TV15MI - TV25MI Projet TER : Réunion du 13/10/23

Compte-rendu

Présents : Sophie LÈBRE, Thomas AYRIVIÉ, Mehdi BELKHITER, Jamila CHERKAOUI, Dina EL HIJJAWI, Magatte LO.

Prochain rendez-vous : Vendredi 21 Octobre 2023 9h30, Université Paul Valéry, bâtiment B.

Travail à effectuer et répartition :

- Description de la liste des variables, l'analyse exploratoire du jeu de données et de la matrice des X avec notamment les cours d'analyse multidimensionnelle (ACP, AFC…).
- Formaliser la planification et la gestion de projet, comme demandé par Mme Teulère et Mr Lafaye.

À aborder au prochain rendez-vous :

- Jointure de la matrice créée avec les données patients (NCBI du package GEOquery).
- Création du Fold Change log de 10.
- Retours sur le rendu d'étape.
- GEOquery : regarder le nombre d'individus et décrire.
- Regarder comment utiliser Tidyverse librairie R qui range les données. Mme Lèbre doit nous envoyer un peu de code Tidyverse.

- Trouver la méthode la plus efficace qui nous donne les résultats les p

Points abordés :

1. Définition date/heure du prochain rendez-vous.
2. Faire la liaison des données dans Rstudio dans R.
3. Discuter des questions de recherche pour le rendu du premier rapport.
    a. Pourquoi le problème est important pour vous ?

Point de vue médical pour développer le traitement. Et d'un point de vue stats condition difficile (corrélation forte) à traiter, Mme lèbre s'interesse à comparer différentes méthodes pour la sélection des variables quand la corrélation est élevée, et proposer d'autres méthodes.

Forces et faiblesses des méthodes LASSO, elastic net (linéaire) + forêts aléatoires

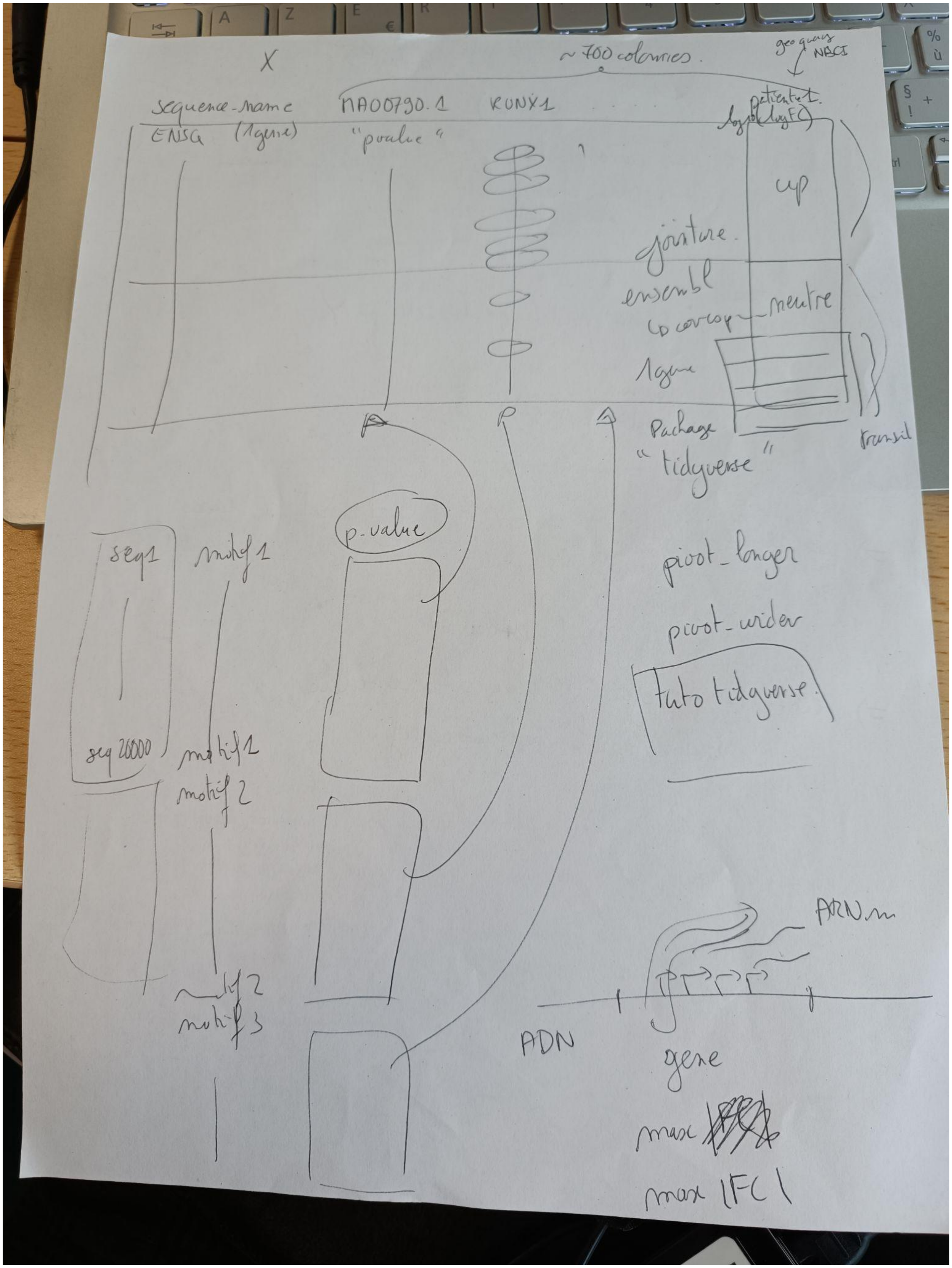
Tous types de cancer, car beaucoup de patients, beaucoup de données…

