Projet Hydrogen

Abdoulaye Diabakhaté

23 mai 2018

Table des matières

1 Multivariate Distance Matrix Regression (MDMR)

1.1 Introduction

L'analyse de régression matricielle à distance multivariée (MDMR) est une technique statistique qui consiste à calculer la distance entre toutes les paires d'individus (N individus) par rapport à P variables d'intérêt et à construire une matrice N*N dont les éléments reflètent ces distances.

Les tests de permutation peuvent être utilisés pour tester des hypothèses linéaires qui considèrent si les facteurs collectés sur les individus peuvent expliquer la variation des distances observées.

Le MDMR n'est pas différent de beaucoup de stratégies de réduction de données, mais il teste directement l'association entre les éléments de la matrice de distance ou de dissimilarité avec les variables auxiliaires et ne nécessite donc pas l'étape de réduction des données, souvent problématique.

Le MDMR peut être utilisé avec toutes les variables résultant d'un essai biologique ou d'un sous-ensemble, ce qui en fait un outil souple et attrayant pour identifier des modèles significatifs.

1.2 Calcul d'une matrice de distance

La formation d'une matrice de distance (ou dissemblance) appropriée est un ingrédient essentiel dans l'analyse du MDMR.

Cependant, il existe un grand nombre de mesures de distances potentilles que l'on pourrait utiliser (Webb,2002) et malheureusement, il y'a très peu de matériel publié pouvant guider un chercheur sur la mesure de distance la plus appropriée pour une situation donnée.

Par exemple,bien que la mesure de la distance euclidienne soit utilisée couramment dans les paramètres d'analyse de clusters traditionnels,les fonctions du coefficient de corrélation sont les mesures de distance les plus largement utilisées dans les analyses d'expression génique de haute dimension(D'Haeseleer,2005).

Nous notons que les mesures de distance avec des propriétés métriques ou non métriques peuvent être utilisées dans les analyses MDMR (Gower and Krzanowski,1999).
En supposant que l'on ait identifié une mesure de distance appropriée,une matrice de distance N*N est construite.

Que cette matrice de distance et ses éléments soient notés :

$$D = d_{ij}(i, j = 1, ..., N)$$

où d_{ij} reflète la distance entre les profils i et j.

1.3 Dérivation de statistiques de test MDMR

Une fois que l'on a calculé une matrice de distance,D,la relation entre M facteurs additionnels (c-à-d « prédicteurs » ou variables « régresseurs ») et la variation des distances entre et parmi les N individus représentés en D peut être explorée.

soit X, une matrice N*M contenant des informations sur les M facteurs qui seront modélisés comme les variables indépendantes ou régressives dont les relations avec les valeurs dans la matrice de distance sont intéressantes.

Calculons la matrice de projection standard :

$$H = X(X'X)^{-1}X'$$

généralement utilisée pour estimer les coefficients reliant les variables prédictives aux variables de résultat dans les contextes de régression multiple.
Ensuite,calculons la matrice :

$$A = (a_{ij}) = (-[\tfrac{1}{2}])d_{ij}^2$$

et centrons cette matrice en utilisant la transformation discutée par Gower (1966) et dénotons cette matrice G :

$$G = (\mathbb{I} - \tfrac{1}{N}11')A(\mathbb{I} - \tfrac{1}{N}11')$$

où, 1 est un vecteur N-dimensionnel de 1.

Une statistique F peut être construite pour tester l'hypothèse que les variables M régresseurs n'ont pas de relation avec la variation de distance ou de dissimilarité des N sujets reflétés des N sujets reflétés dans la matrice distance (McArdle and Anderson, 2001).

$$F = \frac{tr(HGH)/(m-1)}{tr[(I-H)G(I-H)]/(n-m)}$$

1.4 Colinéarité

Un problème fondamental avec toutes les techniques d'analyse basées sur la régression multiple est la colinéarité ou les fortes dépendances parmi les régresseurs.

La colinéarité peut créer des problèmes dans le calcul de la matrice de projection $H = X(X'X)^{-1}X'$ et entrainer des estimations de paramètres instables.

Bien qu'il existe des procédures qui peuvent être utilisées pour surmonter ce problème, comme la régression Ridge et la régression des principales composantes (Mason and Perreault, 1991), on peut utiliser la décomposition orthogonale-triangulaire (Gunst, 1983) pour former la matrice de projection et on peut dire que cela fonctionne bien dans le contexte de l'analyse du MDMR.

