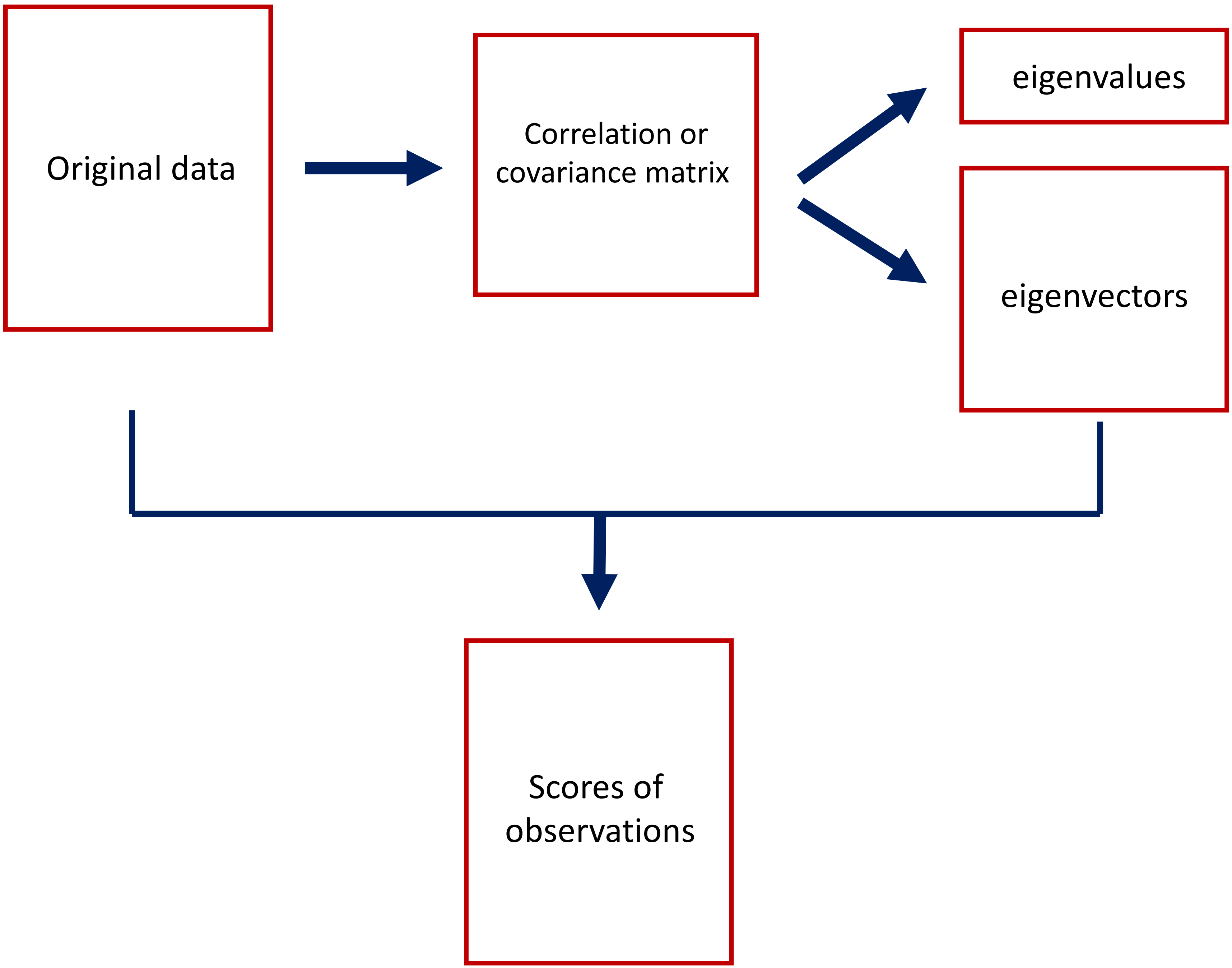
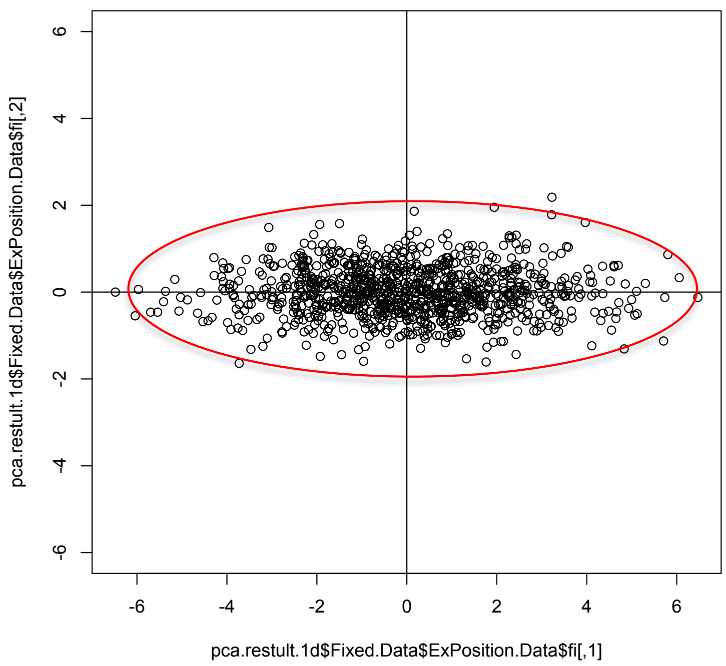
The “mathematics” of Principal Component Analysis (PCA)

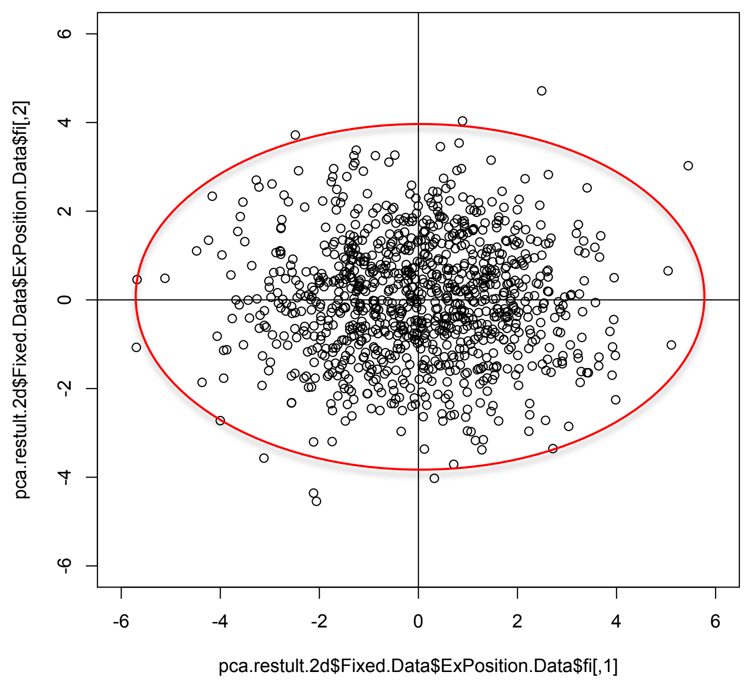


PCA Scores: one versus two dimensions

1D

2D





Lecture 21: How many PCA dimensions? Inferential frameworks for determining number of axes to interpret and the significance of each variable on each axis (lots of work on this area).

1st) determine how many axes to interpret (i.e., how many PCs capture correlated variation in the data?).

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Statistics & Data Analysis 49 (2005) 974–997

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

How many principal components? stopping rules for determining the number of non-trivial axes revisited

Pedro R. Peres-Neto*, Donald A. Jackson, Keith M. Somers

Inferential frameworks for determining number of axes to interpret and the significance of each variable on each axis are usually nor performed.

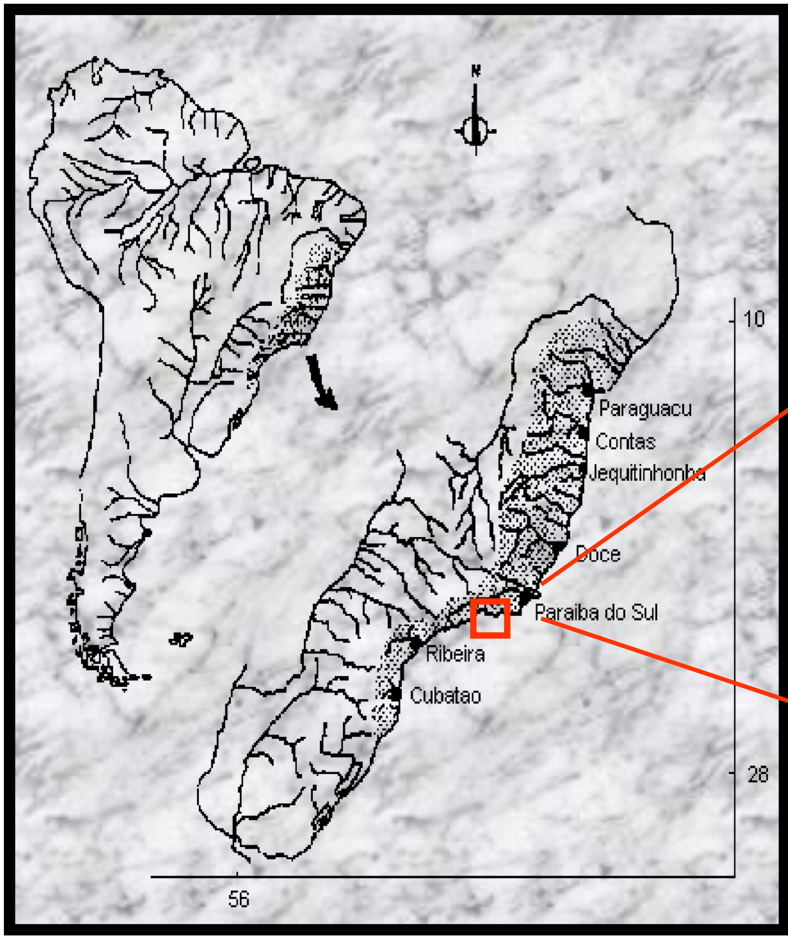
2nd) for each significant axis, determine which variable is significant on each of them.

