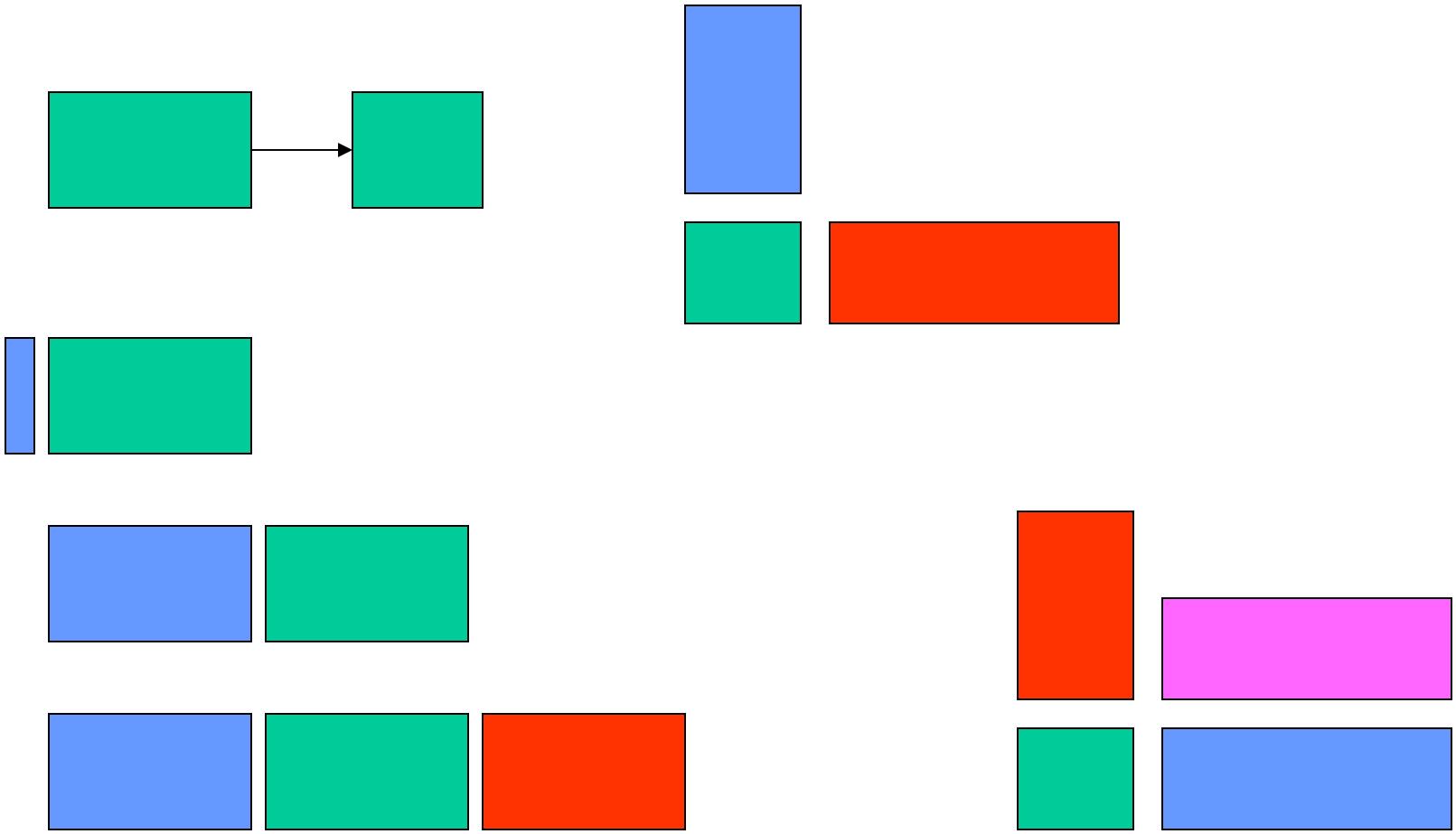
Multiple Blocks of Data

Stan Young(NISS)
Doug Hawkins (U Minn)
Li Liu (Aventis)

Data Mining
Toronto, Canada
Oct 27, 2004

Multiple Blocks



Multiple Blocks

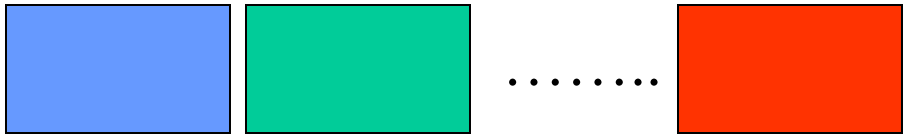
2-way tables of data are ubiquitous.

Two 2-way tables are common.

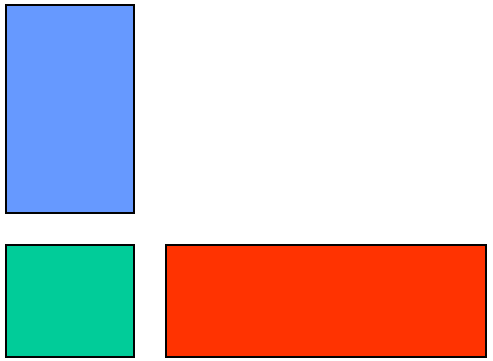
Multiple 2-way tables are becoming important:

ClinChem, metabolism, proteins,gene expression.

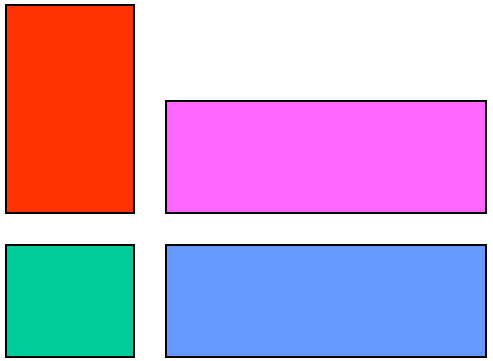
Multiple (Linear) Blocks



Horst. 1965. Factor analysis of data matrices.



Pittman, Sacks, Young. 2001.
3-Way Analysis.



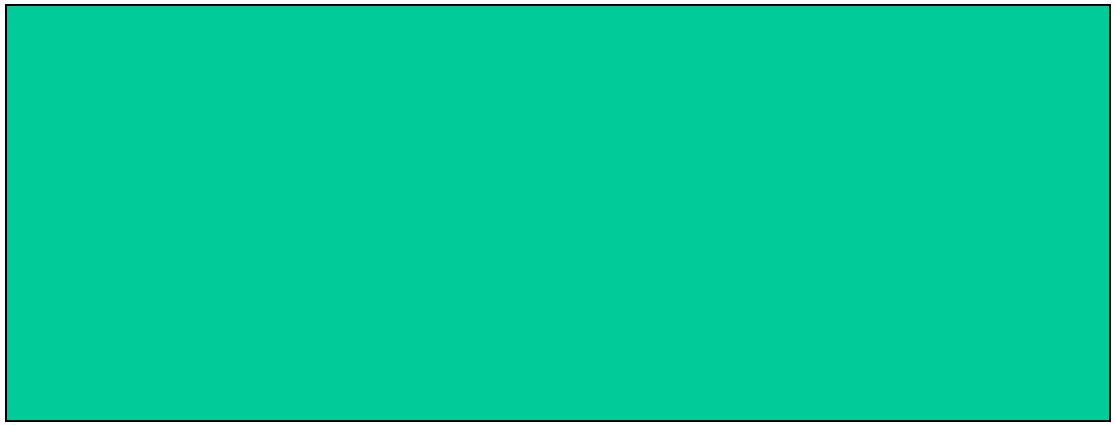
Martens. 2004.
U-Analysis

Outline

- Sketch the rSVD algorithm
- Wine tasting data set
- Forestry species profile.
- Comments

Data Matrix

Goal: permute the rows and columns to find patterns.

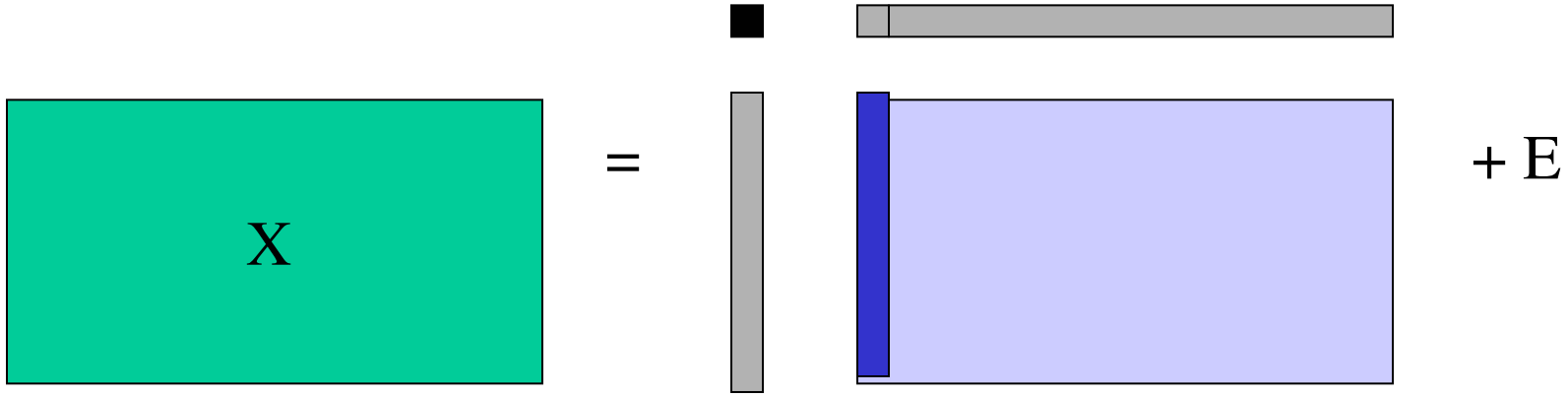


Problems:

1. Large, 10s to 100s of rows and columns.
2. Missing data
3. Non-normal data.
4. Outliers.

Robust SVD

$$X = \lambda * LHE' * RHE + E$$

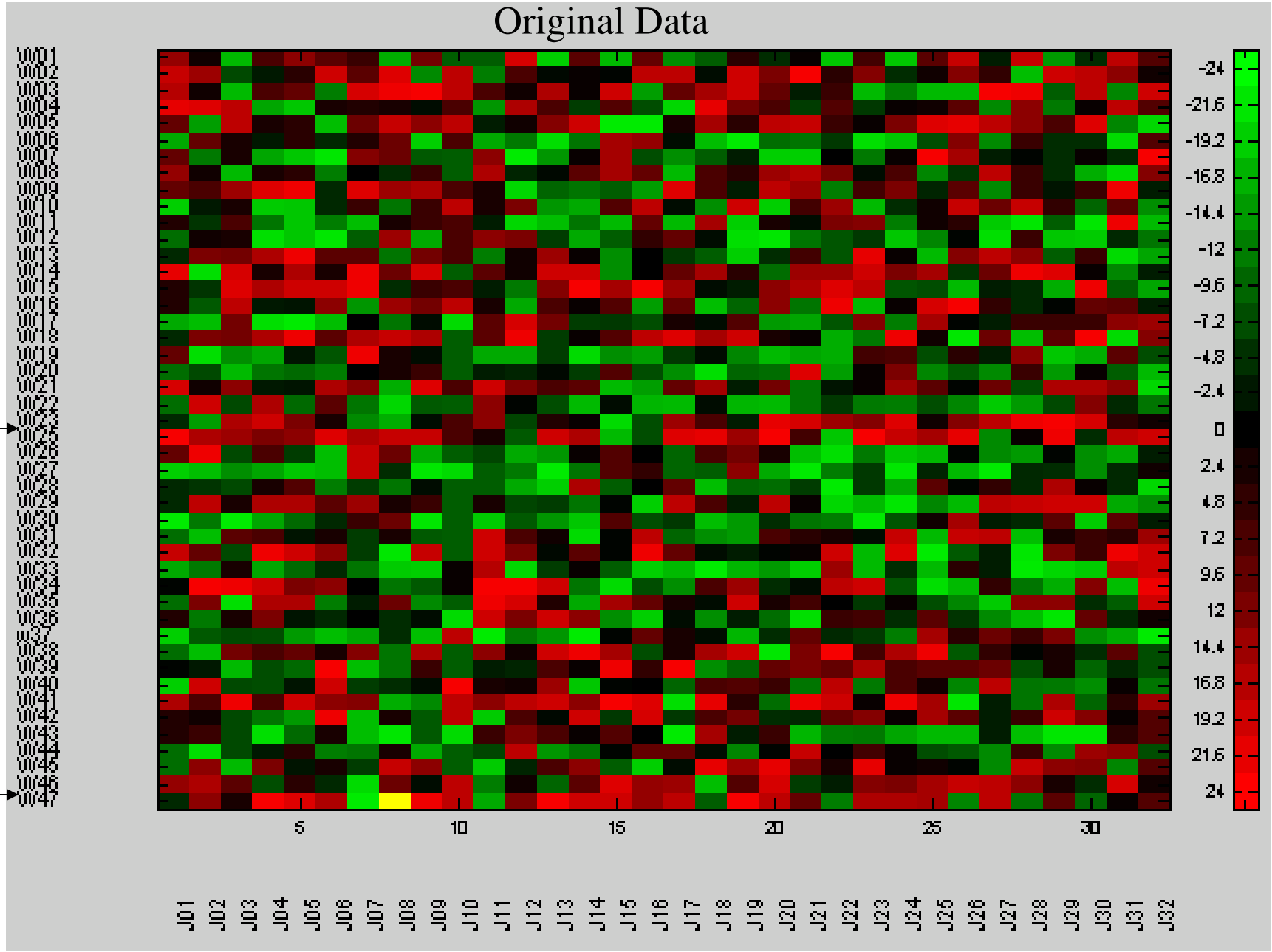


$$y = bx + e$$

1. Guess at LHE.
2. Linear regression of LHE on column of X.
3. Element of RHE is the regression coefficient.
4. Switch LRE and RHE, iterate. Alternating LS regression.
5. Use robust regression method. Least trimmed squares.