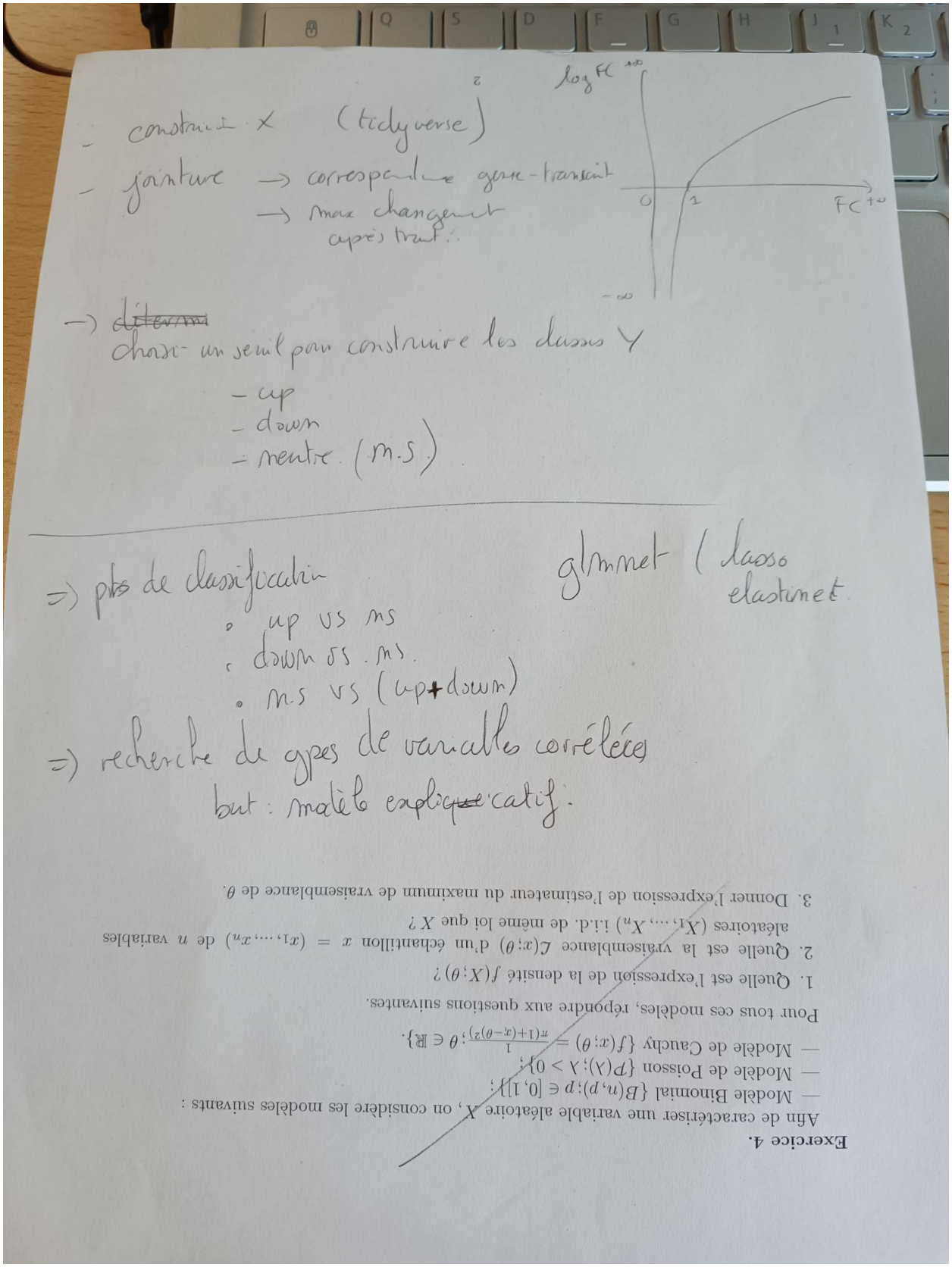
Travail des chercheurs du ircm est sur tous les cancers mais notre jeu de données contient des données du cancer du sein

