−0.233	−0.055	0.997	−0.077	1.000
Phacops	0.224	0.033	0.444	0.997	0.011	0.995
Placopania	0.224	0.083	−0.059	1.000	0.027	1.000
Pricyclopyge	0.223	−0.321	−0.211	0.994	−0.106	0.999
Ptychoparia	0.223	0.359	−0.202	0.992	0.118	0.999
Rhenops	0.224	−0.097	−0.422	0.997	−0.032	0.995
Sphaerexochus	0.223	0.301	0.170	0.995	0.099	0.999
Toxochasmops	0.224	0.174	0.316	0.997	0.057	0.997
Trimerus	0.224	0.035	−0.064	1.000	0.012	1.000
Zacanthoides	0.224	−0.019	−0.144	1.000	−0.006	0.999



Q-Mode Factor Analysis, Fig. 1 Ordination spaces formed by the first two principal coordinates (A) and a two-factor extraction (B) from the three-variable trilobite dataset. Note the similarity of these results in terms of the relative positions of forms projected into the PCoord and factor spaces and the difference in terms of the axis scales

_Q_-Mode Factor Analysis, Table 4 Original cosine θ similarity matrix (upper), cosine θ matrix reproduced on the basis of the first to principal coordinate (middle), and cosine θ matrix reproduced by the two-factor FA solution (lower). Note the bottom two matrices are based on the component/loading values shown in Table 3

Original Similarity Matrix

Genus	*Acaste*	*Balizoma*	*Calymene*	*Ceraurus*	*Cheirurus*	*Cybantyx*	*Cybeloides*	*Dalmanites*	*Deiphon*	*Ormathops*	*Phacopidina*	*Phacops*	*Placopania*	*Pricyclopyge*	*Ptychoparia*	*Rhenops*	*Sphaerexochus*	*Toxochasmops*	*Trimerus*	*Zacanthoides*
Acaste	1.000	0.987	0.998	0.996	0.973	0.985	0.991	0.995	0.966	0.974	0.978	0.991	0.994	0.971	0.999	0.983	0.999	0.997	0.992	0.990
Balizoma	0.987	1.000	0.996	0.998	0.996	1.000	1.000	0.996	0.994	0.997	0.999	0.997	0.998	0.996	0.989	0.997	0.992	0.995	0.999	0.999
Calymene	0.998	0.996	1.000	1.000	0.986	0.994	0.998	0.999	0.980	0.987	0.990	0.996	0.999	0.985	0.998	0.993	0.999	0.999	0.999	0.997
Ceraurus	0.996	0.998	1.000	1.000	0.989	0.997	0.999	0.999	0.984	0.991	0.993	0.997	1.000	0.989	0.997	0.995	0.998	0.998	1.000	0.999
Cheirurus	0.973	0.996	0.986	0.989	1.000	0.995	0.995	0.984	1.000	0.996	0.998	0.994	0.990	0.996	0.973	0.990	0.981	0.988	0.992	0.993
