12.740 Paleoceanography Spring 2002

Problem 2

Factor analysis primer. The things you are expected to include in your answer are indicated numbered and in **boldface,** e.g. **III.(2)**.

I. *Sample space primer*. Given the species composition data:

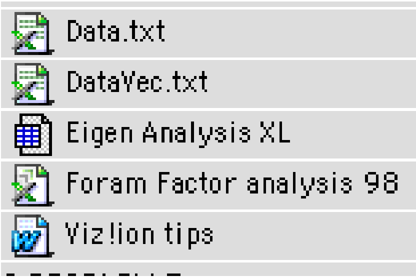
	G.ruber	G.sacculifer	G.truncatulinoides	N.pachyderma(L)
Sample 1	50	30	20	0
Sample 2	30	30	30	10
Sample 3	0	0	30	70
Sample 4	10	20	50	20

I.(1) Calculate the species vectors in sample space, and I.(2) calculate the "correlation coefficient" (cosine of the angle between the species vectors in sample space) of the species relative to each other. Also I.(3) calculate the "similarity coefficient" (cosine of the angle between the sample vectors in species space).

II. *Eigenvalue/eigenvector analysis primer*. You are given a Macintosh disk with 100 x,y,z data points and certain files with which to examine them. The data will be explored using **Viz!on** within **Microsoft Excel 98**. It is assumed that you are minimally proficient at basic Macintosh operations (pointing, clicking, double-clicking, dragging, and quitting programs) or can find someone to tutor you in these. You can do the exercise on any computer that has these applications; however, Viz!on will not work on PCs. The Mac 7500 E34-211 can be used for this exercise.

Also, it should also be possible to do the Excel only portions (sans Viz!on) on **PC Windows** versions of Excel; if you want to try, ask me for the file on a 3.5" PC disk.

The files in your disk are:



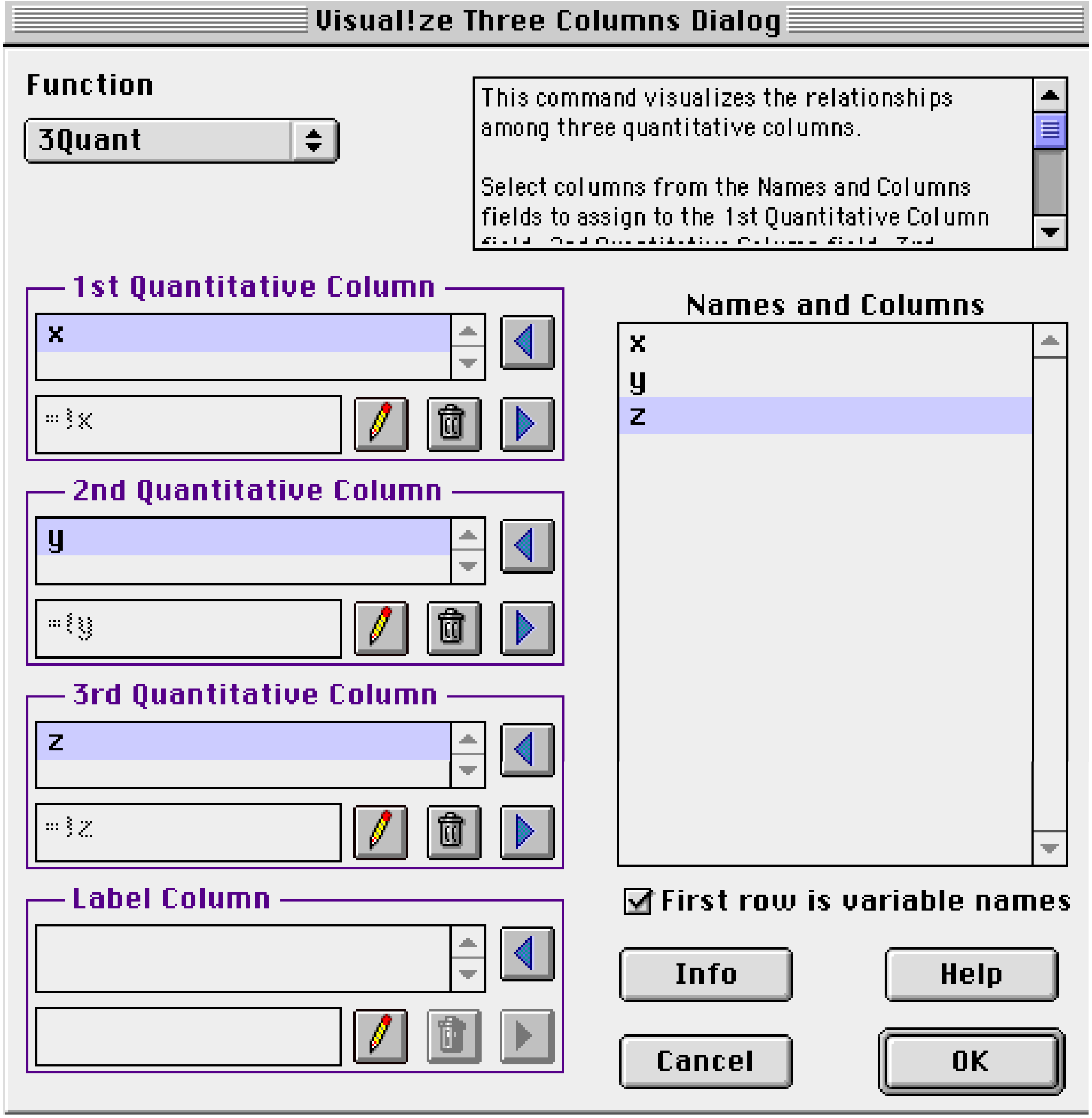
Data.txt
DataVec.txt
Eigen Analysis XL
Foram Factor analysis 98
Viz!on tips

