環境統計学ぷらす

第6回
多変量解析

高木　俊
shun.takagi@sci.toho-u.ac.jp

2013/12/05

予定

- 第1回：　Rの基礎と仮説検定
- 第2回：　分散分析と回帰
- 第3回：　一般線形モデル・交互作用
- 第4.1回：一般化線形モデル
- 第4.2回：モデル選択
- 第5回：　一般化線形混合モデル
- 第6回：　多変量解析

今日やること

- 統計編
  - PCA・DCA
  - RDA・CCA
  - NMDS・クラスター分析

CAUTION

高木自身が今まで研究で使ってない（いじってみた程度）解析が多く含まれています。正直そんなに詳しくないです。
「どういう時に使うのか」＋「どうすれば解析できるか」を話しますが、正確でない表現があるかもしれません。

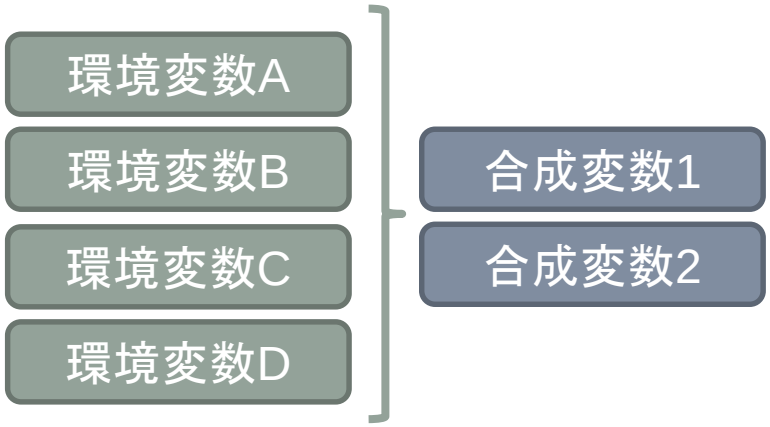
最後に参考文献を出しますので、詳しい部分を知りたい方は参照してください

多変量解析

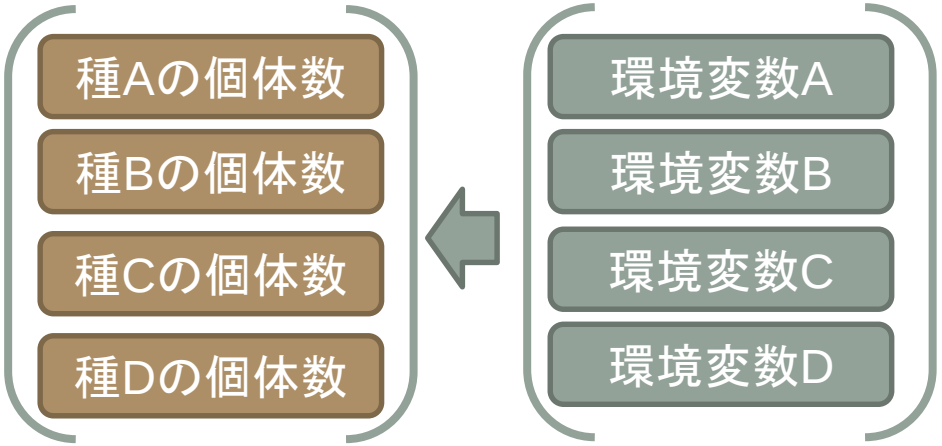
多変量解析って？

- 今まで紹介した統計解析

- 本章で扱う内容



多くの変数をまとめる



多くの変数同士の関係を解析する

今回の内容

- 主成分分析（PCA）・（除歪）対応分析（CA・DCA）

  → 似た変化パターンをする変数同士をまとめる

  （例：10項目の水質に関する変数を3つにまとめる）

- 冗長性分析（RDA）・正準対応分析（CCA）

  → 変数群の変化パターンで別の変数群の応答パターンを説明する

  （例：生息地の環境変数（5項目）で群集組成（10種の出現パターン）を説明する）

- クラスター分析・非計量多次元尺度法（NMDS）

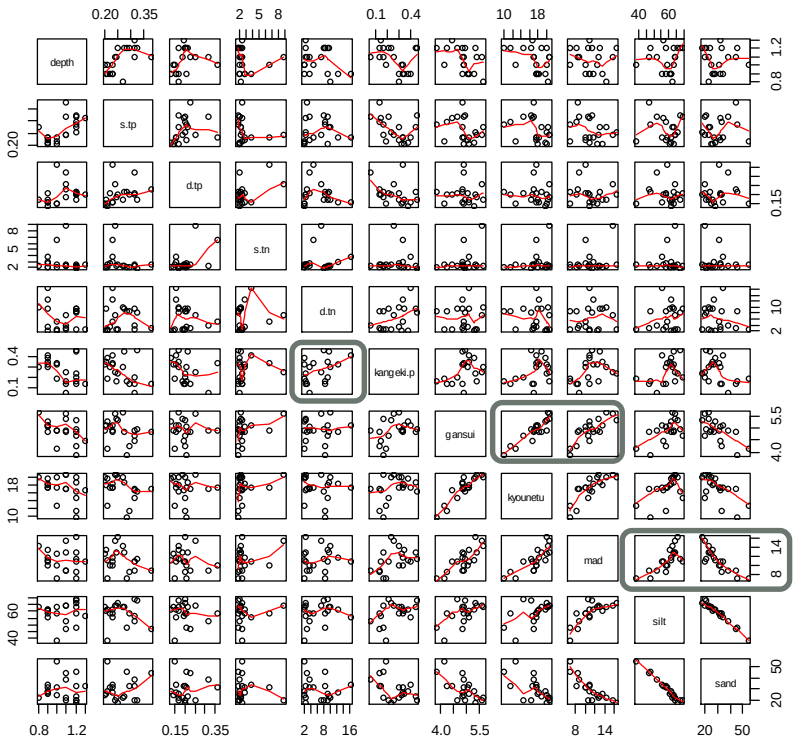
  → 変数群の変化パターンでサンプル同士をまとめる

  （例：群集組成（10種の各個体数）の似た生息地同士をグループにまとめる）

主成分分析（Principal Component Analysis）

- 印旛沼内20地点において11種類の環境項目を測定した

水深・表層TP・低層TP・表層TN・低層TN・土壌間隙水中P・
土壌含水率・強熱減量・粘土率・シルト率・砂率



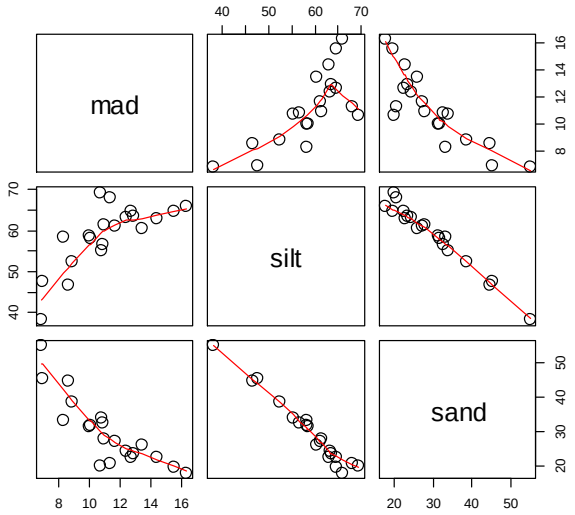
すべての変数は互いに独立とはいえない

間隙水Pは低層TNと正の相関？

含水率は強熱減量と正の相関？

粘土はシルトと正の相関、砂と負の相関？

データのもつばらつきをまとめる

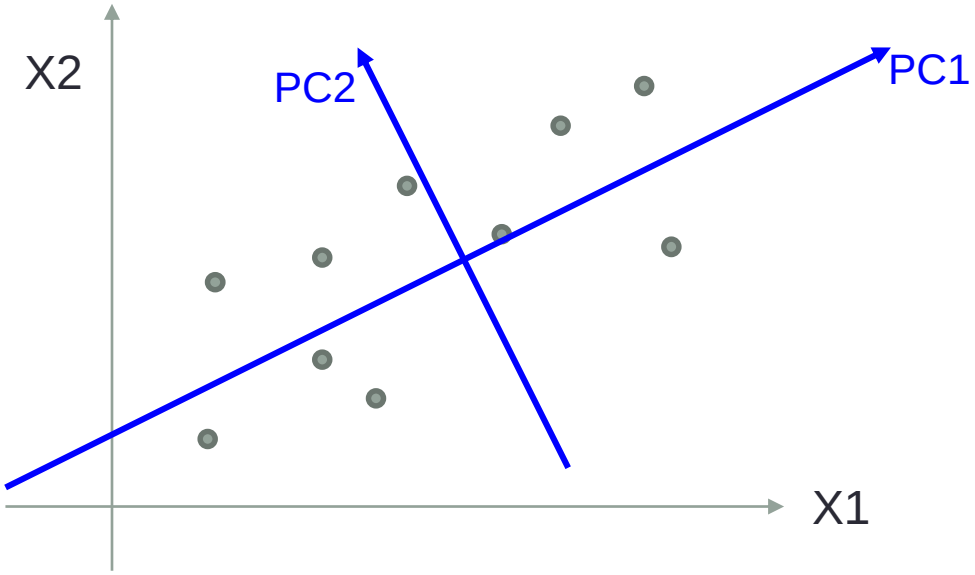


たとえば、

粘土率・シルト率・砂率の3変数は、個別に扱うのではなく、一つにまとめられそう

（イメージ）
新たに「泥率（粘土＋シルト）」という指標を作れば、この値が大きい時、粘土・シルト率が高く、砂率が低いことを1変数で表すことができそう！

主成分分析の原理



- 主成分分析では、データ全体がもつばらつきを最も説明するような軸を第一軸（第一主成分：PC1）とおく。
- さらに、その軸に直交する（独立な）軸のうち、残りのばらつきを最も説明するような軸を第二軸とおき、以下同様に軸を設定していく