1.5 Tests de permutations

Les propriétés distributionnelles de la statistique F seraient compliquées à dériver analytiquement pour différentes mesures de distance non euclidienne, en particulier lorsque ces meures de distance sont calculées pour plus d'une variable.

Des tests basés sur la signification, tels que les tests de permutation, peuvent alors être utilisés pour évaluer la signification statistique de la pseudo statistique F comme alternatives à l'utilisation de tests basés sur la distribution asymptotique de la statistique F (Jockel, 1986 ; Edgington, 1995 ; Manly,1997 ; Good, 2000).

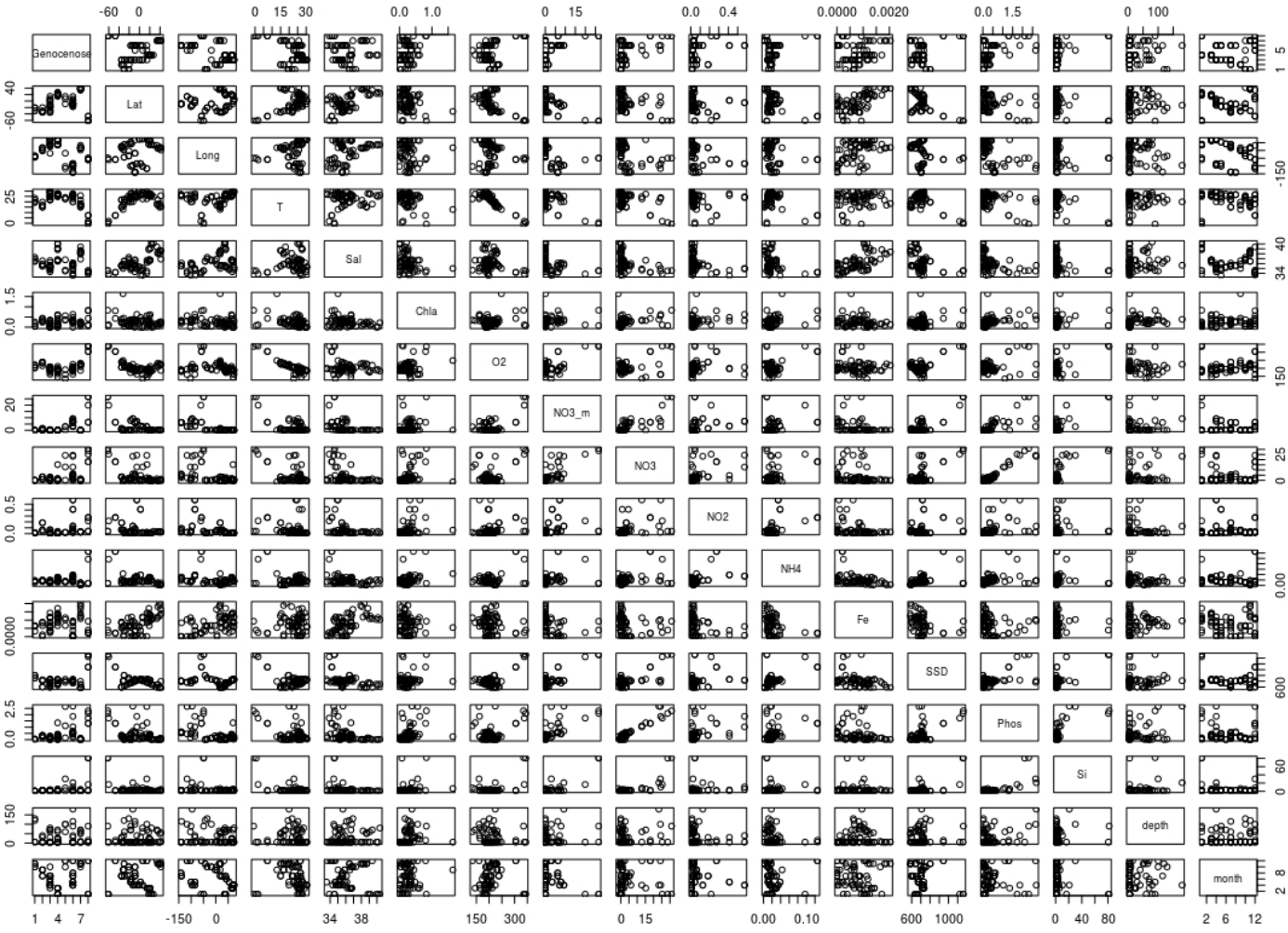
Les tests de permutations peuvent être poursuivis en permutant les variables indépendantes ou prédictives, en recalculant la statistique MDMR, en répétant ce processus et en additionnant le nombre de fois que les statistiques calculées avec les données permutées sont plus grandes que la statistique générée avec les données réelles.