Double-clicking on any of these files will open them up.

1. On the desktop of the Macintosh 7500 in E34-211, double click on the Problem Set 2 icon on the desktop. Double click on the “Launch Viz!on/Excel” icon. After Excel opens, close the blank spreadsheet.

2. Open the file **Data.txt.** Highlight the three columns (ABC). Click on Viz!on in the menu bar and select Vizualize 3 Columns.

3. Click on “Click Me” and select 3Quant. Select the x,y, and z axes by dragging them in turn from their listing on the right to their appropriate box on the left:



Visual!ze Three Columns Dialog

Function: 3Quant

This command visualizes the relationships among three quantitative columns.

Select columns from the Names and Columns fields to assign to the 1st Quantitative Column

1st Quantitative Column: x

2nd Quantitative Column: y

3rd Quantitative Column: z

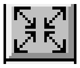
Label Column:

Names and Columns: x, y, z

☑ First row is variable names

(5) Click on "View the data in a rotating plot”. Click on “Add Color by: x” (this makes the plot a little easier to follow when it is spinning)

(6) You can expand the screen either by clicking on the expand bar or using the resizing handle on the lower right. When you expand at first, the plot remains the same size. You can expand or contract the size of the plot by using the tool shown below. This tool works by pointing and clicking near the outer edge if you want to expand, and near the center if you want to contract.