パッケージveganのrdaによる解析

```
library(vegan)
pca6.1<- rda(data6.1[,-1],scale=T)
summary(pca6.1)

Call:
rda(X = data6.1[, -1], scale = T)

Partitioning of correlations:
              Inertia Proportion
Total              11          1
Unconstrained      11          1

Eigenvalues, and their contribution to the correlations

Importance of components:
                          PC1    PC2    PC3    PC4     PC5    PC6     PC7     PC8     PC9    PC10
Eigenvalue            4.6338 1.9230 1.5750 1.2143 0.63616 0.4158 0.32539 0.16773 0.06656 0.04224
Proportion Explained  0.4213 0.1748 0.1432 0.1104 0.05783 0.0378 0.02958 0.01525 0.00605 0.00384
Cumulative Proportion 0.4213 0.5961 0.7393 0.8497 0.90748 0.9453 0.97486 0.99011 0.99616 1.00000

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
* General scaling constant of scores:  3.802214
```

#prcompでもできます

（略）

固有値 Eigenvalue

```
Eigenvalues, and their contribution to the correlations

Importance of components:
                          PC1    PC2    PC3    PC4     PC5    PC6     PC7     PC8     PC9
Eigenvalue             4.6338 1.9230 1.5750 1.2143 0.63616 0.4158 0.32539 0.16773 0.06656
Proportion Explained   0.4213 0.1748 0.1432 0.1104 0.05783 0.0378 0.02958 0.01525 0.00605
Cumulative Proportion  0.4213 0.5961 0.7393 0.8497 0.90748 0.9453 0.97486 0.99011 0.99616
```

Eigenvalue: 固有値。元データのばらつきのうち、各主成分（PC1～PC10）が説明しうるばらつきを示す。これが1以上の主成分は考慮。
PC1の固有値は4.6なので、もとの変数4.6個分のバラ付きを要約したものになっている（イメージ）

