Analyse de données en grande dimension

Céline Lévy-Leduc

Professeur de Statistique

Département MMIP (UFR de Maths)
et UMR INRA MIA Paris (Equipe Statistique & Génome)

Changements en biologie moléculaire

Conséquences de l'apparition de nouvelles techniques de séquençage :

- la résolution la plus fine a été atteinte : le nucléotide et séquençage du génome humain possible (3.5 milliards de nucléotides)
- diversité des données récoltées sur un individu (génome, transcriptome, métabolome,...) $\Rightarrow$ **données multidimensionnelles**
- mesures effectuées sur des cohortes d'individus $\Rightarrow$ **augmentation du nombre d'observations disponibles**
- apparition de **nouvelles structures de données** : données d'interactions pour la capture de la conformation chromosomique (HiC)

$\Rightarrow$ par leur dimension, hétérogénéité de nature et de structure, les données de la biologie moléculaire font partie des "big data".

Changement de paradigme

AVANT :
plus d'observations que de variables

$\Rightarrow$ vif intérêt pour les **statistiques asymptotiques** dans les cas où beaucoup d'observations et peu de variables

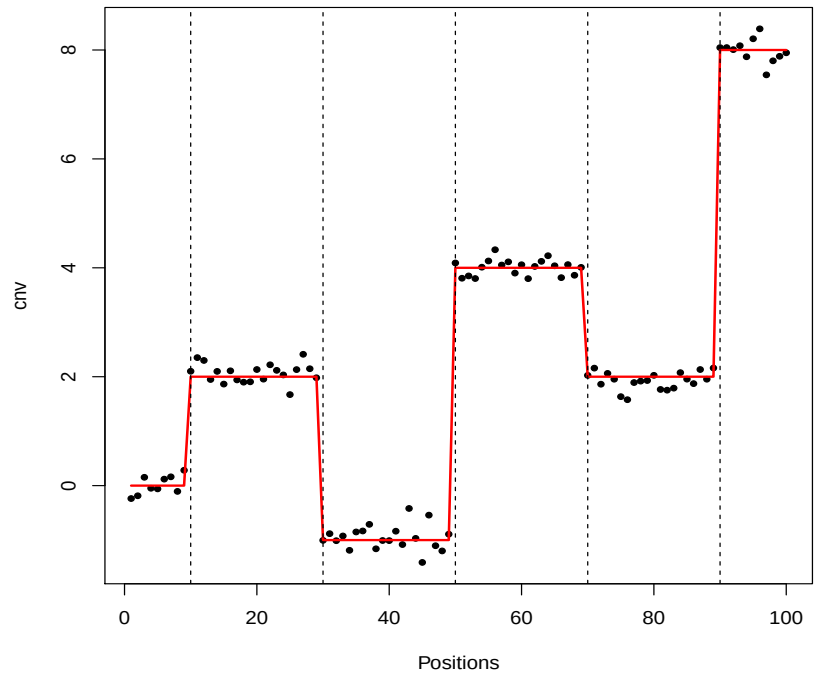
MAINTENANT :
beaucoup plus de variables que d'observations (même si elles peuvent être nombreuses)

$\Rightarrow$ Apparition de **nouvelles méthodes de statistique en grande dimension** pouvant traiter les problèmes "small n large p" : méthodes d'inférence dans les modèles parcimonieux en grande dimension.

Nouvelles thématiques de recherche en statistique

- **mise en place de nouveaux algorithmes efficaces** pour faire de l'inférence dans les modèles parcimonieux en grande dimension grâce à des méthodes dites "régularisées" à l'interface entre les statistiques et l'optimisation
- **questions théoriques** pour donner les conditions sur n, p et le degré de parcimonie sous lesquelles des estimateurs performants existent
- gestion du caractère **hétérogène** des données (données continues, données de comptage,...)
- analyse et inférence de **réseaux biologiques**, ex : inférence du réseau de régulation de gènes de plantes.

Exemple 1 : Détection de CNV



Taille du vecteur des observations $\mathbf{Y}$: $n \approx 10^6$

But : Développer une méthode de segmentation automatique

Modélisation : $\mathbf{Y} = \mathbf{T}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, où

