

Program SPHERIKM for Spherical k-means clustering

by Mark Hill, CEH Fellow affiliated to **Biological Records Centre (BRC)** moh@ceh.ac.uk

Version 1, 25 January 2013 Copyright © Mark Hill 2013 All rights reserved

This version is issued free for the use of persons downloading it from the BRC website www.brc.ac.uk. It is not guaranteed to be error-free, nor does the author guarantee support for users. It is supplied as source code in Fortran 77 and as an executable file that can run under Windows XP or Windows 7. All rights are reserved. The program may not be redistributed, rewritten or sold to third parties without permission of the author.

Contents

1 Data input format	1
2 Analysis options	2
3 Interpretation of results file m.out	3
3.1 Simple listing of key species, ordered by correspondence analysis	3
3.2 Enumeration of cluster members in descending cosines for each cluster	3
3.3 Proximities to clusters - species	3
3.4 Cluster hierarchy	3
3.5 Sample classification	4
3.6 Ordering species and samples according to the hierarchy	4
3.7 Representation as a biclustering	5
3.8 Cluster coincidence statistics	5
3.9 Summary statistics	6
4 Technical parameters affecting only the algorithm	6
5 Statistics of performance within loops	8
5.1 Number of swapping cycles	8
5.2 Cluster hierarchy	8
6 References	8

1 Data input format

Occurrences of species in samples are represented as triplets. There are no headers. The following points should be noted.

1. The order within each triplet is: **sample species quantity**
2. The order of the triplets is irrelevant (as if they were being submitted to a database); but within the triplets, the order **sample species quantity** must be retained.
3. Samples and species are represented as a string of 10 or fewer characters with no intervening blanks. Here, *Achillea millefolium* is represented as Achi_mill with 9 characters and no blanks.
4. There is no other formatting. The number of blanks separating elements within a triplet does not have to be 1. Any number of blanks is acceptable, subject to the constraint that the total number of characters on a line, including blanks, does not exceed 80 (i.e. the number of characters on a traditional IBM data card).
5. Presence-only data should be entered as couplets: **sample species** rather than as triplets. Do not mix presence-only data with quantitative data.

The Danube-valley meadow dataset (Mueller-Dombois & Ellenberg, 1974) is used as the main example in this manual, with an explanation of the output for the CSKM (chord spherical k-means) analysis with 6 species clusters and 8 sample clusters. Values are % biomass.

```
Danube01 Achi_mill      6
Danube01 Arrh_elat    0.1
Danube01 Briz_medi     1
Danube01 Brom_erec    50
Danube01 Camp_glom    0.1
Danube01 Camp_rotu     1
...
Danube25 Tara_offi    0.1
Danube25 Trif_repe    0.1
Danube25 Vero_cham     1
```

The example of a presence-only dataset is for liverworts in hectads (10-km squares) of eastern England. The first row specifies that the floating aquatic *Ricciocarpos natans* was found in hectad TL44, just south of Cambridge.

```
TL44 Ricc_nata
TL44 Ricc_cave
TL45 Loph_bide
TL45 Caly_muel
TL45 Chil_pall
TL45 Cono_coni
TL45 Frul_dila
```

2 Analysis options

Three parameters do not have a default value and always have to be entered:

kmean – analysis parameter, 1 for chord SKM (CSKM), 2 for perpendicular SKM (PSKM)

k₁ – number of species clusters

k₂ – number of sample clusters

Two other parameters, the exponents for weighting (p_1 , p_2), are critical to how the analysis performs. By default they are set to (1.0, 1.0). When both p_1 and p_2 are 1, each vector to be clustered is weighted by its length. Let \mathbf{a}_j be the vector representing species j and let $\|\mathbf{a}_j\|$ be its length. The vector projected on the unit hypersphere is $\mathbf{a}_j / \|\mathbf{a}_j\|$ and is given weight $\|\mathbf{a}_j\|^{p_1}$. For unweighted spherical k-means (not recommended) weights (p_1 , p_2) are (0, 0). The significance of these weights is explained in Hill, Harrower & Preston (2013).