Nous notons également que les variables M régresseurs évaluées dans une analyse MDMR peuvent être testées individuellement ou par étapes (McArdle and Anderson, 2001 ; Zapala and Schork, 2006

2 Résultats statistiques

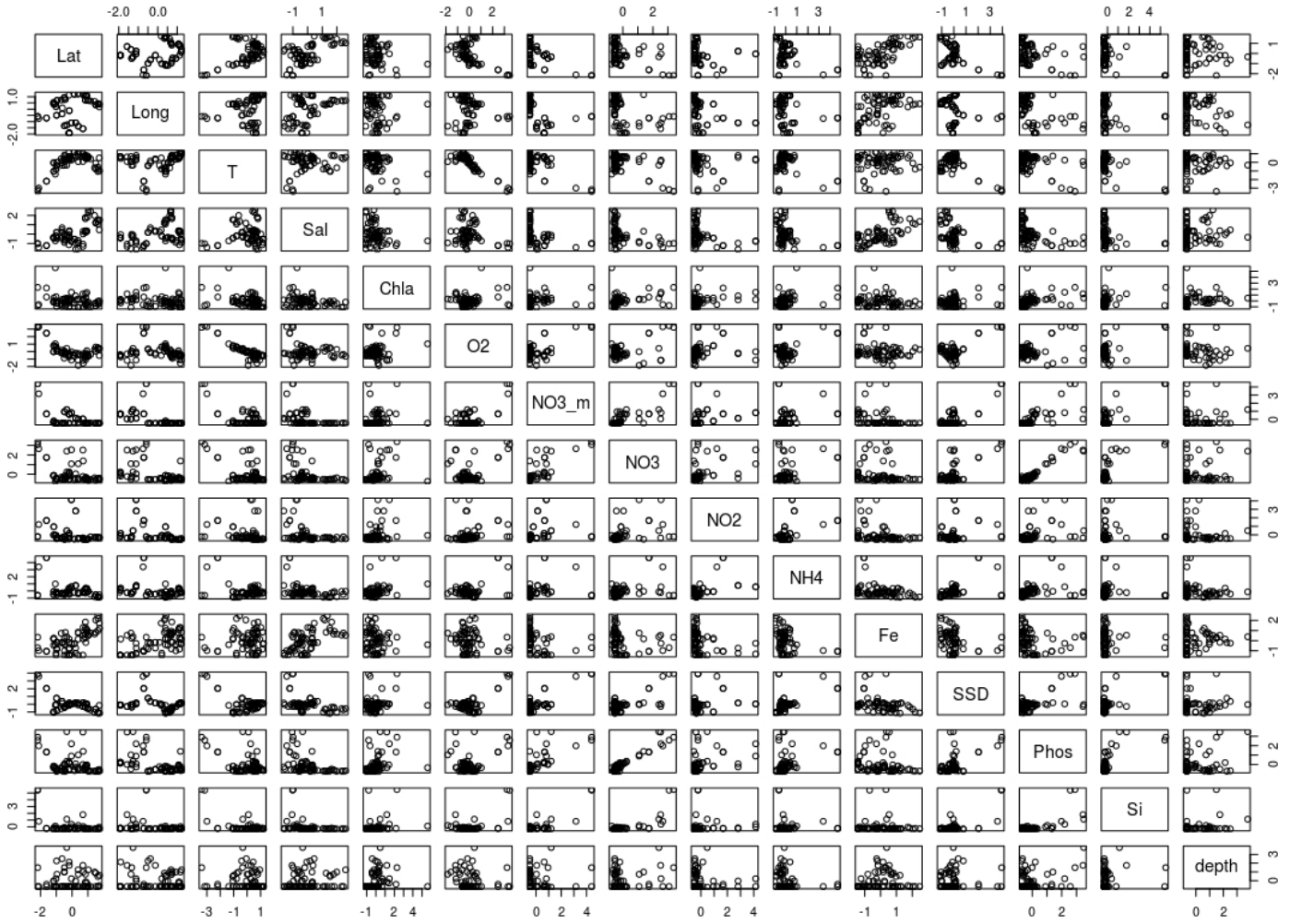
2.1 Résumé statistique avant centrage réduction

```
> summary(new_des[,-c(1,17)]) # Avant Centrage Réduction
      Lat               Long                T                Sal             Chla
 Min.   :-62.038   Min.   :-142.56   Min.   :-0.8266   Min.   :33.54   Min.   :0.0000
 1st Qu.:-20.935   1st Qu.: -60.06   1st Qu.:19.7948   1st Qu.:35.11   1st Qu.:0.1287
 Median : -8.779   Median :  17.56   Median :24.8154   Median :35.75   Median :0.2348
 Mean   : -4.047   Mean   :  -9.53   Mean   :22.1673   Mean   :36.18   Mean   :0.2700
 3rd Qu.: 18.537   3rd Qu.:  40.22   3rd Qu.:26.5390   3rd Qu.:36.75   3rd Qu.:0.3297
 Max.   : 42.173   Max.   :  73.90   Max.   :30.4624   Max.   :40.37   Max.   :1.6421
       O2             NO3_m               NO3               NO2                NH4
 Min.   :134.5   Min.   : 0.001139   Min.   :-1.55908   Min.   :0.004047   Min.   :0.001
