Cluster elastic net pour la régression en grande dimension avec des blocs inconnus de variables

Mémoire de master

Mikhaël Presley KIBINDA-MOUKENGUE
(mikhael.p.k.moukengue@aims-senegal.org)
African Institute for Mathematical Sciences (AIMS), Senegal
Supervisé par : Professeur Ndeye Niang KEITA
Centre d'études et de recherche en informatique et communications (Cedric)
CNAM, France

June 25, 2020

Déclaration

Ce travail a été effectué à AIMS Sénégal en remplissant pleinement les conditions d'obtention d'un Master en Science.

Je déclare par la présente que, sauf mention contraire, ce travail n'a jamais été présenté en tout ou en partie pour l'obtention d'un diplôme à AIMS Sénégal ou dans une autre université. Et que tout travail effectué par d'autres ou par moi-même précédemment a été reconnu et référencé en conséquence.

Étudiant : **Mikhaël Presley KIBINDA-MOUKENGUE** 

Superviseur(e) : **Ndeye Niang KEITA** (scan) supervisor's signature

Tuteur : **Jean Claude Utazirubanda** (scan) tutor's signature

Remerciements

Dédicaces

Résumé

La régression en grande dimension est une régression dans le cas où le nombre de variables est plus grand que celui d'observations. On considère un jeu de donnée **X** de dimension $n \times p$ et une variable réponse **y** de dimension $n \times 1$. Si on veut appliquer un modèle de régression linéaire en présence de grande dimension sur **X**, l'utilisation de la methode des moindres carrées[19] est limitée. Car la matrice **X** est singulière en grande dimension. Pour remédier à ce problème, on utilise les méthodes dites de régularisation à savoir : Ridge, Lasso, Elastic Net etc. Dans ce mémoire, nous étudions la méthode de régularisation du Cluster Elastic Net[30] proposée par Daniela M. Witten et al., (2014). C'est une méthode de régularisation qui réduit sélectivement les coefficients. L'idée de ce mémoire est de comparer les méthodes dans le cas où on a a priori la structuration en blocs des variables et celles dans le cas où c'est la méthode qui trouve les groupes de variables; autrement dit, faire une étude critique entre celles-ci. Ma contribution a été de faire une synthèse de l'article, en éclaircissant les parties qui étaient laissées au lecteur, et j'ai apporté des démonstrations notamment sur les lemmes et propositions présents dans l'article. Ensuite j'ai mis en place une simulation numérique dans laquelle j'ai pu comparer les résultats statistiques donnés par la méthode du cluster elastic net à ceux donnés par d'autres méthodes de regréssion régularisée disposant des groupes de variables a priori.

Contents

1. Introduction

De nos jours, près de 80%[24] des données disponibles sont des données non structurées et viennent continument. En effet, ces données sont très massives et constituent une partie intégrante dans la recherche de l'information et dans l'interprétation. En génomique par exemple, on collecte des grands volumes de données, facteurs de certaines maladies héréditaires pour mieux comprendre et même pour prévenir ou corriger ces maladies. Face à ces données volumineuses, l'étude statistique sur celles-ci rencontre souvent de nombreux défits à savoir : des jeux de données de grandes dimensions, plusieurs gènes présents dans certaines maladies ne sont pas souvent informatifs et il y en a ceux qui sont étroitement voir même fortement liés à d'autres. Lors d'une étude statistique, l'un des objectifs est souvent de prédire les symptômes liés à la maladie en construisant un modèle de régression sur des prédicteurs (gènes). Supposons que ces gènes forment une combinaison linéaire, effectuer une régression linéaire c'est trouver une fonction qui soit combinaison linéaire des prédicteurs (gènes) du modèle, l'enjeu majeur consiste de trouver le bon vecteur de poids qui minimise la fonction coût (fonction de l'erreur) du modèle statistique. Lorsqu'on fait une régression linéaire, la méthode couramment utilisée pour déterminer les coefficients du vecteur des poids c'est la méthode des moindres carrées[19]. Mais en présence d'un jeu de données de grande dimension[5], et lorsque les prédicteurs sont corrélés, cette méthode présente des limites car il y'a une singularité de la matrice du jeu de données. Pour régler ce problème, il existe des techniques de pénalisation (régularisation)[20] qui consistent à contrôler simultanément l'erreur sur le jeu test du modèle et la complexité du modèle, cela afin d'éviter un sur-apprentissage (overfit)[28]. Parmi lesquelles on a :
La régularisation de Ridge[12], cette régularisation utilise la norme $\mathcal{L}^2$ pour pénaliser les coefficients de régression, elle donne toujours une solution analytique et fait une sélection groupée : les variables (prédicteurs) corrélées ont le même coefficient; dans la prédiction, elle donne une variance moins élevée dans l'étude. Par contre c'est une méthode non appropriée pour la sélection des variables. En effet, si les prédicteurs sont fortement corrélés entre eux, leurs coefficients seront très proches les uns des autres; cette méthode ne pénalise pas les variables non pertinentes par des coefficients exactement nuls (pas de parcimonie);