Ecology, 84(9), 2003, pp. 2347–2363

GIVING MEANINGFUL INTERPRETATION TO ORDINATION AXES: ASSESSING LOADING SIGNIFICANCE IN PRINCIPAL COMPONENT ANALYSIS

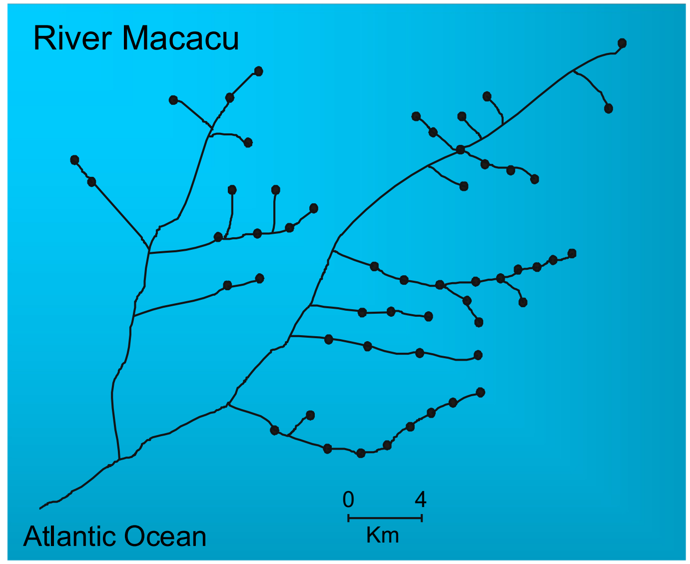
PEDRO R. PERES-NETO,[1] DONALD A. JACKSON, AND KEITH M. SOMERS

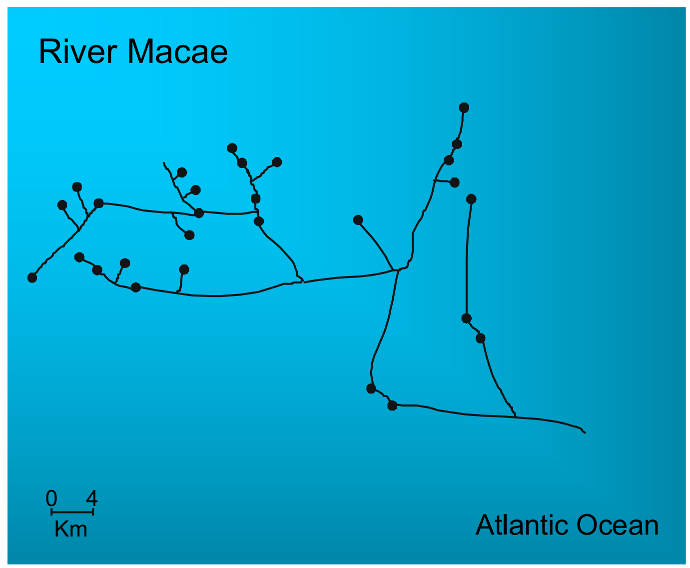
Principal component analysis: a complete example



53 sites

28 sites





What is the correlation structure and differences among streams in terms of their environmental features?

- Depth
- Depth variation
- Current velocity
- Current variation
- Substrate composition: Boulder, rubble, gravel and sand
- Substrate variation (variance in composition)
- Stream width variation (irregularity)
- Area
- Altitude

Oecologia (2004) 140: 352–360
DOI 10.1007/s00442-004-1578-3

COMMUNITY ECOLOGY

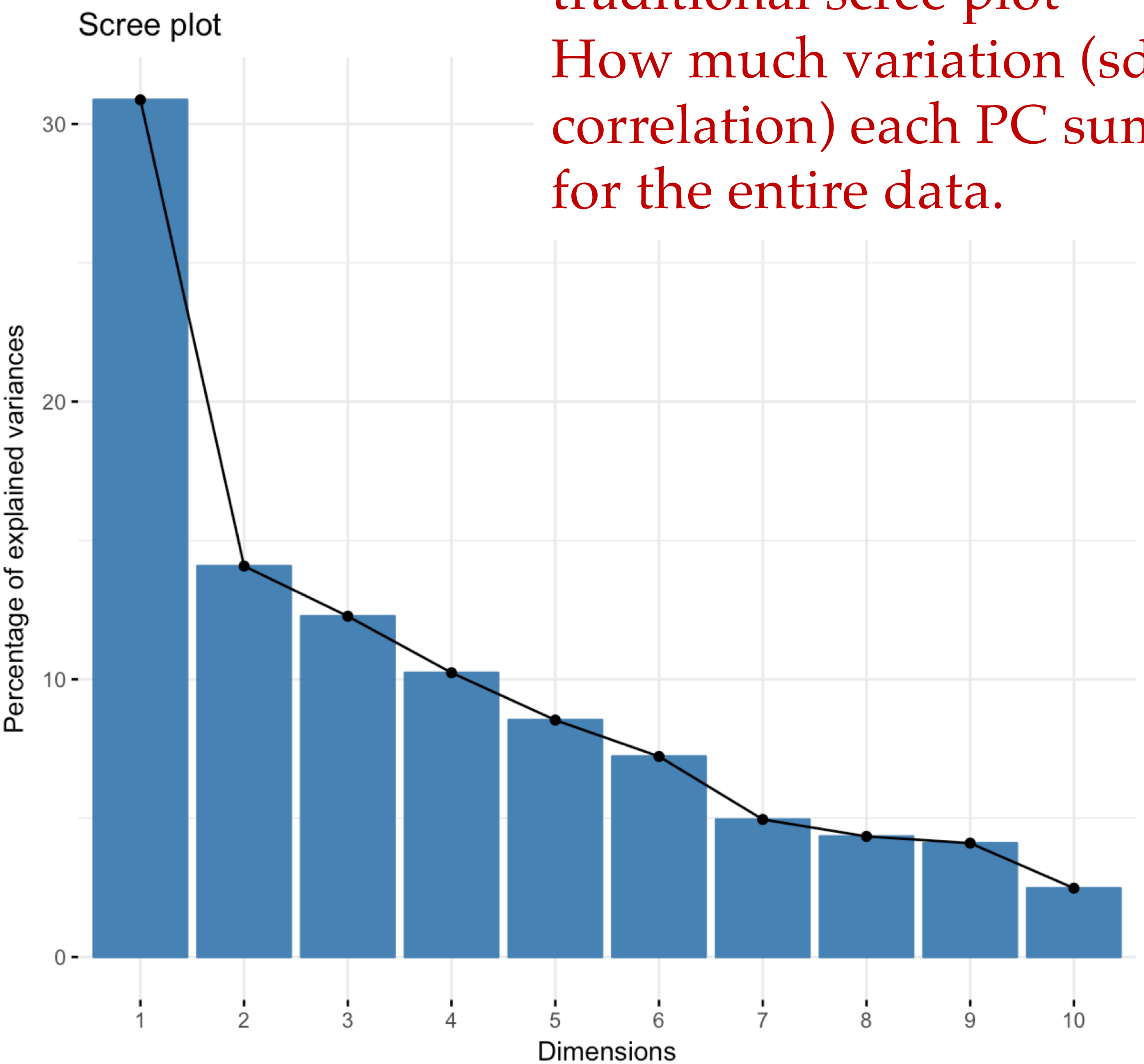
Pedro R. Peres-Neto

Patterns in the co-occurrence of fish species in streams: the role of site suitability, morphology and phylogeny versus species interactions

Correlation matrix

1.00	0.04	0.28	-0.07	0.06	-0.33	-0.02	0.12	-0.02	0.05	0.01	-0.11
0.04	1.00	-0.12	0.85	0.31	0.07	0.07	-0.35	-0.42	0.84	0.86	-0.66
0.28	-0.12	1.00	-0.07	-0.17	-0.08	0.02	0.19	-0.13	-0.13	-0.17	-0.03
-0.07	0.85	-0.07	1.00	0.36	0.17	0.06	-0.44	-0.33	0.71	0.71	-0.57
0.06	0.31	-0.17	0.36	1.00	0.08	-0.33	-0.81	0.33	0.36	0.20	-0.17
-0.33	0.07	-0.08	0.17	0.08	1.00	-0.11	-0.52	0.09	0.00	0.11	0.23
-0.02	0.07	0.02	0.06	-0.33	-0.11	1.00	-0.04	-0.22	0.14	0.05	0.00
0.12	-0.35	0.19	-0.44	-0.81	-0.52	-0.04	1.00	-0.25	-0.39	-0.25	0.04
-0.02	-0.42	-0.13	-0.33	0.33	0.09	-0.22	-0.25	1.00	-0.38	-0.38	0.33
0.05	0.84	-0.13	0.71	0.36	0.00	0.14	-0.39	-0.38	1.00	0.66	-0.58
0.01	0.86	-0.17	0.71	0.20	0.11	0.05	-0.25	-0.38	0.66	1.00	-0.60
-0.11	-0.66	-0.03	-0.57	-0.17	0.23	0.00	0.04	0.33	-0.58	-0.60	1.00

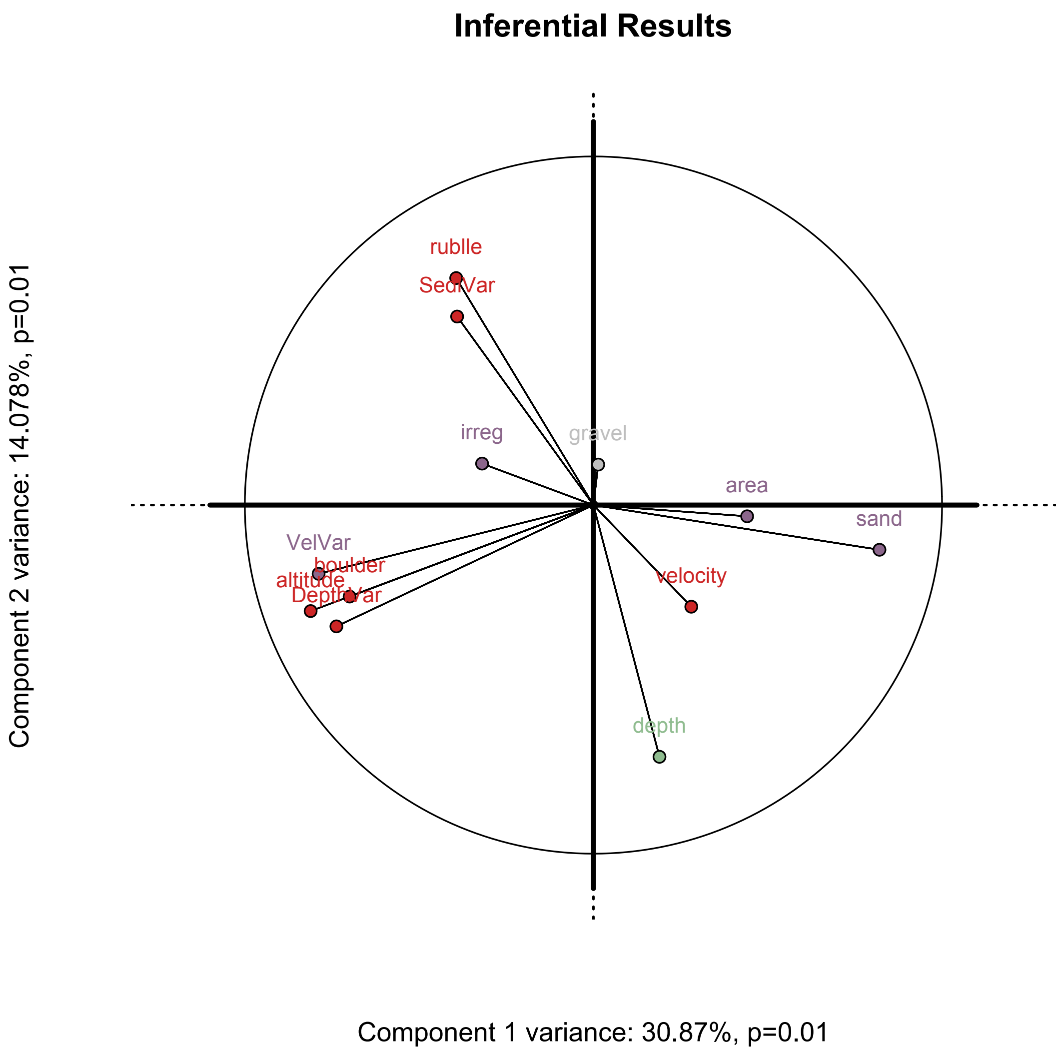
Eigenvalue contribution – the traditional scree plot - How much variation (sd and correlation) each PC summarizes for the entire data.