Log file for SPHERIKM

```
Input limits:
Max number of samples = 10000
Max number of species = 10000
Max length of data array = 2000000
Max number of clusters = 99

Type 1 for CSKM (chord), 2 for PSKM (perpendicular)
1
p1,p2 - exponents for weighting species and sample clusterings
1.000 1.000
Type name of input file [sample species quant] ...
m.txt
Type name of output file ...
mc68.out
Type name of log file ...
mc68.log
Type number of clusters k1 for species clustering...
6
Type number of clusters k2 for sample clustering
(if 0 or 1 then samples are not clustered)...
```

3 Interpretation of results file m.out

3.1 Simple listing of key species, ordered by correspondence analysis

Data input file... m.txt

SPHERIKM CHORD - SPECIES CLASSIFICATION - Exponent for weighting = 1.000

```
Key species in ordination order  1 Brom_erec
Key species in ordination order  2 Poa_prat
Key species in ordination order  3 Gali_moll
Key species in ordination order  4 Arrh_elat
Key species in ordination order  5 Geum_riva
Key species in ordination order  6 Phal_arun
```

3.2 Enumeration of cluster members in descending cosines for each cluster

Headers are abbreviated thus: **Spno** – alphabetic serial number of species; **Sp_count** – count of occurrences of species; **Species_tot** – sum of matrix elements for species; **Spec_wgt** – weight for species; **Species** – name (or code) of species; **Clus** – cluster number as in 3.1; **Key_spec** – key species as in 3.1; **Cosine** – cosine angle between species and cluster centre; **Cl_Cosine** – weighted average cosine for species in cluster; **Mean_cos** – weighted average cosine of all species to cluster centres, averaged over all clusters. (NB with CSKM, the chord option of SKM, these averages are ordinary weighted means, whereas with PSKM they are weighted root-mean-squares.)

Spno_	Sp_count	Species_tot	Spec_wgt_	Species____	_Clus	Key_spec__	Cosine	Cl_Cosine	Mean_cos
11	7	274.000	115.412	Brom_erec	1	Brom_erec	0.9872	0.87151	0.798503
47	4	11.000	5.568	Koel_pyra	1	Brom_erec	0.8414	0.87151	0.798503
52	5	0.500	0.224	Linu_cath	1	Brom_erec	0.7966	0.87151	0.798503
33	15	41.400	17.636	Fest_rubr	1	Brom_erec	0.7507	0.87151	0.798503

3.3 Proximities to clusters - species

Here, as in Section 3.2, species are ordered first by cluster, then by cosine (alignment) to cluster centres, in descending order of alignment.

```
PROXIMITIES TO CLUSTERS - species
Species____  1    2    3    4    5    6
Brom_erec  0.987 0.227 0.243 0.175 0.019 0.001
Koel_pyra  0.841 0.151 0.266 0.114 0.022 0.001
Linu_cath  0.797 0.192 0.250 0.239 0.032 0.001
Fest_rubr  0.751 0.222 0.295 0.251 0.101 0.007
```

3.4 Cluster hierarchy

The cluster hierarchy is easiest to understand when shown as a dendrogram. The dendrogram here was produced by copying the hierarchy in Newick format into Dendroscope (Huson *et al.*, 2007), which is a program available for free download over the internet. The heights of the nodes in the dendrogram represent the increase in mean square within-group dispersal (MS_disp) from the case where the original 6 clusters are kept separate. This can be read off from **CLUSTER HIERARCHY - Species**, which shows how the calculation is made. At the bottom of the hierarchy (row 6), all the clusters are separate. At the next row up (row 5), clusters 3 and 5 have been amalgamated and cluster 5 has been

included in cluster 3. This process is repeated upwards until at row 1, the most distinct cluster (*Bromus erectus*) is united with the rest.

The heights of the nodes in the dendrogram are got by subtracting the value of MS_disp in the bottom row (here row 6) from MS_disp for the node in question. Thus the amalgamation of the *Bromus erectus* cluster with the rest raises MS_disp to 0.926004. The starting value was 0.402995 (row 6), and the value that appears in the Newick formula is

(Brom_erec: 0.523009,

which signifies that the height of this union in the dendrogram is 0.523009, got by subtracting 0.402995 from 0.926004.

The clustering method is Ward's method, which at each level selects the amalgamation that minimally increases the average within-group mean-square dispersion. In this example, the *Phalaris arundinacea* cluster is more distinct in composition from the other clusters than the *Bromus erectus* cluster, but it is based on only one sample. The *Bromus erectus* cluster is based on 6 samples, so has much larger weight. Ward's method was chosen because SKM clustering is based on minimum variance, so is appropriately followed by a hierarchical clustering method based on the same criterion.