La régularisation de LASSO (Least Absolute Shrinkage and Selection Operator)[29], elle permet de réduire de plus en plus les coefficients jusqu'à les annuler afin d'avoir un modèle parcimonieux. Comme son nom l'indique, cette méthode fait une réduction de dimension et sélection des variables. Mais c'est une méthode non appropriée pour la sélection des groupes de variables. En effet, si des prédicteurs sont fortement corrélés entre eux, cette méthode choisit un prédicteur et pénalise les autres avec des coefficients nuls, cette régularisation donne une estimation de la variance beaucoup plus élevée;

La régularisation de Elastic Net[31], cette régularisation combine la régularisation de Ridge et celle de LASSO, elle permet en plus d'avoir un modèle parcimonieux, toutes les variables corrélées ont un coefficient similaire. Elle lève la limite sur Ridge et sur Lasso, notamment en traitant le cas des groupes de variables.

Dans la vie réelle, surtout dans le domaine de la génomique, les gènes présentent souvent des fortes

liaisons entre-eux et forment des groupes de variables en fonction de ces liaisons. Et nous avons souvent une structure des jeux de données $\mathbf{X}$[1] en blocs ou groupes partageant les caractéristiques communes au sein de chaque groupe. Devant ce genre de données, les pénalisations de Ridge, de Lasso et de Elastic Net n'opèrent pas normalement. Plusieurs propositions d'algorithmes ont été faites pour effectuer des regroupements et ensuite effectuer une régression en utilisant les résultats des regroupements notamment par les auteurs (Hastie et al. 2001)[11], (Dettling et Buhlmann 2004)[8] et (Park et al. 2007)[21]...
Dans ce mémoire, nous nous intéresserons principalement à la méthode appelée Cluster Elastic Net avec les groupes non connus, proposée par (Daniela M. Witten et al 2014)[30]. Cette méthode recherche des ensembles de caractéristiques corrélées ayant des associations similaires avec la variable réponse (variable à prédire); cela est particulièrement avantageux si certaines caractéristiques corrélées, mais pas toutes, ont une association similaire avec la réponse $\mathbf{y}$[2]. C'est une méthode qui trouve simultanément les coefficients de régression et des groupes des variables corrélées, celles qui sont similaires avec la variable réponse $\mathbf{y}$. Nous travaillons dans le cas où le nombre de variables p est strictement plus grand que celui d'observations n, $(p > n)$[3].

Enfin, nous ferons une étude statistique sur des données simulées constituées de 120 observations et 500 variables afin de comparer la méthode du Cluster Elastic Net avec groupes non connus à celles qui disposent des groupes de variables a priori.