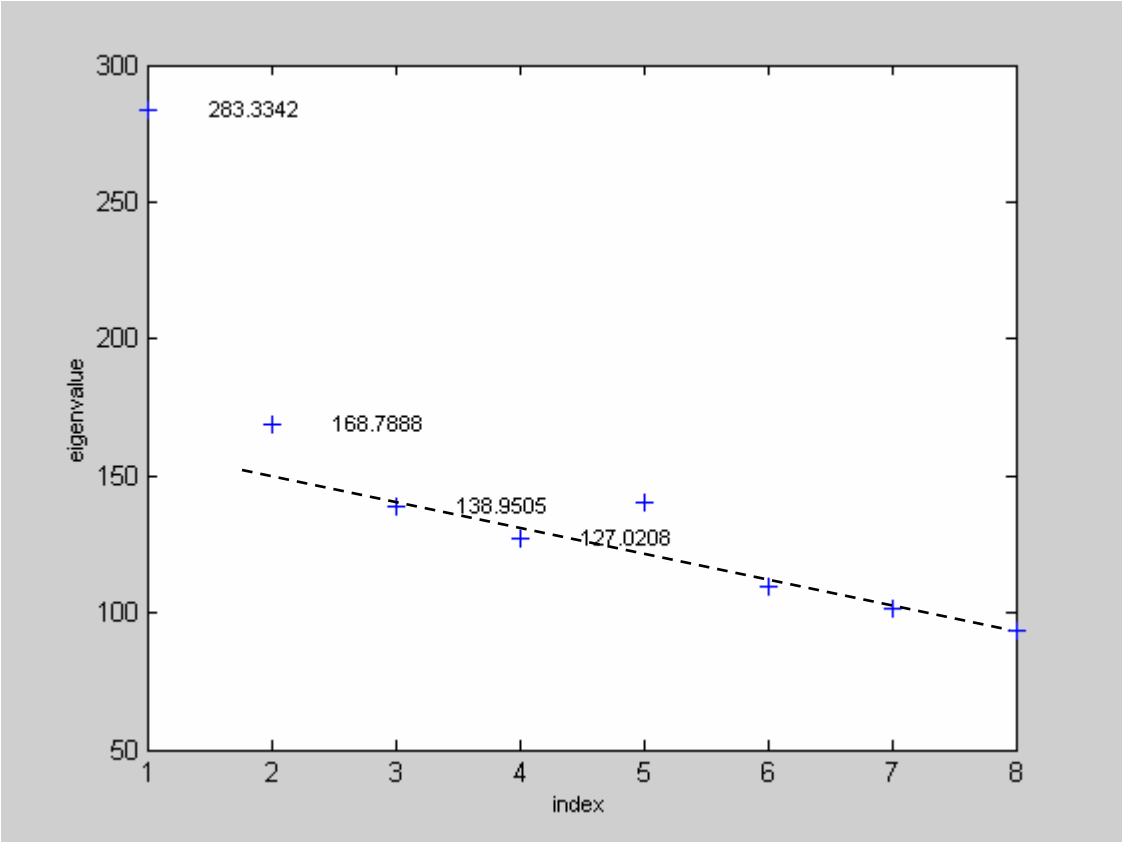
California Versus All Challengers, The 1999 Cabernet Challenge

- 47 wines judged by 32 wine experts
- No data for 1 wine
- One missing data point
- Results are ranks of wine by each judge



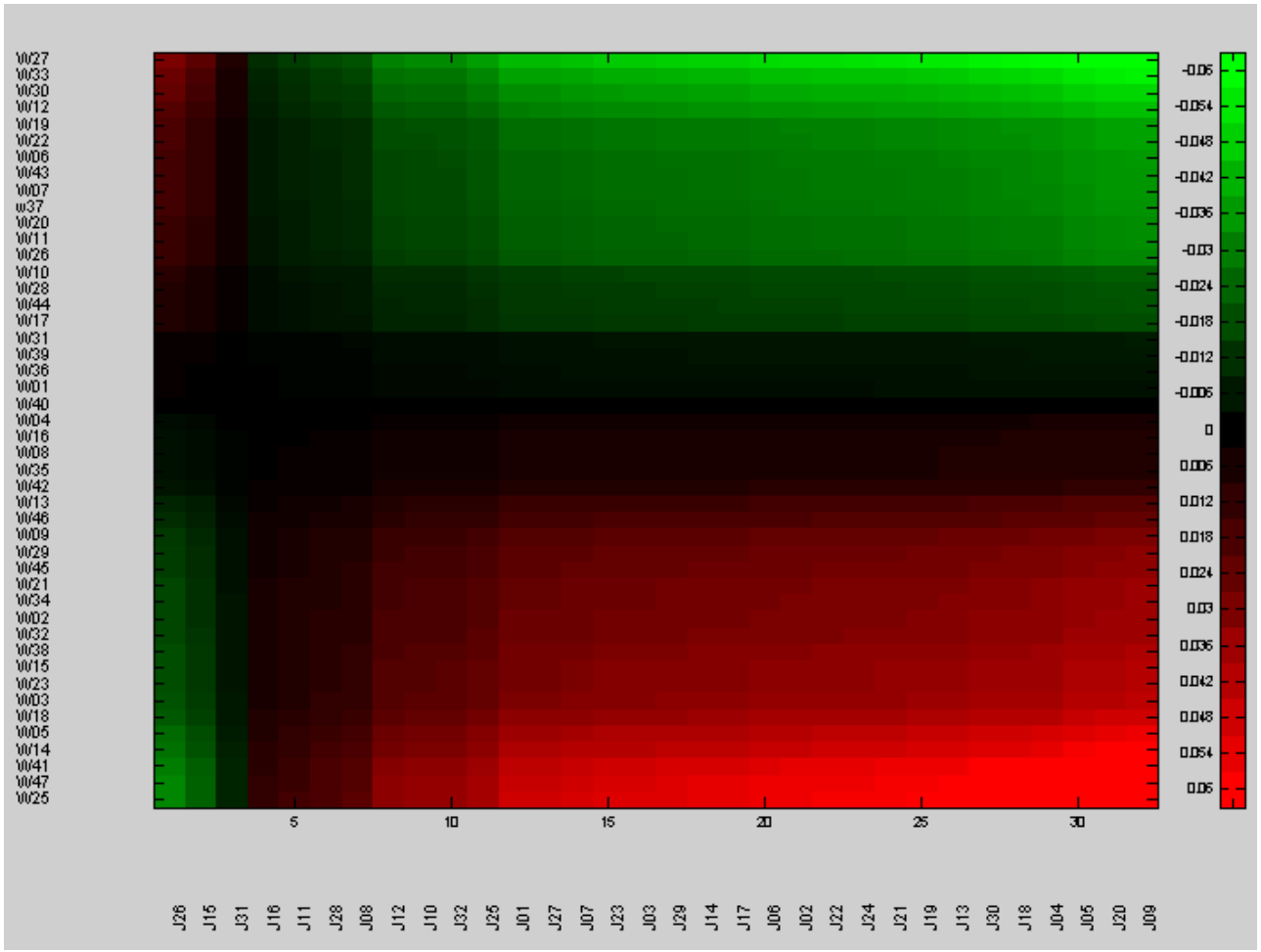
The missing cell is colored yellow.

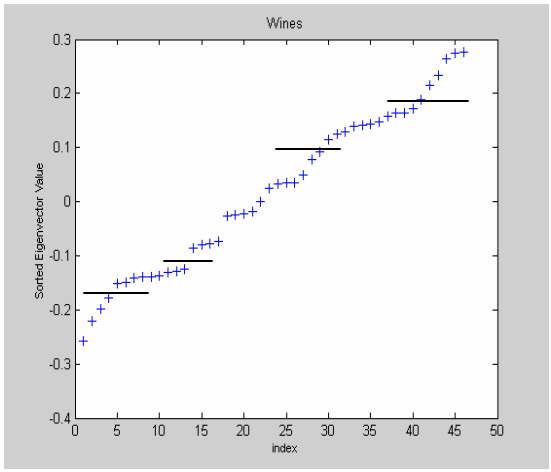
Plot of Eigenvalues

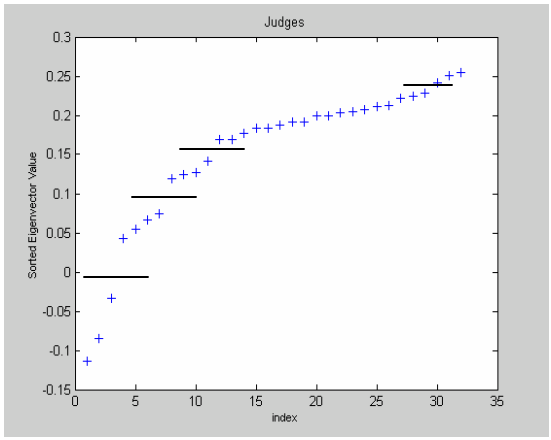


The plot suggests <u>one</u> or two components.

Component 1

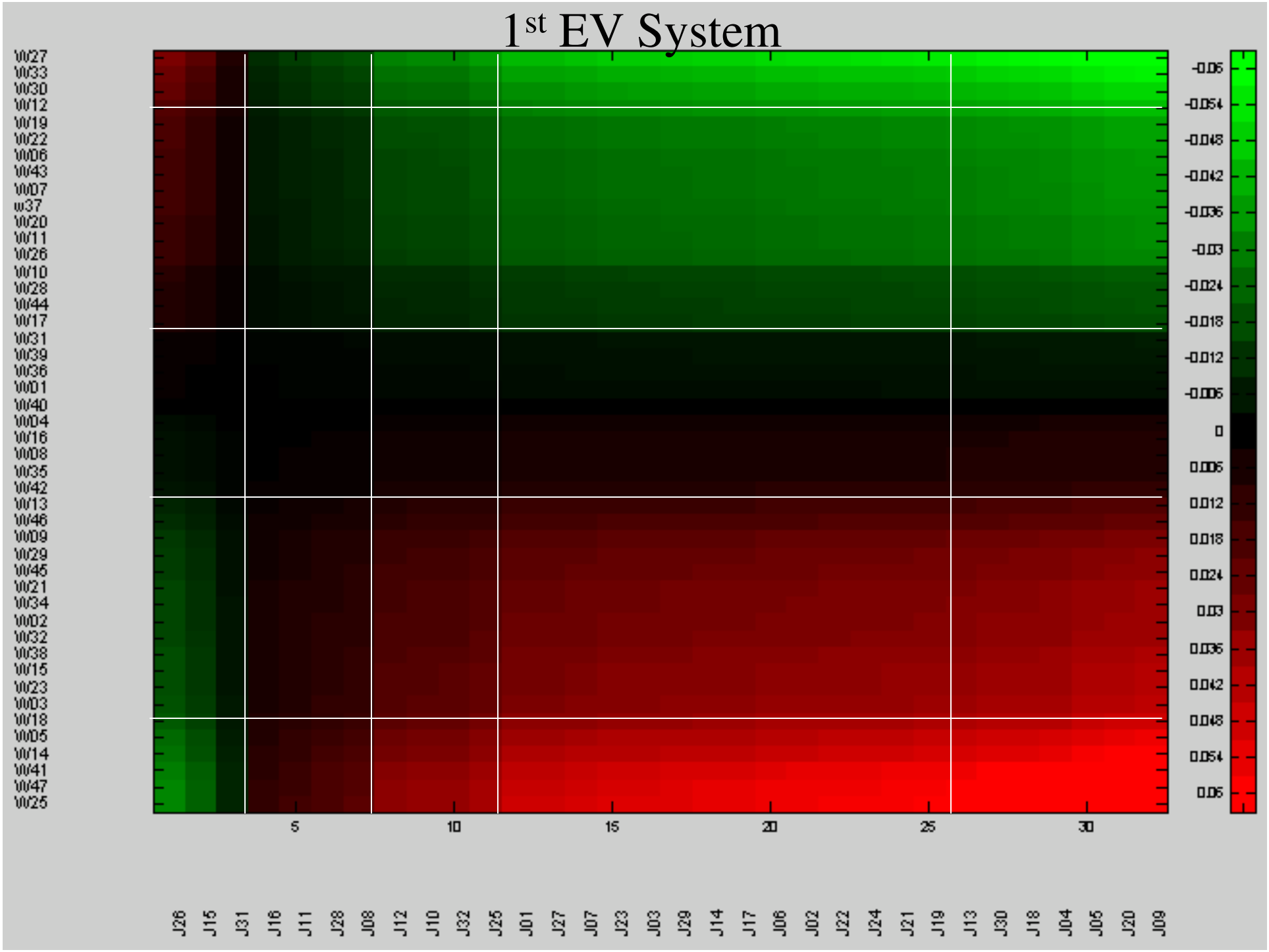






Judges are divided into the following groups: 1-3, 4-7, 8-11, 12-26, 27-32
Wines are divided into the following groups: 1-4, 5-17,18-27,28-41,42-46