Cybantyx	0.985	1.000	0.994	0.997	0.995	1.000	0.999	0.996	0.993	0.998	0.999	0.994	0.998	0.998	0.988	0.999	0.989	0.993	0.999	0.999
Cybeloides	0.991	1.000	0.998	0.999	0.995	0.999	1.000	0.997	0.991	0.995	0.997	0.998	0.999	0.994	0.991	0.996	0.994	0.997	1.000	0.999
Dalmanites	0.995	0.996	0.999	0.999	0.984	0.996	0.997	1.000	0.979	0.990	0.991	0.992	0.999	0.988	0.998	0.996	0.996	0.995	0.999	0.998
Deiphon	0.966	0.994	0.980	0.984	1.000	0.993	0.991	0.979	1.000	0.995	0.997	0.990	0.986	0.996	0.966	0.988	0.974	0.983	0.988	0.989
Ormathops	0.974	0.997	0.987	0.991	0.996	0.998	0.995	0.990	0.995	1.000	1.000	0.989	0.992	1.000	0.978	0.998	0.980	0.985	0.994	0.996
Phacopidina	0.978	0.999	0.990	0.993	0.998	0.999	0.997	0.991	0.997	1.000	1.000	0.993	0.995	0.999	0.981	0.997	0.984	0.989	0.996	0.997
Phacops	0.991	0.997	0.996	0.997	0.994	0.994	0.998	0.992	0.990	0.989	0.993	1.000	0.996	0.987	0.989	0.989	0.995	0.999	0.997	0.995
Placopania	0.994	0.998	0.999	1.000	0.990	0.998	0.999	0.999	0.986	0.992	0.995	0.996	1.000	0.991	0.996	0.996	0.997	0.998	1.000	0.999
Pricyclopyge	0.971	0.996	0.985	0.989	0.996	0.998	0.994	0.988	0.996	1.000	0.999	0.987	0.991	1.000	0.975	0.997	0.977	0.983	0.993	0.995
Ptychoparia	0.999	0.989	0.998	0.997	0.973	0.988	0.991	0.998	0.966	0.978	0.981	0.989	0.996	0.975	1.000	0.988	0.998	0.995	0.994	0.992
Rhenops	0.983	0.997	0.993	0.995	0.990	0.999	0.996	0.996	0.988	0.998	0.997	0.989	0.996	0.997	0.988	1.000	0.987	0.989	0.997	0.999
Sphaerexochus	0.999	0.992	0.999	0.998	0.981	0.989	0.994	0.996	0.974	0.980	0.984	0.995	0.997	0.977	0.998	0.987	1.000	0.999	0.995	0.993
Toxochasmops	0.997	0.995	0.999	0.998	0.988	0.993	0.997	0.995	0.983	0.985	0.989	0.999	0.998	0.983	0.995	0.989	0.999	1.000	0.997	0.995
Trimerus	0.992	0.999	0.999	1.000	0.992	0.999	1.000	0.999	0.988	0.994	0.996	0.997	1.000	0.993	0.994	0.997	0.995	0.997	1.000	1.000
Zacanthoides	0.990	0.999	0.997	0.999	0.993	0.999	0.999	0.998	0.989	0.996	0.997	0.995	0.999	0.995	0.992	0.999	0.993	0.995	1.000	1.000