Proportion Explained: 寄与率。説明されたばらつきを割合にしたもの。全部足すと1。

Cumulative Proportion: 累積寄与率。第x主成分までに説明されたばらつきの合計。
例えば第2主成分までで60%、第5主成分までで元データのばらつきの90%を考慮できる。

主成分のスコア

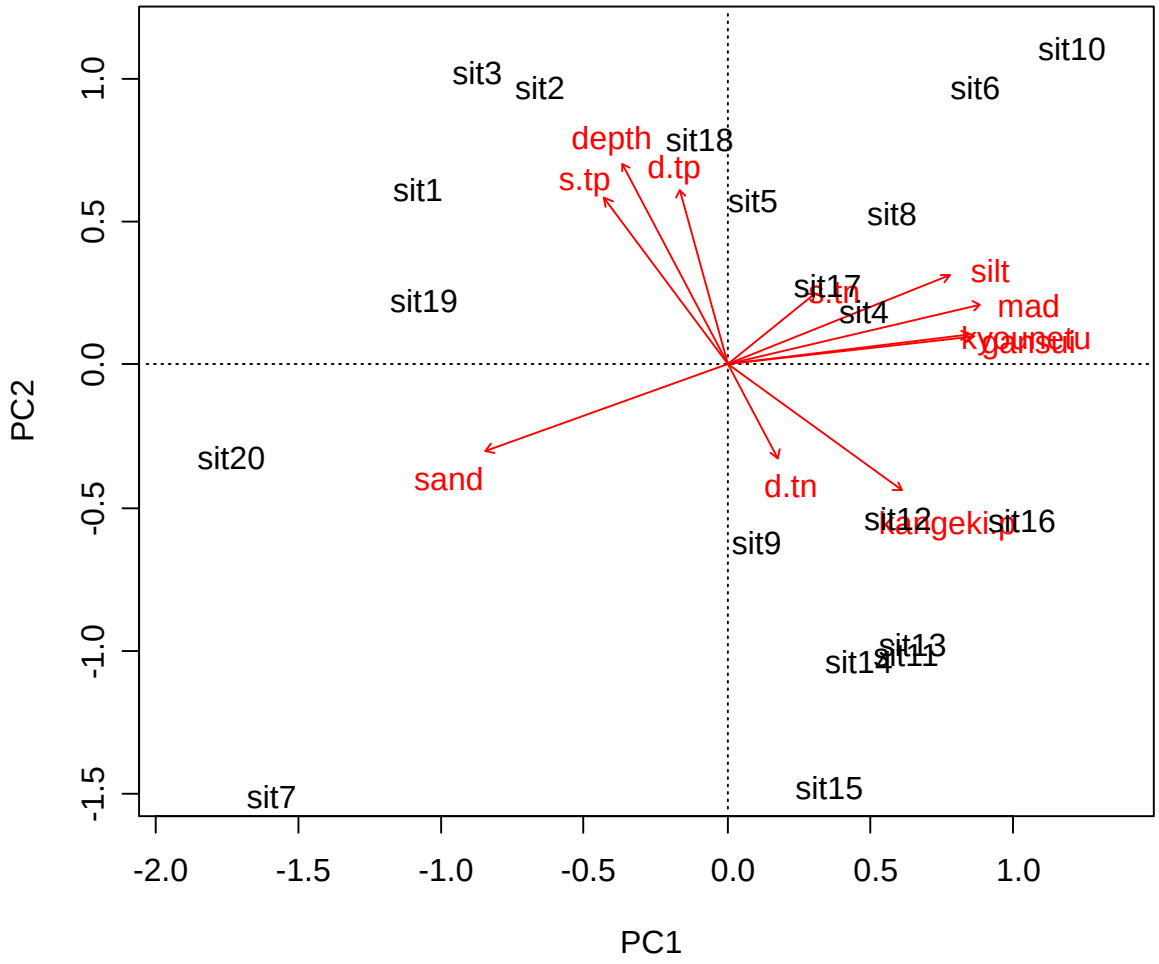
```
Species scores
              PC1     PC2      PC3      PC4       PC5       PC6
depth     -0.4322  0.8276 -0.46144 -0.04179  0.367868 -0.074925
s.tp      -0.5064  0.6862 -0.42616 -0.13240 -0.549990  0.149302
d.tp      -0.1995  0.7219  0.63045 -0.38619 -0.025876  0.320905
s.tn       0.3611  0.2959  0.78951 -0.53999  0.195501 -0.218947
d.tn       0.2106 -0.3823 -0.36256 -0.91361 -0.296498 -0.211950
kangeki.p  0.7154 -0.5154 -0.19846 -0.35217  0.224829  0.495204
gansui     1.0154  0.1187  0.24770  0.16414 -0.375271  0.006443
kyounetu   0.9906  0.1265  0.26056  0.35623 -0.194037  0.092960
mad        1.0378  0.2482 -0.07034  0.10042  0.004139 -0.253976
silt       0.9161  0.3661 -0.45520 -0.10170  0.161201  0.063680
sand      -0.9959 -0.3541  0.37738  0.05368 -0.128163  0.016793
```

環境変数のスコア（主成分係数）

```
Site scores (weighted sums of species scores)
           PC1     PC2      PC3      PC4      PC5     PC6
sit1  -1.08902  0.6524  0.37693  0.55691 -1.97996  0.5730
sit2  -0.66629  1.0053 -0.46297  0.66452  0.41114 -0.4666
sit3  -0.88609  1.0609  0.83194  0.34093 -1.27763  0.5368
sit4   0.46942  0.2287 -0.85159 -0.68841  0.48766  1.1966
sit5   0.08192  0.6112  1.97577 -1.27957  0.92945  1.6407
sit6   0.85839  1.0083 -0.34100  1.06372  0.52572 -0.6778
sit7  -1.60276 -1.4683  0.89721  0.77378  0.80643 -0.5957
sit8   0.56218  0.5707  0.02024  1.27578 -0.01943 -0.5724
sit9   0.09425 -0.5825  0.64905  0.45683  0.16345  0.7775
（略）
```