1st EV System



Species Profile Data Set

Biometrics 60, 543–549
June 2004

Multivariate Regression Trees for Analysis of Abundance Data

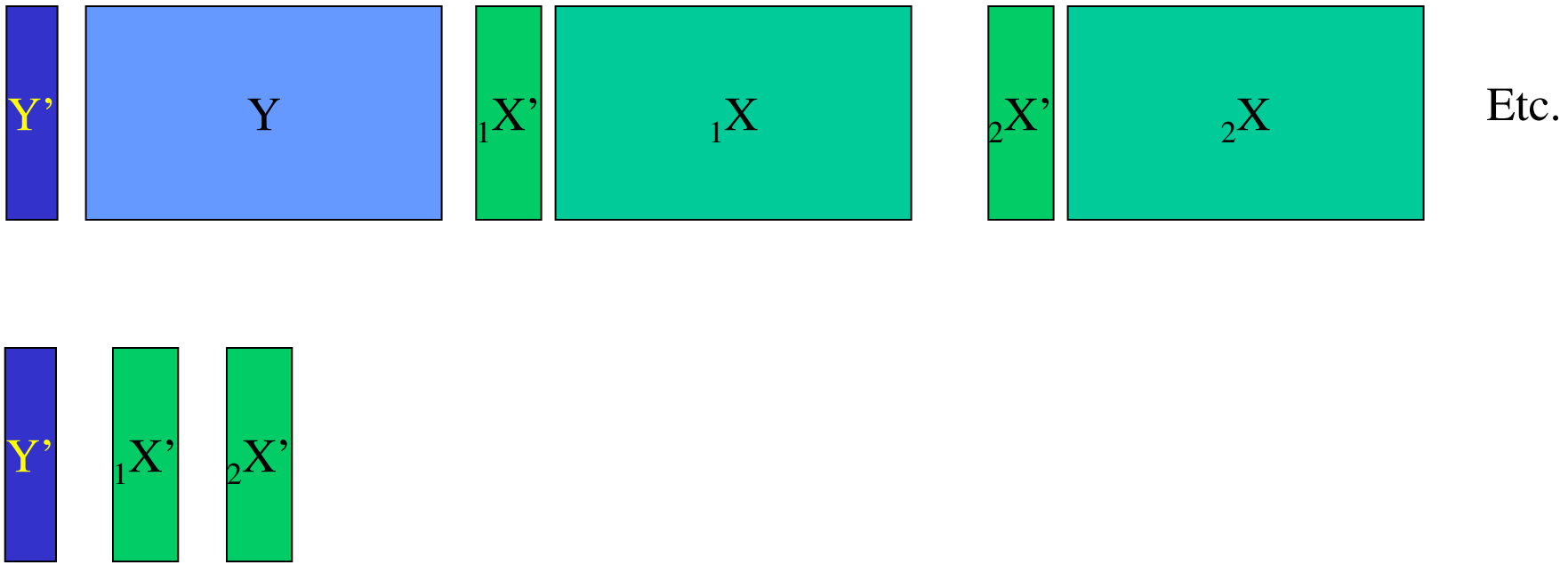
David R. Larsen
Department of Forestry, University of Missouri, Columbia, Missouri 65211, U.S.A.
email: LarsenDR@missouri.edu

and

Paul L. Speckman
Department of Statistics, University of Missouri, Columbia, Missouri 65211, U.S.A.

Y : 12 Species of trees
X : 7 Categorical predictors, 30-7 total categories

PLS Multi-Block Strategy



Data

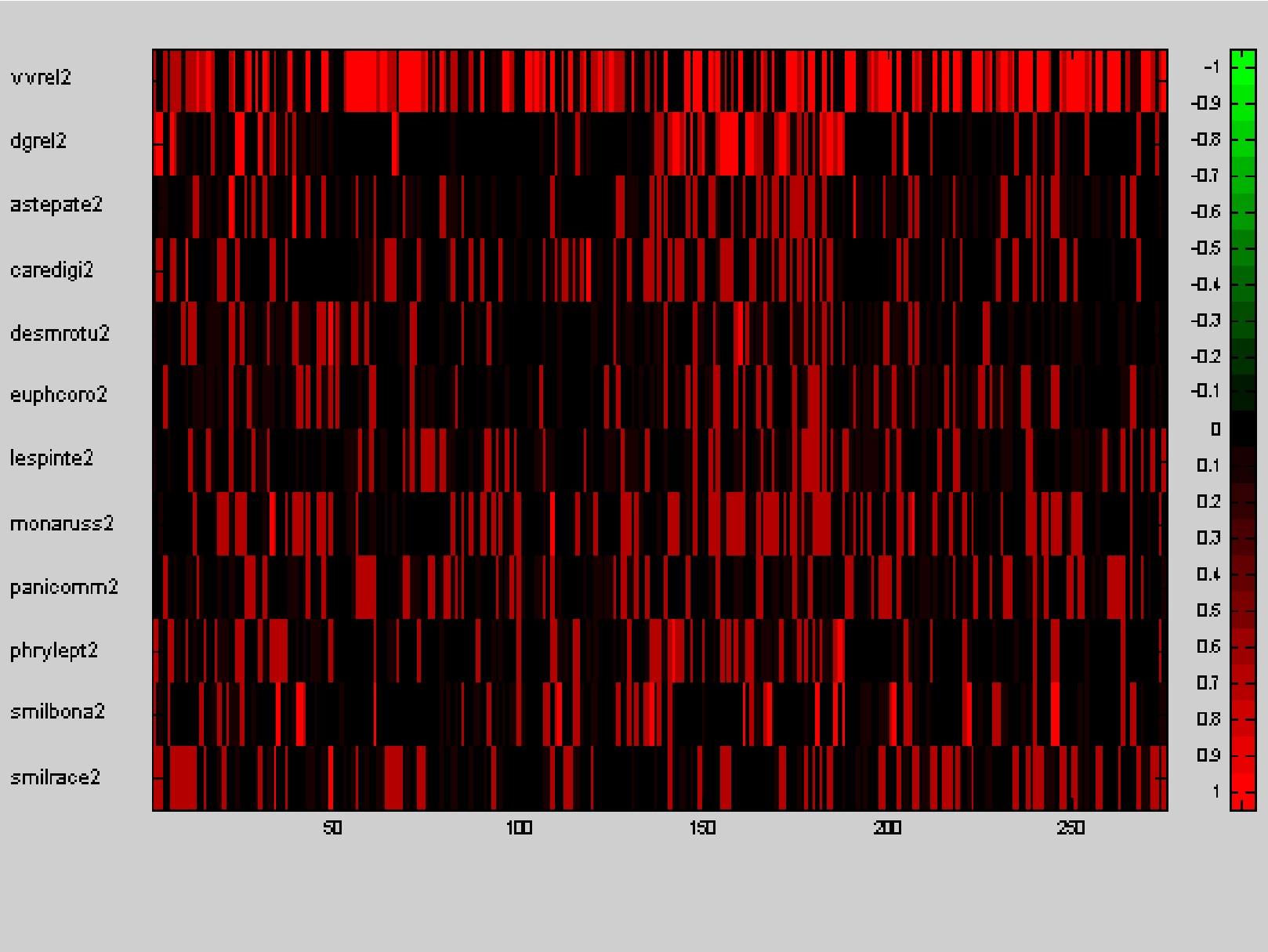
Variable	N^a	$\bar{x}$
Aster patens	118	0.3019
Carex digitalis	100	0.4133
Desmodium glutinosum	101	3.361
Desmodium roundifolium	117	0.2244
Euphorbia corollata	122	0.1948
Lespedeza intermedia	130	0.2098
Monarda russeliana	125	0.4110
Panicum commutatum	133	0.2495
Phryma leptostachya	106	0.3160
Smilax bona-nox	101	1.0650
Smilax racemosa	125	0.3244
Vaccinium vacillans	195	3.3610

Goal: Group species profile based on site characteristics.

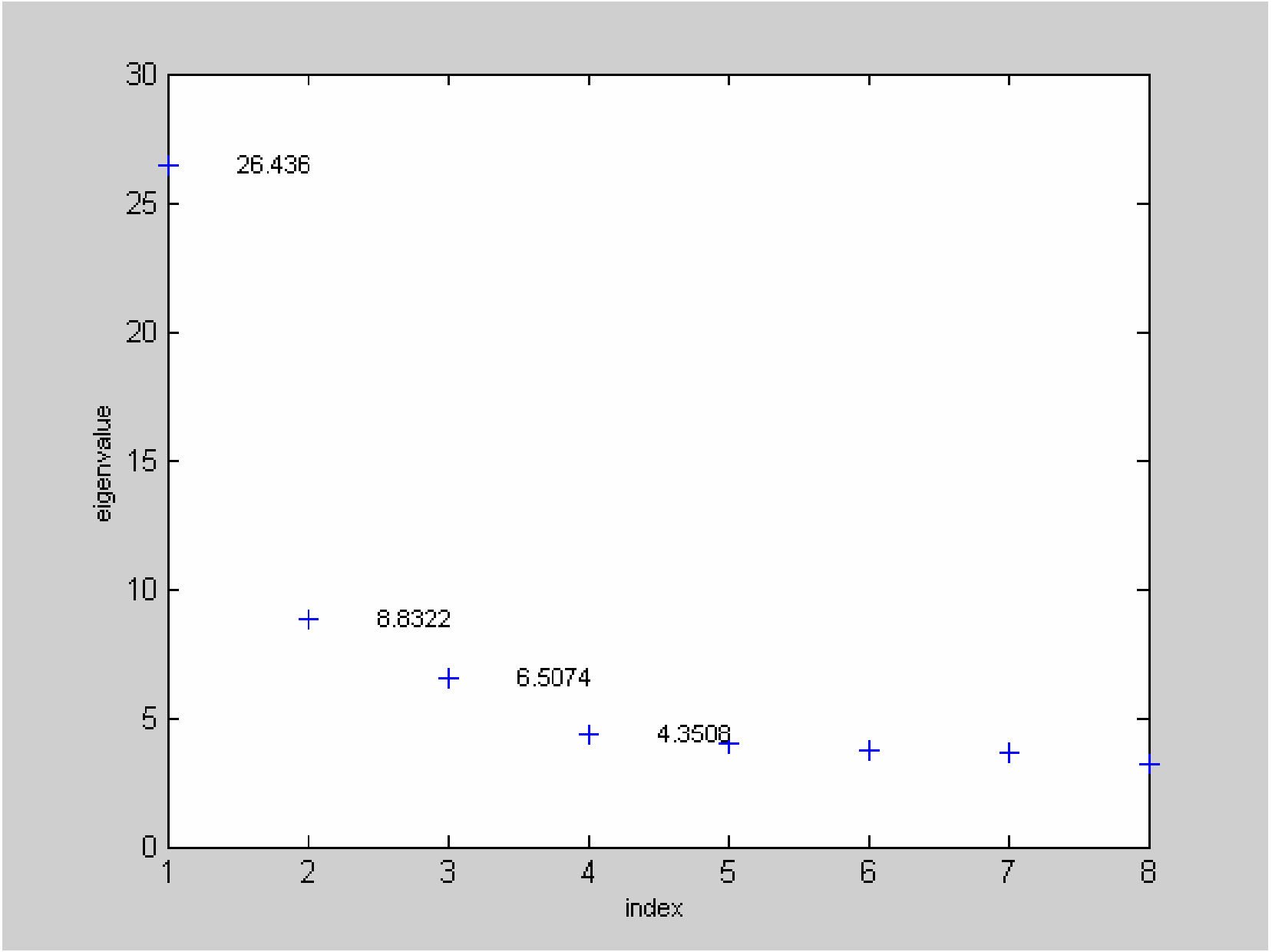
Table 1
Independent categorical variables with the number in each category

Variable	Category	N
Landtype association	Current River Breaks	118
	Current River Hills	76
	Jack Fork, Eminence Breaks	81
Geology	Roubidoux	86
	Upper Gasconade	107
	Lower Gasconade	55
	Gunter	6
	Eminence	15
	Van Buren	6
Landform	Summit	16
	Shoulder ridge	24
	Shoulder	16
	Backslope	195
	Bench	24
Aspect class	Exposed	97
	Neutral east	54
	Neutral west	45
	Protected	79
Phase	Deep	226
	Variable depth	49
Soil order	Alfisol	104
	Mollisol	9
	None	72
	Ultisol	90
Position	Upper	90
	Upper-middle	28
	Middle	51
	Lower-middle	28
	Lower	61
	None	18