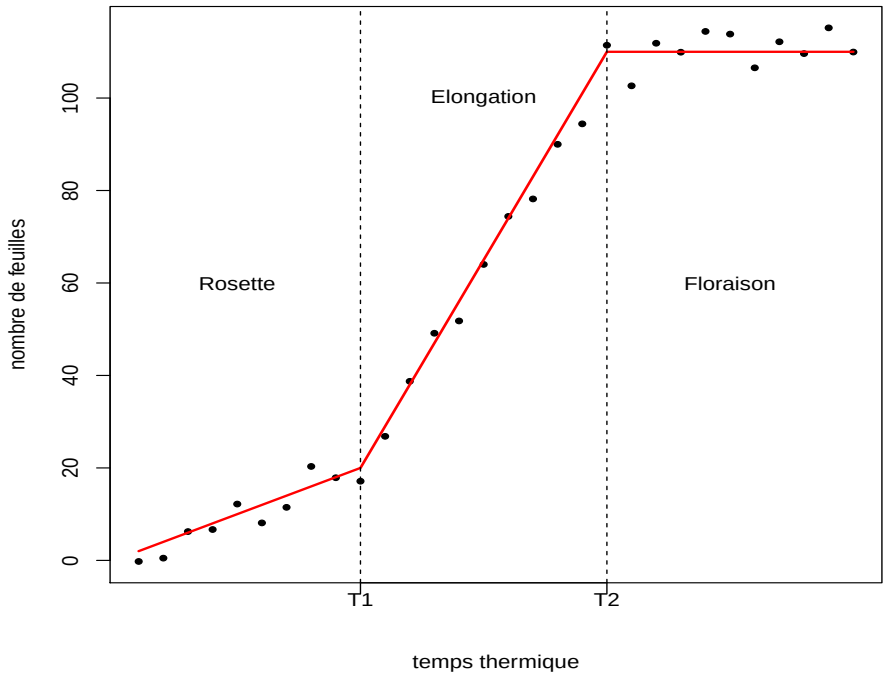
$$\mathbf{T} = \begin{pmatrix} 1 & 0 & & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & 0 \\ 1 & 1 & \cdots & & 1 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \beta_{t_1^\star} \\ 0 \\ \vdots \end{pmatrix}$$

Critère Lasso :

$$\widehat{\boldsymbol{\beta}}(\lambda) = \mathrm{Argmin}_{\boldsymbol{\beta}} \left\{ \|\mathbf{Y} - \mathbf{T}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

Les indices des composantes non nulles de $\widehat{\boldsymbol{\beta}}(\lambda)$ donnent les estimateurs des instants de rupture

Exemple 2 : Agronomie



Modélisation : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, où $\mathbf{X}$ est une matrice spécifique.

Critère Lasso :

$$\widehat{\boldsymbol{\beta}}(\lambda) = \mathrm{Argmin}_{\boldsymbol{\beta}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

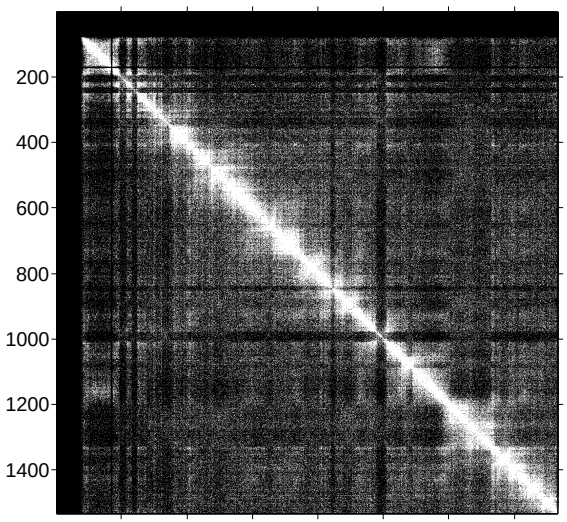
But : Développer une méthode de segmentation automatique pour trouver T_1 et T_2 pour différentes plantes.

Exemple 3 : données HiC

Question : Analyser l'organisation spatiale des chromosomes pour mieux comprendre le processus de régulation des gènes.

Chromosome 19 (mouse cortex)

$n = 1534$



2.3 millions de données à analyser
But : Segmentation automatique de la matrice en blocs.

Modélisation :

$$\mathbf{Y} = \mathbf{TBT}' + \mathbf{E}$$

où

- $\mathbf{B}$ ne contient que des valeurs nulles sauf aux positions des ruptures.
- $\mathbf{T}$: matrice triangulaire inférieure ne contenant que des 1 sous la diagonale

Exemple 4 : Métabolomique et écologie

Question : Quelles sont les métabolites qui caractérisent un écotype donné ?
Modélisation :

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

où

- $\mathbf{Y}$: $n \times q$, où $n \approx 100$ et $q \approx 10000$
- $\mathbf{X}$ est la matrice de design d'une ANOVA à un facteur
- $\mathbf{B}$ contient les effets des différentes modalités de la variable qualitative "Ecotype" sur les différentes métabolites

Exemple 5 : Estimation de l'héritabilité dans les modèles linéaires mixtes en grande dimension

On souhaite estimer l'héritabilité d'un trait quantitatif :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

où :

- $\mathbf{X}\boldsymbol{\beta}$ correspond aux effets fixes
- $\mathbf{Z}$: $n \times N$, $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2}\mathrm{Id}_{\mathbb{R}^N})$, $n \approx 1000$, $N \approx 300000$
- $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n})$

On cherche à estimer :

$$\eta^\star = \frac{Nq\sigma_u^{\star 2}}{Nq\sigma_u^{\star 2} + \sigma_e^{\star 2}},$$

où q est la proportion de composantes non nulles dans $\mathbf{u}$.

Conclusion

- Avalanche de données souvent très hétérogènes dans beaucoup de domaines
- Nécessité de formuler une question d'intérêt et un modèle statistique pour pouvoir y répondre.