Ce mémoire est organisé de la manière suivante : Au chapitre (1), l'introduction. Au chapitre (2), nous présentons les rappels et définitions des concepts utilisés. Au chapitre (3), nous présentons le problème d'optimisation du Cluster Elastic Net ainsi que ses propriétés. Au chapitre (4), nous présentons l'algorithme de résolution du problème Cluster Elastic Net, nous présentons la sélection des paramètres de régularisation ainsi que la technique de la validation croisée. Au chapitre (5), nous faisons une analyse du rétrécissement des variables. Au chapitre (6), nous étudions la relation du Cluster Elastic Net avec d'autres approches. Au chapitre (7), nous faisons une étude sur les données de simulations (120 observations et 500 variables), nous présentons une conclusion et les perspectives au chapitre (8), enfin nous présentons en annexe les démonstrations de quelques lemmes et propositions, ainsi que les références sur des ouvrages utilisés pour l'élaboration de ce mémoire.

[1] Matrix design, de dimension $n \times p$
[2] Variable à prédire, c'est un vecteur colonne de dimension $n \times 1$
[3] Dans le cadre de grande dimension

2. Rappels et définitions

2.1 La régression

Soit X une matrice de dimension $n \times p$ et y un vecteur colonne de dimension $n \times 1$. Faire de la régression sur X c'est trouver une relation entre y et X qui est décrite de la manière suivante :

$$y = f(x) + \epsilon \tag{2.1.1}$$

où f est une fonction inconnue à estimer et ϵ est un bruit ou un terme d'erreur.

2.1.1 Pourquoi estimer la fonction f ?

Dans la littérature, il existe plusieurs approches linéaires et non linéaires, paramétriques et non paramétriques, pour estimer la fonction f. Le modèle d'apprentissage contient les couples $\{(x_i, y_i)\}_{i=1,...,p}$, où y_i est la variable réponse pour l'observation i et $x_i = (x_{i_1}, ..., x_{i_p}) \in \mathbb{R}^p$ sont les p-prédicteurs.

Le but c'est de chercher une fonction $\widehat{f}$, telle que $\widehat{y}_i = \widehat{f}(x_i)$ pour toutes observations (x_i, y_i). Pour obtenir une telle fonction, on minimise une fonction de perte $L\{y_i, f(x_i)\}$[14], qui mesure les écarts entre les observations y_i et le modèle $f(x_i)$. Il existe différentes formes de fonction de perte à savoir la différence en norme $\mathcal{L}^1$, en norme $\mathcal{L}^2$ etc. Comme illustration, la fonction de perte avec la norme $\mathcal{L}^2$ s'écrit de la manière suivante :

$$L(y, f(x)) = \|y - f(x)\|^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2,$$

avec $f(x) = (f(x_1), ..., f(x_n))$.

Il existe deux grandes familles de méthodes qui permettent d'estimer la fonction f : les méthodes paramétriques et les méthodes non paramétriques.

2.1.2 Méthodes paramétriques

Ces méthodes se déroulent en deux étapes : une première étape dans laquelle on spécifie le modèle et une deuxième étape pour estimer les paramètres du modèle.

Etape 1 : On suppose qu'il existe une relation linéaire entre y et les x_i et f est définie par :

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon,$$

avec $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t$.

C'est un modèle de régression linéaire qui suppose que f est linéaire en β_j, avec $j \in \{1, ..., p\}$.

Alors, estimer la fonction f revient à estimer les paramètres $\beta_0, \beta_1, ..., \beta_p$, ainsi que la variance des erreurs $\epsilon_1, \epsilon_2, ..., \epsilon_n$.

Etape 2 : Après le choix du modèle, on a besoin d'une approche qui nous permet d'estimer les paramètres $\beta_0, \beta_1, ..., \beta_p$ qui minimisent la fonction de perte L.

Pour ce genre de problème, l'approche la plus utilisée est celle des moindres carrés[2] où la fonction de perte L est définie par : $L(y, f(x)) = (y - f(x))^2$.

En effet, si la matrice de variance-covariance est de plein rang, cette méthode donne des solutions explicites pour les paramètres du modèle et fonctionne bien si nous avons une relation linéaire entre la variable réponse et les prédicteurs. Dans le cas contraire, il existe plusieurs autres méthodes d'apprentissage statistique.

2.1.3 Méthodes non paramétriques (cas de la régression locale)

Ces méthodes ne font pas d'hypothèses explicites sur la forme fonctionnelle de f. La régression locale est une approche non paramétrique qui consiste à estimer la fonction f en différents points x_0, en tenant compte des observations x_i $(i = 1, 2, ..., n)$ qui sont proches de $x_0 \in \mathbb{R}^p$.