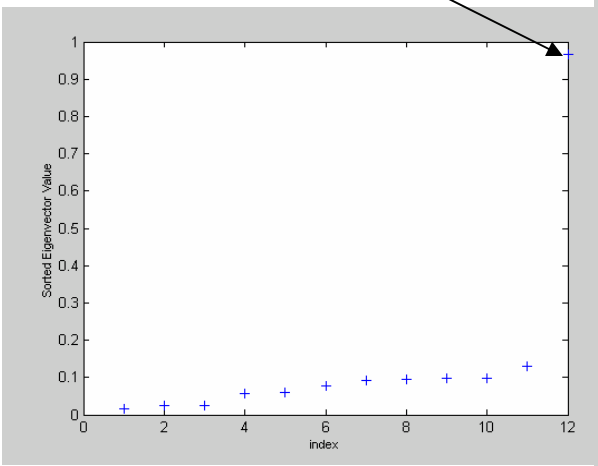
Raw data

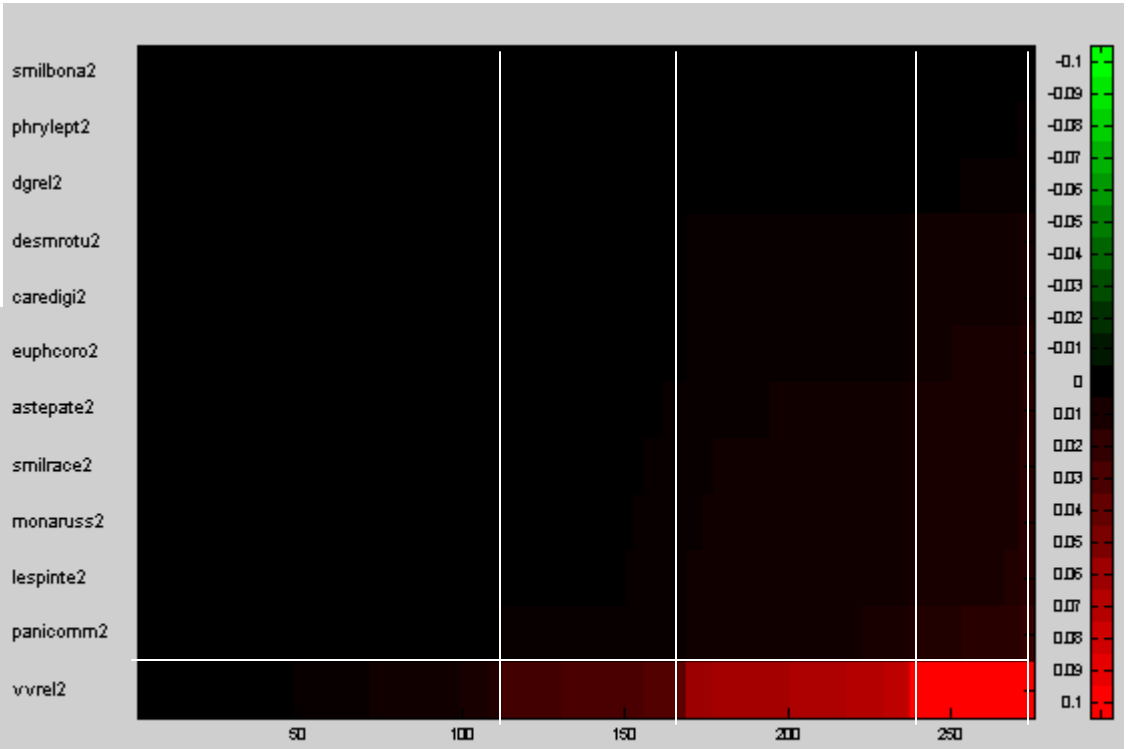

vvrel2
dgrel2
astepate2
caredigi2
desmrotu2
euphcoro2
lespinte2
monaruss2
panicomm2
phrylept2
smilbona2
smilrace2
50
100
150
200
250
-1
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

rSVD Species

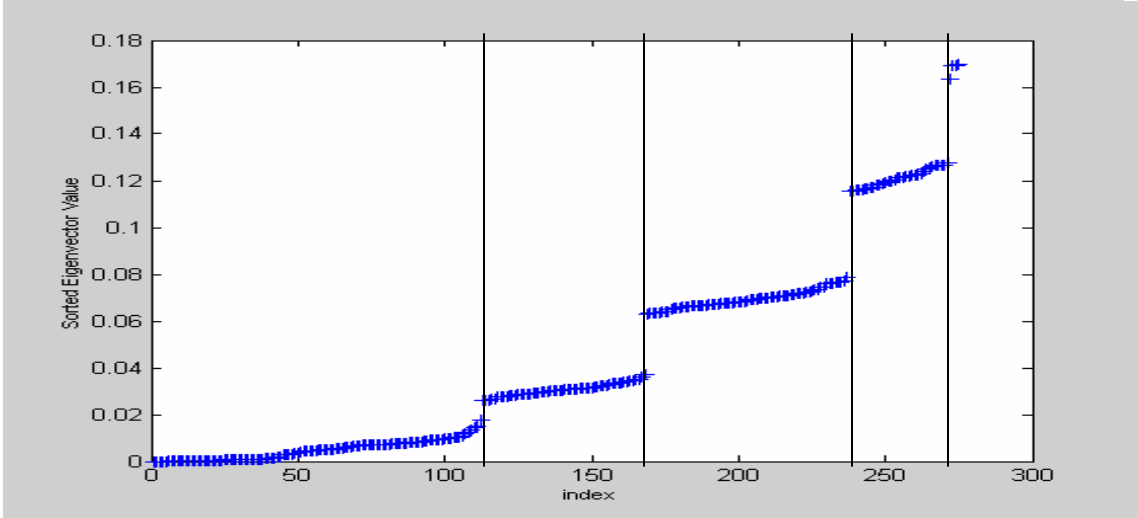


1st EV System

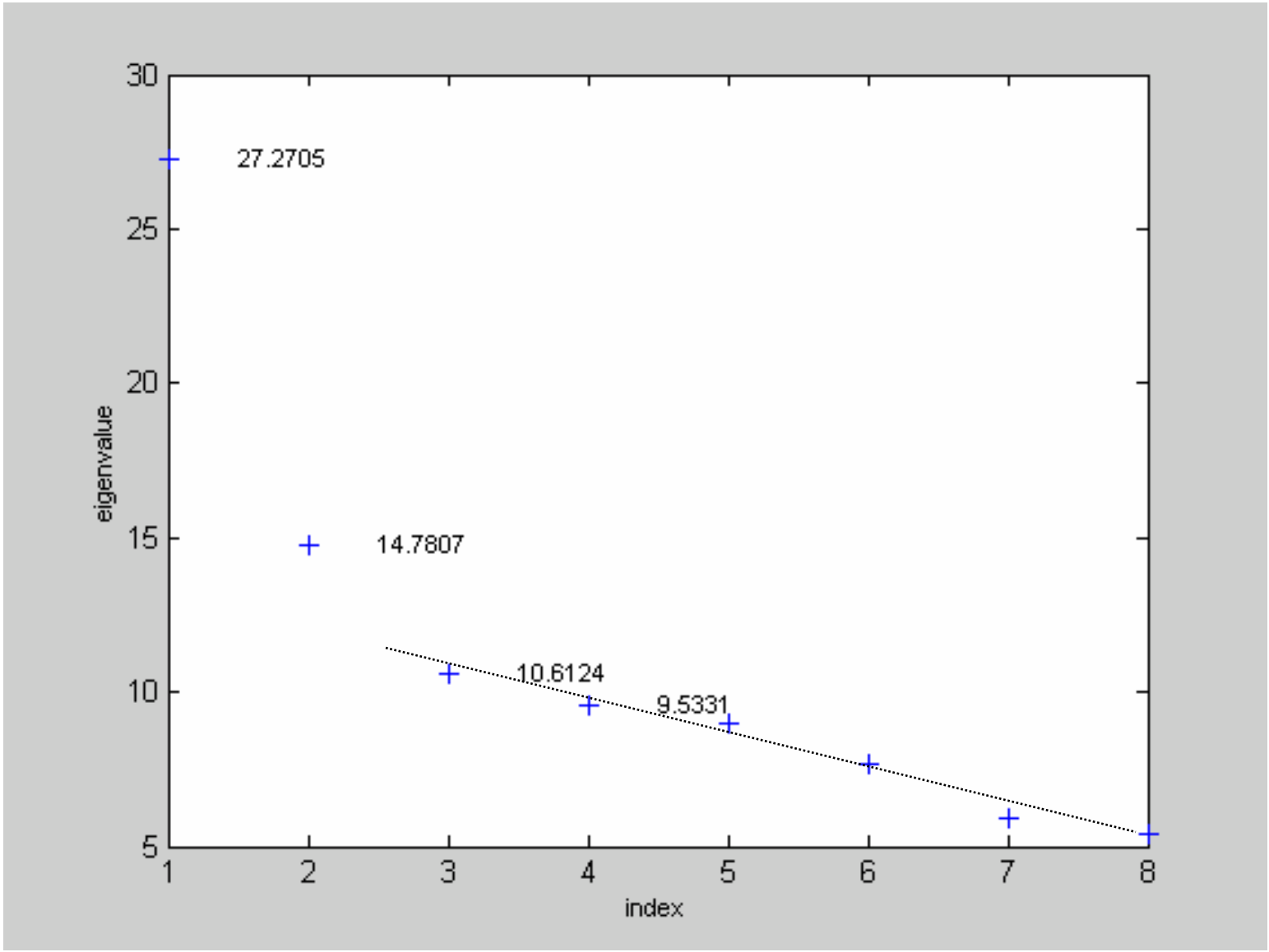




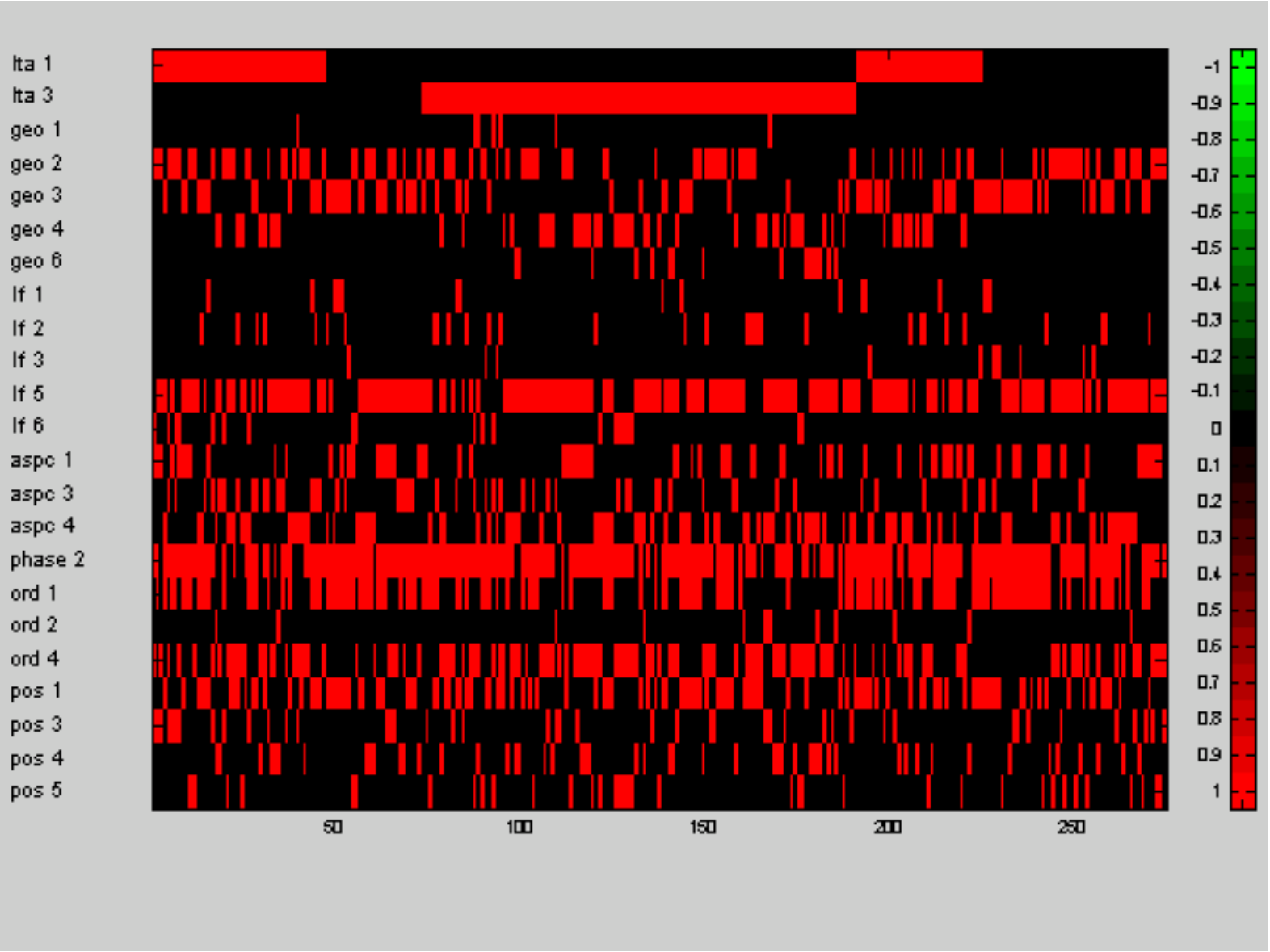
1. 4/5 clusters of sites.
2. 3 outlier sites.
3. 2 species groups.