各地点のスコア（主成分得点）

biplotによる結果の表現



地点は、主成分得点に基づき点で、環境変数は固有ベクトル（係数）にもとづいて矢印で表現

矢印の角度（方向）が近い変数同士は、似た挙動（相関が高い）

PC1の値が大きいほど、

シルト・粘土・含水・強熱が大きく、砂が少ない

Site 10・16などが該当

PC2の値が大きいほど、

深さ・湖底TP・表層TPが大きい

Site 3・10・18などが該当

主成分分析の使い道

- 地点ごとの環境のばらつきや、環境変数間の関係性を記述するのに、どういった特徴が見られるかをまとめる

- 重回帰分析において、相関の強い変数同士を説明変数に加えると、多重共線性の問題が生じる。この問題を解消するために、いくつかの相関のある変数を主成分にまとめ、これを説明変数として用いる

種子密度～水深＋表層TP＋低層TP＋表層TN＋低層TN＋土壌間隙水中P・・・

種子密度～水深＋水質PC1＋水質PC2＋底質PC1＋底質PC2

（除歪）対応分析（Detrended）Correspondence Analysis

- PCAが環境変数をまとめるのに使われるのに対し、複数種の個体数や分布の情報をまとめる際に使われる

```
data(varespec)
varespec

#CA
ca6.1<- cca(varespec)
summary(ca6.1)
plot(ca6.1)
```

```
#DCA
dca6.1<- decorana(varespec)
summary(dca6.1)
plot(dca6.1)
```

- PCAは変数間に線形の関係性を想定しているが、生物の個体数や現存量同士の関係性は非線形な場合も多い→CAやDCAでは一山形の関係を想定
- 「ある生物がいない」地点同士が似通って判断される(アーチ効果)は、DCAで解決できることがある

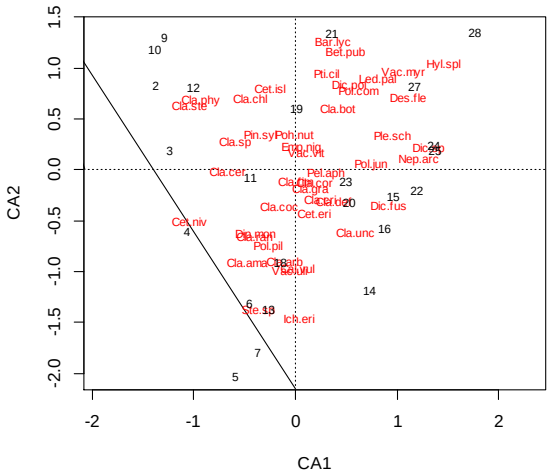
結果CA/DCA

- CA・DCAともに見方はPCAとそれほど変わらない

```
> summary(dca6.1)

Call:
decorana(veg = varespec)

Detrended correspondence analysis with 26 segments.
Rescaling of axes with 4 iterations.

                 DCA1   DCA2    DCA3    DCA4
Eigenvalues     0.5235 0.3253 0.20010 0.19176
Decorana values 0.5249 0.1572 0.09669 0.06075
Axis lengths    2.8161 2.2054 1.54650 1.64864
```

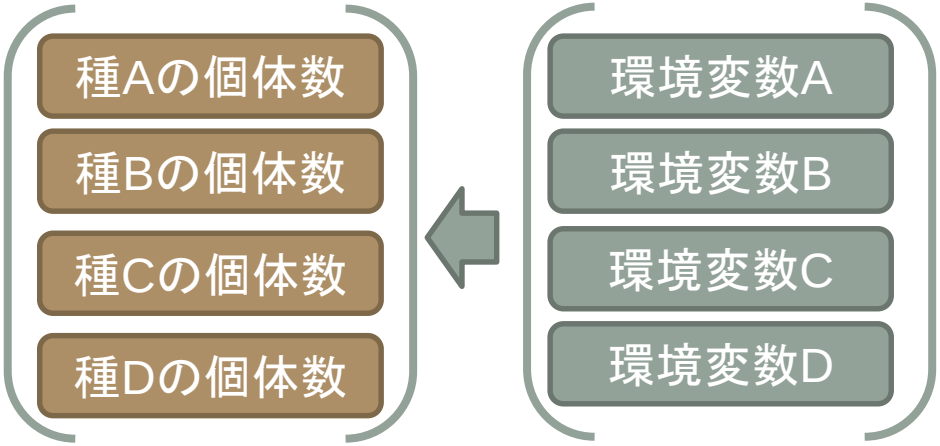
DCAの場合この値が重要。Axis length（Gradient lengthとも）が4を超えている場合、CAよりDCAが良いらしい…

```
> plot(dca6.1)
```



冗長性分析RDA

- 群集構造と環境条件の関係性を見たい。
- PCAと同様に標本スコアを固有ベクトルにより算出するが、固有ベクトルは環境変数によって「制約」される。
- 目的変数群のもつばらつきのうち、説明変数群で説明される割合を「冗長性」と呼ぶ



rdaによる実行

- 砂丘植物の種組成を、環境要因で説明する

```
data(dune)
data(dune.env)
rda6.1<- rda(decostand(dune,“hell”)~A1+Manure,dune.env) #hellinger変換した種構成データ
summary(rda6.1)
```

```
rda(formula = decostand(dune, "hell") ~ A1 + Manure, data = dune.env)

Partitioning of variance:
              Inertia Proportion
Total          0.5559     1.0000
Constrained    0.2313     0.4161
Unconstrained  0.3246     0.5839

Eigenvalues, and their contribution to the variance

Importance of components:
                          RDA1     RDA2     RDA3     RDA4     RDA5      PC1      PC2      PC3      PC4
Eigenvalue            0.1108 0.06862 0.02766 0.01341 0.01082 0.09785 0.05309 0.03838 0.03398
Proportion Explained  0.1994 0.12344 0.04975 0.02411 0.01946 0.17603 0.09550 0.06905 0.06112
Cumulative Proportion 0.1994 0.32281 0.37256 0.39667 0.41613 0.59216 0.68766 0.75671 0.81783
```