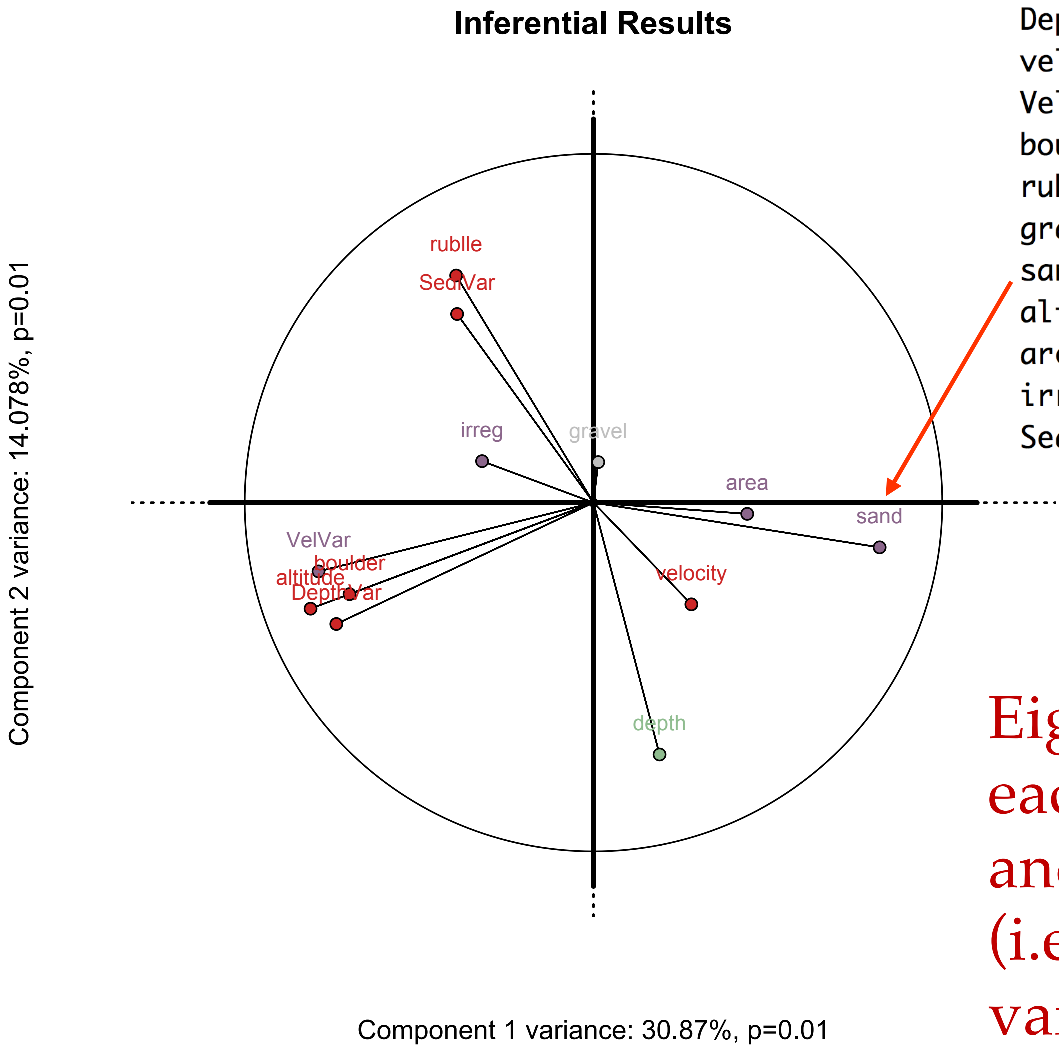
Eigenvector structure (2 first dimensions)

	PC-1	PC-2
depth	0.098416371	-0.55557259
DepthVar	-0.383072589	-0.26772556
velocity	0.145820452	-0.22434910
VelVar	-0.409585483	-0.15169873
boulder	-0.363399847	-0.20189977
rublle	-0.204526467	0.50098773
gravel	0.007091107	0.08935752
sand	0.426264131	-0.09866678
altitude	-0.421467330	-0.23396335
area	0.229031867	-0.02477526
irreg	-0.165951470	0.09149688
SediVar	-0.203159109	0.41607768

Eigenvector plot :



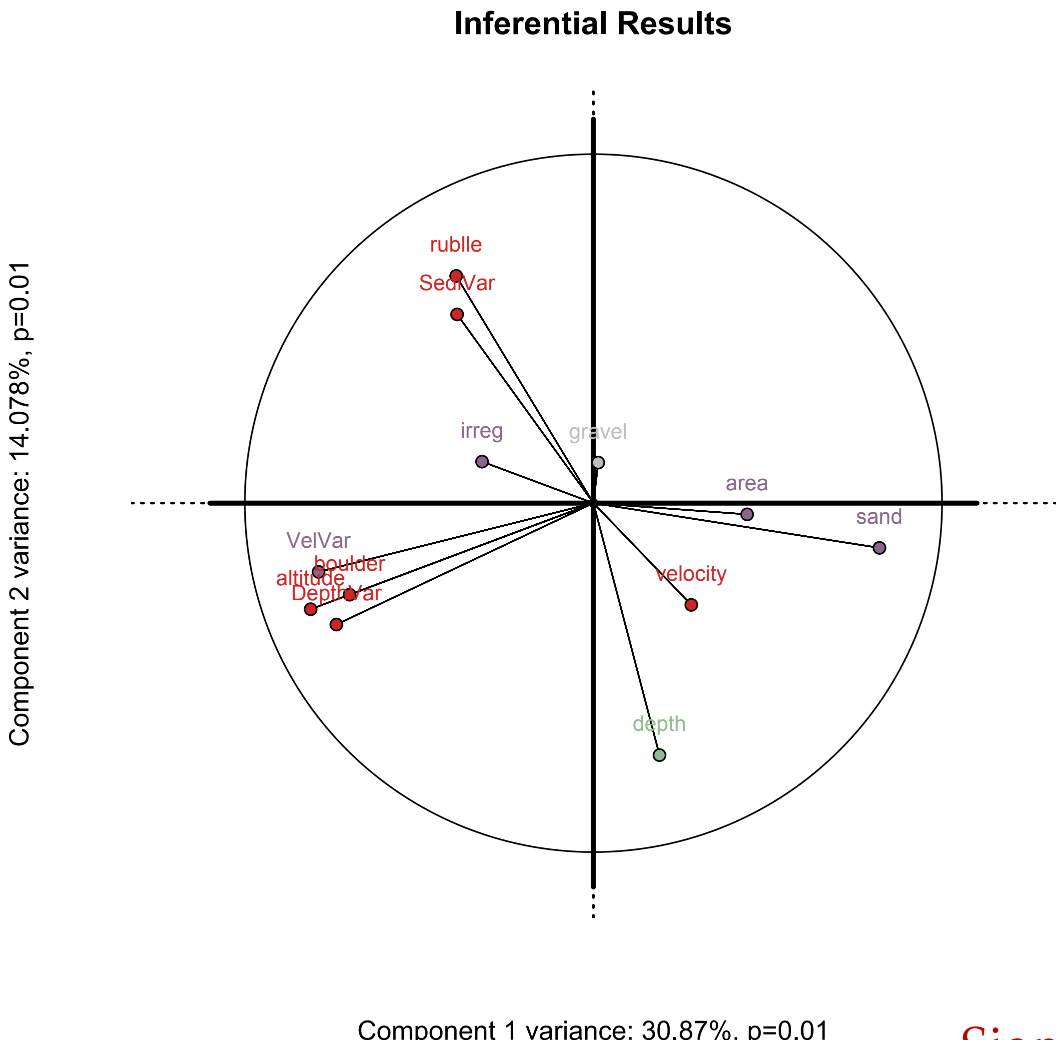
Eigenvector plot :

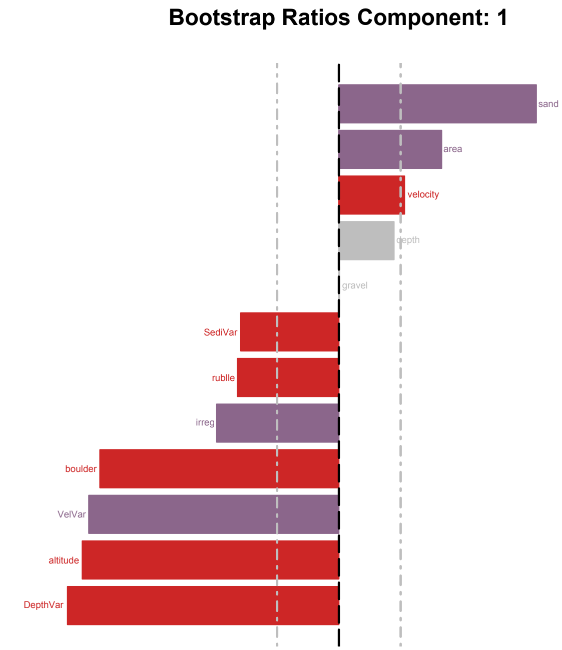


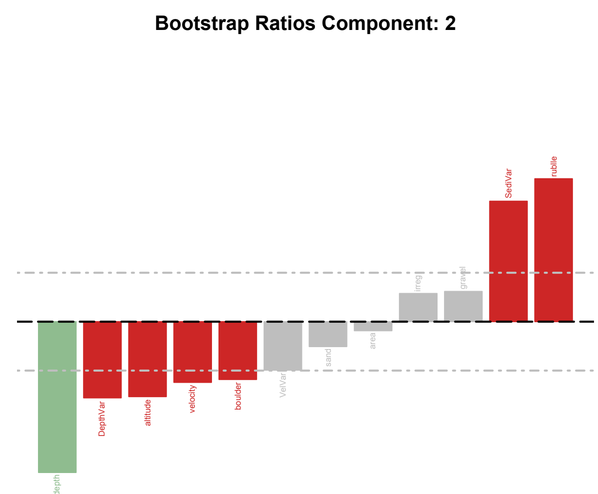
depth	0.098416371	-0.55557259
DepthVar	-0.383072589	-0.26772556
velocity	0.145820452	-0.22434910
VelVar	-0.409585483	-0.15169873
boulder	-0.363399847	-0.20189977
rublle	-0.204526467	0.50098773
gravel	0.007091107	0.08935752
sand	0.426264131	-0.09866678
altitude	-0.421467330	-0.23396335
area	0.229031867	-0.02477526
irreg	-0.165951470	0.09149688
SediVar	-0.203159109	0.41607768

Eigenvector coefficients for each variable are plotted and presented as arrows (i.e., correlation of each variable with each principal component).