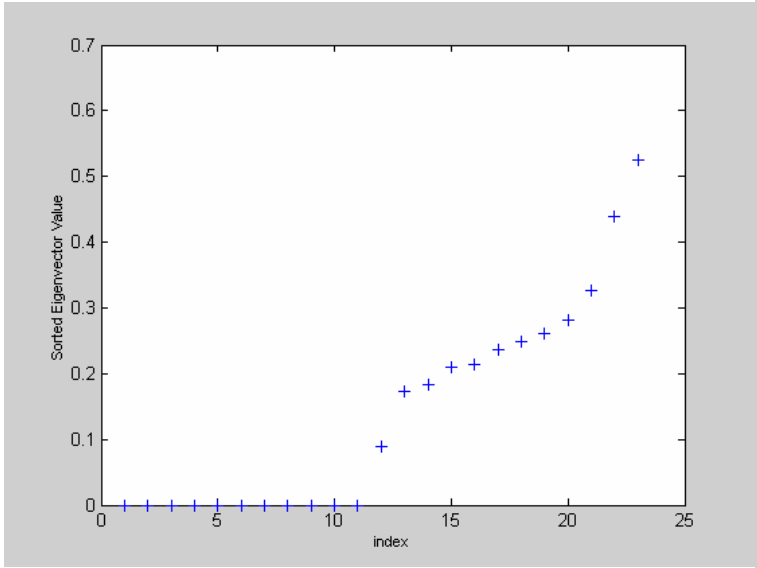
Y Eigenvalues

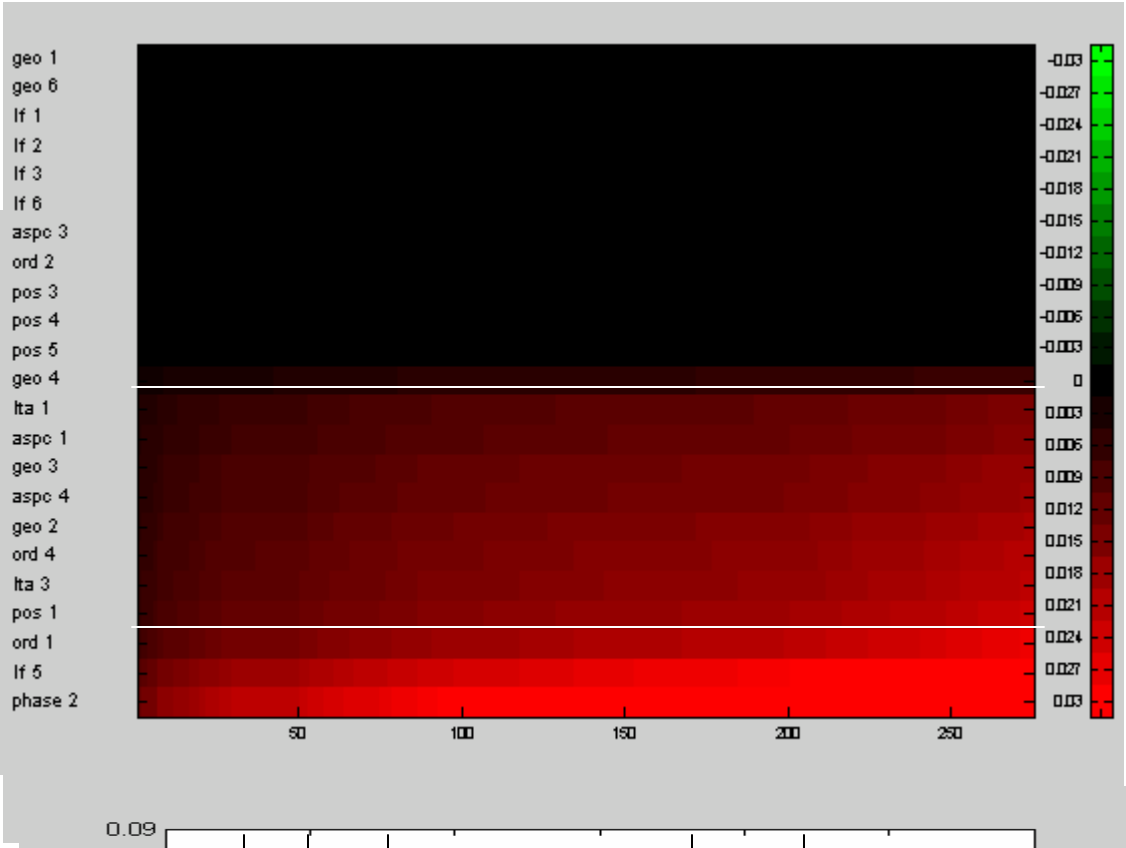


Raw data

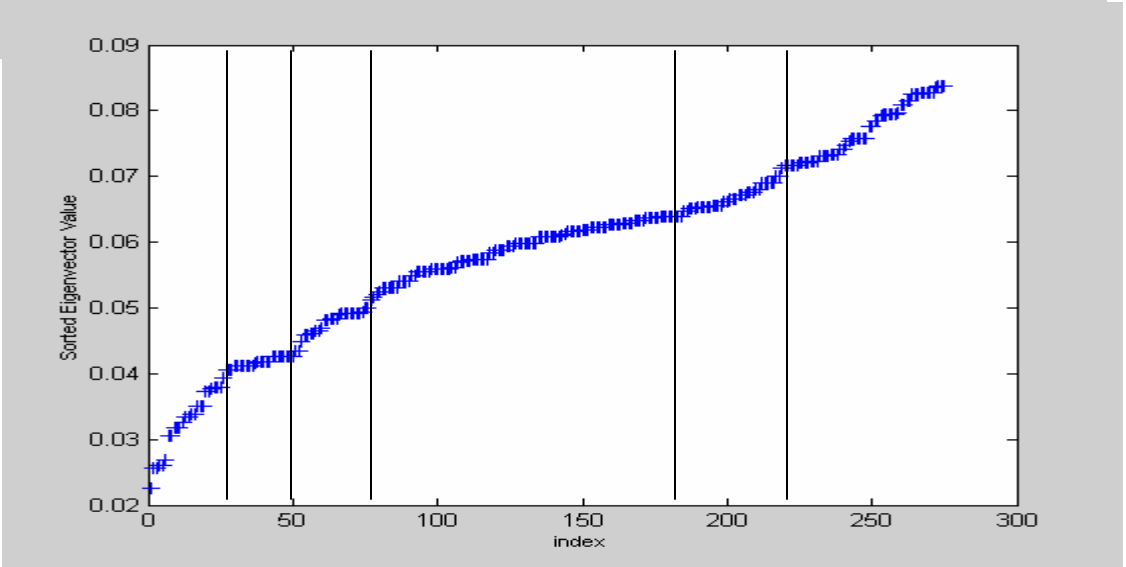

lta 1
lta 3
geo 1
geo 2
geo 3
geo 4
geo 6
lf 1
lf 2
lf 3
lf 5
lf 6
aspc 1
aspc 3
aspc 4
phase 2
ord 1
ord 2
ord 4
pos 1
pos 3
pos 4
pos 5
50
100
150
200
250
-1
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Component 1

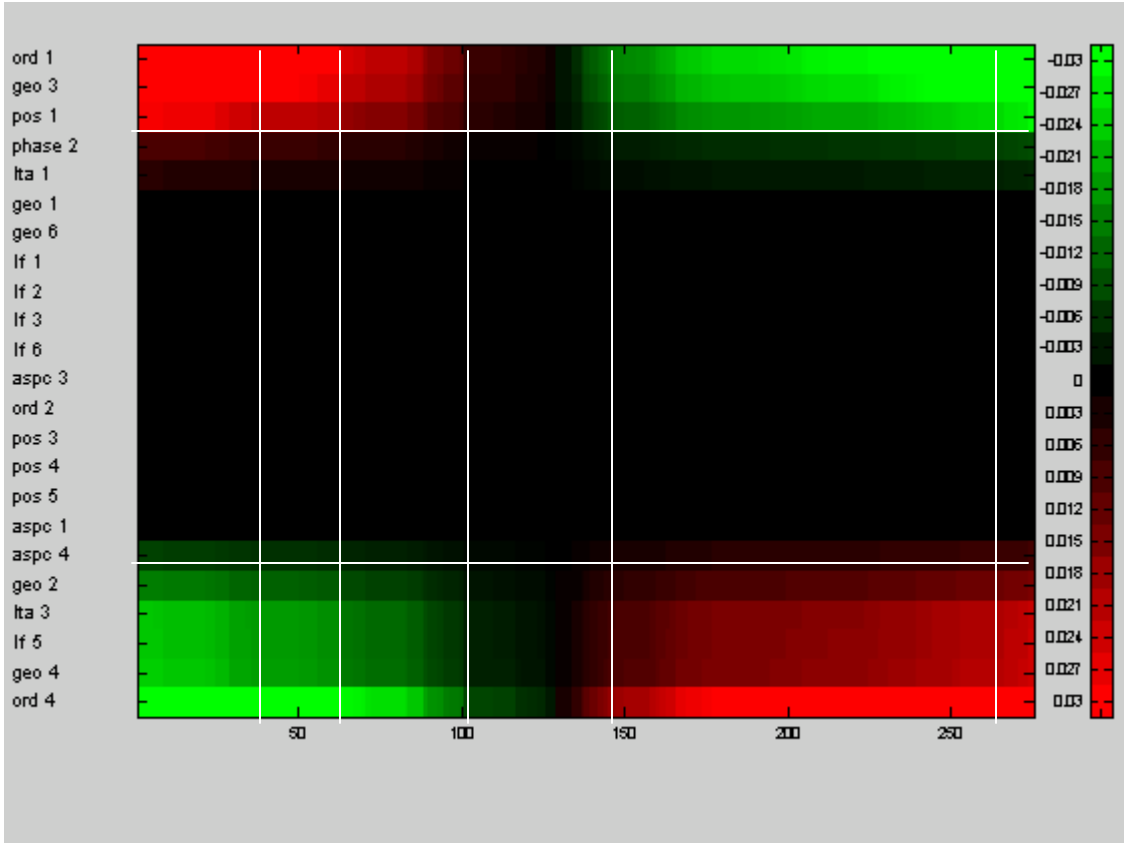


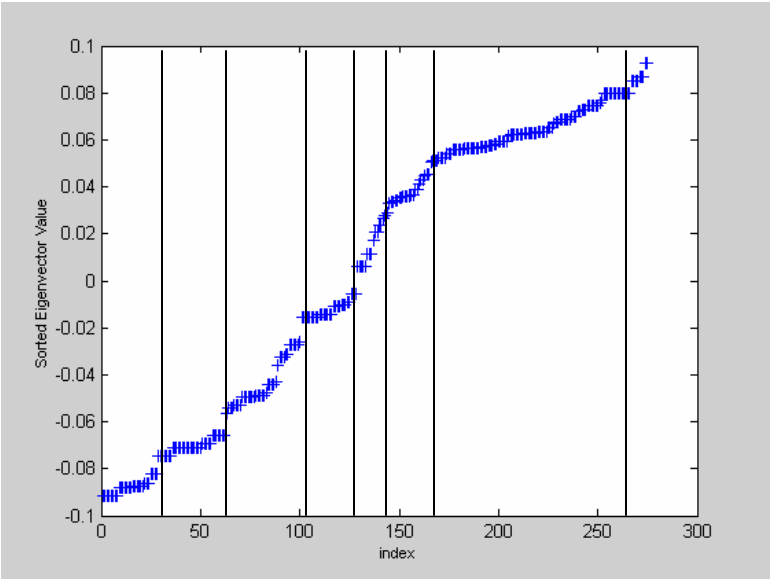


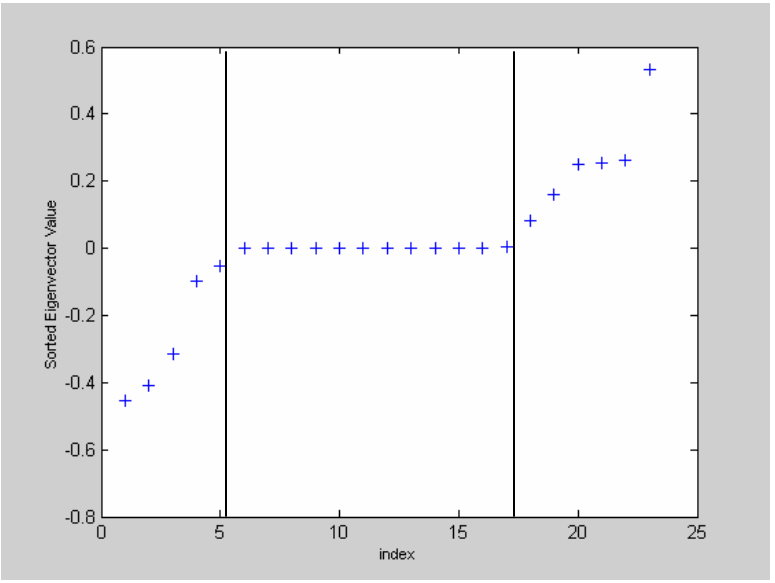
Quite a bit of imagination!