←群集の持つばらつきの41.61%が環境によって説明された

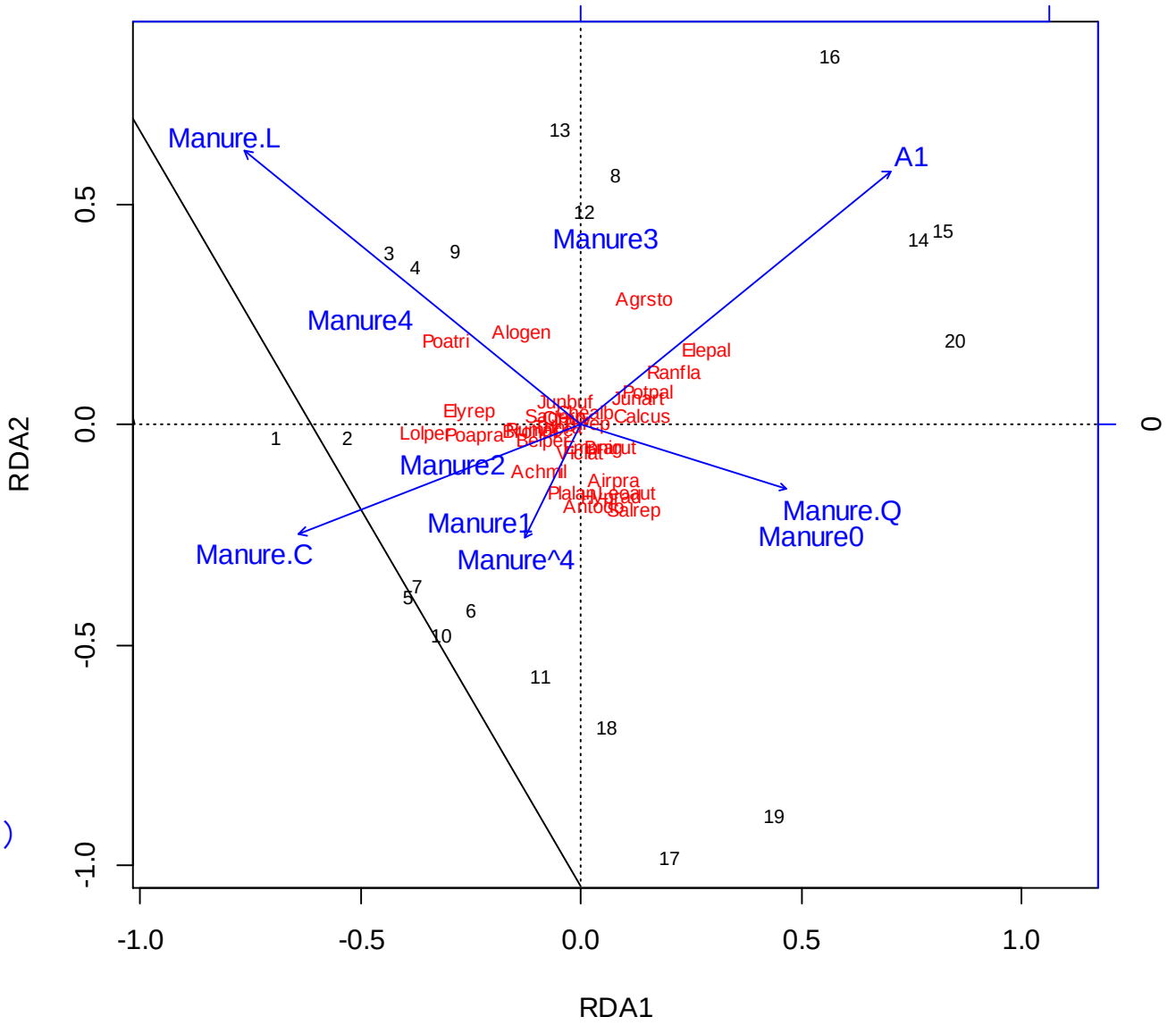
この部分が説明変数により制約されている軸

残りのばらつきはPCAと同じ形で軸が設定される

基本的にPCAと似た形式の結果表示

triplot

plot(rda6.1)



説明変数（環境変数）
目的変数（種）
サンプル（調査地）
の3者の関係を図示

検定など

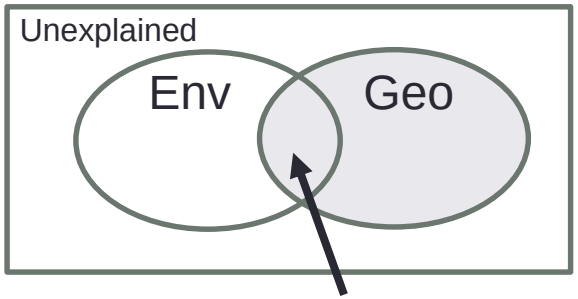
＊詳細はveganのヘルプ参照

検定

```
> anova(rda6.1,by="margin")
Permutation test for rda under reduced model
Marginal effects of terms

Model: rda(formula = decostand(dune, "hell") ~ A1 + Manure, data = dune.env)
          Df     Var      F N.Perm Pr(>F)
A1         1 0.05452 2.3516    299   0.02 *
Manure     4 0.15918 1.7165    299   0.02 *
Residual 14 0.32457
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

補正R^2