 1st Qu.:187.1   1st Qu.: 0.006881   1st Qu.: 0.03762   1st Qu.:0.009464   1st Qu.:0.011
 Median :199.6   Median : 0.051411   Median : 0.77125   Median :0.026783   Median :0.015
 Mean   :207.4   Mean   : 2.707273   Mean   : 4.29566   Mean   :0.069475   Mean   :0.021
 3rd Qu.:218.2   3rd Qu.: 3.319110   3rd Qu.: 3.33247   3rd Qu.:0.058897   3rd Qu.:0.022
 Max.   :343.7   Max.   :26.624906   Max.   :31.00000   Max.   :0.573160   Max.   :0.125
       Fe                 SSD              Phos               Si              depth
 Min.   :5.035e-06   Min.   : 582.0   Min.   :0.00000   Min.   : 0.319   Min.   :  2.00
 1st Qu.:3.295e-04   1st Qu.: 657.8   1st Qu.:0.07591   1st Qu.: 1.056   1st Qu.:  5.00
 Median :6.855e-04   Median : 706.0   Median :0.22517   Median : 1.756   Median :  5.00
 Mean   :7.535e-04   Mean   : 713.7   Mean   :0.45307   Mean   : 5.219   Mean   : 31.09
 3rd Qu.:1.139e-03   3rd Qu.: 726.8   3rd Qu.:0.50356   3rd Qu.: 2.660   3rd Qu.: 58.75
 Max.   :2.036e-03   Max.   :1161.0   Max.   :2.68675   Max.   :81.820   Max.   :180.00
```


