12.740 Paleoceanography
Spring 2008

Spring 2006 Lecture 04b

## Fundamentals of Factor Analysis: Satellite Image

I. Correlation

A. Variance and Covariance

1. Variance (of **x** is denoted $\mathbf{S_x^2}$; variance of **y** is denoted $\mathbf{S_y^2}$) is a measure of the scatter of values of a variable about its mean:

$$S_x^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 / n$$

2. Covariance (of **x** and **y**) expresses the relationship between two variables (a measure of the scatter of values of points in a plane relative to the centroid of the data set):

$$S_{xy}^2 = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) / n$$

B. Correlation coefficient

$$r = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - \overline{x}}{S_x})(\frac{y_i - \overline{y}}{S_y}) = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$$

–

where $\mathbf{S_x}$ = std. dev. of $\mathbf{x} = (\Sigma(\mathbf{x_i} - \overline{\mathbf{x}}\ \ )^2/\mathbf{n-1})^{1/2}$

and $\mathbf{S_y}$ = std. dev. of $\mathbf{y} = (\Sigma(\mathbf{y_i} - \overline{\mathbf{y}}\ \ )^2/\mathbf{n-1})^{1/2}$

i.e. $\mathbf{S_x}$ and $\mathbf{S_y}$ relate the deviations of points from the average relative to the "range" (actually std. dev.) of the observations.
In other words, the Covariance divided by the Variance.

3. $\mathbf{r^2}$ is "the variance of **Y** accounted for by its covariance with **x**" (usually expressed in % units).

C. Relation to linear regressions (x on y; y on x; others)

1. Common linear regression of **y** on **x**: **y = A + Bx**

Let $\mathbf{S = \Sigma\ (y_i - A - B\ x_i)^2}$

Set $\mathbf{\partial S/\partial A = 0}$; $\mathbf{\partial S/\partial B = 0}$; solve for **A** and **B**.

2. Matrix math solution of Linear Regression:

for eq'n $\mathbf{A\underline{x} = \underline{b}}$ (m eq'ns, n unknowns),
<u>if</u> columns of A are <u>linearly independent</u>,
then:

$$\underline{x} = (A^T A)^{-1} A^T \underline{b}$$

a. For example, for the simple linear regression

$$y = C + Dx,$$

where we want to fit pairs of data

$$x_i, \ y_i$$

we want to find

$$C, \ D$$

that mimimize

$$\Sigma \ [y_i - (C + Dx_i)]^2$$

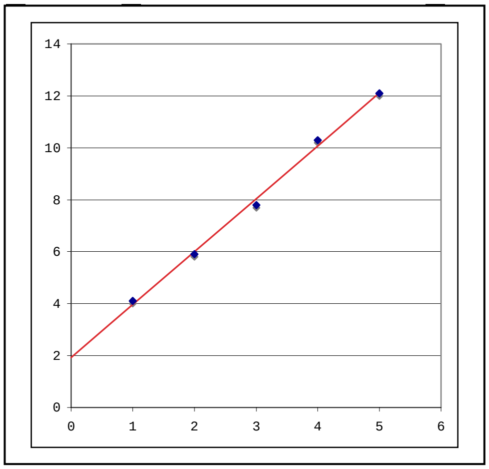
In matrix form, we write the equation **y = C + Dx** as:

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

$$\text{i.e.} \qquad \mathbf{A} \qquad \mathbf{x} \quad = \quad \mathbf{b}$$

Numerical example:

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 4.1 \\ 5.9 \\ 7.8 \\ 10.3 \\ 12.1 \end{bmatrix}$$



gives C = 1.92 and D = 2.04

Similarly, to solve the equation $y = A + Bx + Cx^2$:

$$\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & . & \\ . & . & \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

b. Simple matrix formulas also allow you to compute the estimated uncertainties of the regression coefficients and the correlation coefficients.