```
> RsquareAdj(rda6.1)
$r.squared
[1] 0.4161343

$adj.r.squared
[1] 0.2076108
```



その他

partialRDA

空間の効果を除いた後、
環境の影響を見たい

rda(sp,env,geo)

Variation Partitioning（分散分割）

環境・空間の独立、共通で説明
される割合を分割したい

varpart(sp,env,geo)

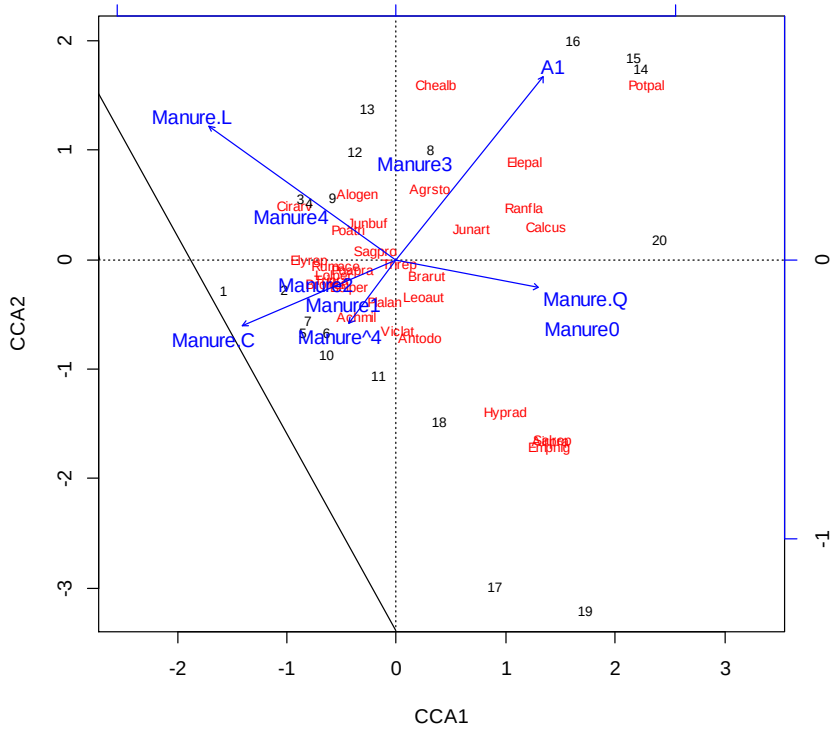
正準判別分析Canonical Correspondence Analysis

- 群集構造と環境条件の関係性を見たい。のCAバージョン
- PCA⇔RDAの関係と同様にCA⇔CCAの関係

Euclid距離ベース
線形モデル

X^2距離ベース
一山形分布仮定

```
data(dune)
data(dune.env)
cca6.1<- cca(dune~A1+Manure,dune.env)
summary(cca6.1)
plot(cca6.1)
anova(cca6.1,by="margin")
```



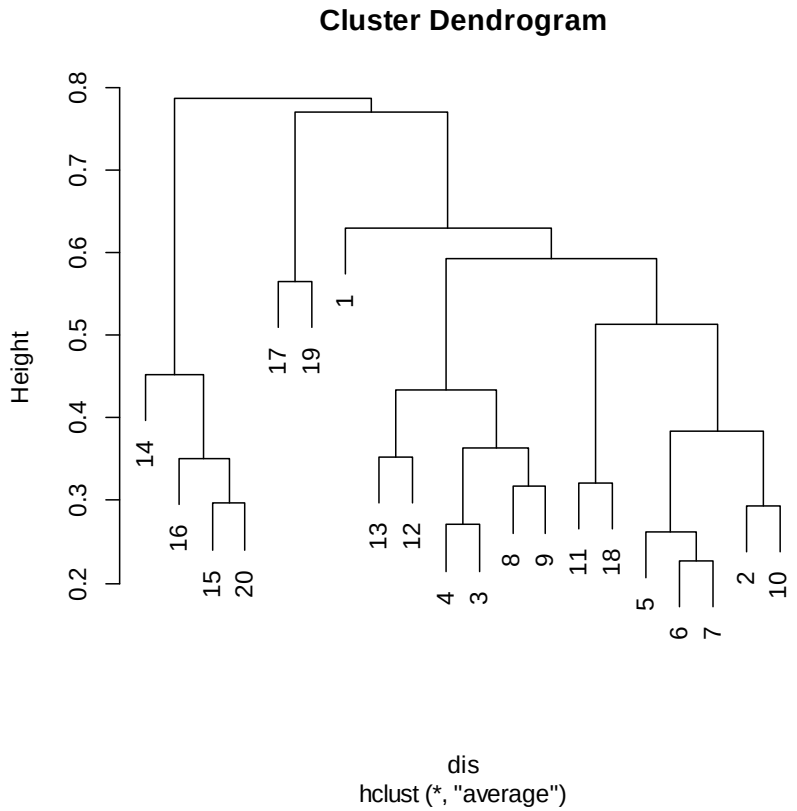
＊基本的にRDAと同様の関数が使え、類似した表示がなされる
（Variation PartitioningはRDAのみ使える）

クラスター分析

- 群集の組成に応じて似た地点ごとをグルーピングしたい、地点同士の関係性を知りたい場合に使われる

群集組成に基づいて、
地点間の（非）類似度を
総当りで計算

（非）類似度を基準に似
たもの同士をまとめる



(非)類似度指数

- いろいろあります(似た解析をしている論文を参考にしてください…)

- Bray-Curtis指数　　比較的よく使われる方法

