De la biologie à l'algèbre linéaire ... en passant par R
Expérimenter la notion de projection

A.B. Dufour[a], **S. Dray**[a], **J.R. Lobry**[b] and **J. Thioulouse**[a]

[a]Laboratoire de Biométrie et Biologie Evolutive, UMR 5558
CNRS, Université Claude Bernard, FST, Département de Biologie
43, Bd du 11 novembre 1918, 69622 Villeurbanne cedex
anne-beatrice.dufour@univ-lyon1.fr

[b]Institut National de la Police Scientifique
Ministère de l'Intérieur
31, av. Franklin Roosevelt, BP 30169 69134 Ecully cedex
jean.lobry@interieur.gouv.fr

Mots clefs : Enseignement, Biologie Humaine, Algèbre linéaire, Projection.

L'enseignement de la statistique auprès des biologistes se confond, à Lyon, avec l'histoire du laboratoire de Biométrie et de son fondateur J.M. Legay. Dès le début des années soixante, ce dernier réunit biologistes et mathématiciens pour initier un dialogue et développer de nouvelles méthodes [1]. Cette impulsion conduit à la création d'enseignements intégrés liant biologie, statistique et informatique. L'objectif de cette communication est de montrer l'apport du logiciel R dans cette relation triangulaire [2].

Une partie de la statistique comme l'analyse des données relève de l'algèbre linéaire et plus particulièrement de la notion de projection. Celle-ci prend des formes différentes selon la problématique posée : (i) les variables étudiées jouent des rôles asymétriques (*i.e.* explicatives ou à expliquer) comme en régression linéaire simple, (2) les variables étudiées jouent un rôle identique comme dans l'analyse en composantes principales. Cette projection est l'essence même de la méthode mais son formalisme mathématique peut la rendre difficile à comprendre.

C'est pourquoi la méthode et la notion de projection sont exposées à partir de la donnée brute. Elles se visualisent, s'expérimentent, se révèlent. L'exemple proposé porte sur les mesures de la stature (taille, en cm), de l'empan de la main dominante (empan1, en cm) et de l'empan de la main non dominante (empan2, en cm) réalisées sur 168 étudiants (`data(survey)` de la librairie `MASS`). L'empan est la distance entre l'auriculaire et le pouce, le poignet et la main étant posés, à plat sur une table, les doigts écartés au maximum.

Il existe une relation linéaire entre l'empan et la taille d'un individu. Cette relation peut être modélisée par une droite dite droite de régression. L'objectif est de trouver les valeurs de l'ordonnée à l'origine et de la pente qui minimisent la somme des carrés des écarts entre la valeur de l'empan et son estimation par le modèle (sa projection parallèlement à l'empan, variable à expliquer). Avec le logiciel R, une fonction des deux paramètres (pente et ordonnée à l'origine) peut être construite et l'étudiant peut essayer de rechercher la droite optimum (Figure 1, [3]).

Mais il est rare de n'avoir que deux mesures à mettre en relation. Un des objectifs de la morphométrie est de séparer la taille globale d'un individu de sa forme. L'idée est alors de prendre l'ensemble des mesures, de ne privilégier aucune variable et de rechercher cette taille globale. La solution est le premier axe d'une analyse en composantes principales c'est-à-dire de la droite qui maximise la variance projetée ou qui minimise la somme des carrés des écarts entre un individu et sa projection orthogonale. Avec le logiciel R et la fonction `plot3d`, l'étudiant peut faire tourner le nuage de points (liant taille et empans) jusqu'à faire apparaître une direction la plus étendue possible (Figure 2, [4]).

Enseigner la statistique par le formalisme mathématique éloigne de la donnée. Pour que l'étudiant comprenne et s'approprie une méthode, l'enseignant se doit d'être pragmatique. L'échange s'opère autour de la visualisation de la donnée et de la méthode. Les outils proposés par le logiciel R permettent d'initier ce dialogue.

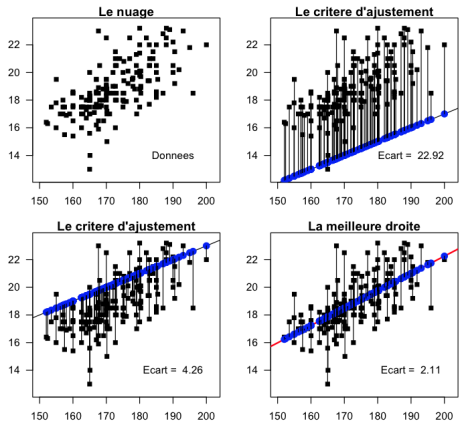

Figure 1 : Différentes expressions de la relation liant la taille et l'empan de la main dominante

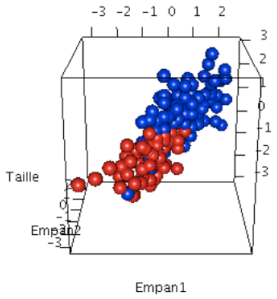

Figure 2 : Une représentation du nuage des 168 étudiants en dimension 3

Références

[1] Legay, J.M. (2004) L'interdisciplinarité vue et pratiquée par les chercheurs en sciences de la vie. *Nature, Sciences et Sociétés*, **12**, 63-74

[2] Dufour, A.B. (2012) La part du logiciel R dans l'enseignement de la statistique en biologie. Le site Web de Lyon. *Statistique et Enseignement*, **2**(2), 41-47

[3] Dufour, A.B., Lobry, J.R., Chessel, D., Rochette, N. (2012) De la stature chez l'Homme ... à la taille des cerveaux chez les mammifères. Réversion, Régression, Corrélation.
`http://pbil.univ-lyon1.fr/R_svn/pdf/bem3.pdf`

[4] Dufour, A.B., Lobry, J.R. (2011) Initiation à l'analyse en composantes principales.
`http://pbil.univ-lyon1.fr/R_svn/pdf/tdr601.pdf`