## II. Correlation in n dimensions

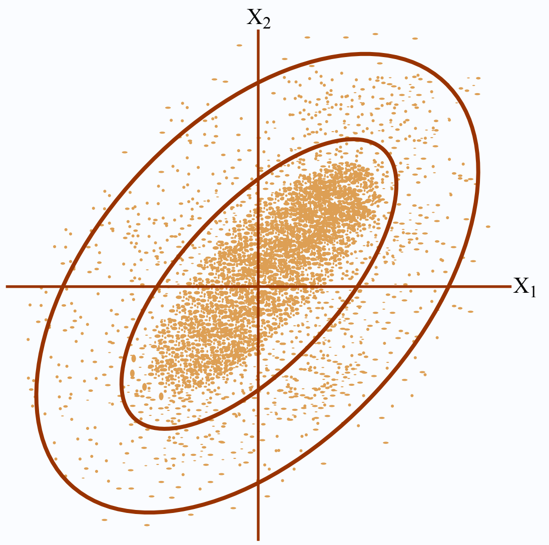
A. Multiple linear regression (e.g. $y = a + bx_1 + cx_2$, the equation for a plane in 3D space)

B. r-matrix (later we will refer to this as the matrix $\Sigma$)

| Property | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1.00 | 0.86 | 0.45 | 0.83 | 0.45 |
| 2 | 0.86 | 1.00 | 0.74 | 0.23 | 0.64 |
| 3 | 0.45 | 0.74 | 1.00 | 0.78 | 0.57 |
| 4 | 0.83 | 0.23 | 0.78 | 1.00 | 0.39 |
| 5 | 0.45 | 0.64 | 0.57 | 0.39 | 1.00 |

$(\rho_{ij})$

C. Ellipsoids and eigenvectors



Bivariate scatter diagram with equal-density contour lines. The inner ellipse encloses 66% of the data points, The outer ellipse 95% of the data points.

Figure by MIT OpenCourseWare. Adapted from source: Joreskog et al. Geological Factor Analysis (1976).

$$\begin{array}{cccc} \mathbf{A} & \underline{\mathbf{x}} = & \lambda & \underline{\mathbf{x}} \\ \text{square matrix} & \text{eigen-vector} & \text{eigen-value} & \end{array}$$

One way to find eigenvectors:

1st eigenvector = major axis of ellipsoid
2nd eigenvector = largest minor axis of ellipsoid
etc.

This works more or less as if we did a regression to get a line that "explains" most of the variance (the dominant linear trend of the data in n-dimensional space), subtracted that regression from the data, then perform another regression to find the next most important contribution to the variance, and so forth.

1. This procedure works because we can rewrite the equation (II.C.) as:

$$(\mathbf{A} - \lambda \mathbf{I})\ \underline{\mathbf{x}} = 0 \qquad \textbf{(Equation II.C.1)}$$

In other words, the unknown vector **x** is orthogonal to all row vectors of **($\mathbf{A}$-$\lambda\mathbf{I}$)**.

The expression for the determinant of **$\mathbf{A}$-$\lambda\mathbf{I}$** is a polynomial of degree equal to the number of rows and columns of the square matrix **A**. The roots of the polynomial are the eigenvalues. If **A** is a real, square, and symmetric matrix, the roots are always real. However, these **p** eigenvalues may not always be different and some may equal zero. If two or more eigenvalues are equal, we say that they are multiple eigenvalues; otherwise we say the eigenvalue is distinct.

Example: consider a 2x2 symmetrical singular (i.e., determinant = 0) matrix; equation II.C.1 is then:

$$\begin{vmatrix} r_{11}-\lambda & r_{12} \\ r_{21} & r_{22}-\lambda \end{vmatrix} = 0$$

$$(r_{11}-\lambda)(r_{22}-\lambda) - r_{12}r_{21} = 0$$

$$\lambda^2 - \lambda(r_{11}+r_{22}) + (r_{11}r_{22} - r_{12}r_{21}) = 0$$

i.e., a simple quadratic equation (also: remember, by assumption of symmetry, $r_{12}=r_{21}$)

Things get out of hand quickly as the matrix gets bigger (that's what computers are for!).

2. Once the eigenvalues are known, the eigenvectors can be calculated from (a.)

   a. A unique solution cannot be obtained for an eigenvector. If **x** is a solution, so is **cx**, where **c** is a scalar. By convention, eigenvectors are therefore always normalized (unit length).

   b. Eigenvectors associated with different eigenvalues are orthogonal. Multiple solutions are possible for multiple eigenvalues, but it is always possible to choose an eigenvector which is orthogonal to all other eigenvectors.