```
vegdist(dune,"bray")
```

- Sørensen指数　　個体数ではなく1/0データを扱う

```
dist.binary(dune,method=5)
```

- Chao指数　　希少種のサンプリング過程を考慮

```
vegdist(dune,“chao")
```

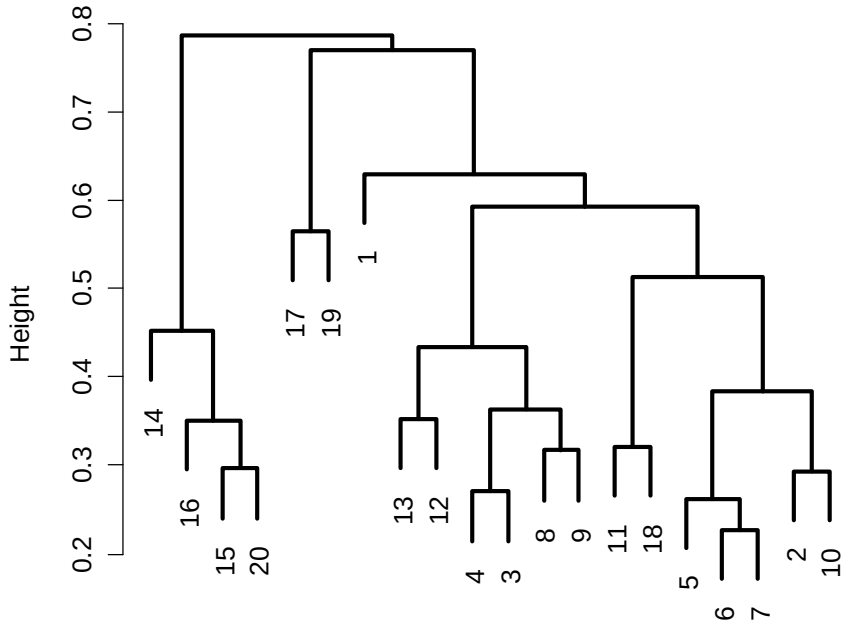
詳しくは、

土居・岡村(2001)「生物群集解析のための類似度とその応用:Rを使った類似度の算出、グラフ化、検定」日本生態学会誌61:3-20

Rによる解析

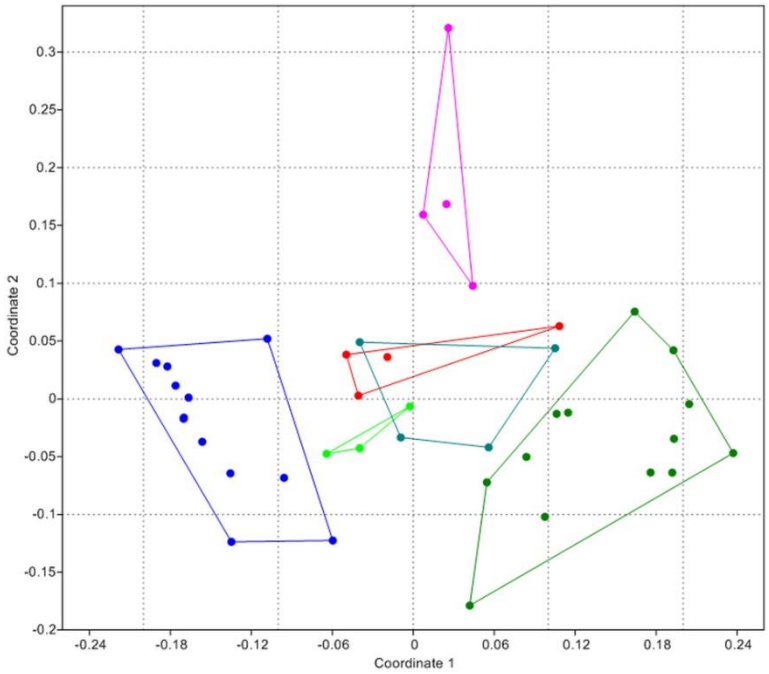
```
dis.bray<- vegdist(dune,"bray")
clus6.1b<- hclust(dis.bray,"average")
plot(clus6.1b)
```

#UPGMA法によるクラスタリング
#樹形図を書く



（非計量）多次元尺度法 （Nonmetric）Multi-Dimensional Scaling

- 群集間距離（非類似度）と二次元平面上の距離が同じような関係になるように、群集を配置
- NMDSでは非線形性の強いデータ（生物群集データ）にも対応できることから、群集間の関係を図示する際に使われる



Rによる実行（NMDS）

```
> nmds6.1<- metaMDS(dis.chao)
> nmds6.1

Call:
metaMDS(comm = dis.chao)

global Multidimensional Scaling using monoMDS

Data:     dis.chao
Distance: chao

Dimensions: 2
Stress:     0.1016694
Stress type 1, weak ties
Two convergent solutions found after 1 tries
Scaling: centring, PC rotation
Species: scores missing
```

非類似度行列を入れる
metaMDS(データ,”類似度指数”) でも

Stress値：当てはまりの尺度
以下の様な基準で解釈

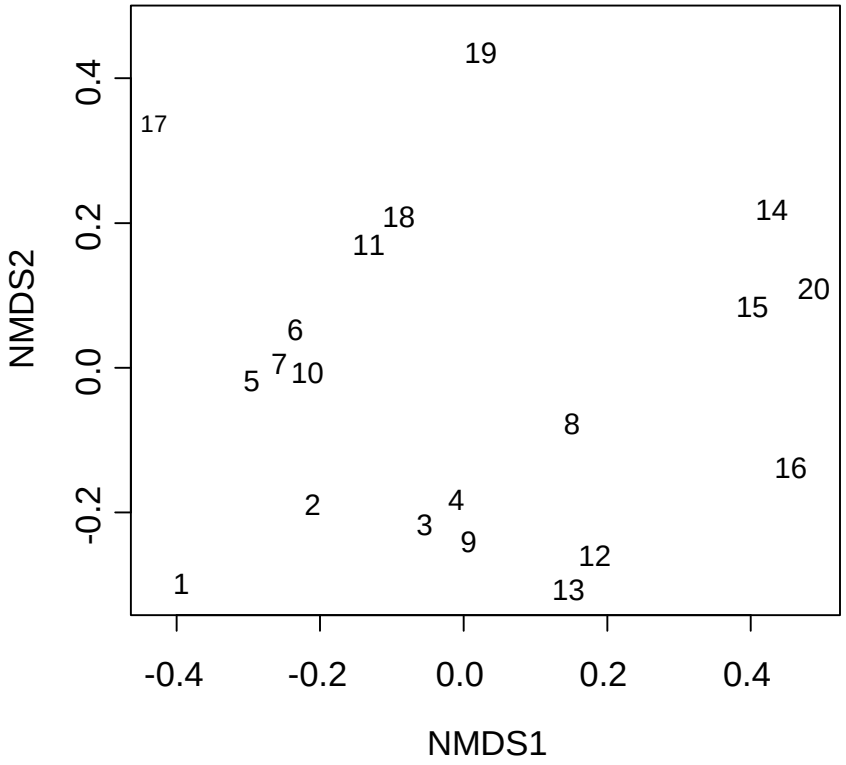
↓

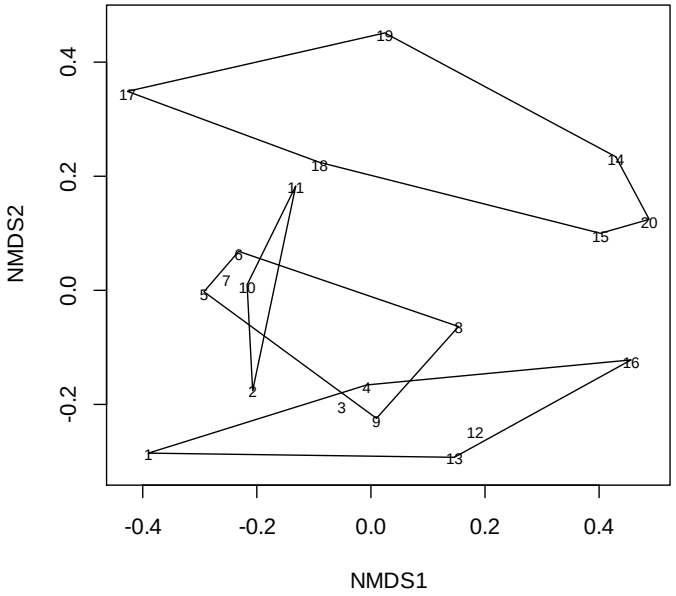
0.4	Poor
0.2	Fair
0.1	Good
0.05	Excellent

NMDSの結果の図示