Comme exemple, on pose $p = 1$ (un seul prédicteur), on peut définir la régression locale comme suit :
On considère un prédicteur $x = (x_1, x_2, ..., x_n)$ et une variable réponse $y = (y_1, y_2, ..., y_n)$.
Pour estimer la fonctionnelle f au point x_0, on affecte des poids aux observations x_i, qui sont proches de x_0. De tels points sont notés $K_i = K(x_i, x_0)$ et pour les points qui sont éloignés de x_0 on leur affecte un poids égal à zéro. Un exemple pour les poids K_i est le noyau uniforme défini par :

$$K_i = \frac{1}{2} \mathbb{1}_{\left\{ \left| \frac{x_0 - x_i}{h} \right| \leqslant 1 \right\}}, \quad i = 1, 2, ..., n.$$

Avec $h > 0$, le pas de l'intervalle à choisir de manière judicieuse. La prochaine étape consiste à appliquer la méthode des moindres carrés pondérés pour estimer les paramètres du modèle.

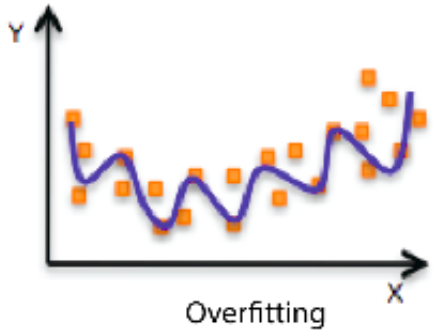
En effet, on cherche le couple $(\widehat{\beta}_0, \widehat{\beta}_1)$ qui minimise la somme suivante :

$$\sum_{i=1}^{n} K_i (y_i - \beta_0 - \beta_1 x_i)^2,$$

et l'estimé de la fonction f en x_0 est donné par la forme :

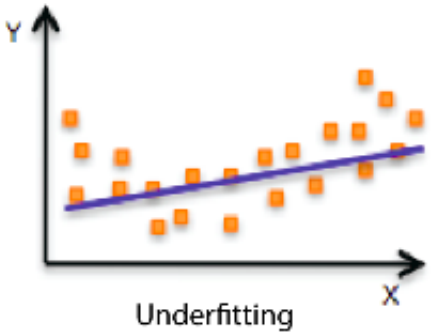
$$\widehat{f}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

2.1.4 Overfitting ou surapprentissage



On dit qu'il y'a overfitting (ou surapprentissage) lorsque le modèle fonctionne bien sur les données d'entrainement mais ne fonctionne pas bien sur les données d'évaluation. Cela s'explique par le fait que le modèle mémorise les données qu'il a vues et qu'il est incapable de généraliser sur les nouvelles données.

2.1.5 Underfitting ou sous-apprentissage



On dit qu'il y'a underfitting (ou sous-apprentissage) lorsque le modèle fonctionne mal sur les données d'apprentissage. Cela est dû au fait que le modèle est incapable de saisir la relation entre les prédicteurs X et les variables cibles Y.

2.2 Méthodes de regroupement

L'objectif c'est de trouver une partition des données en groupes distincts de sorte que les observations au sein de chaque groupe soient plus homogènes possibles les unes aux autres.

Il existe deux grandes familles des méthodes de regroupements[13]: les méthodes basées sur la distance et les méthodes basées sur des modèles.

2.2.1 Méthodes basées sur la distance

✓ **Le K-means clustering** (partitionnement direct ou non hiérarchique) : On cherche à partitionner les observations en un nombre prédéfini K, de clusters ou groupes. L'idée derrière le clustering K-means est qu'un bon clustering est celui pour lequel l'inertie intra-groupe qui est la moyenne des inerties locales aux classes soit le faible possible, c'est à dire la minimisation de l'inertie intra-groupe.