3. If the eigenvalues $\lambda$ are placed as the elements of a diagonal matrix $\Lambda$, and the eigenvectors are collected as columns into the matrix **U**, then eq. (II.C.1) becomes:

$$\mathbf{A}\ \mathbf{U} = \mathbf{U}\ \Lambda$$

$$\begin{array}{c|c|}
 & 1\;2\;3\ldots\quad n \\
1 & \\
 & \\
2 & n \times n \\
 & \text{Square} \\
3 & \text{Matrix} \\
 & \\
 & \\
n & 
\end{array}
\;
\begin{array}{|ccc|}
1 & 2 & n \\
EV_1 & EV_2 \ldots & \\
\downarrow & \downarrow & \\
. & . & \\
. & . & \\
. & . & \\
. & . & \\
. & . & \\
. & . & 
\end{array}
=
\begin{array}{|ccc|}
1\;2\;3\ldots & & n \\
EV_1 & EV_2 \ldots & \\
\downarrow & \downarrow & \\
. & . & \\
. & . & \\
. & . & \\
. & . & \\
. & . & \\
. & . & 
\end{array}
\;
\begin{array}{|cccc|}
1 & 2 & 3\ldots & n \\
\lambda_1 & 0 & 0\ldots & \\
0 & \lambda_2 & & \\
 & & . & \\
 & & & . \\
 & & & \\
 & & & \\
 & & & \lambda_{\underline{n}}
\end{array}$$

4. The matrix U (the eigenvector matrix) is square orthonormal (the matrix is nxn and the eigenvectors are of unit length), so $\mathbf{U\ U^T = U^T\ U = I}$, therefore:

$$\mathbf{A = U\ \Lambda\ U^T}$$

Therefore any symmetric matrix such as the correlation coefficient table can be expressed in this form:

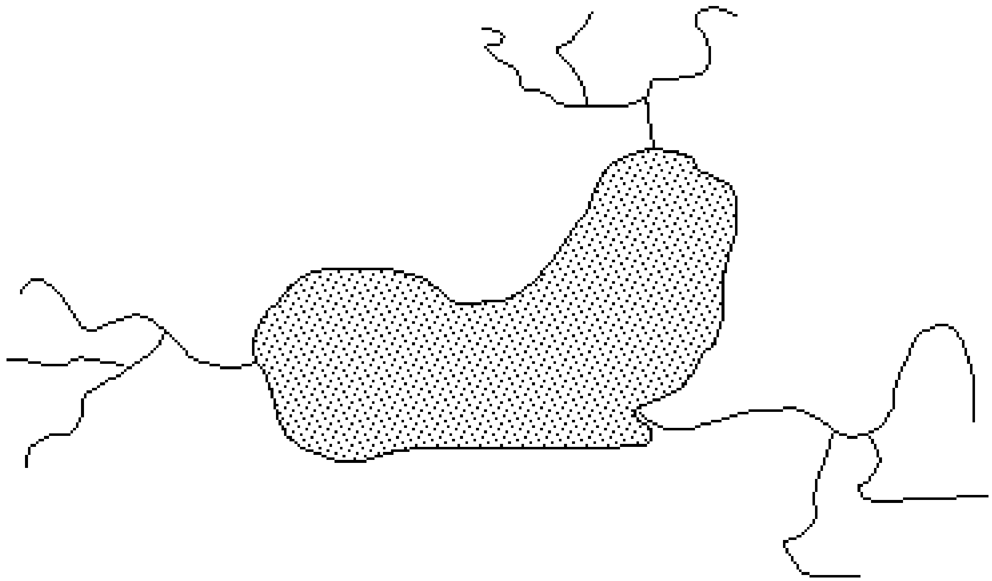
$$\begin{array}{llll}
\mathbf{\Sigma\ =} & \mathbf{U} & \mathbf{\Phi} & \mathbf{U^T} \\
\mathbf{m \times m} & \mathbf{m \times N} & \mathbf{N \times N} & \mathbf{N \times m} \\
\text{r-table} & & & \\
 & \text{eigenvectors} & & \\
 & & \text{eigenvalues} & \\
 & & & \text{eigenvectors(transpose)}
\end{array}$$

III. Principle Components Analysis: basically, the <u>principle components</u> are the eigenvectors as outlined above:

A. PCA is inherently variance-oriented; it accounts for maximum variance of all the observed variables. In other words, PCA accepts that large part of the total variance of a variable is important and common to other observed variables.