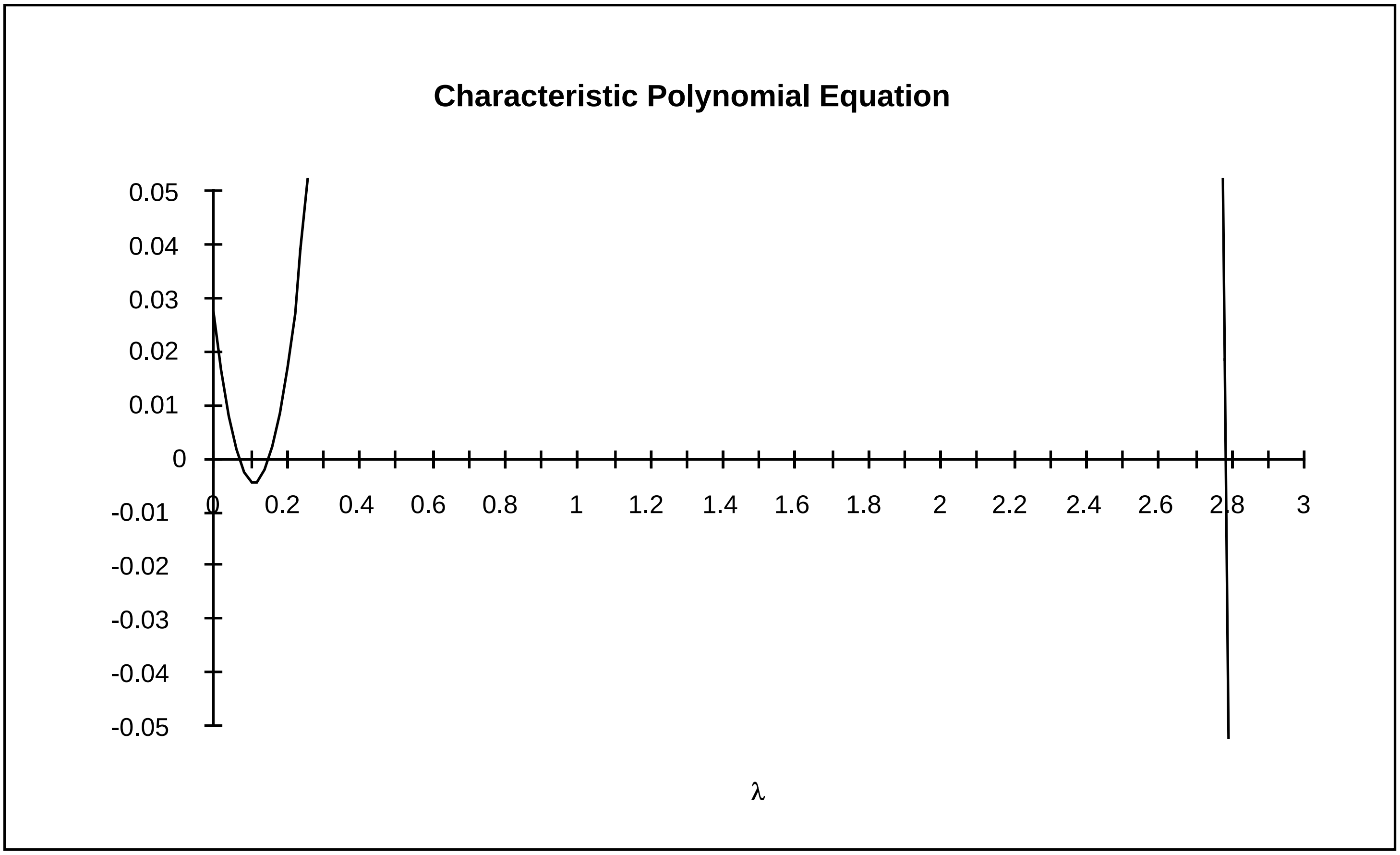


(7) You start the plot spinning by using the (arrow) hand tool to click and drag (letting go) to give the plot a spin. You can also click and drag without letting go to move it as you please. **II.(1)Examine the data set and describe the characteristics you notice**.

2. Open the file **Eigen-analysis (XL)**. Near the top of the screen you will see the correlation matrix:

$$\begin{array}{c|ccc|} & \multicolumn{3}{c}{\mathbf{A}} \\ \hline & x & y & z \\ x & 1 & 0.93 & 0.86 \\ y & 0.93 & 1 & 0.88 \\ z & 0.86 & 0.88 & 1 \end{array}$$

Immediately below that matrix is a section where guesses for the eigenvalues can be entered (**bold** font), where the determinant of the matrix $A-\lambda I$ is calculated (outline font), and where the pre-computed eigenvectors are shown (shadow font). You can change the guesses for the eigenvalues until you arrive at a solution (determinant = 0.00). There are three solutions (roots) to the characteristic cubic polynomial that arises from the $A-\lambda I$ determinant. You can arrive at your guesses to the roots by examining the graph that is located below the eigenvector section:



If you would like to see this equation plotted on a different scale, you may change the scale axes of this graph by double-clicking on an axis and selecting the scale option, and entering the new values you wish for the scale. By interacting with this graph and the calculation section above, **II.(2) find the eigenvalues of the data set (to three decimal places) and describe how you found them**(*). Notice that when you find an eigenvalue, the numbers to the right of the pre-computed eigenvector x (which are the elements of the column vector $(A-\lambda I)x$) approach zero [note: this assumes that you put the smallest eigenvalue in the top group and the largest in the bottom group].. What is the