Reproduced Similarity Matrix (PCoord)

Genus	*Acaste*	*Balizoma*	*Calymene*	*Ceraurus*	*Cheirurus*	*Cybantyx*	*Cybeloides*	*Dalmanites*	*Deiphon*	*Ormathops*	*Phacopidina*	*Phacops*	*Placopania*	*Pricyclopyge*	*Ptychoparia*	*Rhenops*	*Sphaerexochus*	*Toxochasmops*	*Trimerus*	*Zacanthoides*
Acaste	0.209	0.016	0.127	0.099	-0.068	0.002	0.043	0.111	-0.106	-0.063	-0.043	0.063	0.083	-0.079	0.193	0.011	0.170	0.119	0.064	0.042
Balizoma	0.016	0.057	0.034	0.040	0.075	0.060	0.052	0.037	0.083	0.074	0.070	0.047	0.043	0.077	0.020	0.058	0.025	0.035	0.047	0.052
Calymene	0.127	0.034	0.087	0.074	-0.007	0.027	0.047	0.079	-0.025	-0.004	0.005	0.056	0.066	-0.012	0.119	0.032	0.108	0.083	0.057	0.047
Ceraurus	0.099	0.040	0.074	0.066	0.014	0.036	0.048	0.069	0.002	0.015	0.021	0.054	0.061	0.010	0.094	0.038	0.087	0.072	0.055	0.048
Cheirurus	-0.068	0.075	-0.007	0.014	0.136	0.085	0.055	0.005	0.164	0.133	0.118	0.040	0.025	0.144	-0.056	0.078	-0.039	-0.001	0.040	0.056
Cybantyx	0.002	0.060	0.027	0.036	0.085	0.064	0.052	0.032	0.096	0.084	0.078	0.046	0.040	0.088	0.007	0.062	0.014	0.029	0.046	0.053
Cybeloides	0.043	0.052	0.047	0.048	0.055	0.052	0.051	0.048	0.056	0.055	0.054	0.050	0.049	0.055	0.044	0.052	0.045	0.047	0.050	0.051
Dalmanites	0.111	0.037	0.079	0.069	0.005	0.032	0.048	0.073	-0.009	0.007	0.015	0.055	0.063	0.001	0.104	0.035	0.096	0.077	0.056	0.047
Deiphon	-0.106	0.083	-0.025	0.002	0.164	0.096	0.056	-0.009	0.200	0.160	0.140	0.037	0.017	0.174	-0.090	0.087	-0.067	-0.018	0.036	0.057
Ormathops	-0.063	0.074	-0.004	0.015	0.133	0.084	0.055	0.007	0.160	0.130	0.116	0.041	0.026	0.141	-0.052	0.077	-0.035	0.001	0.040	0.055
Phacopidina	-0.043	0.070	0.005	0.021	0.118	0.078	0.054	0.015	0.140	0.116	0.104	0.042	0.031	0.125	-0.034	0.073	-0.020	0.009	0.042	0.055
Phacops	0.063	0.047	0.056	0.054	0.040	0.046	0.050	0.055	0.037	0.041	0.042	0.051	0.053	0.039	0.062	0.047	0.060	0.056	0.051	0.050
Placopania	0.083	0.043	0.066	0.061	0.025	0.040	0.049	0.063	0.017	0.026	0.031	0.053	0.057	0.023	0.080	0.042	0.075	0.065	0.053	0.049
Pricyclopyge	-0.079	0.077	-0.012	0.010	0.144	0.088	0.055	0.001	0.174	0.141	0.125	0.039	0.023	0.153	-0.065	0.081	-0.047	-0.006	0.039	0.056
Ptychoparia	0.193	0.020	0.119	0.094	-0.056	0.007	0.044	0.104	-0.090	-0.052	-0.034	0.062	0.080	-0.065	0.178	0.015	0.158	0.112	0.063	0.043
Rhenops	0.011	0.058	0.032	0.038	0.078	0.062	0.052	0.035	0.087	0.077	0.073	0.047	0.042	0.081	0.015	0.059	0.021	0.033	0.047	0.052
Sphaerexochus	0.170	0.025	0.108	0.087	-0.039	0.014	0.045	0.096	-0.067	-0.035	-0.020	0.060	0.075	-0.047	0.158	0.021	0.140	0.102	0.061	0.044
Toxochasmops	0.119	0.035	0.083	0.072	-0.001	0.029	0.047	0.077	-0.018	0.001	0.009	0.056	0.065	-0.006	0.112	0.033	0.102	0.080	0.056	0.047
Trimerus	0.064	0.047	0.057	0.055	0.040	0.046	0.050	0.056	0.036	0.040	0.042	0.051	0.053	0.039	0.063	0.047	0.061	0.056	0.052	0.050
Zacanthoides	0.042	0.052	0.047	0.048	0.056	0.053	0.051	0.047	0.057	0.055	0.055	0.050	0.049	0.056	0.043	0.052	0.044	0.047	0.050	0.051

(continued)

Q-Mode Factor Analysis, Table 4 (continued)

Reproduced Similarity Matrix (QFA)