B. Factor analysis (below) is correlation-oriented; it accounts for the maximum intercorrelation of variables. In other words, factor analysis allows for a considerable "amount of uniqueness" to be present in the data and utilizes only that part of a variable that takes part in correlation with other variables; i.e., to account for covariance rather than variance.

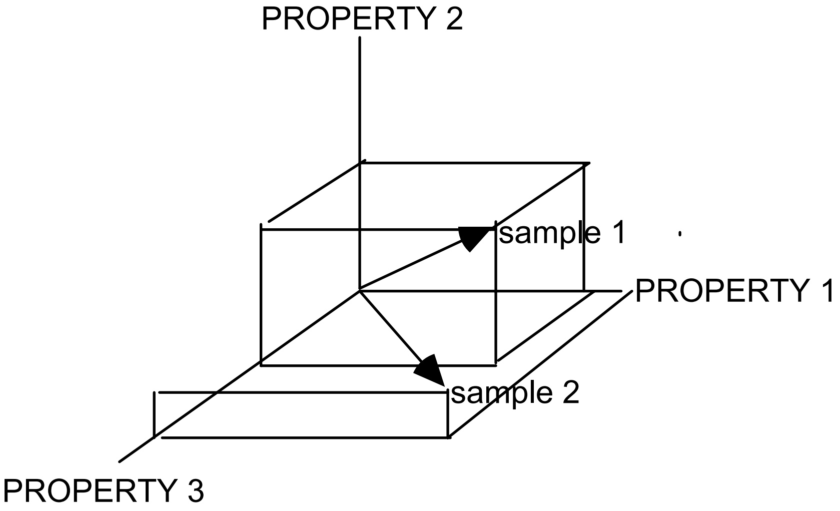
Example: imagine a lake with three stream systems draining into it. Suppose that each of the drainage basins has a distinctive (average) sediment composition, and that the sediments within the lake can be described as linear combinations of the sediments derived from each stream system. What is interesting in this case is the distribution of the three stream system component sediments, <u>not</u> the mineralogy or chemical composition per se (i.e., we don't care if Al correlates with Si). So we adopt a framework that lets us describe the raw data (chemical or mineralogical composition) in terms of sums of source components, and map the distribution of the components.



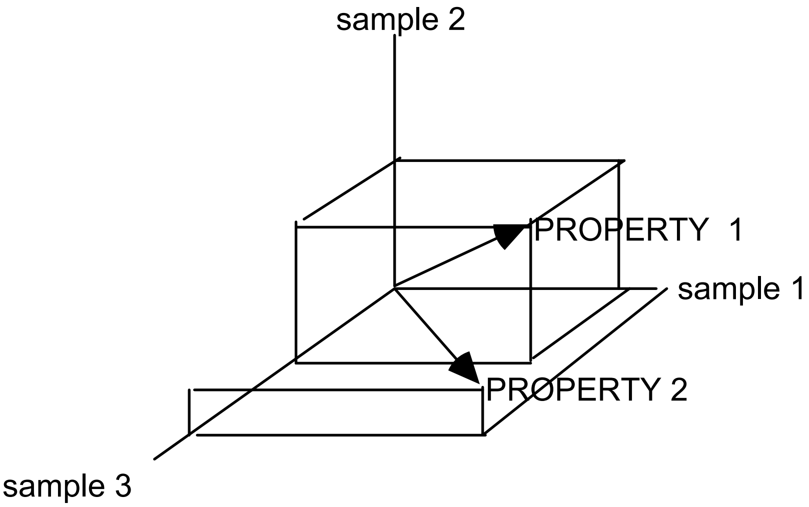
Source of example: Joreskog et al.