Algorithme de K-means

1. Sélectionnez k points au hasard comme centres de groupes.

2. Affecter les individus à leur centre de groupe le plus proche en utilisant la distance euclidienne, cosinus carré ou la corrélation etc.

3. Calculer le centroïde ou la moyenne de tous les individus de chaque groupe.

4. Répétez les étapes 2 et 3 jusqu'à atteindre la convergence.

La convergence est atteinte si aucun individu ne change de classe, si l'inertie intra-classe ne diminue plus, ou bien lorsque les centroïdes sont stables d'une itération à une autre.

✓ **Le regroupement hiérarchique** : Dans ce partionnement, on ne sait pas à l'avance combien de regroupements nous voulons. On obtient une représentation visuelle arborescente des observations, appelée dendrogramme, qui permet de visualiser immédiatement les regroupements obtenus pour chaque nombre possible de grappes, de 1 à n. Il existe deux méthodes de regroupement hiérarchique : les méthodes ascendantes et les méthodes descendantes.

- **La méthode de classification ascendante**, construit une hiérarchie entière qui prend progressivement la forme d'un arbre ou d'un dendrogramme en respectant un ordre ascendant. L'analyse débute en considérant chaque individu comme une classe et tente de fusionner ensuite deux ou plusieurs classes de manière appropriée pour former une nouvelle classe (Boullé et al., 2012). Des petites classes ne comprenant que des individus très semblables sont constituées, puis des classes de moins en moins homogènes sont construites jusqu'à obtenir la classe tout entière, c'est-à-dire l'échantillon total. L'arbre qui en résulte peut potentiellement être coupé à différents niveaux. Il résulte de ce choix un nombre de classes plus ou moins important. Si le principe général reste identique, le processus est toutefois inversé lors d'une classification descendante hiérarchique;

- **La méthode de classification descendante**, dans cette classification, une seule classe regroupant tous les individus est divisée pas à pas en classes de moindres effectifs jusqu'à l'obtention d'une classe par individu ou bien du nombre de classes souhaité a priori.

2.2.2 Méthodes basées sur des modèles

Ces méthodes sont basées sur une fonction de densité ou de connectivité. Par exemple, modèle de mélange, regroupement probabiliste, etc.

La figure (2.1) explique brièvement les familles des méthodes de regroupement et certaines techniques spécifiques.

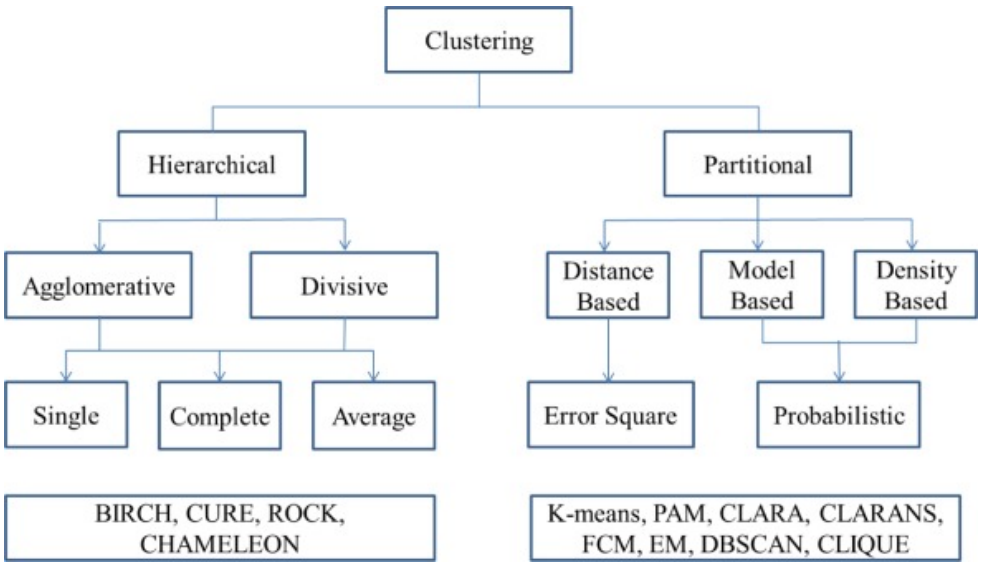


Figure 2.1: méthodes de regroupement[18]