Genus	*Acaste*	*Balizoma*	*Calymene*	*Ceraurus*	*Cheirurus*	*Cybantyx*	*Cybeloides*	*Dalmanites*	*Deiphon*	*Ormathops*	*Phacopidina*	*Phacops*	*Placopania*	*Pricyclopyge*	*Ptychoparia*	*Rhenops*	*Sphaerexochus*	*Toxochasmops*	*Trimerus*	*Zacanthoides*
Acaste	1.000	0.987	0.998	0.996	0.972	0.985	0.990	0.996	0.965	0.974	0.978	0.990	0.994	0.971	0.999	0.984	0.999	0.996	0.993	0.990
Balizoma	0.987	1.000	0.996	0.998	0.996	1.000	1.000	0.996	0.993	0.997	0.999	0.997	0.998	0.996	0.989	0.998	0.992	0.995	0.999	0.999
Calymene	0.998	0.996	1.000	1.000	0.986	0.995	0.998	0.999	0.980	0.987	0.990	0.996	0.999	0.985	0.998	0.993	0.999	0.999	0.999	0.997
Ceraurus	0.996	0.998	1.000	1.000	0.989	0.997	0.999	0.999	0.984	0.990	0.993	0.997	1.000	0.989	0.996	0.995	0.998	0.999	1.000	0.999
Cheirurus	0.972	0.996	0.986	0.989	0.997	0.997	0.994	0.987	0.997	0.998	0.998	0.990	0.991	0.998	0.975	0.994	0.979	0.985	0.993	0.994
Cybantyx	0.985	1.000	0.995	0.997	0.997	0.999	0.999	0.995	0.994	0.998	0.999	0.996	0.997	0.997	0.987	0.997	0.990	0.994	0.998	0.999
Cybeloides	0.990	1.000	0.998	0.999	0.994	0.999	1.000	0.997	0.991	0.995	0.997	0.997	0.999	0.994	0.992	0.997	0.994	0.997	1.000	1.000
Dalmanites	0.996	0.996	0.999	0.999	0.987	0.995	0.997	0.998	0.982	0.988	0.991	0.996	0.999	0.986	0.996	0.993	0.997	0.998	0.998	0.997
Deiphon	0.965	0.993	0.980	0.984	0.997	0.994	0.991	0.982	0.997	0.997	0.997	0.986	0.986	0.998	0.968	0.991	0.972	0.980	0.989	0.991
Ormathops	0.974	0.997	0.987	0.990	0.998	0.998	0.995	0.988	0.997	0.999	0.999	0.991	0.992	0.999	0.977	0.995	0.981	0.987	0.994	0.995
Phacopidina	0.978	0.999	0.990	0.993	0.998	0.999	0.997	0.991	0.997	0.999	1.000	0.993	0.995	0.999	0.980	0.997	0.984	0.990	0.996	0.997
Phacops	0.990	0.997	0.996	0.997	0.990	0.996	0.997	0.996	0.986	0.991	0.993	0.995	0.997	0.990	0.991	0.994	0.993	0.995	0.997	0.997
Placopania	0.994	0.998	0.999	1.000	0.991	0.997	0.999	0.999	0.986	0.992	0.995	0.997	1.000	0.991	0.995	0.996	0.997	0.998	1.000	0.999
Pricyclopyge	0.971	0.996	0.985	0.989	0.998	0.997	0.994	0.986	0.998	0.999	0.999	0.990	0.991	0.999	0.974	0.994	0.978	0.985	0.992	0.994
Ptychoparia	0.999	0.989	0.998	0.996	0.975	0.987	0.992	0.996	0.968	0.977	0.980	0.991	0.995	0.974	0.999	0.986	0.999	0.996	0.994	0.991
Rhenops	0.984	0.998	0.993	0.995	0.994	0.997	0.997	0.993	0.991	0.995	0.997	0.994	0.996	0.994	0.986	0.995	0.989	0.992	0.997	0.997
Sphaerexochus	0.999	0.992	0.999	0.998	0.979	0.990	0.994	0.997	0.972	0.981	0.984	0.993	0.997	0.978	0.999	0.989	0.999	0.997	0.996	0.994
Toxochasmops	0.996	0.995	0.999	0.999	0.985	0.994	0.997	0.998	0.980	0.987	0.990	0.995	0.998	0.985	0.996	0.992	0.997	0.997	0.998	0.996
Trimerus	0.993	0.999	0.999	1.000	0.993	0.998	1.000	0.998	0.989	0.994	0.996	0.997	1.000	0.992	0.994	0.997	0.996	0.998	1.000	1.000
Zacanthoides	0.990	0.999	0.997	0.999	0.994	0.999	1.000	0.997	0.991	0.995	0.997	0.997	0.999	0.994	0.991	0.997	0.994	0.996	1.000	0.999

sense to mean center the case rows across a set of variables of very different types and magnitude ranges.

When QFA ordinations are examined the loadings (= ordination scores) on the first factor (F1) will often all exhibit comparably high positive values. This is usually is a reflection of the "size" or combined magnitudes of all the variables included in the analysis. Since there often seems to be little difference among the cases along F1 it is often regarded as an irrelevant or "nuisance" factor, especially when compared to the oftentimes more informative distinctions between cases revealed in the plots of higher-level QFA ordination spaces. Nonetheless, as uniformly high correlations among variables are not characteristic of all earth science datasets, it is always a good idea to inspect the F1 vs F2 and (if necessary) F1 vs F3 ordination plots to determine whether any useful insight can be gained into the structure of between-case similarity/distance relations from an interpretation of F1.

In some instances the variables being analyzed might be expressed as either compositional proportions or percentages such that the sum of each row is constrained to add up to 1.0 or 100.0, respectively. Such datasets are said to have been artificially "closed". In such cases, application of a centered log-ratio transformation prior to analysis is usually recommended in order to mitigate the effects of the closure constraint.