IV. Species space and sample space

A. Samples in species space:



B. R Mode (species in sample space):



## C. What is sample space?

1. Consider the species percentage data:

| | *PaL* | *PaR* | *Gbu* | *Ndu* |
|---|---|---|---|---|
| *Sample 1* | 17.4 | 17.4 | 39.1 | 26.1 |
| *Sample 2* | 36.4 | 22.7 | 27.3 | 13.6 |
| *Average* | 26.9 | 20.1 | 33.2 | 19.9 |

2. Express each species percentage in terms of its deviation from the average for the samples:

| | *PaL* | *PaR* | *Gbu* | *Ndu* |
|---|---|---|---|---|
| *Sample 1* | -9.5 | -2.7 | +5.9 | +6.2 |
| *Sample 2* | +9.5 | +2.7 | -5.9 | -6.2 |

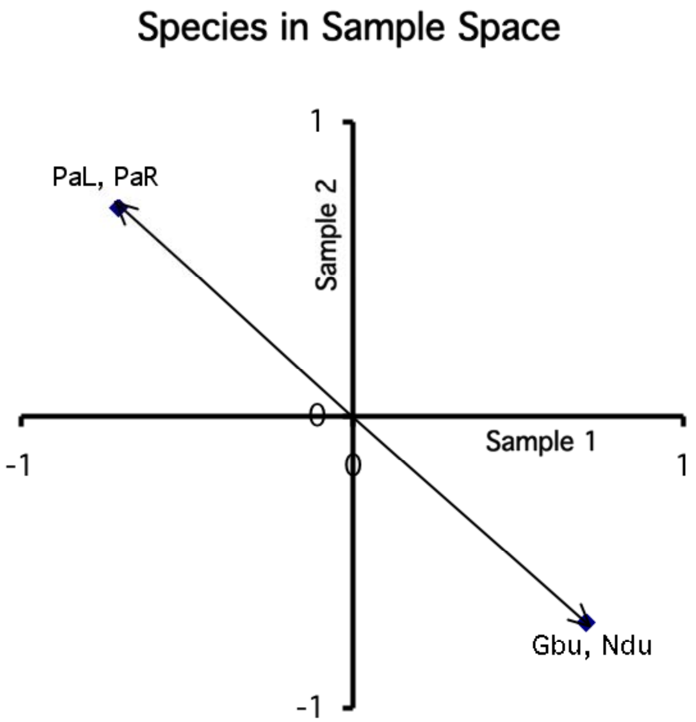
3. Transpose:

| | *Sample 1* | *Sample 2* | *length* |
|---|---|---|---|
| *PaL* | -9.5 | +9.5 | 13.42 |
| *PaR* | -2.7 | +2.7 | 3.77 |
| *Gbu* | +5.9 | -5.9 | 8.38 |
| *Ndu* | +6.2 | -6.2 | 8.80 |

4. Row-normalize (to unit vector length):

| | *Sample 1* | *Sample 2* |
|---|---|---|
| *PaL* | -0.71 | +0.71 |
| *PaR* | -0.71 | +0.71 |
| *Gbu* | +0.71 | -0.71 |
| *Ndu* | +0.71 | -0.71 |

5. Plotting the species as vectors in sample space:



Note that the PaL and PaR vectors coincide, as do the Gbu and Ndu vectors.

6. In sample space, species that are highly correlated fall in the same vector region, i.e., there is a very small angle between highly correlated species.

a. The correlation coefficient is the cosine of the angle between the two PROPERTY vectors in sample space: Let a and b any two PROPERTY vectors (with a common origin). Then

$$\rho = \cos\theta = \frac{\mathbf{a}^T\mathbf{b}}{|\mathbf{a}|\ |\mathbf{b}|}$$

So in this case, the correlation coefficient table is:

| | *PaL* | *PaR* | *Gbu* | *Ndu* |
|---|---|---|---|---|
| *PaL* | +1 | +1 | -1 | -1 |
| *PaR* | +1 | +1 | -1 | -1 |
| *Gbu* | -1 | -1 | +1 | +1 |
| *Ndu* | -1 | -1 | +1 | +1 |

So (PaL and PaR) and (Gbu and Ndu) are positively correlated whereas (PaL and GBu), (PaL and Ndu), (PaR and Gbu) and (PaR and Ndu) are negatively correlated.