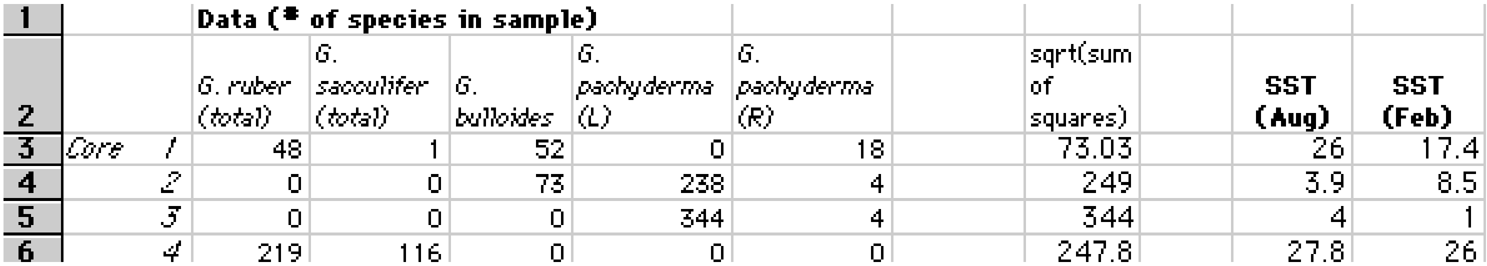
sum of the eigenvalues? This result is not a coincidence - the sum of the eigenvalues corresponds to the rank of the matrix (the minimum number of orthogonal vectors required to span the sample space). **II.(3) Divide the eigenvalues by the sum of the eigenvalues.** These numbers are equal to the fraction of the total variance "accounted" for by the eigenvector associated with that eigenvalue. **II.(4) Look at the eigenvectors and describe the spatial location of each** (i.e., where in the four x-y quadrants and what angle from the x-y plane (1° accuracy); e.g.: "x positive, y negative, 30° above the horizontal").
(*)[Note: if you are familiar with the "Solver" in Microsoft Excel, you can use it to find the roots - but note that this Newton-Rapson solver requires that your initial guess be close to the correct answer).

Close this file.

3. Double click-on **DataVec.txt.** Open up the columns using Viz!on. This file illustrates the raw data set and normalized eigenvectors from step 2. Notice that the eigenvectors are orthogonal and follow the principle axes of elongation of the data set. Which eigenvalues are associated with each of these eigenvectors?

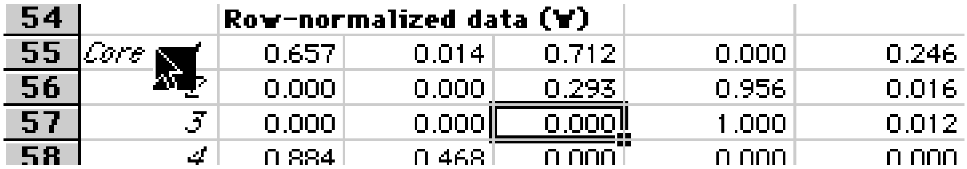
III. *Factor analysis example*. You are given a data set consisting of a subset of the Atlantic core top foram census (50 samples selected at random with data for *G. ruber, G. sacculifer, G. bulloides, N. pachyderma (R), N. pachyderma (L)*. You will follow the Q-mode factor analysis of this data set into three factors, and verify its correctness.

1. The raw data (# of specimens of each species in a sample is given at the top of the spreadsheet. The sum of squares (for unit-length normalization) is calculated. You are also given estimates for sea surface temperature.



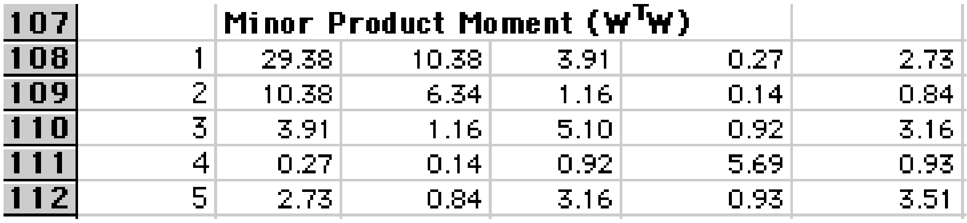
1		Data (# of species in sample)									
2		G. ruber (total)	G. sacculifer (total)	G. bulloides	G. pachyderma (L)	G. pachyderma (R)		sqrt(sum of squares)		SST (Aug)	SST (Feb)
3	Core 1	48	1	52	0	18		73.03		26	17.4
4	2	0	0	73	238	4		249		3.9	8.5
5	3	0	0	0	344	4		344		4	1
6	4	219	116	0	0	0		247.8		27.8	26