CLUSTER HIERARCHY - Species Note: MS_disp = 2*(1-Mean_cos) (CSKM case) or = 1-Mean_cos**2 (PSKM case)

Num	C1	C2	C1_N	C2_N	MS_disp	Change	1	2	3	4	5	6
1	1	0	94	0	0.926004	0.000000	1	1	1	1	1	1
2	1	2	16	78	0.733346	0.192657	1	2	2	2	2	2
3	2	5	53	25	0.627063	0.106283	1	2	2	2	5	5
4	2	3	8	45	0.539268	0.087795	1	2	3	3	5	5
5	5	6	17	8	0.458607	0.080661	1	2	3	3	5	6
6	3	4	22	23	0.402995	0.055613	1	2	3	4	5	6

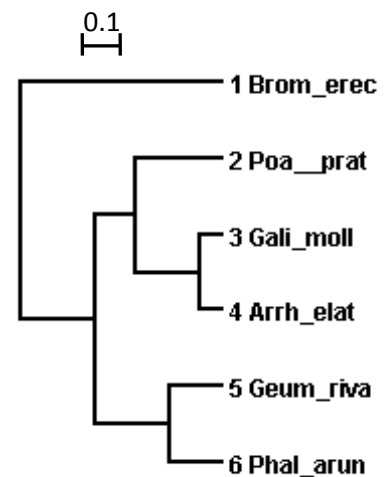
HIERARCHY IN NEWICK PHYLOGRAM FORMAT

```
(Brom_erec:0.523009,
((Poa_prat:0.224069,
(Gali_moll:0.055613,
Arrh_elat:0.055613):0.168456):0.106283,
(Geum_riva:0.136274,
Phal_arun:0.136274):0.194078):0.192657);
```

*** END OF SPECIES CLASSIFICATION ***

3.5 Sample classification

This uses equivalent headings to those of species classification.



3.6 Ordering species and samples according to the hierarchy

The hierarchy may define an ordering of the clusters other than that shown in Section 1 above. Clusters 3 and 4 are amalgamated first, so that cluster 4 (*Arrhenatherum elatius*) should come immediately before or after cluster 3 in the dendrogram. Here, the ordination order is retained, but often it is not. If you want to display the matrix in order, it is probably best to use the order of the hierarchy. To help in doing this, the species and samples are additionally printed out in the order of the hierarchy.

SPECIES IN ORDER OF HIERARCHY

```
Brom_erec      1
Koel_pyra      1
...
Rume_cris      6
Lath_prat      6
```

SAMPLES IN ORDER OF HIERARCHY

```
Danube01       1
Danube04       1
...
Danube11       7
Danube07       7
```

3.7 Representation as a biclustering

The remaining tables show the biclustering. **CLUSTER TOTALS** shows the sum of matrix elements in each species and sample cluster. The sum of values in **CLUSTER TOTALS** is equal to sum of matrix elements. In the example, the largest total, 560.4, is for species cluster 4 (the *Arrhenatherum elatius* cluster) and sample cluster 6. **CLUSTER MEANS** shows the same information, but with the values divided by the number of contributing samples. In this case the values are approximately percent biomass, so that the interpretation is that sample cluster 8 (comprising the single sample Danube14) has 53% of its biomass in species cluster 6 and 40% of its biomass in species cluster 4.

CLUSTER TOTALS

Rows are species clusters, Columns are sample clusters

Ordering is that used to display cluster hierarchy

Spclus	1	2	3	4	6	8	5	7
1	220.8	130.7	2.1	21.5	11.9	0.1	4.0	2.2
2	20.1	33.3	78.1	66.3	55.0	1.0	9.1	13.1
3	22.1	61.3	4.4	56.5	102.1	2.3	132.1	45.8
4	35.4	75.4	14.2	145.8	560.4	39.6	98.2	76.1
5	3.4	2.5	2.2	13.7	76.5	6.1	58.0	165.7
6	0.1	0.0	0.0	0.1	1.3	53.4	1.1	0.3