Component 2



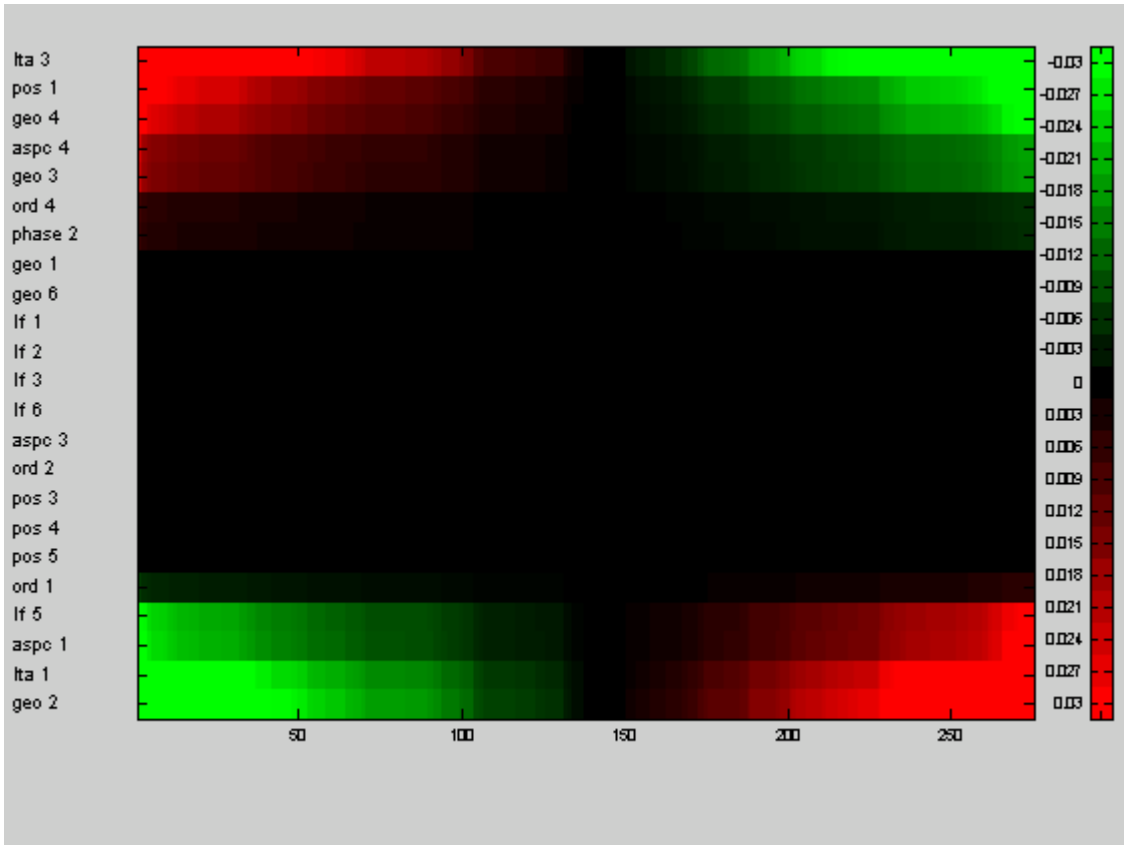


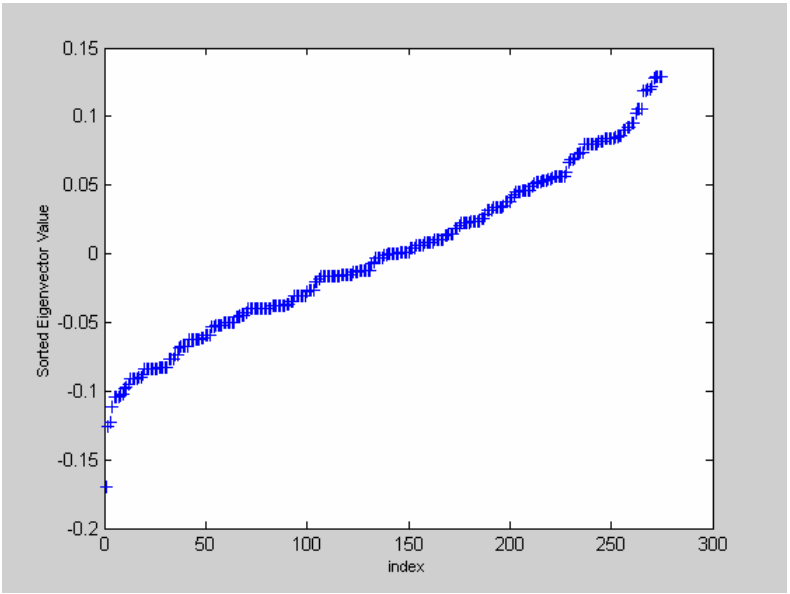


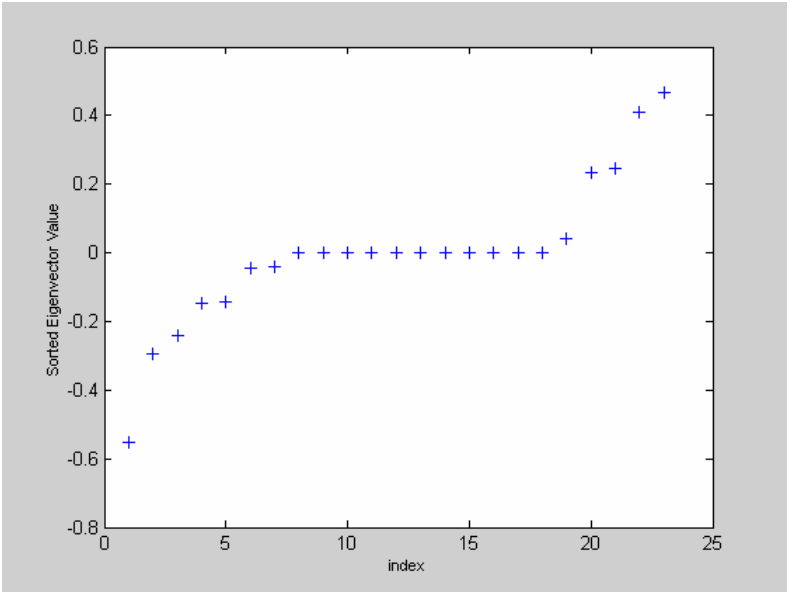
NB: 2nd level clustering.

Real? Not sure. EV looks real.

Component 3

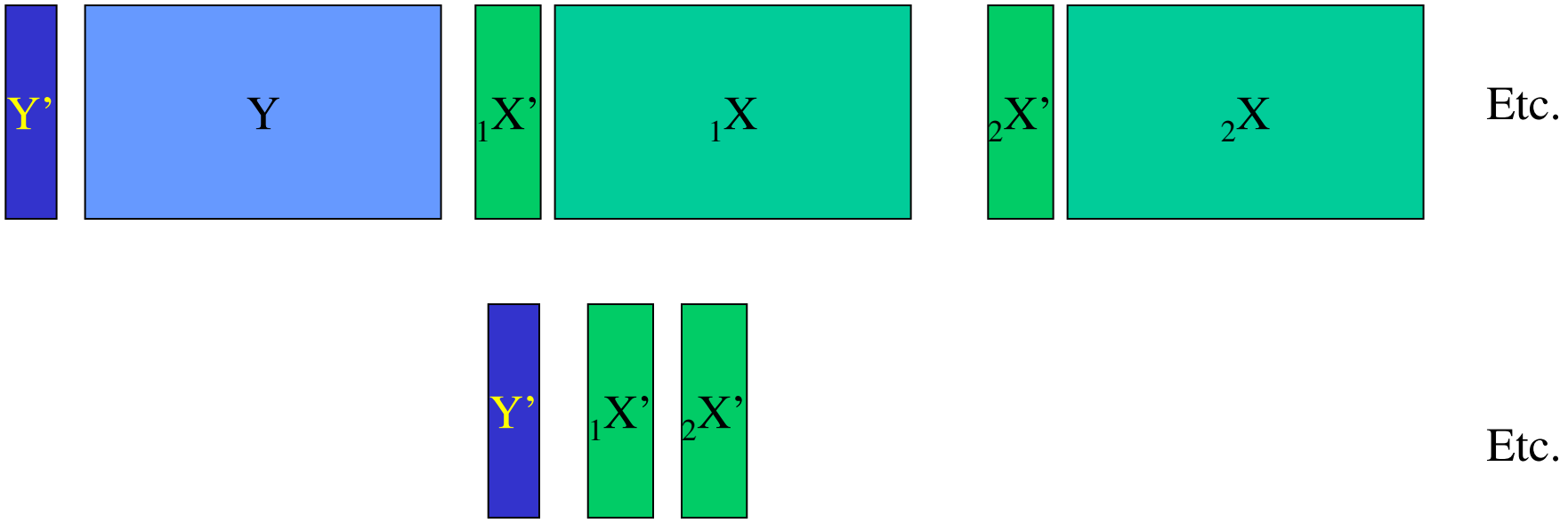






Looks random.

(PLS) Multi-Block Strategy

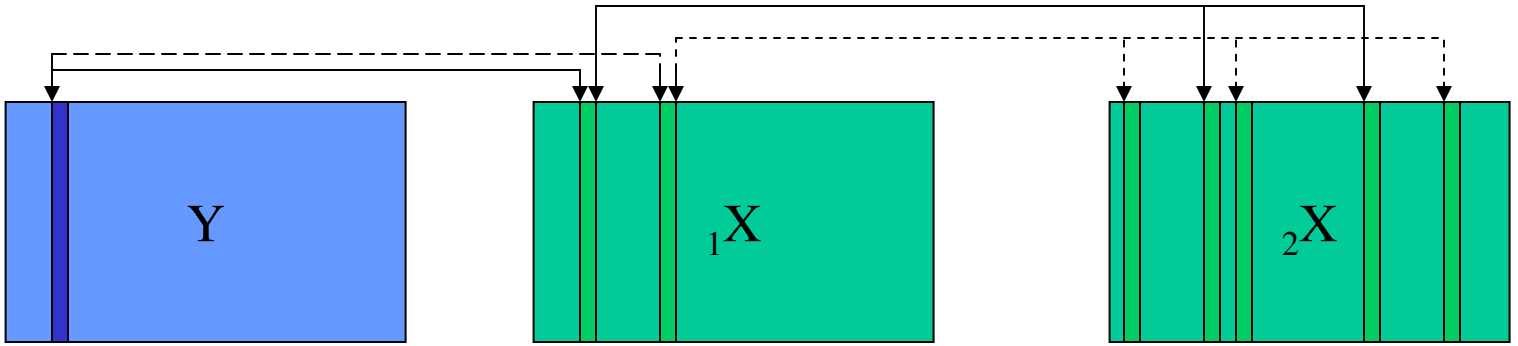


Problems:

1. Linear.
2. Complex analysis and visualizations.
3. No direct focus on individual variables.

Horst, P. (1961), Kettering, J. (1971), PLS guys.

Staged (Interactive) Recursive Partitioning



Advantages

1. Non-linear.
2. Easy to interpret.
3. Focus is on specific response and predictors.

See also, CS Multi Relational Data Mining.