2.3 Optimisation numérique

Un problème d'optimisation c'est un problème qui met en évidence des variables d'état (paramètres) et des contraintes. Le but est de minimiser ou maximiser une fonction $f : \mathcal{C} \longrightarrow \mathbb{R}$ appelée fonction objectif.
Un problème d'optimisation peut être exprimé de la manière suivante[6] :

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{sous la contrainte : } x \in \mathcal{C},$$

où $\mathcal{C}$ (appelé ensemble des contraintes) est un sous-ensemble de $\mathbb{R}^n$.

On peut écrire aussi :

$$\min_{x \in \mathcal{C}} f(x)$$

2.3.1 Convexité

Soit X un espace de Hilbert[4] et $f : X \longrightarrow \mathbb{R}$, une fonction donnée.

i) On dit que f est convexe si : $\forall t \in [0,1],\ \forall x, y \in X,\ f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$,

ii) On dit que f est strictement convexe si :

$\forall t \in (0,1),\ \forall x, y \in X,\ x \neq y \Longrightarrow f(tx + (1-t)y) < tf(x) + (1-t)f(y)$.

Remarque : Si la fonction f n'obéit pas aux conditions i) et ii), alors f est une fonction non convexe.

3. Cluster Elastic Net

3.1 Problème d'optimisation du Cluster Elastic Net

Soit X une matrice de dimension $n \times p$ et $y \in \mathbb{R}^n$ un vecteur de variable réponse, avec n le nombre d'observations et p le nombre de variables de notre jeu de donnée X.

Le problème de régression est défini par l'équation (3.1.1) ci-dessus :

$$Y = \beta_1 \mathbf{x}_1 + \ldots + \beta_p \mathbf{x}_p + \epsilon \tag{3.1.1}$$

ou bien,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_1 \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \ldots + \beta_p \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

où β est un vecteur coefficient de régression et ϵ est un vecteur aléatoire de termes de bruit non corrélé avec une moyenne de 0 et une variance commune σ^2. On suppose que la variable réponse y est centrée donc de moyenne nulle et de variance commune σ^2.
$X_j \in \mathbb{R}^n$ représente la j-ième colonne de la matrice X, et on supposera aussi que les colonnes de la matrice X ont été toutes normalisées c'est à dire ont une moyenne 0 et une norme de $\mathcal{L}^2$ égale à 1:

$$\sum_i X_{ij} = 0 \text{ et } \sum_i X_{ij}^2 = 1$$

En général, nous allons supposer que nous nous trouvons dans le cadre de grande dimension et à faible densité dans lequel $p > n$, mais la majorité des covariables ne sont pas associées au résultat, c'est-à-dire $\beta_j = 0$ pour $j = \overline{1, p}$. Nous annonçons donc les hypothèses supplémentaires suivantes :

Hypothèse 3.1.1. Il existe des groupes de variables inconnus et K groupes de variables distincts inconnus, avec des niveaux modérés ou élevés de corrélation (absolue) entre les variables au sein d'un groupe, et peu ou pas de corrélation entre les groupes.

Hypothèse 3.1.2. Les variables qui se trouvent dans le même groupe ont une association similaire avec la réponse. Si X_j et X_l appartiennent au même groupe, alors $X_j\beta_j$ et $X_l\beta_l$ prennent des valeurs similaires.

Les hypothèses (3.1.1)-(3.1.2) indiquent qu'il existe des groupes inconnus parmi les variables et que la connaissance de ces groupes nous permettrait d'estimer plus précisément le coefficient de régression β.

Le cluster elastic net est défini comme solution au problème d'optimisation (3.1.2) suivant:

$$\min_{C_1,...,C_K,\beta}\{\|y - X\beta\|^2 + \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2\} \tag{3.1.2}$$

Ici δ et λ sont des paramètres ou coefficients de régularisation, et $C_1, ..., C_K$ désignent une partition des p caractéristiques en $K - groupes$, tels que:

$$C_k \cap C_l = \emptyset \; et \; C_1 \cup C_2 \cup ... \cup C_k = \{1, ..., p\},$$

$\|\beta\|_1$ est le terme de pénalisation de lasso, qui rend le modèle parcimonieux quand δ est grand.