CLUSTER MEANS

Rows are species clusters, Columns are sample clusters

Ordering is that used to display cluster hierarchy

Spclus	1	2	3	4	6	8	5	7
1	73.6	43.6	2.1	7.2	1.5	0.1	1.3	0.7
2	6.7	11.1	78.1	22.1	6.9	1.0	3.0	4.4
3	7.4	20.4	4.4	18.8	12.8	2.3	44.0	15.3
4	11.8	25.1	14.2	48.6	70.1	39.6	32.7	25.4
5	1.1	0.8	2.2	4.6	9.6	6.1	19.3	55.2
6	0.0	0.0	0.0	0.0	0.2	53.4	0.4	0.1

3.8 Cluster coincidence statistics

Cluster coincidence statistics compare the values in **CLUSTER TOTALS** with those that would be expected if the species occurred at random in the samples. The highest values are 23.37 for the *Phalaris arundinacea* cluster in sample cluster 8, and 4.70 for the *Bromus erectus* cluster in sample cluster 1. The interpretation is that the biomass of *Phalaris arundinacea* species (chiefly *Glyceria fluitans* and *Phalaris arundinacea*) exceeds the biomass that would be expected from random occurrence by a factor of 23.37, and the *Bromus erectus* species

(chiefly *Bromus erectus*, *Koeleria pyramidata* and *Festuca rubra*) exceed their expectation in sample cluster 1 by a factor of 4.70.

CLUSTER COINCIDENCE SUMMARY

Rows are species clusters, Columns are sample clusters

Ordering is that used to display cluster hierarchy

Spclus	1	2	3	4	6	8	5	7
1	4.70	2.77	0.13	0.45	0.09	0.01	0.08	0.05
2	0.61	1.00	7.08	2.00	0.62	0.09	0.28	0.40
3	0.43	1.20	0.26	1.10	0.75	0.13	2.59	0.89
4	0.28	0.60	0.34	1.16	1.68	0.93	0.78	0.61
5	0.09	0.06	0.17	0.35	0.73	0.46	1.48	4.21
6	0.01	0.00	0.00	0.01	0.07	23.37	0.16	0.04

3.9 Summary statistics

The equivalent number N_2 is calculated as

$$N_2 = (\sum \sum a_{ij})^2 / \sum \sum a_{ij}^2$$

summed over all matrix elements a_{ij} . N_2 is a measure of the apparent number of elements, taking account of the fact that they may vary greatly in size; if the elements are all 1 or 0 then N_2 is simply the number of non-zero elements.

The geometric mean concentration ratio is calculated as the geometric mean of values in **CLUSTER COINCIDENCE SUMMARY**, weighted by the values in **CLUSTER TOTALS**. If the elements in the original data matrix are presences and absences, this value is simply $\exp(\text{Entropy of biclustering})$. Where, as here, the matrix elements are quantities, the concentration ratio has an analogous interpretation and is also basically an entropy value – i.e. a measure of the ‘lumpiness’ of **CLUSTER TOTALS**.

Finally, AIC is the quasi-Akaike information criterion, which can be used to select the number of row and column clusters. The solution with the minimum AIC should be the one with most information. Solutions with more numerous clusters are over-fitted. (For a fuller explanation, refer to Hill, Harrower & Preston, 2013).

Number of clusters k1,k2	6	8
Number of species n and samples m	94	25
Number of parameters for AIC	153	- calculated as (k1-1)*(k2-1)+m+n-1
Total sum of matrix elements	2525.408	
Number of non-zero items	788	
Equivalent number N_2	141.3	
Geometric mean concentration ratio	1.56470	
Akaike information criterion (N_2)	430.8	

4 Technical parameters affecting only the algorithm

You can select these values yourself, but you are recommended to use the default values unless you suspect that you are not in fact reaching the optimum (or, for large problems, a sufficiently good solution) or if you find that 10 loops take intolerably long.

Randomizations *irmax*

The default value of *irmax*, the number of randomizations in the inner loop, is 50. This means, for example, that if k seeds are to be selected, then the process of selecting them

requires *imax* randomizations to select each seed. However, as explained below, the actual number of randomizations may be substantially larger, because solutions of insufficient quality are rejected. In any cycle of randomizations, *imax* is the number of good-quality solutions that are required.

Shortlist factor *over*

This parameter defines the number of shortlisted species from which the seeds are to be selected. The default value of *over* is 3.00, so that if *k* clusters are required, then by default 3*k* species are shortlisted.