Data

Variable	N^a	$\bar{x}$
Aster patens	118	0.3019
Carex digitalis	100	0.4133
Desmodium glutinosum	101	3.361
Desmodium roundifolium	117	0.2244
Euphorbia corollata	122	0.1948
Lespedeza intermedia	130	0.2098
Monarda russeliana	125	0.4110
Panicum commutatum	133	0.2495
Phryma leptostachya	106	0.3160
Smilax bona-nox	101	1.0650
Smilax racemosa	125	0.3244
Vaccinium vacillans	195	3.3610

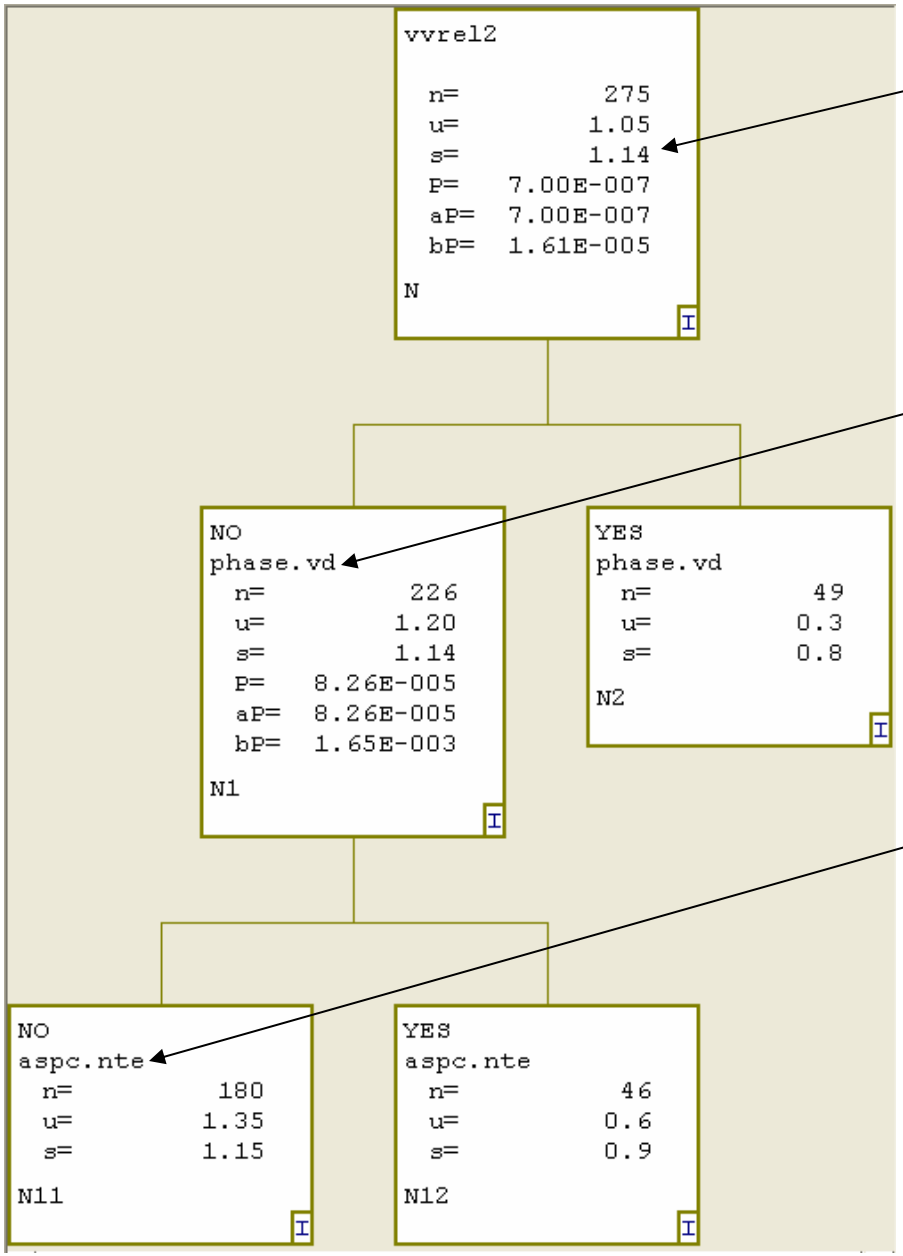
Target Response.

1. Most common tree.
2. Highest average.
3. Selected by rSVD.

Table 1
Independent categorical variables with the number in each category

Variable	Category	N
Landtype association	Current River Breaks	118
	Current River Hills	76
	Jack Fork, Eminence Breaks	81
Geology	Roubidoux	86
	Upper Gasconade	107
	Lower Gasconade	55
	Gunter	6
	Eminence	15
	Van Buren	6
Landform	Summit	16
	Shoulder ridge	24
	Shoulder	16
	Backslope	195
	Bench	24
Aspect class	Exposed	97
	Neutral east	54
	Neutral west	45
	Protected	79
Phase	Deep	226
	Variable depth	49
Soil order	Alfisol	104
	Mollisol	9
	None	72
	Ultisol	90
Position	Upper	90
	Upper-middle	28
	Middle	51
	Lower-middle	28
	Lower	61
	None	18

RP on Target Response



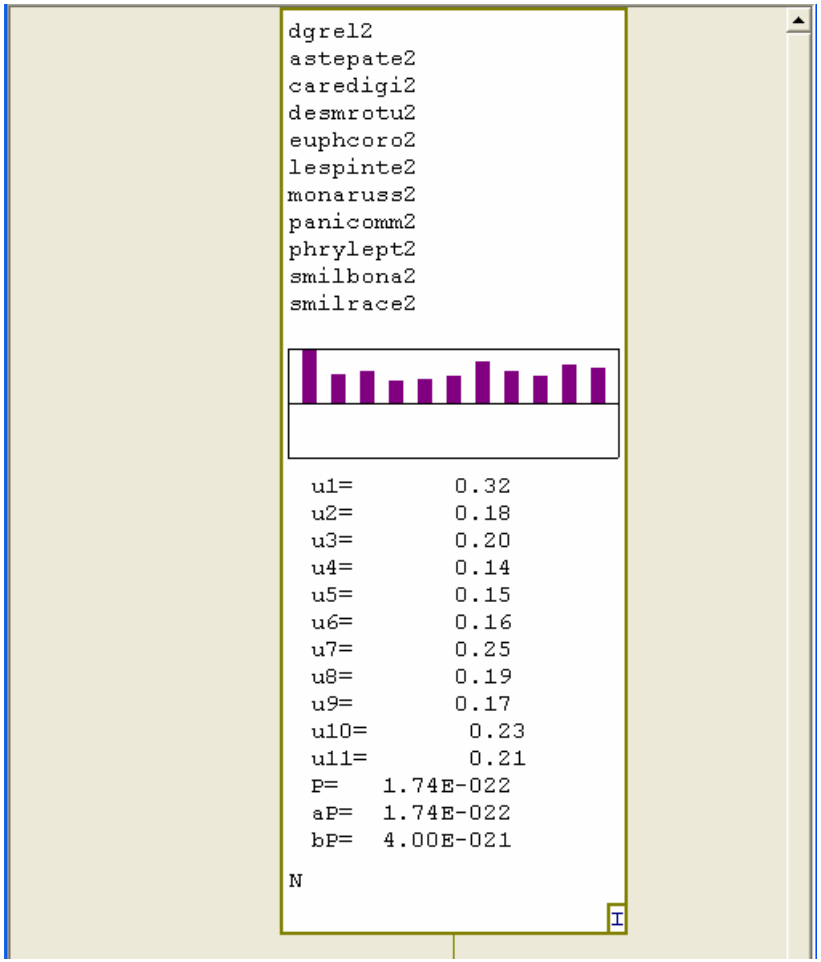
Response with largest variance.

Soil depth.

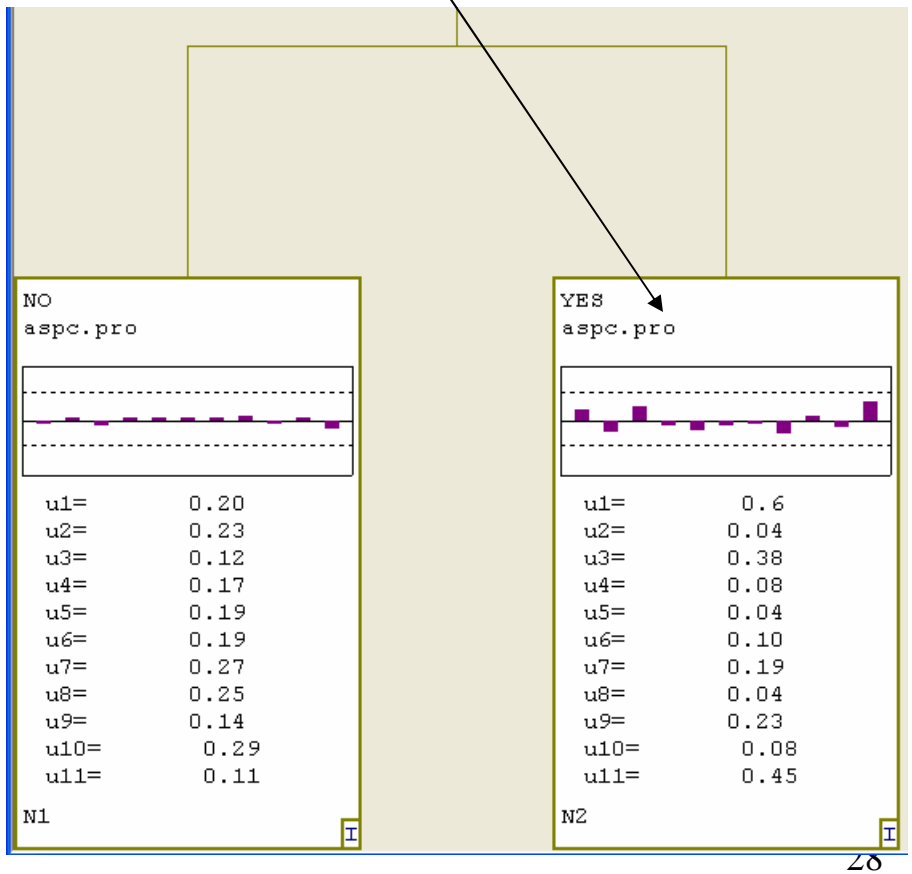
Sun exposure.

Hypothesis testing based recursive partitioning.

Multivariate Recursive Partitioning



Protected, North



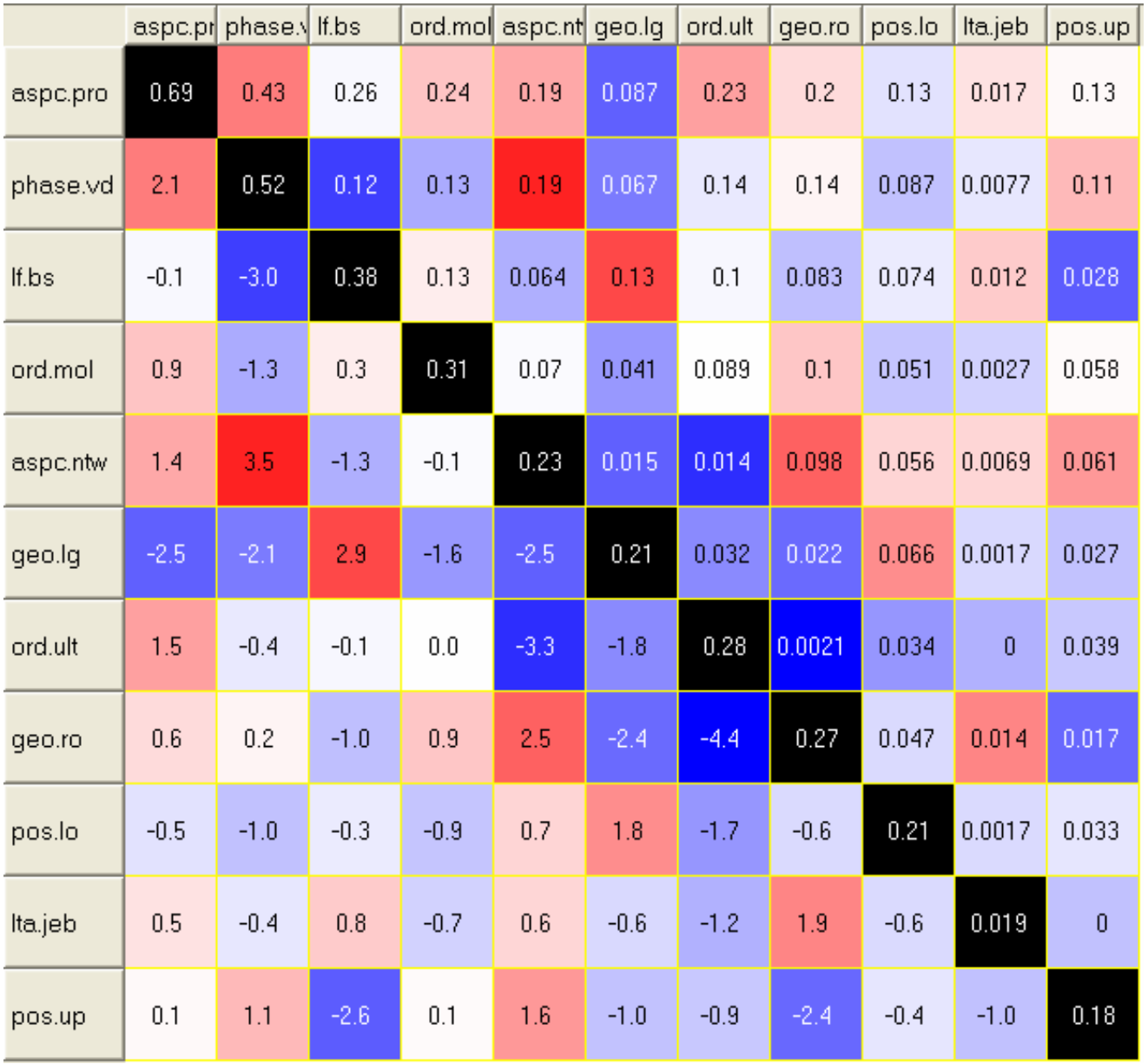
Multivariate, Multiple-Tree

Recursive Partitioning

Relationships among Predictors

Red
Synergism

Blue
Correlated



	aspc.pr	phase.v	lf.bs	ord.mol	aspc.nt	geo.lg	ord.ult	geo.ro	pos.lo	lta.jeb	pos.up
aspc.pro	0.69	0.43	0.26	0.24	0.19	0.087	0.23	0.2	0.13	0.017	0.13
phase.vd	2.1	0.52	0.12	0.13	0.19	0.067	0.14	0.14	0.087	0.0077	0.11
lf.bs	-0.1	-3.0	0.38	0.13	0.064	0.13	0.1	0.083	0.074	0.012	0.028
ord.mol	0.9	-1.3	0.3	0.31	0.07	0.041	0.089	0.1	0.051	0.0027	0.058
aspc.ntw	1.4	3.5	-1.3	-0.1	0.23	0.015	0.014	0.098	0.056	0.0069	0.061
geo.lg	-2.5	-2.1	2.9	-1.6	-2.5	0.21	0.032	0.022	0.066	0.0017	0.027
ord.ult	1.5	-0.4	-0.1	0.0	-3.3	-1.8	0.28	0.0021	0.034	0	0.039
geo.ro	0.6	0.2	-1.0	0.9	2.5	-2.4	-4.4	0.27	0.047	0.014	0.017
pos.lo	-0.5	-1.0	-0.3	-0.9	0.7	1.8	-1.7	-0.6	0.21	0.0017	0.033
lta.jeb	0.5	-0.4	0.8	-0.7	0.6	-0.6	-1.2	1.9	-0.6	0.019	0
pos.up	0.1	1.1	-2.6	0.1	1.6	-1.0	-0.9	-2.4	-0.4	-1.0	0.18

Summary

rSVD : Identify variable relationships.