Le terme de cluster pénalité peut aussi être écrit sous la forme:

$$\frac{1}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2 = \sum_{k=1}^{K}\sum_{j\in C_k}\|X_j\beta_j - \frac{1}{|C_k|}\sum_{l\in C_k}X_l\beta_l\|^2 \tag{3.1.3}$$

Remarque 1. Dans l'expression (3.1.2), si $\delta = 0$ on a le cluster ridge régression (cas special du cluster elastic net) en raison des similitudes entre la pénalité de ridge et la pénalité de cluster et si $\lambda = 0$ on a simplement une régularisation de lasso.

3.2 Propriétés du Cluster Elastic Net

Propriété 3.2.1. Si $K = p$, de sorte que chaque variable constitue son propre groupe, alors le cluster elastic net se réduit exactement au lasso.

Preuve : Soit $P(\beta) = \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2$, la pénalité du cluster elastic net.
Alors pour $K = p$,

$$P(\beta) = \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{p}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2,$$

Du fait que chaque prédicteur constitue son propre groupe, nous aurons au total p-groupes ce qui équivaut à p-prédicteurs. Comme chaque variable constitue son propre groupe, alors le CEN[1] se réduit exactement au $LASSO$.

Les indices j et l étant pris dans C_k, $(k = \overline{1,p})$ et du fait que chaque variable constitue son

[1]CEN : Cluster Elastic Net

propre groupe, alors j et l coïncident. Ceci dit: $X_j\beta_j - X_l\beta_l = 0$, alors le terme:

$$\frac{\lambda}{2}\sum_{k=1}^{p}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2 = 0.$$

Il en résulte que $P(\beta) = \delta\|\beta\|_1$.

On vient de montrer que pour $K = p$, faire du CEN est équivalent à faire du $LASSO$. ■

Propriété 3.2.2. Si $K = 1$, de sorte que toutes les caractéristiques soient dans le même groupe, alors Cluster Elastic Net est équivalent à elastic net.

Preuve : Soit $P(\beta) = \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2$, la pénalité du CEN.

En utilisant la relation (3.1.3), $P(\beta)$ peut s'écrire sous la forme:

$$P(\beta) = \delta\|\beta\|_1 + \lambda\sum_{k=1}^{K}\sum_{j\in C_k}\|X_j\beta_j - \frac{1}{|C_k|}\sum_{l\in C_k}X_l\beta_l\|^2.$$

Pour $K = 1$, on aura:

$$P(\beta) = \delta\|\beta\|_1 + \lambda\sum_{j\in C_1}\|X_j\beta_j - \frac{1}{|C_1|}\sum_{l\in C_1}X_l\beta_l\|^2$$

On fait un changement de variable, passage à la version à l'échelle de X et y. On suppose que $\lambda < p$, on pose : $\widetilde{X} = X/\sqrt{1-\frac{\lambda}{p}}$ et $\widetilde{y} = y/\sqrt{1-\frac{\lambda}{p}}$ on a :

$$P(\beta) = \delta\|\beta\|_1 + \lambda(1-\frac{\lambda}{p})\sum_{j\in C_1}\|\widetilde{X}_j\beta_j - \frac{1}{|C_1|}\sum_{l\in C_1}\widetilde{X}_l\beta_l\|^2, \tag{3.2.1}$$