Quality relaxation factor *degrad*

In each cycle of *imax* randomizations, only those solutions whose mean-square cosine equals or exceeds the average for the previous cycle are selected as of high enough quality to be worth retaining. Lower-quality solutions are rejected. This criterion may, however, set the bar too high, resulting in an inordinate number of rejections and making the algorithm inefficient. Each time that *imax* randomizations are performed, the bar is – so to speak – lowered by multiplying by *degrad*, whose default value is 0.999.

Seed number reduction factor *downjt*

To make the shortlist of species from which to select seeds, the algorithm starts with the full list and then gradually whittles the number down, rejecting solutions of insufficient quality as explained above. By default, *downjt* = 0.80. In the example given here, there are initially 94 species to consider. However, in the first 50 randomizations, only 31 of these ever qualify as key species. From this pool of 31 species, 24 are selected for the next cycle, based on 50 good-quality solutions and 49 solutions that are rejected. The numbers are then whittled down to 19 (rejecting 241 solutions) and finally 17 (rejecting 368 solutions). The target length of the shortlist was 18 (i.e. *over* × *k* = 3.0 × 6), but by chance, two of the 19 previous candidates were never selected as key species at this round.

Number of loops *nouter*

The entire process of selecting *k* seed species is repeated *nouter* times, in an outer loop, and the best of the solutions is selected. With small datasets such as the Danube meadow data, each one of the inner solutions is identical. With larger datasets there is more variation, and the *nouter* could in principle be increased beyond its default value of 10.

Random start number *irinit*

The pseudo-random number generator is normally initiated by the value 0. If a different randomization is required, then *irinit* (standing for Integer Random-number INITiator) can be given a different value. This will allow you to find out whether the number of loops *nouter* is in fact adequate to reach an optimum.

```
PARAMETERS
Exponent weightings  1.000  1.000
Clusters             6    8
Randomizations       50
Shortlist factor     3.00
Quality relaxation factor 0.999000
Seed number reduction factor 0.80
Number of outer loops   10
Random start number    0
```

5 Statistics of performance within loops

The log file gives some statistics of performance within loops. In particular, it gives the following information.

5.1 Number of swapping cycles

When seed species have been selected by the sequential addition process, each selected seed is swapped with each unselected member of the shortlist. If a better solution is then found, the chosen but previously unselected species is substituted, and the best such solution is selected as the next trial solution. Each of these sequences of swapping is called a swapping cycle. For each loop, the number of swapping cycles before such substitutions no longer produce a better solution is reported here, along with the resulting average cosine (or for PSKM the mean-square cosine) and a list of key species.

```
Number of swapping cycles 0      Mean cosine 0.79850
Key species 1 Arrh_elat
Key species 2 Brom_erec
Key species 3 Chen_albu
Key species 4 Gali_moll
Key species 5 Geum_riva
Key species 6 Poa__prat
```

5.2 Cluster hierarchy

The cluster hierarchy is printed out in the log file as it is produced. In the example, the first amalgamation (Level 1) unites clusters 3 and 4 (Clustr1 and Clustr2) and the second (Level 2) unites clusters 5 and 6. The number of members in each amalgamated cluster is shown. Thus at Level 1, cluster 3 has 22 members and cluster 4 has 23 members, producing a new cluster with 45 members. This is in turn amalgamated at level 3. Remember that with CSKM, the mean-square dispersion MS_disp is $2 \times (1 - \text{Mean cosine})$. In this case, $0.402995 = 2 \times (1 - 0.79850)$.

```
CLUSTER HIERARCHY
Source_____ Level      Num Clustr1 Clustr2 N_Clus1 N_Clus_2 MS_disp_
From_k-means    0         6      0      0      94      0 0.402995
Amalgamation    1         5      3      4      22      23 0.458607
Amalgamation    2         4      5      6      17      8 0.539268
Amalgamation    3         3      2      3       8     45 0.627063
Amalgamation    4         2      2      5     53     25 0.733346
Amalgamation    5         1      1      2     16     78 0.926004
```

6 References

- Hill, M.O., Harrower, C.A. & Preston, C.D. (2013) Spherical k-means clustering is good for interpreting multivariate species occurrence data. *Methods in Ecology and Evolution* (in press).
- Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M. & Rupp, R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, **8**, 460. doi:10.1186/1471-2105-8-460.
- Mueller-Dombois, D. & Ellenberg, H. (1974) *Aims and methods of vegetation ecology*. John Wiley & Sons, New York.