b. Pourquoi est-il important de résoudre ce problème maintenant ?

- Mieux on comprends le fonctionnement du traitement, mieux on va pouvoir le donner au patient
- Plus de plus de cancers…
    c. Qu'est-ce qui vous est demandé exactement de résoudre ? Comment les cellules cancérigènes ~~résistent à la~~ chimiothérapie ?
- Identifier des ensembles de variables (scores de motifs) qui permettent de différencer les classes de gènes (ACTIVE, INIBéé, NEUTRE).
- 99 patientes différentes, un modèle pour chaque patiente
- Quelles sont les variables associées au phénomène ? On veut pas prédire quels sont les gènes pour telle patiente.
- Mots clés : CLASSIFICATION, SELECTION DE VARIABLES
- On veut un ensemble de groupes de variables associées. Ouverture : regarder ce qui est commun, différent entre les variables (Apprentissage multi-tâche)

4. Quelles sont les étapes suivantes (prochaines et plus lointaines) ?
    a. Décrire les données de séance. Décrire les scores.
    b. Fichier Jamila. Construire une matrice avec pour chaque séance (nom des colonnes), variables = p-valeurs


X
~ 700 colonnes
geo query NBCI
sequence_name
MA00790.1
RUNX1
patient 1
ENSG (1 gene)
"pvalue"
log(logFC)
up
jointure
ensemble
neutre
1 gene
Package
"tidyverse"
p-value
seq1
motif 1
seq 20000
motif 1
motif 2
motif 2
motif 3
pivot_longer
pivot_wider
tuto tidyverse
ARN m
ADN
gene
max |FC|



Exercice 4.

Afin de caractériser une variable aléatoire X, on considère les modèles suivants :

— Modèle Binomial $\{B(n,p); p \in [0,1]\}$;
— Modèle de Poisson $\{\mathcal{P}(\lambda); \lambda > 0\}$;
— Modèle de Cauchy $\{f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}; \theta \in \mathbb{R}\}$.

Pour tous ces modèles, répondre aux questions suivantes.

1. Quelle est l'expression de la densité $f(X;\theta)$?
2. Quelle est la vraisemblance $\mathcal{L}(x;\theta)$ d'un échantillon $x = (x_1,...,x_n)$ de n variables aléatoires $(X_1,...,X_n)$ i.i.d. de même loi que X ?
3. Donner l'expression de l'estimateur du maximum de vraisemblance de θ.

- Fimo construit des séances,
- Mlgl est choisi car il est récent,
- Projet pluridisciplinaire : Objectif partie stats, partie biologie (médicale),
- Quels sont les facteurs de transcriptions actifs dans le développement de certaines tumeurs ?

Classification

Chaque patiente des 89 répondent d'une maniere precise et on doit trouver l'ensemble de variable pour chacune

Les gènes up down et neutre pas de grand changement up/neutre on essaie de trouver les var associées à up et à neutre

Classif sélections trouver le groupe de variable qui ont la même info et qui sont associés.