This method of calculating the correlation coefficient is mathematically equivalent to the "normal" definition of the correlation coefficient given at the beginning; i.e., this sample space diagram is the basis for the definition of the correlation coefficient!

When there are only two samples as in this example, this definition seems as there are only three possibilities: r = -1, 0, +1. This is because any two data pairs define a straight line (unless they coincide, in which case no line is defined) and this line has either a positive slope or a negative slope. But when there are more samples (i.e. multidimensional sample space), the data pairs no longer have to define a line, the vectors can occur in any region of sample space, and r can have any value $-1 \leq r \leq 1$.

If A is the matrix with the row-normalized (i.e., unit length so that |a| = |b| =1) vectors of species data (relative to the mean for each species) in sample space, then the correlation coefficient table is the inner product of A:

$$\Sigma = A\,A^T$$

7. Likewise, one can define a similarity coefficient as the cosine of the angle between samples in PROPERTY space; it is a measure of how similar two samples are to one another in their PROPERTY composition. We will call the similarity coefficient matrix H.

8. What we are going to do next is reduce the number of "species" by combining them into "factors". In other words, we are going to reduce our system from having (say) 31 species counts for each sample to having (say) 6 "factor loadings" for each sample.

Continuing on from equation II.C.4., a math theorem says that a symmetric matrix like the correlation matrix Σ can be expressed as:

$$\Sigma = \mathbf{A}\;\Phi\;\mathbf{A}^{\mathbf{T}} + \Psi$$

where "Σ is the **p** by **p** population covariance matrix of the observed variables" (the symmetrical correlation coefficient matrix), "**A** is the **p** by **k** matrix of factor loadings, Φ is the **k** by **k** covariance matrix for the factors (if the factors are in standardized form, this is a correlation matrix with ones in the diagonal), and Ψ is the **p** by **p** residual covariance matrix."
Source: Joreskog et al.

9. If we force the solution to be orthogonal (factors are uncorrelated), then $\Phi = \mathbf{I}$ and hence:

$$\Sigma = \mathbf{A}\;\mathbf{A}^{\mathbf{T}} + \Psi$$

$\Psi$ is now a diagonal matrix (i.e. there is covariance between the factors)

10. A math theorem tells us how to calculate the **Singular Value Decomposition** (SVD) of a matrix:

Let us assume that we have an $N \times p$ ($N>p$) data matrix **X**. The <u>product moments</u> are defined as:

major product moment (MPM) $\mathbf{V} = \mathbf{X}\,\mathbf{X}^T$
minor product moment (mPM) $\mathbf{U} = \mathbf{X}^T\,\mathbf{X}$

Then we can decompose the data matrix as:

$\mathbf{X} = \mathbf{V}\,\Gamma\,\mathbf{U}$

where

**V** is an $N$ by $r$ matrix with orthonormal columns
**U** is a $p$ by $r$ matrix with orthonormal columns, and
$\Gamma$ is a diagonal matrix of order $r$ by $r$ with positive diagonal elements $\gamma_1$, $\gamma_2$, ...$\gamma_r$ called <u>singular values</u> of **X.**

The major product moment $\mathbf{XX}^T$, which is square, symmetric, and of order $N$ by $N$ has $r$ positive eigenvalues and ($N$-$r$) zero eigenvalues. The positive eigenvalues are $\gamma_1^2$, $\gamma_2^2$, ..., $\gamma_r^2$ and the corresponding eigenvectors are $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_r$.

The minor product moment $\mathbf{X}^T\mathbf{X}$, which is square, symmetric, and of order $p$ by $p$ has $r$ positive eigenvalues and ($p$-$r$) zero eigenvalues. The positive eigenvalues are $\gamma_1^2$, $\gamma_2^2$, ..., $\gamma_r^2$ and the corresponding eigenvectors are $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_r$.