Immediately below the raw data data is the row-normalized data, **W**:



54		Row-normalized data (W)				
55	Core 1	0.657	0.014	0.712	0.000	0.246
56	2	0.000	0.000	0.293	0.956	0.016
57	3	0.000	0.000	0.000	1.000	0.012
58	4	0.884	0.468	0.000	0.000	0.000

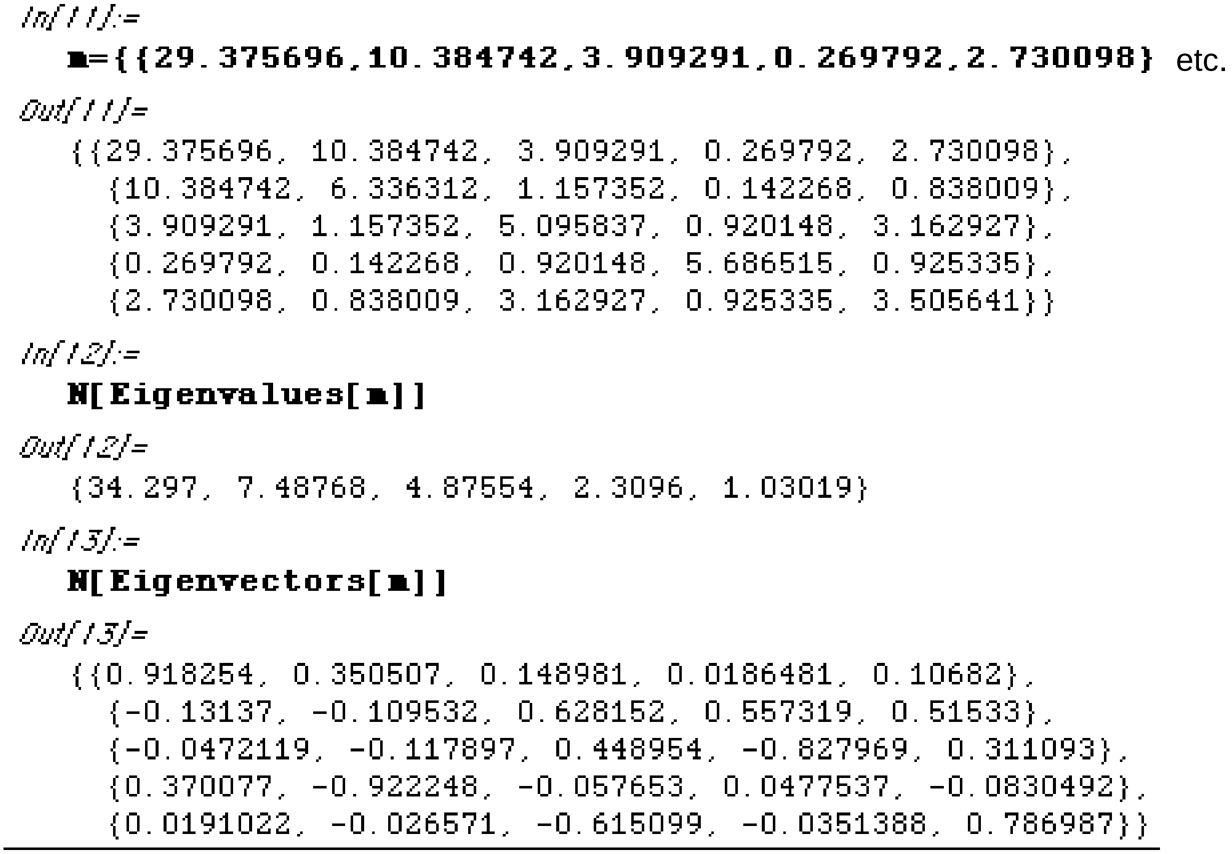
The analysis then proceeds by calculation of the 5 x 5 minor Product Moment (mPM). Actually, you could also use the Major Product Moment (50 x 50), but the calculations are obviously easier for the mPM.