```
plot(nmds6.1,type="text")
```

数字は地点の番号
近いところにプロットされるほど群集も似ている





#管理のタイプでグループをくくる↑

```
plot(nmds6.1,type="t")
ordihull(nmds6.1,dune.env$Management)
```

類似度行列を用いた検定

- Mantel検定

  （非）類似度行列同士の相関を検定。
  例） 群集の類似度と環境の類似度は関係あるか？

```
dis.env<- dist(dune.env$A1)   #この例は一変量ですが、多変量の環境データも扱えます
mantel(dis.chao,dis.env)
```

- ANOSIM（類似性分析）

  グループ内類似度とグループ間類似度の差を検定。
  例） ヒシ帯の群集と開放水面の群集で差があるか？

```
anosim(dis.chao,dune.env$Management)
```

- PERMANOVA（ノンパラメトリック多変量分散分析）

  ANOSIMと同様の場面で使われる。ANOSIMの欠点をカバーできているとか

```
adonis(dis.chao~dune.env$Management)
```

R以外のソフト

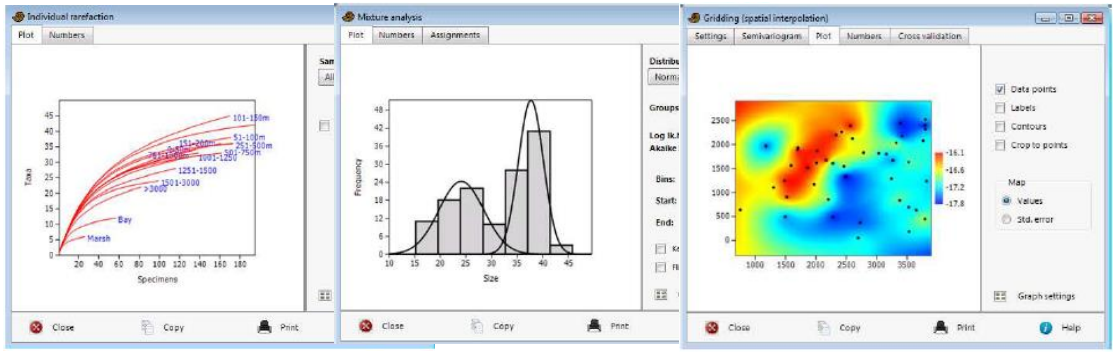
PRIMER

多変量解析をやっている論文では世界標準でよく使われている。有料です。
（6階のPD部屋のパソコンに入っています。使い方は柚原さんにでも・・・）

PAST

古生物データの解析用に開発されたらしい。無料でダウンロードできる。
やたらと充実した機能。（多変量以外もなにかと揃っている。グラフィックも綺麗）
実はできる子かもしれないと思っている。知名度は低い

GUIなので、エクセルデータ
貼り付けてボタンをポチれば
解析できる



多変量解析まとめ

非常に大雑把ですが・・・

多変量データをまとめたい

PCA　CA/DCA

多変量データ同士の関係を見たい

RDA　CCA

クラスター分析
NMDS

多変量データをもとに似たサンプル同士をまとめたい

参考資料

- vegan
  http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf
  でpdfのマニュアルが見られる。作者のOksanenのページには100ページ超えの講義資料なんかもおいてある（もちろん英語）

- 序列化（PCA・DCAなど）
  色々とネット上に情報あり。
  加藤（1995）「生物群集分析のための序列化手法の比較研究」環境科学会誌8:339-352

- 類似度・NMDS
  土居・岡村(2001)「生物群集解析のための類似度とその応用：Rを使った類似度の算出、グラフ化、検定」日本生態学会誌61:3-20

- 本　持ってますので困ったらどうぞ
  Borcard, Gillet, Legendre「Numerical Ecology with R」Springer

おしまい

おつかれさま！
あとは個別に相談してください

苦労してとったデータは、せっかくなので美味しく料理しましょう

まずい素材は頑張ってもあまり美味しくはできないですし、いい素材もありきたりの調理法では物足りないです

素材を活かして好きなように料理しましょう。もちろん素材で勝負もOK