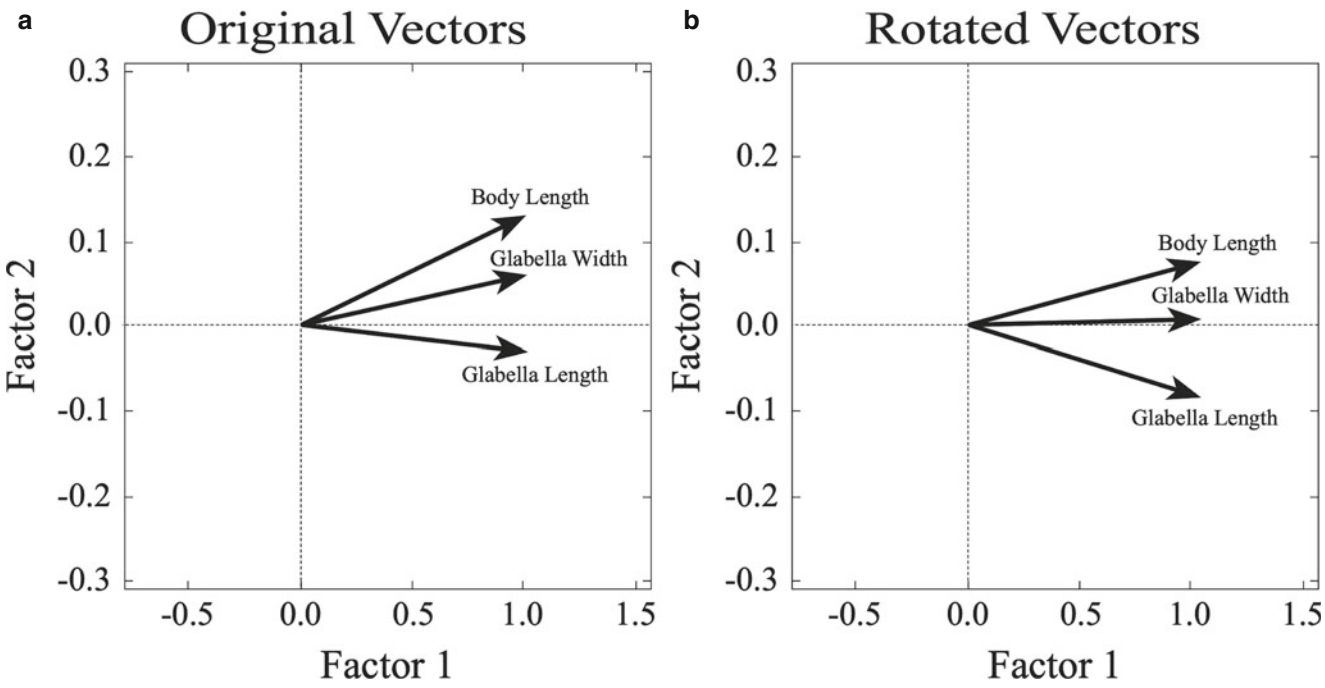
Factor Rotation

As with FA, QFA factor axes may be rotated relative to the variable axes in order to improve their interpretability. Axis rotation is less commonly applied in QFA because the emphasis is often on identifying the structure of similarity relations among cases rather than interpretation of the factors in terms of the original variables. Nonetheless, such interpretations are possible and factor-axis rotations often make this interpretation easier. Operationally this amounts to adjusting the set of factor loadings, rigidly in a geometric sense, until all variables exhibit loadings at or as close to 0.0 or $\pm$ 1.0 across all factors as possible.

Neuhaus and Wrigley's (1954) quartimax, and Kaiser's (1958) varimax orthogonal axis-rotation procedures can be used to perform this operation. If even more simplification is desired the constraint of the factor axes maintaining orthogonal relations can be relaxed using procedures such Carroll's (1953) quartimin or Imbrie's (1963) oblique axis rotation, in which case the factor axes will coincide with the orientation of the extreme variable vectors. Figure 2 illustrates the unrotated and varimax-rotated solutions for the example trilobite dataset.

