Analyse multivariée non paramétrique à partir de matrice de distance

Abdoulaye Diabakhaté

2 juillet 2018

Plan de la présentation

Notations

- Soit un échantillon de taille n d'observations individuelles, indicées par i=1,...,n réalisations de variables aléatoires (y_i, x_i).
  - y_i variable continue prenant ses valeurs dans $\mathbb{R}$.
  - x_i variables en nombre m, de type quelconque.
- On confond les variables aléatoires et leurs réalisations et on réserve les majuscules pour des vecteurs
- La variable dépendante y_i s'écrit comme : $y_i = x_i\beta + u_i$
  - β est un paramètre à estimer
  - Le modèle est linéaire en β

$Y = X\beta + U$

Avec X= $\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$

$Y = (y_1, ..., y_n)'$

$U = (u_1, ..., u_n)$

Propriétés

D'après le théorème de Gauss-Markov :

$$\hat{\beta} = \underset{\beta}{\text{argmin}} ||Y - X\beta|| = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

Donc le meilleur estimateur linéaire en Y et sans biais est :

$$\hat{\beta} = (X'X)^{-1}X'Y$$

La matrice des valeurs prédites est :

$\hat{Y} = X\hat{\beta} = HY$

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

La matrice des résidus est :

$R = Y - \hat{Y} = (I - H)Y$

La matrice totale SSCP est décomposée par les matrices SSCP prédites et résiduelles :

$$Y'Y = \hat{Y}'\hat{Y} + R'R$$

où $S_T = tr(Y'Y)$; $S_H = tr(\hat{Y'}\hat{Y})$; $S_R = tr(R'R)$

Une statistique appropriée pour tester l'hypothèse nulle de l'absence d'effet des paramètres du modèle est la pseudo statistique F :

$$F = \frac{tr(\hat{Y'}\hat{Y})/(m-1)}{tr(R'R)/(n-m)}$$

- Soit $D = (d_{ij})$, une matrice de distance de taille n*n.
- Posons $A = (a_{ij}) = (\frac{-1}{2} d_{ij}^2)$

Nous pouvons alors calculer la matrice centrée de Gower G en centrant les éléments de A :

$G = (\mathbb{I} - \frac{1}{n} 1 1^{'}) A (\mathbb{I} - \frac{1}{n} 1 1^{'})$.
Avec 1 est une colonne de taille n, contenant uniquement des 1

Ainsi en remplaçant $(YY^{'})$ par G, nous avons $S_T = tr(G)$ et la pseudo statistique F est :

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

Distances entre valeurs prédites

- Soit Y le jeu de données centré contenant toutes nos observations.
- Nous nous plaçons dans un cadre linéaire, c'est-à-dire : $Y = X\theta + \epsilon$, où X est la matrice de design contenant les covariables.
- Nous disposons uniquement de la matrice de distances $D^2 = (||Y_{i,.} - Y_{j,.}||^2)_{i,j}$
- Supposons que D est une matrice de distances euclidiennes
- Notre objectif est de calculer une matrice de distances entre valeurs prédites(ou ajustées)à partir de la prédication $\hat{Y}$ de Y : $\hat{Y} = X\hat{\theta} = HY$

On peut noter, une telle matrice par :

$$\hat{D}^2 = (||\hat{Y}_{i,.} - \hat{Y}_{j,.}||^2)_{i,j}$$

Distances entre résidus

Cette fois-ci nous cherchons à calculer la matrices de distances entre résidus,c'est-à dire :

$$D_R^2 = (||R_{i,.} - R_{j,.}||^2)_{i,j}$$

Avec la matrice des résidus qui est :