RP : Organized and applied in stages
based on subject matter organization.

Obvious : go after genes, proteins, metabolism.

References

rSVD : Liu, Hawkins, Young. PNAS 2003.

Recursive Partitioning. www.goldenhelix.com

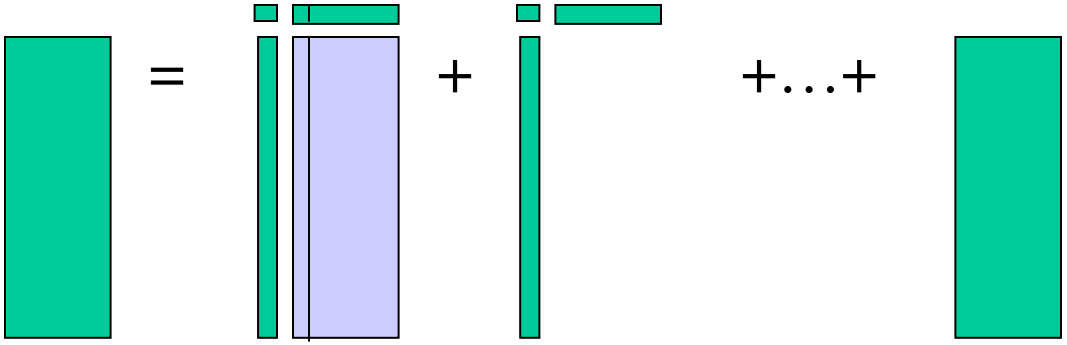
www.niss.org/PowerMV (Jack Liu, Jun Feng)

Post docs at NISS and SAMSI - Apply!

Needed research

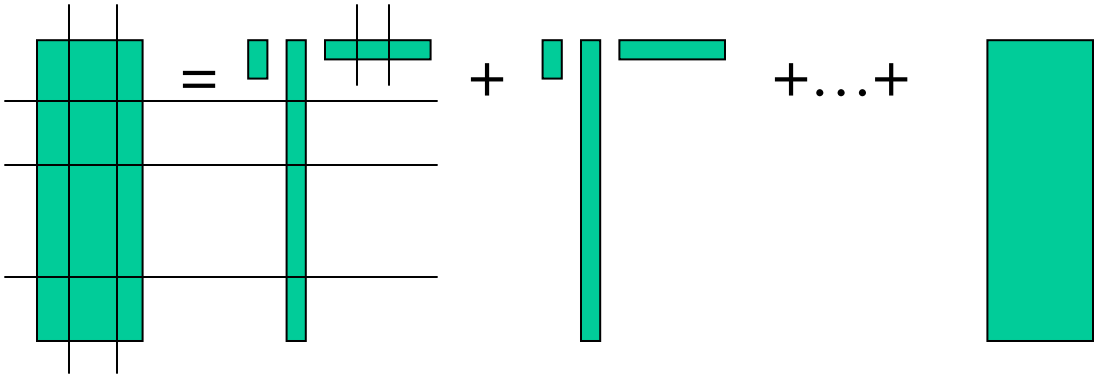
- Analytical stopping rule for rSVD
- Scale rSVD (hundreds*thousands) Done
- Good visualization methods Done
- Benchmarking (works well at drug companies)
- Linking to row and column annotations

Review Computation of rSVD



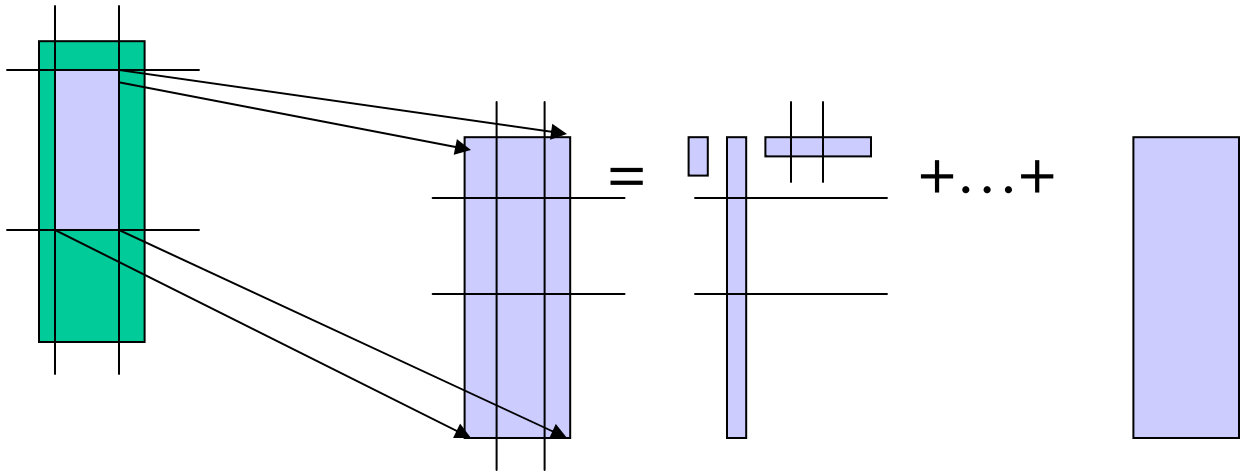
Segmentation

1. Rank R and L eigenvector elements
2. Reorder rows and cols of data matrix
3. Segment



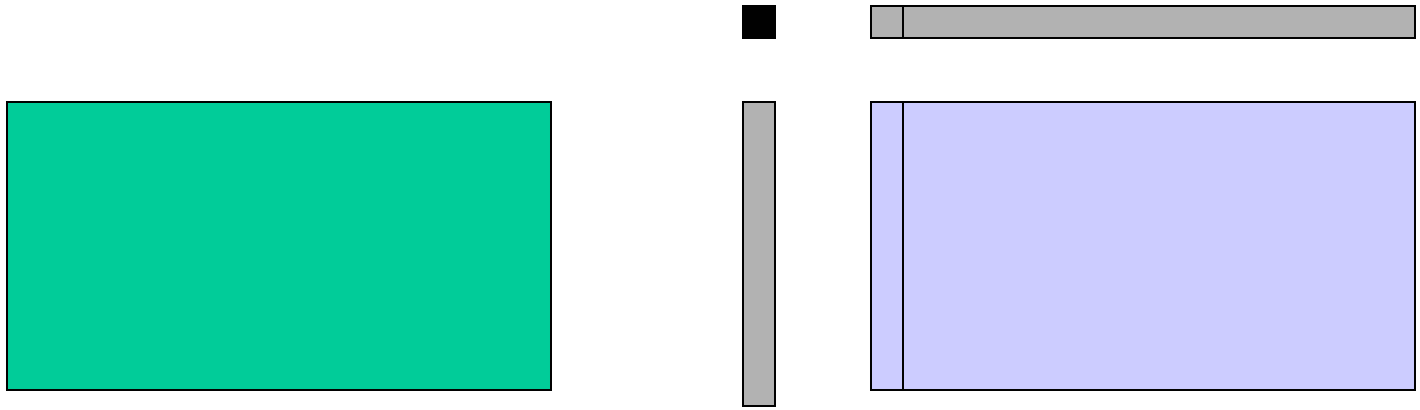
Deep Clustering

1. Select a 2D segment from the original matrix.
3. Re-compute rSVD.
2. Reorder rows and cols of data matrix
3. Re-segment



Continue, as makes sense.

Y = eigenvalue * LHE ‘ * RHE + E



rSVD :

1. Clusters rows and columns at same time.
2. Automatic weighting of observations.
3. Automatically deals with missing data.
4. Robust to outliers.

References:

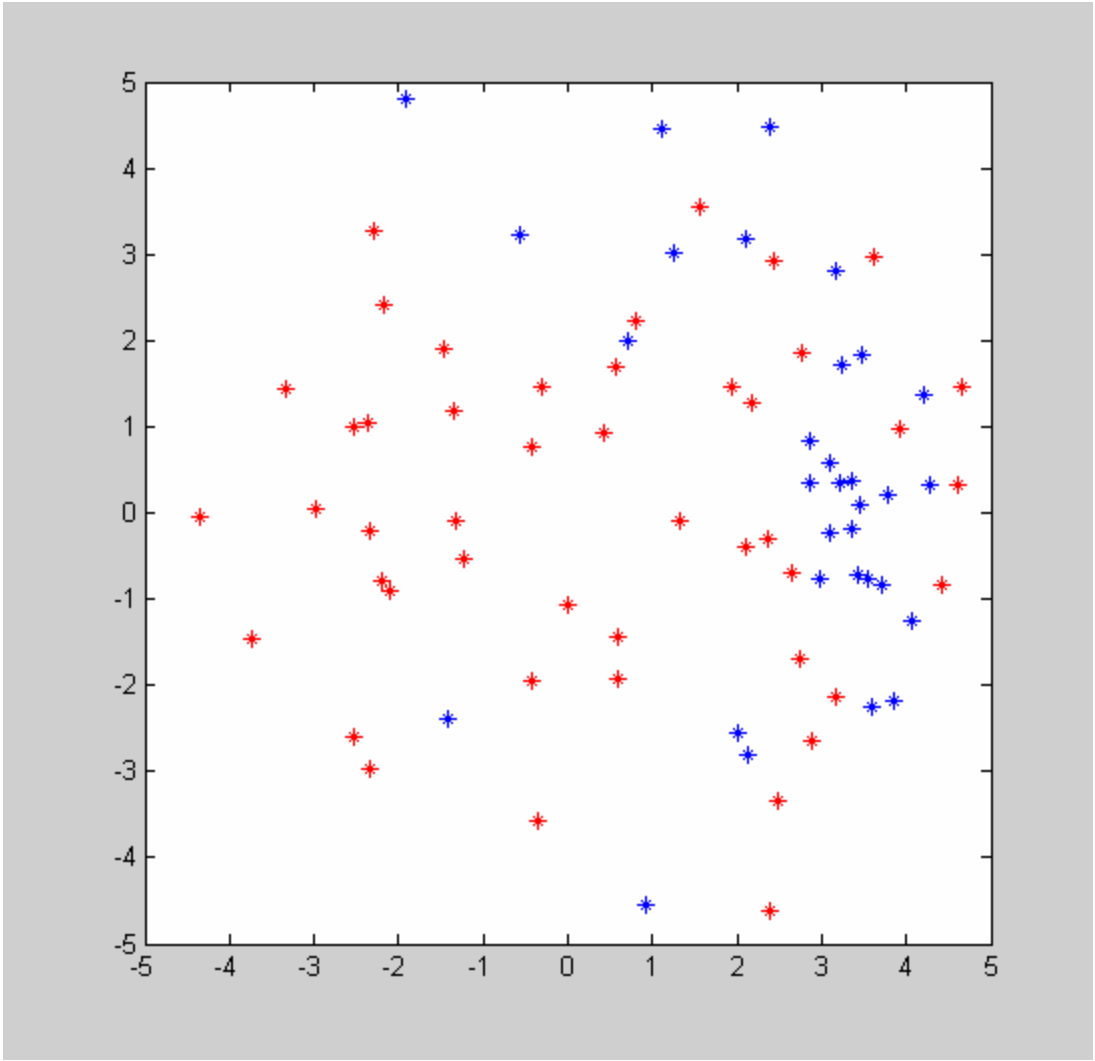
Papers 122 and 123 www.niss.org

Li, Hawkins, Young (2003) PNAS

Orley Ashenfelter, California Versus All Challengers: The 1999 Cabernet Challenge

Coming: PowerMV from NISS.

Biplot



Goal

Group species profile based on site characteristics.