Conclusion

Q-mode factor analysis is often used to array cases along a spectrum defined by opposing case extrema. In some instances the purpose is to identify these extreme end-members while, in others, it is to gauge which of these spectral end-members particular cases are closest (= most similar) to. Such analyses can also be accomplished, to some extent, by *Q*-mode cluster analysis. However, in most



Q-Mode Factor Analysis, Fig. 2 Original (**a**) and rotated (**b**) variable-axis orientations with respect to the fixed, orthogonal factor axes for the three-variable trilobite data. The rotation of these variables is centered on Factor 1 owing to their high mutual correlations (reflected in the high Factor 1 eigenvalue, see Table 3). More complex datasets will typically exhibit larger variable-vector rotations

cases cluster analysis represents the structure of between-case similarity relations as a hierarchy despite the fact that the generative processes that give rise to between-case structural relations might not be organized hierarchically. For systems in which similarity-distance relations are known, or suspected, not to be structured hierarchically, QFA may be the more appropriate conceptual model to employ in addition to one that will minimize the distortions that often accompany attempts to portray non-hierarchically structured data using hierarchical models.

Q-mode factor analysis has proven especially attractive to geochemists, mineralogists petrologists, hydrologists, paleoecologists, and biostratigraphers who often deal with compositional data and find themselves in need of a procedure to non-hierarchically define both end-members and gradient orderings of cases in which a number of influences are mixed in varying proportions. In perhaps one of the more scientifically significant applications of the method, Imbrie and Kipp (1971) employed QFA to develop a sets of linear transfer functions whereby sea surface temperatures for ancient ocean basins could be inferred from the relative abundances of Cenozoic planktonic foraminifer species. Sepkoski (1981) also used QFA to summarize the diversity history marine life across the Phanerozoic and used the results generated through its application to recognize the three great evolutionary faunas that, together, constitute fundamental macroevolutionary historio-structural units of life on Earth. These are but two prominent examples of QFAs potential for making important data analysis-based advances in the study of earth history.

Cross-References

- Association Coefficients
- Biology
- Cluster Analysis
- Correlation
- Cosine θ Similarity Index
- Covariance
- Dice Association Coefficient
- Dissimilarity
- Distance
- Eigenanalysis
- Eigenvalues
- Eigenvectors
- Euclidian
- Gower Distance
- Index of Proportional Similarity
- Interval Variable
- Jaccard Association Coefficient
- Mahalanobis Distance
- Manhatten Distance
- Multivariate Data Analysis
- Ordination
- Otsuka Association Coefficient

- Outlier
- Pelycosaur
- Phylogeny
- Planktonic Foraminifera
- Principal Component Analysis
- Principal Coordinate Analysis
- Psychology
- Ratio Variable
- *R*-mode Analysis
- *R*-mode Factor Analysis
- Shape
- Similarity
- Simpson Association Coefficient
- Size
- Variance
- Vector

Bibliography

Carroll JB (1953) An analytical solution for approximating simple structure in factor analysis. Psychom Theory 18:23–38

Cheetham AH, Hazel JE (2012) Binary (presence-absence) similarity coefficients. J Paleontol 43:1130–1136

Davis JC (2002) Statistics and data analysis in geology, 3rd edn. Wiley, New York

Gould SJ (1967) Evolutionary patterns in pelycosaurian reptiles: a factor analytic study. Evolution 21:385–401

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871

Imbrie J (1963) Factor and vector analysis programs for analyzing geologic data. Office of Naval Research, Technical Branch, Report 6:83

Imbrie J, Kipp NG (1971) A new micropaleontological method for quantitative paleoclimatology: application to a Pleistocene Caribbean core. In: Turekian KK (ed) The Late Cenozoic Glacial Ages. Yale University Press, New Haven, pp 72–181

Imbrie J, Purdy E (1962) Classification of modern Bahamian carbonate sediments. AAPG Mem 7:253–272

Kaiser HF (1958) The varimax criterion for analytic rotations in factor analysis. Psychometrika 23:187–200

MacLeod N (2006) Minding your Rs and Qs. Palaeontological Assoc Newsl 61:42–60

Neuhaus JO, Wrigley C (1954) The quartimax method: an analytical approach to orthogonal simple structure. Br J Stat Psychol 7:81–91

Pearson K (1901) On lines and planes of closest fit to a system of points in space. Philos Mag 2:557–572

Sepkoski JJ (1981) A factor analytic description of the Phanerozoic marine fossil record. Paleobiology 7:36–53

Spearman C (1904) "General intelligence", objectively determined and measured. Am J Psychol 15:201–293