Genocenose
Lat
Long
T
Sal
Chla
O2
NO3_m
NO3
NO2
NH4
Fe
SSD
Phos
Si
depth
month

2.2 Résumé statistique après centrage réduction

```
> summary(new_des_CR) # Après Centrage Réduction
      Lat                Long                 T                  Sal                Chla
 Min.   :-2.1818   Min.   :-1.9564   Min.   :-3.3847   Min.   :-1.6290   Min.   :-1.1010
 1st Qu.:-0.6354   1st Qu.:-0.7431   1st Qu.:-0.3492   1st Qu.:-0.6595   1st Qu.:-0.5764
 Median :-0.1780   Median : 0.3984   Median : 0.3898   Median :-0.2662   Median :-0.1436
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.8497   3rd Qu.: 0.7317   3rd Qu.: 0.6435   3rd Qu.: 0.3556   3rd Qu.: 0.2433
 Max.   : 1.7390   Max.   : 1.2269   Max.   : 1.2210   Max.   : 2.5906   Max.   : 5.5941
       O2                NO3_m               NO3                NO2                 NH4
 Min.   :-1.8461   Min.   :-0.4984   Min.   :-0.7406   Min.   :-0.53771   Min.   :-0.892
 1st Qu.:-0.5156   1st Qu.:-0.4974   1st Qu.:-0.5386   1st Qu.:-0.49319   1st Qu.:-0.452
 Median :-0.1984   Median :-0.4892   Median :-0.4458   Median :-0.35086   Median :-0.243
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.000
 3rd Qu.: 0.2731   3rd Qu.: 0.1127   3rd Qu.:-0.1218   3rd Qu.:-0.08693   3rd Qu.: 0.050
 Max.   : 3.4494   Max.   : 4.4052   Max.   : 3.3780   Max.   : 4.13946   Max.   : 4.685
       Fe                 SSD                Phos                Si                depth
 Min.   :-1.4401   Min.   :-1.15910   Min.   :-0.70730   Min.   :-0.3518   Min.   :-0.73
 1st Qu.:-0.8158   1st Qu.:-0.49235   1st Qu.:-0.58880   1st Qu.:-0.2988   1st Qu.:-0.66
 Median :-0.1307   Median :-0.06765   Median :-0.35578   Median :-0.2486   Median :-0.66
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00
 3rd Qu.: 0.7419   3rd Qu.: 0.11499   3rd Qu.: 0.07883   3rd Qu.:-0.1837   3rd Qu.: 0.70
 Max.   : 2.4670   Max.   : 3.93727   Max.   : 3.48709   Max.   : 5.4995   Max.   : 3.77
```


Lat
Long
T
Sal
Chla
O2
NO3_m
NO3
NO2
NH4
Fe
SSD
Phos
Si
depth

3 Comparaison de deux méthodes

3.1 Méthode Adonis

```
>adonis(as.dist(jaccard_abundance)~Lat+Long+T+Sal+Chla+O2+NO3_m+NO2+NH4+
Fe+SSD+Phos+Si+depth,
data =as.data.frame(new_des_CR))

Call:
adonis(formula = as.dist(jaccard_abundance) ~ Lat + Long
+T+Sal + Chla + O2 + NO3_m + NO2 + NH4
+ Fe + SSD + Phos+Si+depth,
data = as.data.frame(new_des_CR))

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
Lat        1     12903 12903.4  4.7293 0.04614  0.001 ***
Long       1     19900 19900.1  7.2938 0.07116  0.001 ***
T          1     22267 22266.9  8.1612 0.07962  0.001 ***
Sal        1      8741  8741.2  3.2038 0.03126  0.001 ***
Chla       1      6559  6559.3  2.4041 0.02345  0.001 ***
O2         1      8420  8419.8  3.0860 0.03011  0.001 ***
NO3_m      1      9918  9917.8  3.6351 0.03546  0.001 ***
NO2        1      6812  6812.0  2.4967 0.02436  0.001 ***
NH4        1      8685  8685.1  3.1833 0.03106  0.001 ***
Fe         1      3872  3872.2  1.4192 0.01385  0.037 *
SSD        1      7485  7485.2  2.7435 0.02676  0.001 ***
Phos       1      3801  3800.9  1.3931 0.01359  0.036 *
Si         1      5553  5552.5  2.0351 0.01985  0.002 **
depth      1      4688  4688.4  1.7184 0.01676  0.005 **
Residuals 55    150060  2728.4         0.53657
Total     69    279665                 1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3.2 Méthode MDMR

```
> mdmr.res <- mdmr(X = new_des_CR,D)
100 % of permutation test statistics computed.
> summary(mdmr.res)
```

```
          Statistic Numer.DF Pseudo.R2 Permutation.p.value
(Omnibus)    1.4281       15    0.5881               <0.002 ***
Lat          0.0920        1    0.0379               <0.002 ***
Long         0.0884        1    0.0364               <0.002 ***
T            0.0525        1    0.0216               <0.002 ***
Sal          0.0611        1    0.0252               <0.002 ***
Chla         0.0364        1    0.0150                0.012 *
O2           0.0332        1    0.0137                0.024 *
NO3_m        0.0654        1    0.0269               <0.002 ***
NO3          0.0571        1    0.0235                0.002 **
NO2          0.0624        1    0.0257               <0.002 ***
NH4          0.0553        1    0.0228               <0.002 ***
Fe           0.0248        1    0.0102                0.150
SSD          0.0521        1    0.0215               <0.002 ***
Phos         0.0679        1    0.0280               <0.002 ***
Si           0.0377        1    0.0155                0.008 **
depth        0.0372        1    0.0153                0.016 *
---
Signif. Codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```