The positive eigenvalues of $\mathbf{X}\,\mathbf{X}^T$ and $\mathbf{X}^T\,\mathbf{X}$ are the same, namely $\gamma_1^2$, $\gamma_2^2$, ..., $\gamma_r^2$ . Furthermore, if $\mathbf{v}_m$ is an eigenvector of $\mathbf{X}\,\mathbf{X}^T$ and $\mathbf{u}_m$ and eigenvector of $\mathbf{X}^T\,\mathbf{X}$ corresponding to one of these eigenvalues $\gamma_m^2$, then the following relationships hold between $\mathbf{u}_m$ and $\mathbf{v}_m$ :

$\mathbf{v}_m = (1/\gamma_m)\,\mathbf{X}\,\mathbf{u}_m$ **and** $\mathbf{u}_m = (1/\gamma_m)\,\mathbf{X}^T\,\mathbf{v}_m$

These relationships make it possible to compute $\mathbf{v}_m$ from $\mathbf{u}_m$ and vice-versa, i.e.

$\mathbf{V} = \mathbf{X}\,\mathbf{U}\,\Gamma^{-1}$ and $\mathbf{U} = \mathbf{X}^T\,\mathbf{V}\,\Gamma^{-1.}$

The analysis of the minor product moment $\mathbf{X}^T\mathbf{X}$ is referred to as R-mode analysis and that of the major product moment $\mathbf{XX}^T$ as Q-mode analysis.

11. Major problem with this analysis: there are an infinite number of solutions! If we find a solution to the problem (a set of orthogonal vectors that describe the data), then any <u>rotation</u> of that set of vectors is also a solution! How do we choose any one of these?

a. Rotation of a set of row vectors in the matrix X can simply be done through the operation:

$$\mathbf{Y} = \mathbf{X}\,\mathbf{R}$$

where R is a rotation (transformation) matrix conforming to the requirement that $\mathbf{R}\,\mathbf{R}^T = \mathbf{I}$ for a rigid rotation.

b. In the end, the choice is arbitrary (this is one of the problems with factor analysis). However, certain more-or-less reasonable choices can be made:

i. As much as possible, try to make the factors simple, i.e., have a few high loadings and many zero or near-zero loadings.

ii. Rotate the axes so as to put as much variance as possible into the factors (VARIMAX criterion). i.e., we try to "explain" as much of the variance as possible with the fewest possible factors. This solution is often favored because it is an objective solution (i.e., it arises untouched by human hands). That doesn't necessarily make it a better solution, however.

12. Brief summary of factor analysis steps:

a. Start with a data table with rows of samples and columns of species
b. Row-normalize the data table so that each row has unit length. (it is now referred to as **W**).
c. Calculate the minor product moment (mpm) = $\mathbf{W^T W}$
d. Calculate the eigenvalues ($\Lambda$) and eigenvectors (**U**) of the mpM.
e. We can then express each sample in terms of the eigenvectors by multiplying: **W U** (the result of this calculation is a list of the eigenvector composition of each sample; e.g., each sample can be expressed as a linear combination of the eigenvectors
f. We now choose to simplify the properties of our data set by throwing out all but the first few (dominant) eigenvectors, and then describe our data set in terms of these. This is now the factor matrix **V**, which describes the loadings of each species to each factor.
g. Because we have thrown out some information, the factor loading matrix **A = W V** does not fit the original data exactly.
h. In order to improve the fit somewhat, we will now rotate the factor matrix V so that it accounts for as much of the variance as possible. This is done through an iterative process: we rotate around axis one to find the maximum, then rotate around axis two to find the maximum, etc. for all of the axes (= # of factors). This process is repeated until the result converges (no further improvement in fit is found).
i. Our final rotated factor matrix (the Varimax solution) is **F**, a list of the composition of each species described as a linear combination of the individual factors. The samples are now described as

13. One problem with this type of analysis is that the number of factors is arbitrary! The geologist's approach to choosing number of factors: make maps. Keep adding factors until additional factors become "unmappable".

14. Negative factor problem: the mathematics don't specify that you only can add things; the math is just as happy subtracting a factor. In chemical and physical reality, a sample can only be the sum of positive components (philosophically, it may be possible to subtract components, e.g., incongruently dissolve something away to create a "negative" component of a different composition than the pre-existing components). Various approaches to this problem: (a) ignore the problem (most common), (b) rotate solution so that factors are (mainly) positive (but which rotation should you choose?)

15. Another problem: summation to 100% and the "...and everything else factor"

Supplementary Reading:

Joreskog, Klovan, and Reyment, Geological Factor Analysis, Elsevier, Amsterdam, 1976, 178 p.