Eigenvector plot:







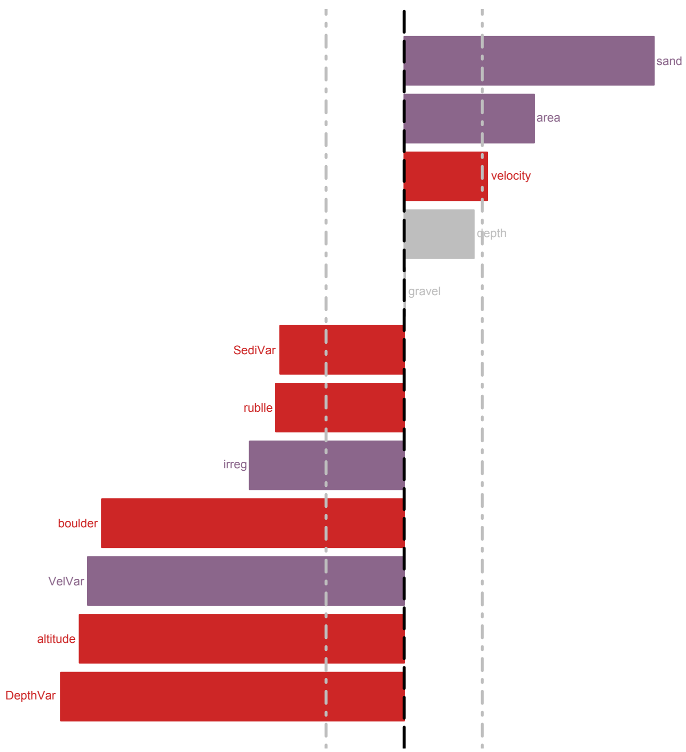
Significant variable contributions are determined if the eigenvector coefficient for the variables are beyond the confidence interval.

Ecology, 84(9), 2003, pp. 2347–2363

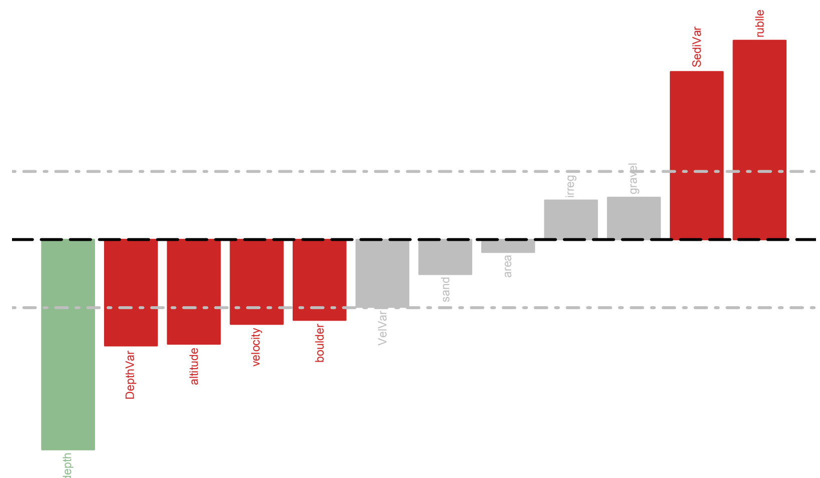
GIVING MEANINGFUL INTERPRETATION TO ORDINATION AXES: ASSESSING LOADING SIGNIFICANCE IN PRINCIPAL COMPONENT ANALYSIS

