Projet d'étude

4 ModIA
UFs "Analyse de données" & "Eléments de modélisation statistique"
Intervenantes : Cathy Maugis-Rabusseau et Béatrice Laurent-Bonneau

2024-2025

Organisation du projet et documents à rendre

- Le projet sera réalisé par groupe de 3 étudiant-e-s. La constitution des groupes sera donnée lors de la première séance.
- 6 séances de 2h30 sont dédiées dans votre emploi du temps au travail du projet. L'une de nous sera présente lors de chacune de ces séances pour répondre à vos questions.
- Livrables : vous devrez rendre (en déposant sous Moodle) au plus tard le **jeudi 30 janvier 2025 minuit** les 2 documents suivants :
  1. un fichier Rmarkdown (*nom1-nom2-nom3-Rapport.Rmd*) contenant les codes R et générant le rapport au format pdf.
  2. un rapport au format .pdf (*nom1-nom2-nom3-Rapport.pdf*) généré par la compilation du fichier .Rmd précédent.
     Attention : le rapport est limité à 25 pages, figures incluses.
- Un dossier "ModeleRapport", disponible sur Moodle, vous donne un exemple avec des consignes pour la rédaction de votre rapport. Il est important d'en prendre connaissance dès la première séance !

Evaluation du projet

Pour chaque UF, la note de projet compte pour un tiers de la note finale de l'UF. Elle sera issue de l'évaluation des critères suivants :

Critère	**UF EMS**	**UF AD**
Utilisation pertinente des méthodes d'exploration de données		x
Utilisation pertinente de méthodes de clustering adaptées à la question traitée		x
Utilisation pertinente de l'analyse discriminante linéaire		x
Choix des modélisations ML et MLG adaptées à la question traitée	x	
Ecriture mathématique des modèles considérés	x	
Choix des procédures de tests adaptées à la question traitée	x	
Utilisation de méthodes de sélection de variables	x	
Aller-retour exploration $\leftrightarrow$ modélisation	x	x
Analyse ($\neq$ lecture !) des résultats obtenus	x	x
Choix et rendu des graphiques illustratifs	x	x
Rédaction d'un document en Rmarkdown		x
Programmation en R	x	x
Rédaction générale du document	x	x
Bonus pour des choix originaux adaptés	x	x

Jeu de données étudié

On observe pour $G = 1615$ gènes d'une plante modèle les valeurs suivantes :

$$Y_{gtsr} = \log_2(X_{gtsr} + 1) - \log_2(X_{gt_0} + 1)$$

où

- X_{gtsr} est la mesure d'expression du gène $g \in \{G1, \ldots, G1615\}$ pour le traitement $t \in \{T1, T2, T3\}$ pour le réplicat $r \in \{R1, R2\}$ et au temps $s \in \{1h, 2h, 3h, 4h, 5h, 6h\}$
- X_{gt_0} est l'expression du gène g pour un traitement de référence t_0

Dans la suite, on dira qu'un gène g est sur-exprimé si $Y_{gtsr} > 1$, sous-exprimé si $Y_{gtsr} < -1$ et non-exprimé sinon. A noter que le traitement $T3$ est une combinaison des traitements $T1$ et $T2$. Des explications plus précises du contexte seront données lors de la première séance.

Questions à aborder

Dans votre rapport final, vous devez avoir abordé par une/des méthodes adaptées les questions suivantes :

- Analyse descriptive du jeu de données
  - Faites quelques statistiques descriptives du jeu de données, étudiez les corrélations entre les différentes variables.
  - Analyse des variables Tt_sH_Rr
    - Visualisez les variables Tt_sH_Rr dans un espace de faible dimension (interprétez en particulier selon l'aspect réplicat biologique, l'effet traitement et l'effet temps)
    - Faites un clustering pour regrouper les Tt_sH_Rr en classes homogènes.
  - Analyse des gènes
    - Visualisez les données dans un espace de faible dimension
    - Réalisez un clustering des gènes ayant des profils d'expression similaires (co-exprimés) dans les différentes conditions et comparez.
- Etude de la différence entre les deux réplicats.
  - Les lois de probabilités associées aux valeurs observées pour les deux réplicats sont-elles significativement différentes ? Même question en se concentrant sur chaque traitement pris séparément.
  - Etudier l'effet combiné du temps et du traitement sur la différence des réplicats à l'aide d'un modèle linéaire.

  Dans la suite, on se concentre sur la moyenne des deux réplicats.
- Etude de la dynamique de l'expression des gènes :
  - Peut-on prévoir l'expression des gènes à 6h à partir de celle observée à 1h et du traitement considéré ? Commenter la qualité de l'ajustement et la visualiser graphiquement.
  - Reprendre la question précédente en remplaçant 1h par 3h et comparer les résultats obtenus dans les deux cas.
- Etude de l'expression des gènes pour le traitement T3 à 6h :
  - Quelles sont les variables prédictives pour le traitement T3 à 6h parmi les différents temps observés pour les traitements T1 et T2 ?
  - Peut-on prédire les gènes sur-exprimés (codés 1) et les gènes sous-exprimés (codés 0) à 6h pour le traitement T3 à partir des observations pour les traitements T1 et T2 et les heures 1 à 3 pour ces mêmes gènes ?
- Le caractère sur exprimé/sous exprimé/non exprimé des gènes à 6h dépend-il du traitement ? Même question si on se limite aux traitements T2 et T3.