TP : Etude détaillée des données `cookies` et des données `copals`

1 Étude des données `cookies`

On s'intéresse à un problème de contrôle de qualité sur une chaîne de fabrication de cookies. Il est nécessaire de contrôler le mélange des ingrédients avant cuisson pour s'assurer que les proportions en lipides, sucre, farine et eau sont bien respectées, c'est-à-dire proches des valeurs nominales de la recette qui a fait la réputation de l'entreprise. Il s'agit de savoir s'il est possible de dépister au plus tôt une dérive afin d'intervenir sur les équipements concernés. Les mesures et analyses, faites dans un laboratoire classique de chimie, sont relativement longues et coûteuses. Elles ne peuvent être entreprises pour un suivi régulier ou continu de la production. Il a donc été décidé l'utilisation d'un spectromètre en proche infrarouge (NIR). L'appareil mesure l'absorbance c'est-à-dire les spectres dans les longueurs d'ondes du proche infra-rouge.

Les données se nomment `cookies` et sont issues du package `ppls`. Notre étude se restreint au taux de sucre. Les objectif sont i) de détecter les positions qui influent sur la quantité de sucre, ii) de prédire la quantité de sucre à partir des positions sélectionnées.

1. Faire une analyse descriptive des données (ACP, boxplots, heatmap, ...).
2. Mettre en place les méthodes de régression régularisée vues en cours (Ridge, Lasso et éventuellement Elastic Net).
3. Discuter et comparer les méhodes.

2 Étude des données `copals`

Le copal est une résine semi-fossile ou sub-fossile (proche de l'ambre) que l'on trouve principalement en Afrique et en Inde. L'étude a été conduite en Afrique sur $n = 30$ arbres de deux générations différentes (variable `genus` : 1960 et 1967). Pour l'année 1967, deux zones géographiques ont été considérées (variable `origin` : East et West) alors que pour l'année 1960, seule la zone Est a été considérée (variable `origin` : Class 0). Pour chaque arbre, on dispose des données de métabolomiques obtenues par chromatographie en phase liquide couplée à la spectrométrie de masse (LC-MS).

Les données se nomment `copals_camera` et sont issues du package `MultiVarSel`.

Attention ! On utilisera une version antérieure du package : `MultiVarSel_1.0.tar.gz` avec la commande `install.packages("MultiVarSel_1.0.tar.gz", repos = NULL, type = "source")`.

Un traitement des données est nécessaire au préalable pour les analyser (voir la vignette du package). On souhaite étudier l'influence de l'origine sur les métabolites (i) sans prendre en compte une éventuelle dépendance entre elles, puis (ii) en prenant en compte une potentielle dépendance. On cherchera à identifier les métabolites pour lesquelles il existe un effet.

1. Faire une analyse descriptive des données (ACP, boxplots, heatmap, ...).
2. Mettre en place la méthode vue en cours dans le cadre de modèles multivariés en détaillant toutes les étapes (test de blanchiment, choix du blanchiment, sélection de variable).