PEDRO R. PERES-NETO,[1] DONALD A. JACKSON, AND KEITH M. SOMERS

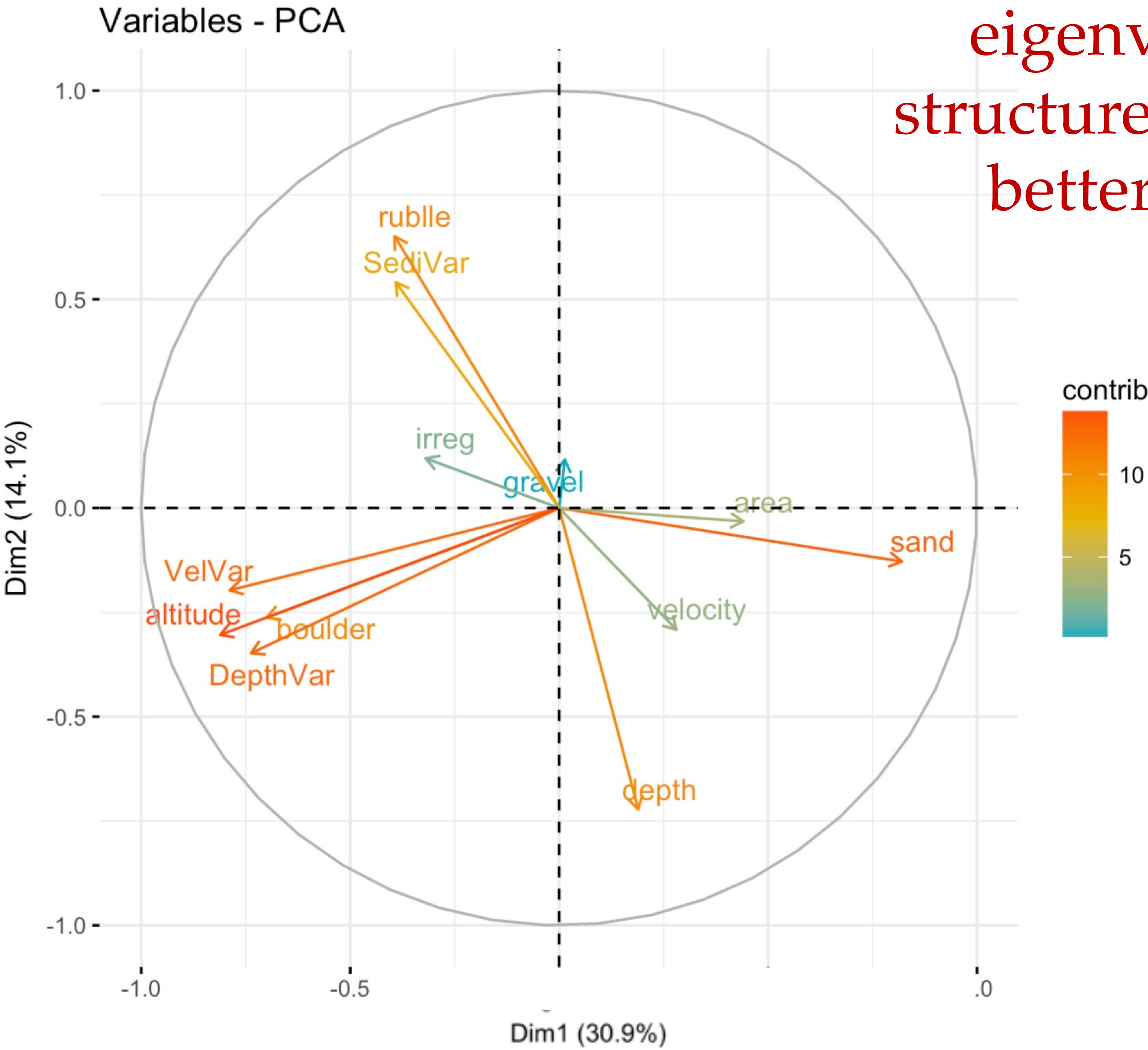
Bootstrap Ratios Component: 1



Bootstrap Ratios Component: 2

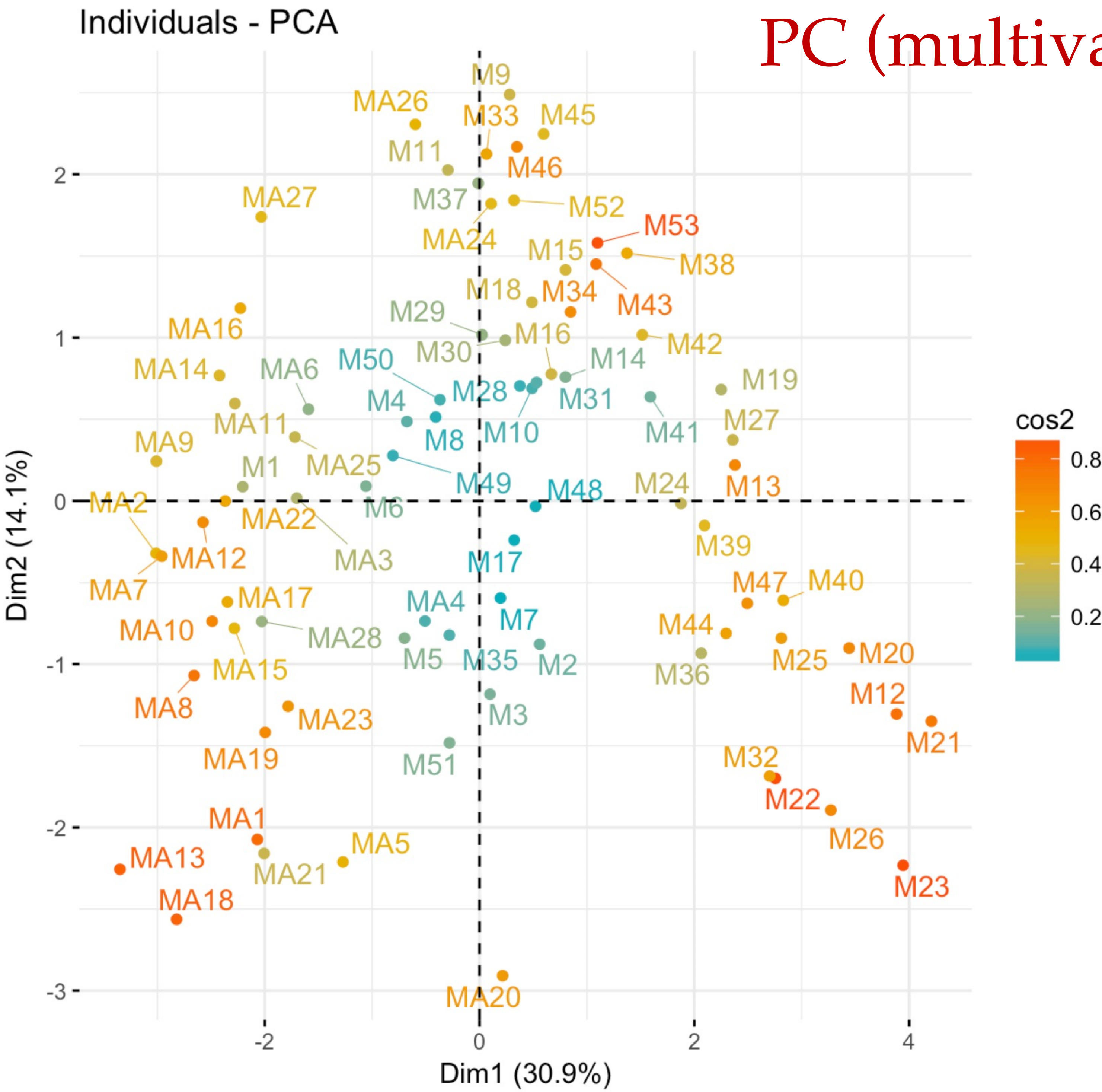


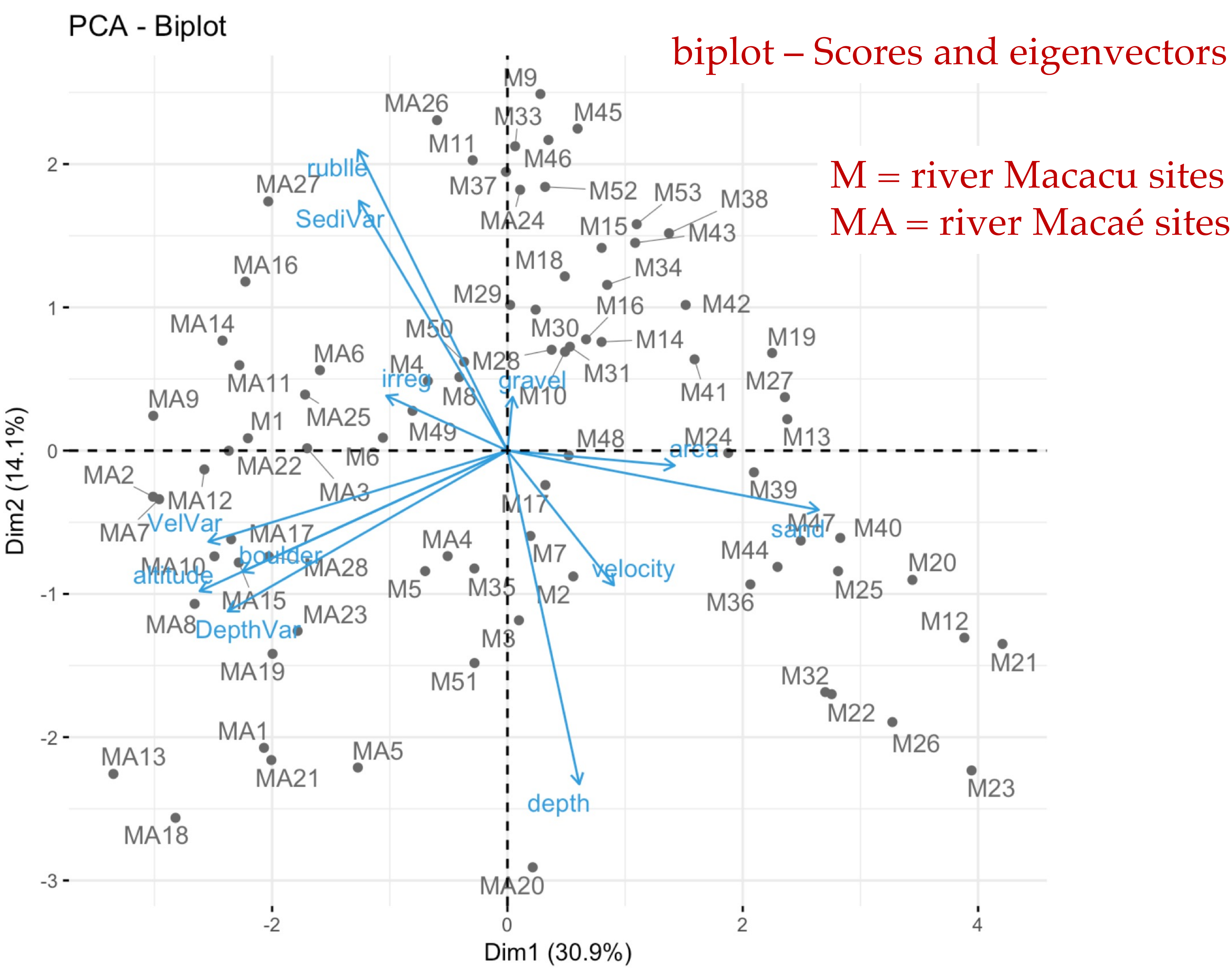
Significant variable contributions are determined if the eigenvector coefficient for the variables are beyond the confidence interval.

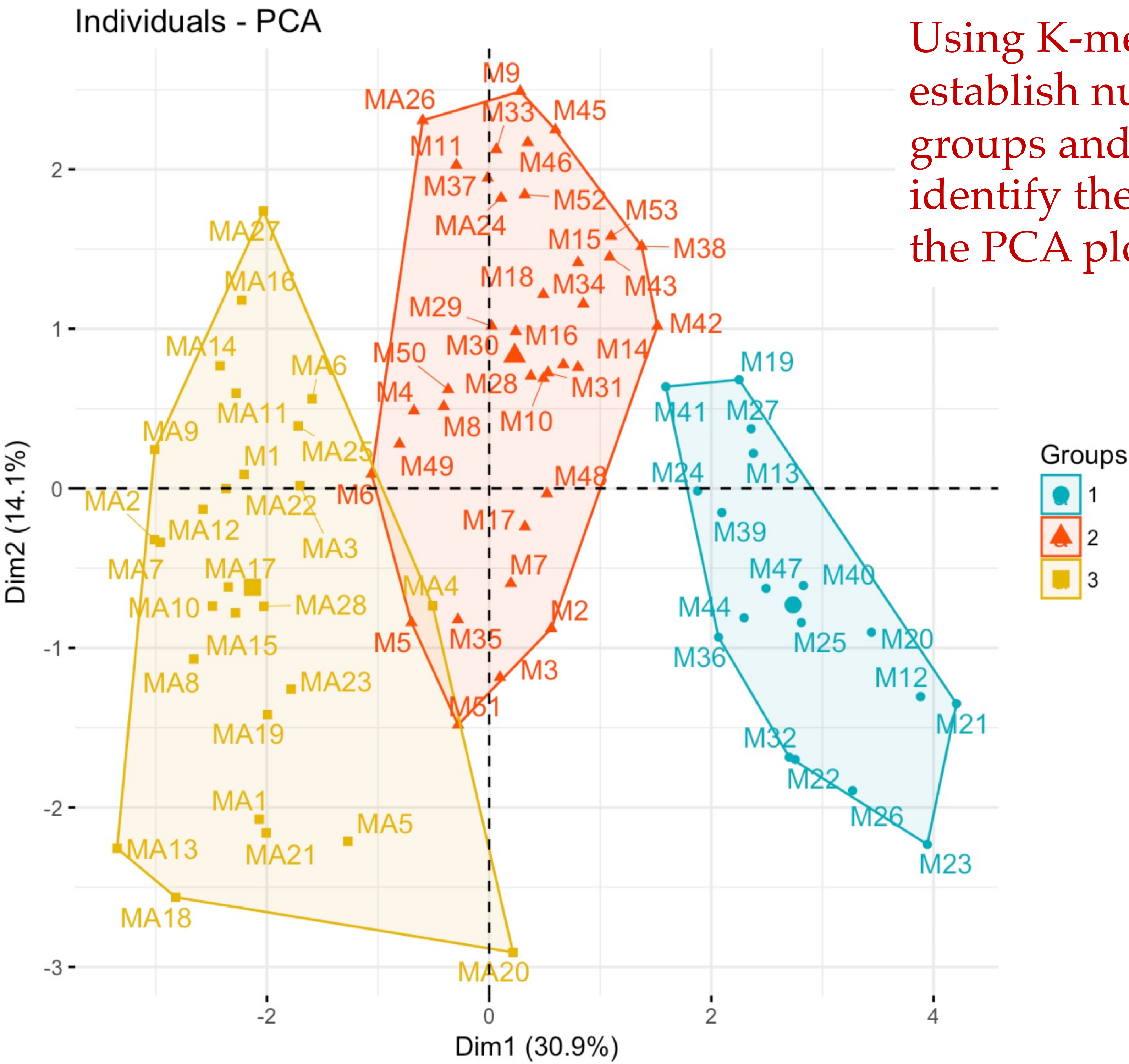


eigenvector structure – much better plot

PC (multivariate) scores




PCA - Biplot
biplot – Scores and eigenvectors
M = river Macacu sites
MA = river Macaé sites
Dim1 (30.9%)
Dim2 (14.1%)
rublle
SediVar
irreg
gravel
area
sand
velocity
depth
VelVar
boulder
altitude
DepthVar



Using K-means to establish number of groups and then identify them the PCA plot

Further applications and interpretations of Principal component axes (representing data in their lower dimensionalities)

Variation in personality and fitness in wild female baboons

Robert M. Seyfarth[a,1], Joan B. Silk[b], and Dorothy L. Cheney[c]

16980–16985 | PNAS | October 16, 2012 | vol. 109 | no. 42

Studies of personality in nonhuman primates have usually relied on assessments by humans and seldom considered the function of the resulting "trait" classifications. In contrast, we applied exploratory principal component analysis to seven behaviors among 45 wild female baboons over 7 y to determine whether the personality dimensions that emerged were associated with measures of reproductive success. We identified three relatively stable personality dimensions, each characterized by a distinct suite of behaviors that were not redundant with dominance rank or the availability of kin. Females

Study details!

Behavioral and Hormonal Data. Ten-minute focal animal observations (37) were conducted almost daily using a common protocol. All approaches, vocalizations, and friendly and aggressive interactions were recorded on a continuous basis. We also noted all grooming interactions and their durations (22).