$$R = Y - \hat{Y} = (I - H)Y$$

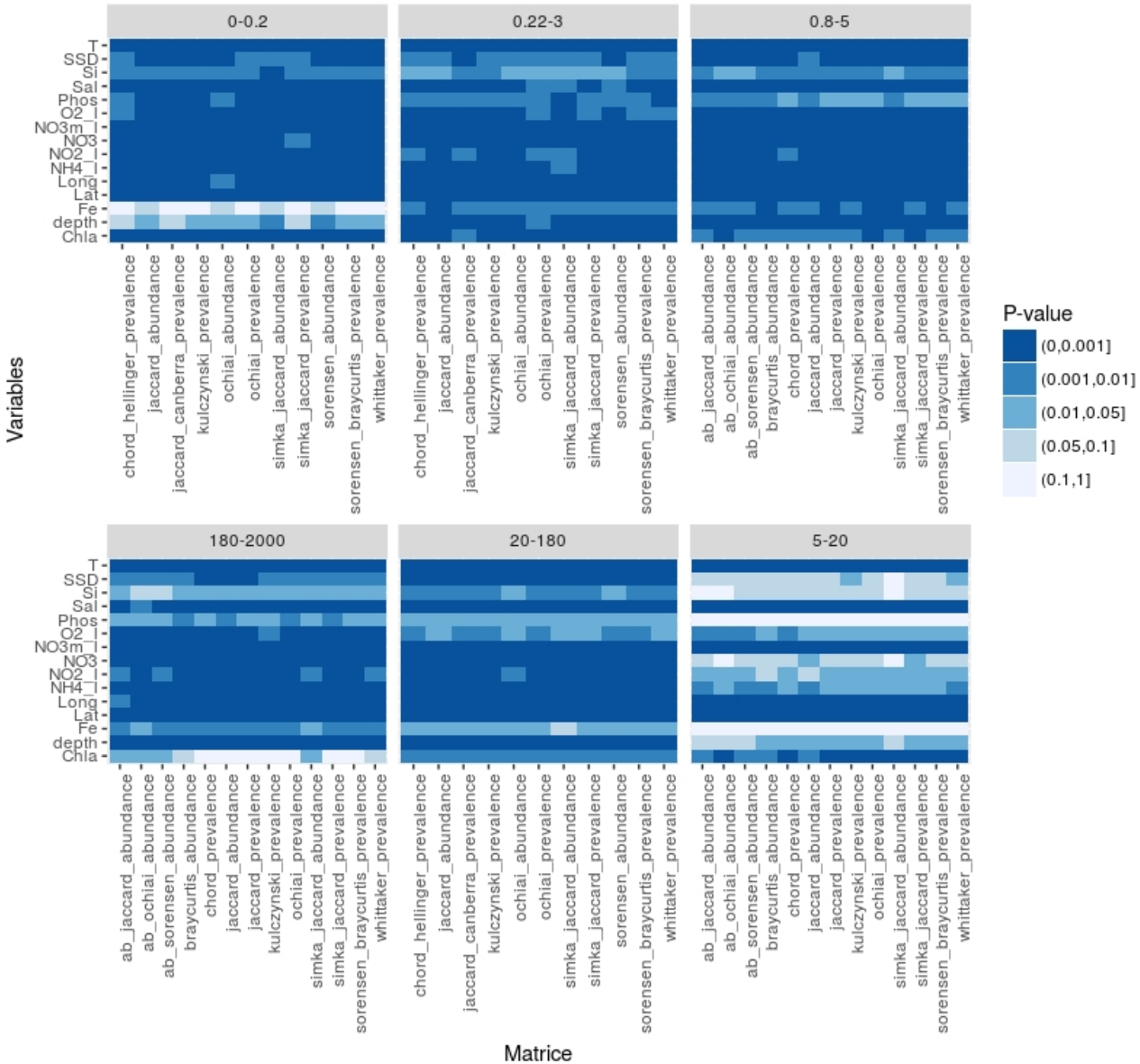
Introduction sur la méthode Adonis

La méthode Adonis a été appliquée sur 15 variables :

- Lat,Long,T,Sal,chla,O2_I,NO3m_I,NO3,NO2_I,NH4_I,SSD, Phos,Si,depth,Fe.
- Les fractions de taille concernées sont au nombre de 6 :

- 3 fractions composées chacunes de 11 matrices de distance : 0_0.2, 0.22_3 et 20_180
- 3 autres fractions composées chacune de 13 matrices de distance : 5_20, 180_2000 et 0.8_5

Résultats de Adonis



Best Subset Of Environmental Variables With Maximum (Rank) Correlation With Community Dissimilarities

Introduction

La fonction bioenv calcule une matrice de dissimilarité de communauté en utilisant vegdist.
Ensuite, elle sélectionne tous les sous-ensembles possibles de variables environnementales, met à l'échelle les variables et calcule les distances euclidiennes pour ce sous-ensemble en utilisant dist.
Ensuite,elle trouve la corrélation entre les dissimilarités communautaires et les distances environnementales, et pour chaque taille de sous-ensembles, enregistre le meilleur résultat.
Il y'a $2^p - 1$ sous-ensembles de p-variables.

En effet, un coefficient de corrélation (typiquement le coefficient de corrélation de rang de Spearman) est calculé entre les deux matrices et le meilleur sous-ensemble de variables environnementales peut alors être identifié et soumis ensuite à un test de permutation pour déterminer la signification.
La méthode est également largement acceptée par la communauté scientifique en raison de sa flexibilité à travers une grande variété de données et est complètement non paramétrique ; l'article Clarke et Ainsworth (1993) décrivant la méthode compte 674 citations sur Google Scholar au moment de cette publication.

Résultats de bioenv