107		Minor Product Moment (W^TW)				
108	1	29.38	10.38	3.91	0.27	2.73
109	2	10.38	6.34	1.16	0.14	0.84
110	3	3.91	1.16	5.10	0.92	3.16
111	4	0.27	0.14	0.92	5.69	0.93
112	5	2.73	0.84	3.16	0.93	3.51

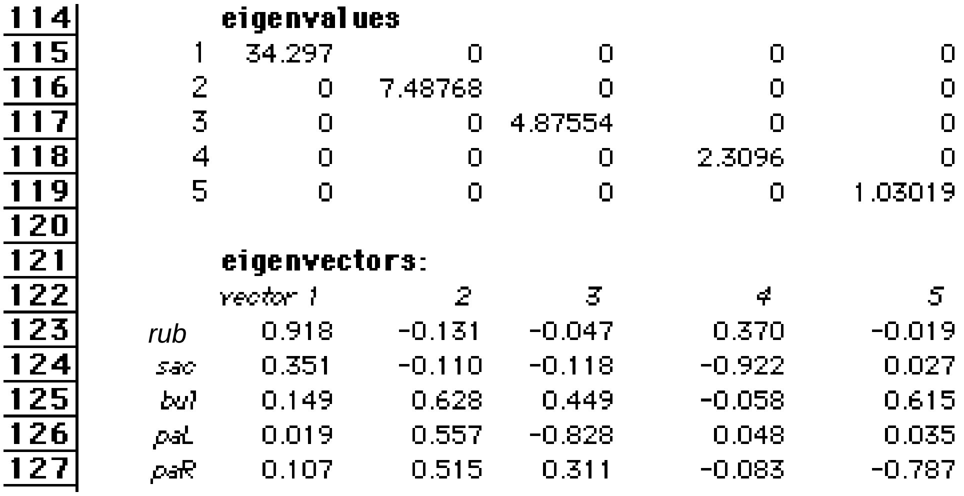
Note: If you move the cursor to any cell of a matrix, it will show the matrix formula used to generate the matrix.

We then calculate the eigenvalues and eigenvectors. Excel 5.0 will not do this calculation, so the mPM was transferred from Excel into Mathematica. Mathematica gives eigenvalues and eigenvectors as a result of the commands:



```
In[11]:=
   m={{29.375696,10.384742,3.909291,0.269792,2.730098} etc.
Out[11]=
   {{29.375696, 10.384742, 3.909291, 0.269792, 2.730098},
     {10.384742, 6.336312, 1.157352, 0.142268, 0.838009},
     {3.909291, 1.157352, 5.095837, 0.920148, 3.162927},
     {0.269792, 0.142268, 0.920148, 5.686515, 0.925335},
     {2.730098, 0.838009, 3.162927, 0.925335, 3.505641}}
In[12]:=
   N[Eigenvalues[m]]
Out[12]=
   {34.297, 7.48768, 4.87554, 2.3096, 1.03019}
In[13]:=
   N[Eigenvectors[m]]
Out[13]=
   {{0.918254, 0.350507, 0.148981, 0.0186481, 0.10682},
     {-0.13137, -0.109532, 0.628152, 0.557319, 0.51533},
     {-0.0472119, -0.117897, 0.448954, -0.827969, 0.311093},
     {0.370077, -0.922248, -0.057653, 0.0477537, -0.0830492},
     {0.0191022, -0.026571, -0.615099, -0.0351388, 0.786987}}
```

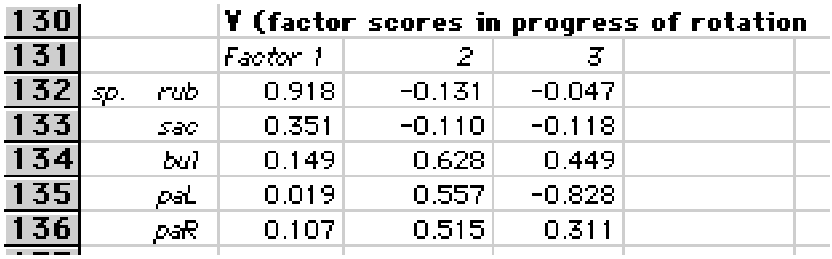
The results were then copied back into Excel:



114		eigenvalues				
115	1	34.297	0	0	0	0
116	2	0	7.48768	0	0	0
117	3	0	0	4.87554	0	0
118	4	0	0	0	2.3096	0
119	5	0	0	0	0	1.03019
120						
121		eigenvectors:				
122		vector 1	2	3	4	5
123	rub	0.918	-0.131	-0.047	0.370	-0.019
124	sac	0.351	-0.110	-0.118	-0.922	0.027
125	bul	0.149	0.628	0.449	-0.058	0.615
126	paL	0.019	0.557	-0.828	0.048	0.035
127	paR	0.107	0.515	0.311	-0.083	-0.787

III.(1) Verify that these are the eigenvalues and eigenvectors of the MPM (i.e., demonstrate orthogonality of (A-λ) and I using the determinant of their product, as in section 2. above).