We used seven independent behavioral variables, calculated annually for each female, to construct the components of personality styles: (*i*) Alone: the proportion of focal samples in which a female did not interact with any other individual (excluding dependent infants) for the entire 10-min period. (*ii*) Friendly: the rate at which a female touched or embraced other females. (*iii*) Aggression: the rate at which a female behaved aggressively (head bobs, lunges, chases, bites) toward other females. Because high-ranking females have more available targets than do low-ranking females, we corrected each female's annual aggression rate for the proportion of females who ranked below her. Thus, a low-ranking female might score higher on 'aggression' than a high-ranking one. (*iv*–*vii*) Grunts: the frequency with which a female grunted when approaching (*iv*) a higher-ranking female who had a young infant ($<$ 3 mo), (*v*) a higher-ranking female who did not have an infant, (*vi*) a lower-ranking female who had a young infant, and (*vii*) a lower-ranking female who did not have an infant.

Table 1. Loadings of behaviors onto three principal components

Behavior	Component 1: Aloof	Component 2: Loner	Component 3: Nice
Alone	0.14	**0.67**	**−0.32**
Aggression	**0.64**	0.04	−0.06
Friendly	0.03	−0.17	**0.76**
Grunting			
HR no inf	−0.03	**0.71**	0.19
HR inf	**0.71**	0.17	0.30
LR no inf	**−0.55**	**0.34**	**0.37**
LR inf	−0.00	0.10	**0.51**

Loadings > 0.32 are in boldface (26). HR inf, higher-ranking with infant; HR no inf, higher-ranking with no infant; LR inf, lower-ranking with infant; LR no inf, lower-ranking with no infant.

Often post-hoc interpretation to each axis is produced.

HR = High-ranking females
LR = Low-ranking females
Inf = infant

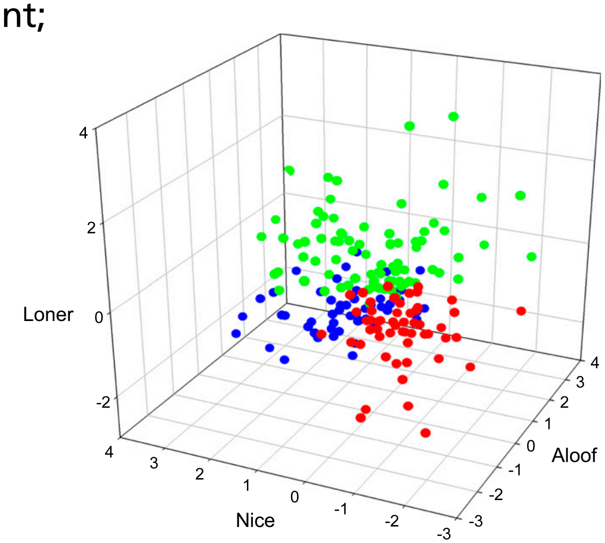


Fig. 1. The distribution of scores on three principal components or personality dimensions. Each point represents scores on Aloof, Loner, and Nice components (*n* = 45 females observed for 1–7 y, for a total of 189 female-years). Colors depict the assignment of a female to a particular personality cluster in that year. The choice of three clusters and cluster assignments were made using the partitioning around medoids program.

Table 1. Loadings of behaviors onto three principal components

Behavior	Component 1: Aloof	Component 2: Loner	Component 3: Nice
Alone	0.14	**0.67**	**−0.32**
Aggression	**0.64**	0.04	−0.06
Friendly	0.03	−0.17	**0.76**
Grunting			
HR no inf	−0.03	**0.71**	0.19
HR inf	**0.71**	0.17	0.30
LR no inf	**−0.55**	**0.34**	**0.37**
LR inf	−0.00	0.10	**0.51**

Loadings > 0.32 are in boldface (26). HR inf, higher-ranking with infant; HR no inf, higher-ranking with no infant; LR inf, lower-ranking with infant; LR no inf, lower-ranking with no infant.

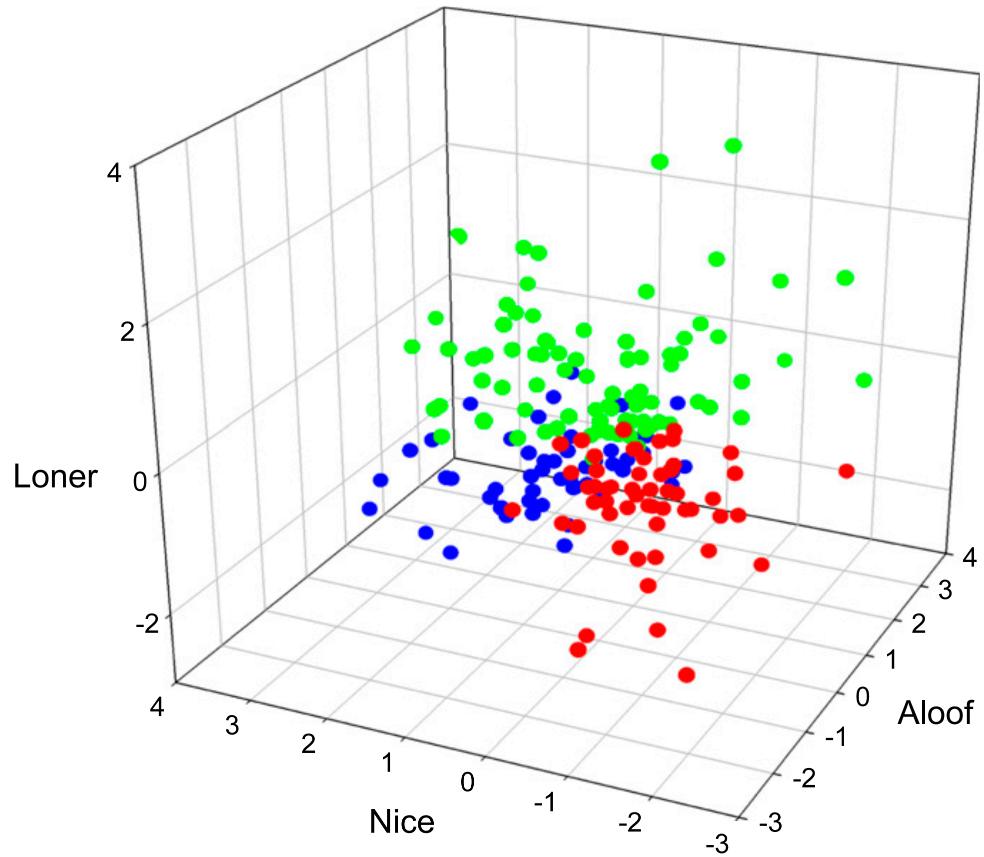


Fig. 1. The distribution of scores on three principal components or personality dimensions. Each point represents scores on Aloof, Loner, and Nice components (n = 45 females observed for 1–7 y, for a total of 189 female-years). Colors depict the assignment of a female to a particular personality cluster in that year. The choice of three clusters and cluster assignments were made using the partitioning around medoids program.

Mapping the environment of our planet – a very “small” example

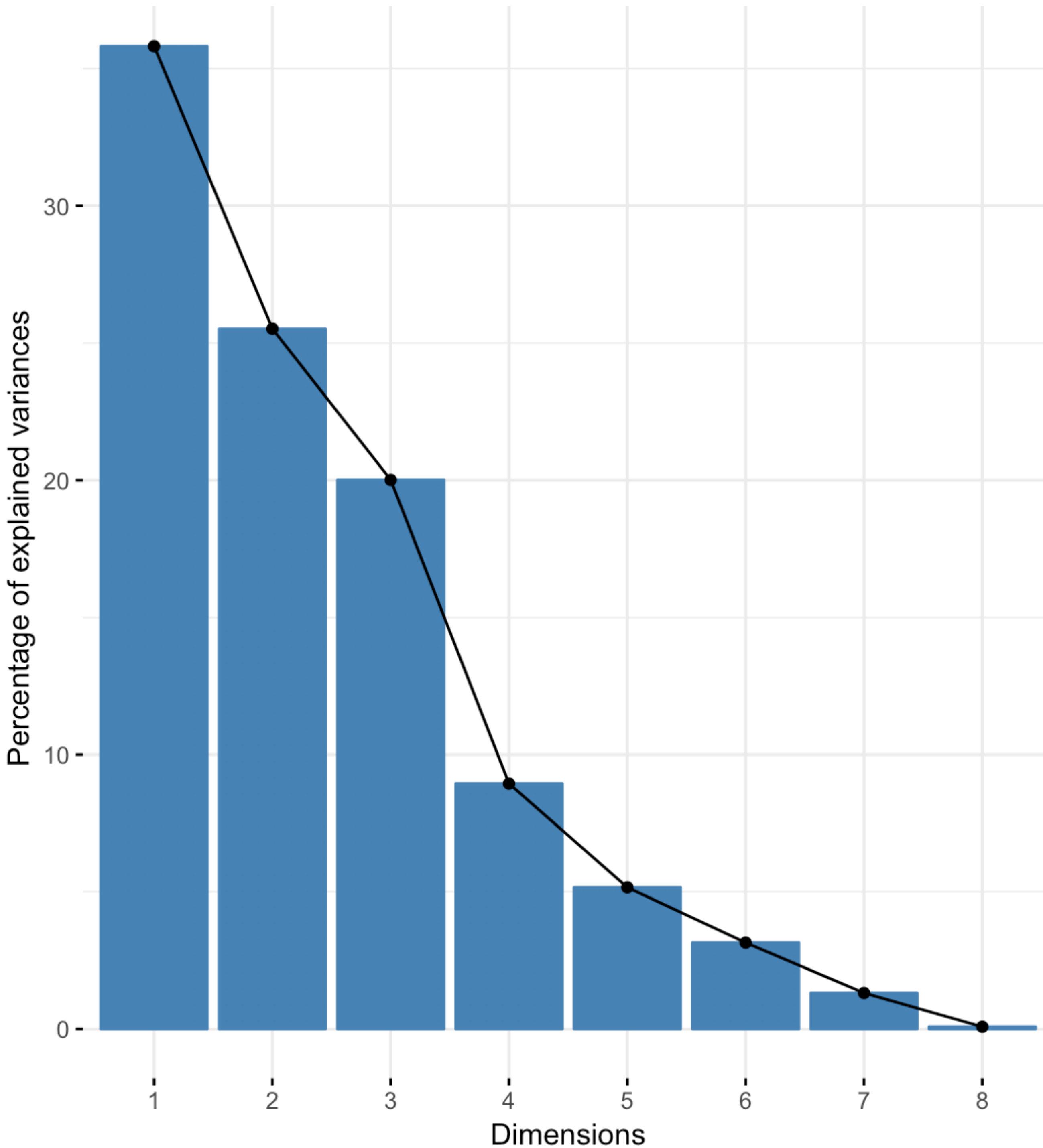
Long	Lat	avg_prec	avg_ET	avg_VI	avg_Alt	range_Alt	avg_temp	seas_temp	seas_prec
-70.5	-55.344	89.49	222.167	421.958	370.806	2160	38.501	13.97282505	258.8423462
-69.5	-55.344	68.95	241.5	482.354	472.088	2470	32.631	15.6611433	282.420105
-68.5	-55.344	50.23	229	599.458	348.21	1258	38.392	16.68809319	316.756958
-67.5	-55.344	37.9	222.333	623.583	222.572	1047	44.807	22.50125504	323.7763367
-66.5	-55.344	38.94	170.167	498.25	176.965	833	45.774	23.47930336	300.9299011
-70.5	-54.046	47.71	222.167	421.958	174.06	763	54.352	16.45620728	315.1140137
-69.5	-54.046	36.37	241.5	482.354	186.163	786	53.772	17.72191429	331.6278076
-68.5	-54.046	29.06	229	599.458	83.993	342	56.596	20.50442696	354.6487122
-67.5	-54.046	28.69	222.333	623.583	42.762	224	53.089	22.89826965	380.9208679
-73.5	-52.788	149.71	447.167	571.81	287.777	1590	54.122	11.40344238	291.9087524
-72.5	-52.788	54.23	415.333	778.905	267.908	1190	53.38	14.44744396	341.0378723
-71.5	-52.788	26.4	315	867.833	214.992	691	57.626	14.78347588	374.8889771
-70.5	-52.788	18.47	285.167	742.786	148.188	355	61.686	20.54120636	392.6990967
-69.5	-52.788	21.24	158.833	697.405	69.194	263	65.235	23.69093513	386.2383728
-73.5	-51.564	108.16	447.167	562.875	720.814	2785	40.757	11.40344238	291.9087524
-72.5	-51.564	33.43	415.333	761.104	605.359	2000	46.611	14.44744396	341.0378723
-71.5	-51.564	15.04	315	655.229	416.289	861	57.116	14.78347588	374.8889771
-70.5	-51.564	13.15	285.167	622.688	247.667	418	66.142	20.54120636	392.6990967
-69.5	-51.564	15.56	158.833	607.313	161.634	367	69.428	23.69093513	386.2383728
-74.5	-50.373	266.93	512.833	540.292	427.273	2164	60.755	12.20302963	278.6205139
-73.5	-50.373	108.13	353.333	562.875	1127.405	3405	29.841	14.35677052	314.4966431
-72.5	-50.373	31.43	296.333	761.104	568.539	1756	61.505	25.55801201	418.8746643
-71.5	-50.373	13.61	275	655.229	500.653	1080	65.597	26.75129128	476.8866882
-70.5	-50.373	11.79	207.833	622.688	340.328	775	74.516	22.04994202	503.4150391
-69.5	-50.373	13.93	198.167	607.313	171.078	549	83.211	20.19462204	514.8027954
-68.5	-50.373	16.74	134.333	448.188	96.374	463	87.676	22.06239319	502.4994507
-74.5	-49.21	252.75	512.833	547.905	383.44	1567	65.613	8.876086235	275.4307251
-73.5	-49.21	110.05	353.333	497.929	1054.021	3435	36.699	12.14703465	320.0220642
-72.5	-49.21	39.53	296.333	589.905	931.375	1961	48.015	20.98670387	388.9425964
-71.5	-49.21	16.23	275	582.929	768.218	1447	61.56	27.47363091	454.5921631

14909 geographic cells (110Km by 110Km)

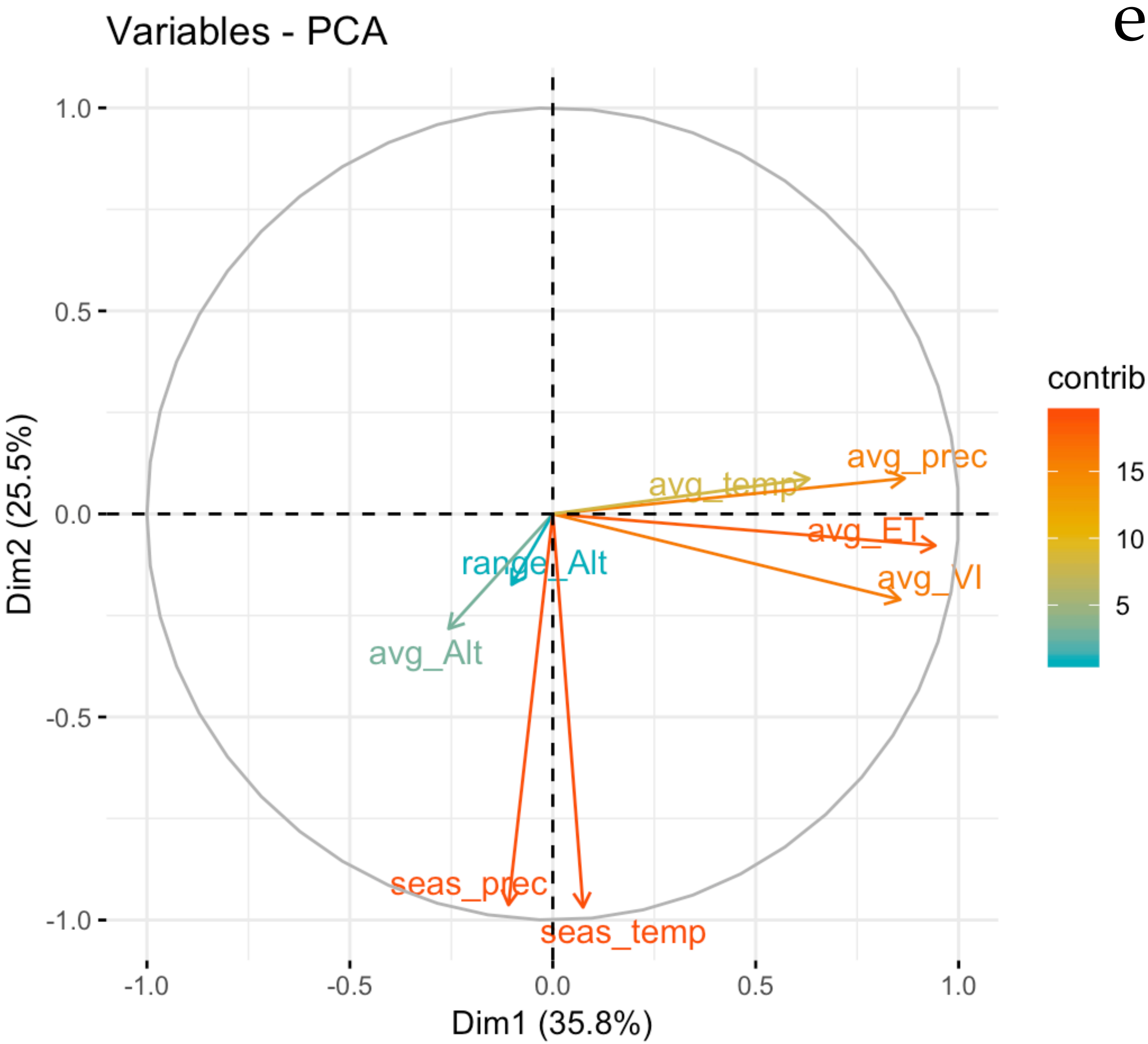
Long	Lat	avg_prec	avg_ET	avg_VI	avg_Alt	range_Alt	avg_temp	seas_temp	seas_prec
-70.5	-55.344	89.49	222.167	421.958	370.806	2160	38.501	13.97282505	258.8423462
-69.5	-55.344	68.95	241.5	482.354	472.088	2470	32.631	15.6611433	282.420105
-68.5	-55.344	50.23	229	599.458	348.21	1258	38.392	16.68809319	316.756958
-67.5	-55.344	37.9	222.333	623.583	222.572	1047	44.807	22.50125504	323.7763367
-66.5	-55.344	38.94	170.167	498.25	176.965	833	45.774	23.47930336	300.9299011
-70.5	-54.046	47.71	222.167	421.958	174.06	763	54.352	16.45620728	315.1140137
-69.5	-54.046	36.37	241.5	482.354	186.163	786	53.772	17.72191429	331.6278076
-68.5	-54.046	29.06	229	599.458	83.993	342	56.596	20.50442696	354.6487122
-67.5	-54.046	28.69	222.333	623.583	42.762	224	53.089	22.89826965	380.9208679
-73.5	-52.788	149.71	447.167	571.81	287.777	1590	54.122	11.40344238	291.9087524
-72.5	-52.788	54.23	415.333	778.905	267.908	1190	53.38	14.44744396	341.0378723
-71.5	-52.788	26.4	315	867.833	214.992	691	57.626	14.78347588	374.8889771
-70.5	-52.788	18.47	285.167	742.786	148.188	355	61.686	20.54120636	392.6990967
-69.5	-52.788	21.24	158.833	697.405	69.194	263	65.235	23.69093513	386.2383728
-73.5	-51.564	108.16	447.167	562.875	720.814	2785	40.757	11.40344238	291.9087524
-72.5	-51.564	33.43	415.333	761.104	605.359	2000	46.611	14.44744396	341.0378723
-71.5	-51.564	15.04	315	655.229	416.289	861	57.116	14.78347588	374.8889771
-70.5	-51.564	13.15	285.167	622.688	247.667	418	66.142	20.54120636	392.6990967
-69.5	-51.564	15.56	158.833	607.313	161.634	367	69.428	23.69093513	386.2383728
-74.5	-50.373	266.93	512.833	540.292	427.273	2164	60.755	12.20302963	278.6205139
-73.5	-50.373	108.13	353.333	562.875	1127.405	3405	29.841	14.35677052	314.4966431
-72.5	-50.373	31.43	296.333	761.104	568.539	1756	61.505	25.55801201	418.8746643
-71.5	-50.373	13.61	275	655.229	500.653	1080	65.597	26.75129128	476.8866882
-70.5	-50.373	11.79	207.833	622.688	340.328	775	74.516	22.04994202	503.4150391
-69.5	-50.373	13.93	198.167	607.313	171.078	549	83.211	20.19462204	514.8027954
-68.5	-50.373	16.74	134.333	448.188	96.374	463	87.676	22.06239319	502.4994507
-74.5	-49.21	252.75	512.833	547.905	383.44	1567	65.613	8.876086235	275.4307251
-73.5	-49.21	110.05	353.333	497.929	1054.021	3435	36.699	12.14703465	320.0220642
-72.5	-49.21	39.53	296.333	589.905	931.375	1961	48.015	20.98670387	388.9425964
-71.5	-49.21	16.23	275	582.929	768.218	1447	61.56	27.47363091	454.5921631