The next step is to throw away the lesser eigenvectors and retain the more important. This is now the matrix V:



130			Y (factor scores in progress of rotation		
131			Factor 1	2	3
132	sp.	rub	0.918	-0.131	-0.047
133		sac	0.351	-0.110	-0.118
134		bul	0.149	0.628	0.449
135		paL	0.019	0.557	-0.828
136		paR	0.107	0.515	0.311

At this point, we can calculate the factor loadings **A=WF** (remember the Singular Value Decomposition). These are the unrotated solutions, and we wish to simplify the factors by rotating the vectors so has to maximize the variance of the columns of the squares of each element of **A** (this is the Varimax criterion). This is done as a stepwise procedure for each of the three planes defined by the axes (plane 1-2, plane 1-3, and plane 2-3). A rotation of angle ϕ_{jl} in the j-l plane is accomplished by multiplying **A** by **R**, where

$R_{1\text{-}2}$:

$$\begin{matrix} \cos\phi_{12} & -\sin\phi_{12} & 0 \\ \sin\phi_{12} & \cos\phi_{12} & 0 \\ 0 & 0 & 1 \end{matrix}$$

$R_{1\text{-}3}$:

$$\begin{matrix} \cos\phi_{13} & 0 & -\sin\phi_{13} \\ 0 & 1 & 0 \\ \sin\phi_{13} & 0 & \cos\phi_{13} \end{matrix}$$

$R_{2\text{-}3}$:

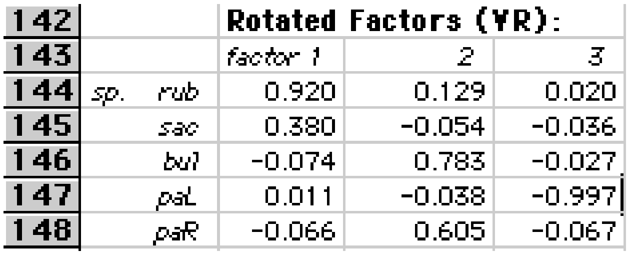
$$\begin{matrix} 1 & 0 & 0 \\ 0 & \cos\phi_{23} & -\sin\phi_{23} \\ 0 & \sin\phi_{23} & \cos\phi_{23} \end{matrix}$$

First we find the angle that maximizes the variance by

maximize by rotating in the 1-2 plane;
maximize by rotating in the 1-3 plane;
maximize by rotating in the 2-3 plane ;

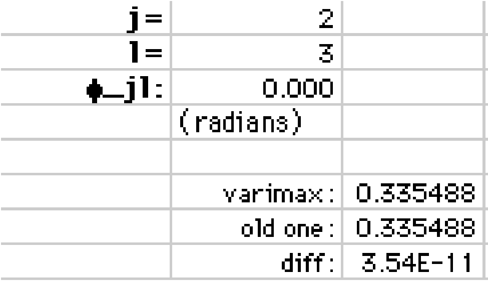
(repeat until solution converges...).

This has already been done in the spreadsheet, so you are looking at the optimally rotated factors:



142			Rotated Factors (YR):		
143			factor 1	2	3
144	sp.	rub	0.920	0.129	0.020
145		sac	0.380	-0.054	-0.036
146		bul	-0.074	0.783	-0.027
147		paL	0.011	-0.038	-0.997
148		paR	-0.066	0.605	-0.067

III.(2) Verify that you can't improve the fit by rotating the factors any more:
You can change j and l and try putting in various angles of rotation. The current variance is indicated in the varimax cell, which you can compare to the one which was previously optimized.



j=	2	
l=	3	
ϕ_jl:	0.000	
	(radians)	
	varimax:	0.335488
	old one:	0.335488
	diff:	3.54E-11

You may examine the relationships between the factors and summer SST by clicking on the tabs:

↓ ↓ ↓

And you can examine the degree to which the three factors "fit" the row-normalized data by clicking on the tabs:

↓ ↓ ↓ ↓ ↓

You may have to use the forward/backward keys:

III.(3) In a few sentences, discuss the outcome of the factor analysis in relationship to fitting the raw data and the relationship with temperature.