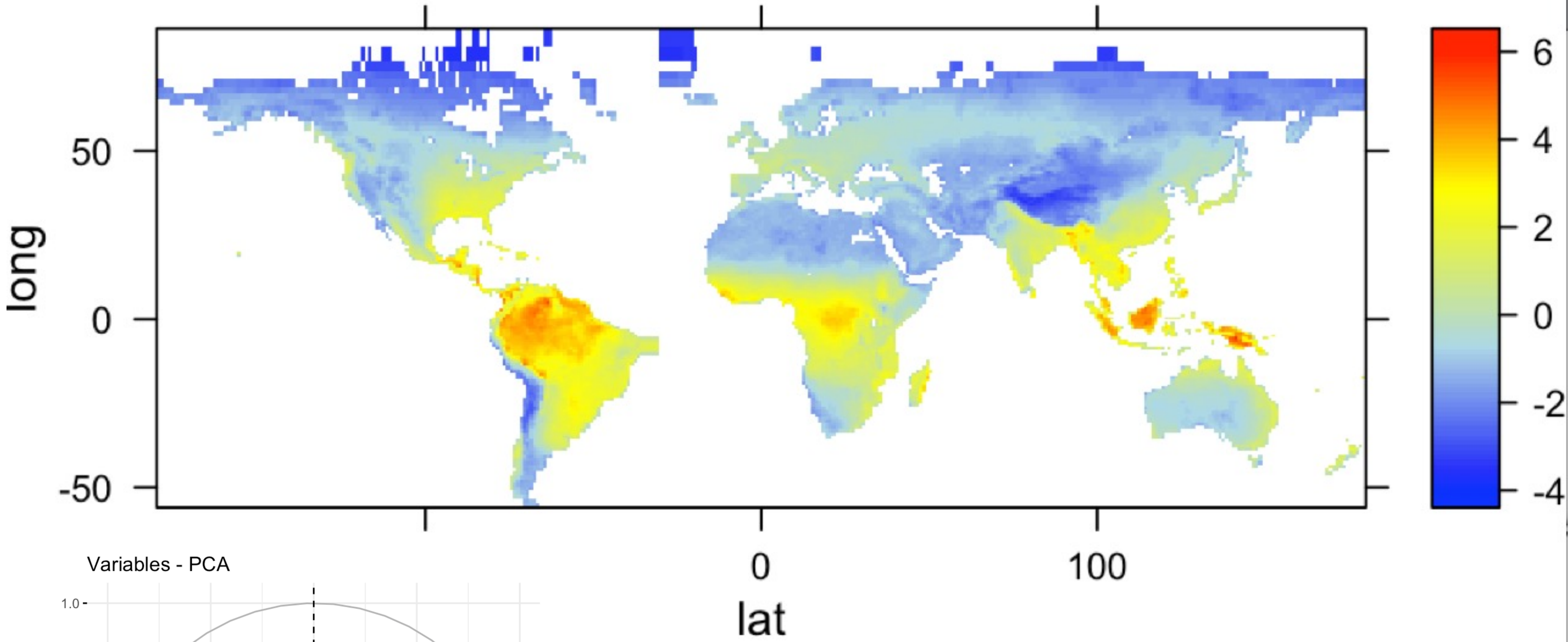
1) Latitude (Lat) and Longitude (Long) at the centre of geographic cell.
2) Average precipitation (last 40 years; avg_prec)
3) Average actual evapotranspiration (avg_ET, a proxy of productivity)
4) Average vegetation index (avg_VI)
5) Mean altitude (avg_Alt)
6) Maximum altitude minus minimum altitude (altitudinal range; range_Alt)
7) Average temperature (avg_temp)
8) Seasonal temperature (annual range in temperature; seas_temp)
9) Seasonal precipitation (annual range in precipitation; seas_prec)

Eigenvalue contribution – the traditional scree plot

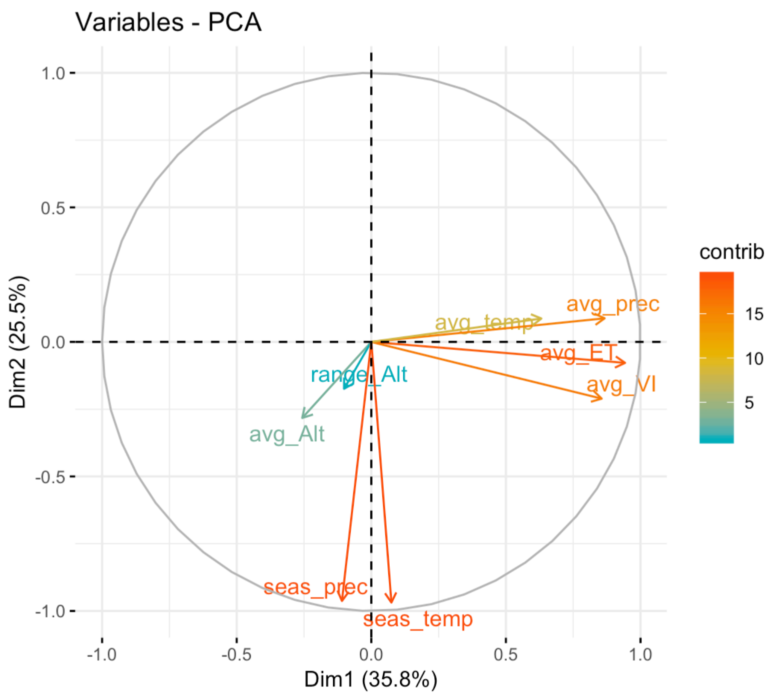


eigenvector structure



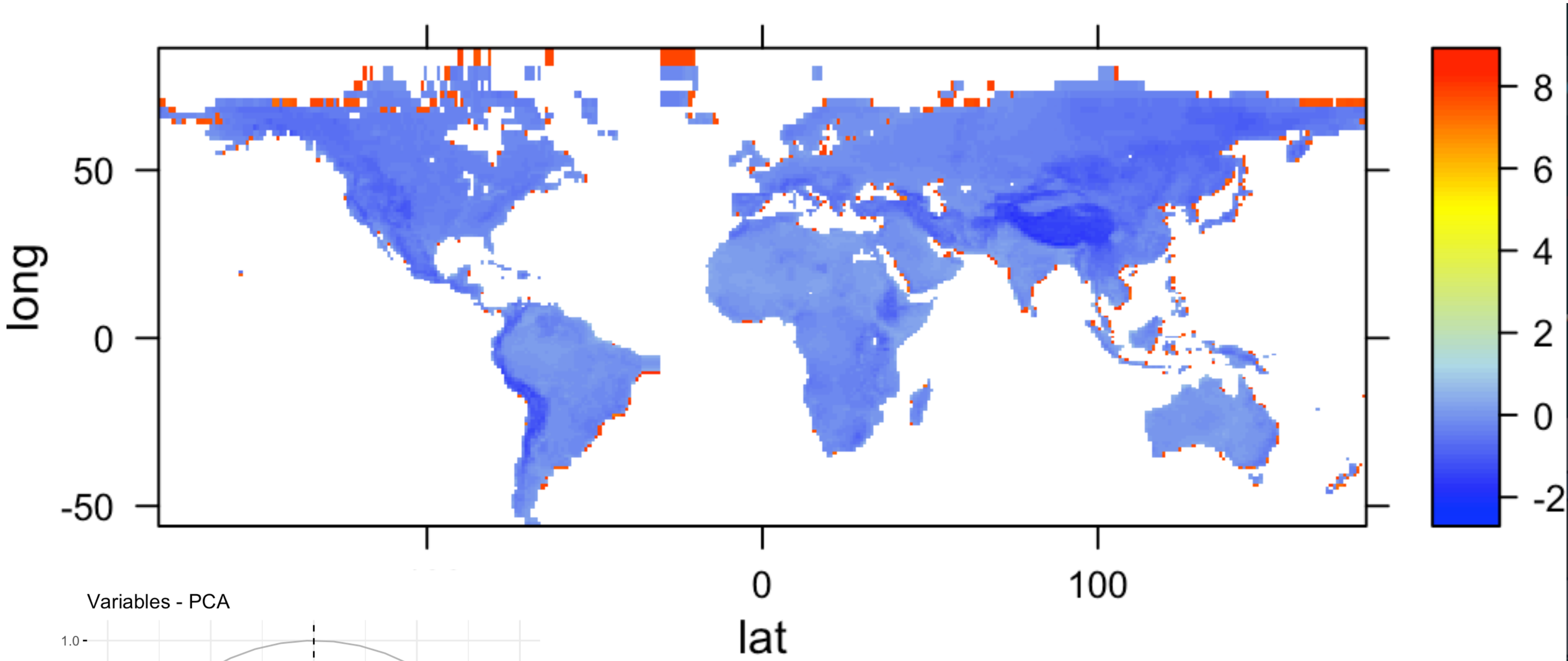


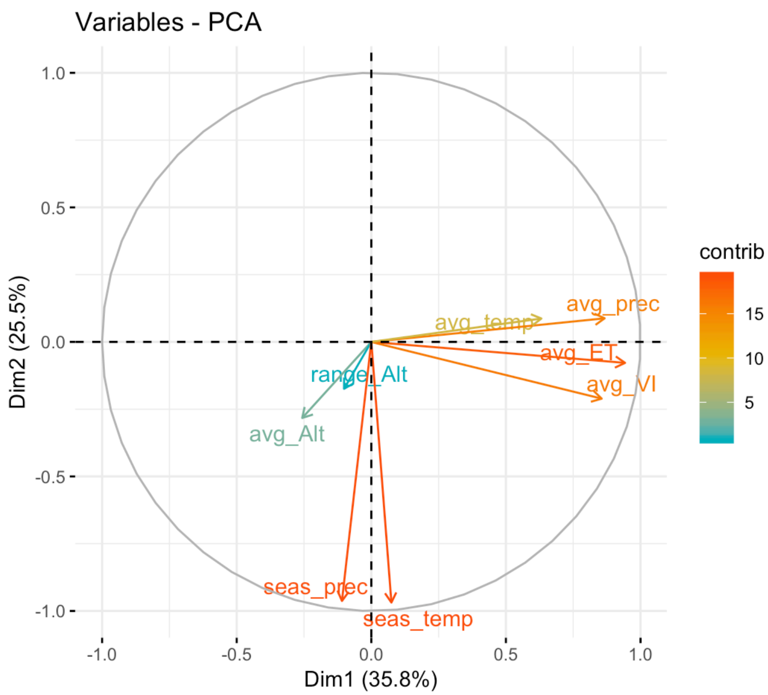
long
lat
50
0
-50
0
100
6
4
2
0
-2
-4



Variables - PCA
Dim2 (25.5%)
Dim1 (35.8%)
avg_prec
avg_temp
avg_ET
avg_VI
range_Alt
avg_Alt
seas_prec
seas_temp
contrib
15
10
5
1.0
0.5
0.0
-0.5
-1.0

Map of PC-1


50
0
-50
long
0
100
lat
8
6
4
2
0
-2


Variables - PCA
1.0
0.5
0.0
-0.5
-1.0
Dim2 (25.5%)
-1.0
-0.5
0.0
0.5
1.0
Dim1 (35.8%)
avg_prec
avg_temp
avg_ET
avg_VI
range_Alt
avg_Alt
seas_prec
seas_temp
contrib
15
10
5

Map of PC-2