Or,

$$\begin{aligned}\sum_{j\in C_1}\|\widetilde{X}_j\beta_j - \frac{1}{|C_1|}\sum_{l\in C_1}\widetilde{X}_l\beta_l\|^2 &= \sum_{j\in C_1}(\widetilde{X}_j\beta_j - \frac{1}{|C_1|}\sum_{l\in C_1}\widetilde{X}_l\beta_l)^T(\widetilde{X}_j\beta_j - \frac{1}{|C_1|}\sum_{l\in C_1}\widetilde{X}_l\beta_l),\\ &= \sum_{j\in C_1}\beta_j^T\widetilde{X}_j^T\widetilde{X}_j\beta_j - \frac{1}{|C_1|}\sum_{j,l\in C_1}\beta_j^T\widetilde{X}_j^T\widetilde{X}_l\beta_l - \frac{1}{|C_1|}\sum_{j,l\in C_1}\beta_l^T\widetilde{X}_l^T\widetilde{X}_j\beta_j\\ &\quad + \frac{1}{|C_1|}\sum_{l\in C_1}\beta_l^T\widetilde{X}_l^T\widetilde{X}_l\beta_l,\end{aligned}$$

Les variables sont corrélées et sont toutes dans un même groupe, on est bien dans le même contexte que Elastic Net, qui favorise l'effet groupé sur les données. Sur ce, comme les colonnes

de la matrice X ont été toutes normalisées, il en est de même pour la matrice $\widetilde{X}$. Alors, pour $j, l \in C_1$; $\widetilde{X}_l^T \widetilde{X}_l = 1$ et $\widetilde{X}_j^T \widetilde{X}_j = 1$. De plus, comme toutes les caractéristiques sont dans le même groupe, on peut dire que :

$$\sum_{j \in C_1} \|\widetilde{X}_j \beta_j - \frac{1}{|C_1|} \sum_{l \in C_1} \widetilde{X}_l \beta_l\|^2 \approx \sum_{j \in C_1} \beta_j^2 + \frac{1}{|C_1|} \sum_{l \in C_1} \beta_l^2 \tag{3.2.2}$$

En remplaçant le terme (3.2.2) dans (3.2.1) on a :

$$P(\beta) = \delta \|\beta\|_1 + \lambda(1 - \frac{\lambda}{p}) \sum_{j \in C_1} \beta_j^2 + \frac{\lambda}{|C_1|}(1 - \frac{\lambda}{p}) \sum_{l \in C_1} \beta_l^2,$$

Le terme $\sum_{j \in C_1} \beta_j^2 + \sum_{l \in C_1} \beta_l^2$, est un terme équivaut à une pénalité de Ridge sur un sous-ensemble des variables. On pose : $\lambda(1 - \frac{\lambda}{p}) \sum_{j \in C_1} \beta_j^2 + \frac{\lambda}{|C_1|}(1 - \frac{\lambda}{p}) \sum_{l \in C_1} \beta_l^2 = \gamma \|\beta\|^2$, avec :

$$\gamma = max(\lambda(1 - \frac{\lambda}{p}), \frac{\lambda}{|C_1|}(1 - \frac{\lambda}{p})) > 0.$$

Alors nous obtenons, $P(\beta) = \delta \|\beta\|_1 + \gamma \|\beta\|^2$ c'est la pénalité de Elastic Net.

D'où on vient de montrer que si $K = 1$, le Cluster Elastic Net est équivalent à Elastic Net.

On montre aussi que pour les valeurs intermédiaires de K; $1 < K < p$ on a :

$$P(\beta) = \delta \|\beta\|_1 + \frac{\lambda}{2} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{j,l \in C_k} \|X_j \beta_j - X_l \beta_l\|^2,$$

Or,

$$\begin{aligned} \sum_{j,l \in C_k} \|X_j \beta_j - X_l \beta_l\|^2 &= \sum_{j,l \in C_k} (X_j \beta_j - X_l \beta_l)^T (X_j \beta_j - X_l \beta_l), \\ &= \sum_{j,l \in C_k} [\beta_j^T X_j^T X_j \beta_j - \beta_j^T X_j^T X_l \beta_l - \beta_l^T X_l^T X_j \beta_j + \beta_l^T X_l^T X_l \beta_l] \end{aligned}$$

Le cluster elastic net aboutira à la mise en commun des coefficients de régression des variables au sein d'un même groupe à condition que ces variables soient corrélées.

On pose r_{jl}[2] $= X_j^T X_l = X_l^T X_j, \forall j, l \in C_k$,
Alors on a:

$$\sum_{j,l \in C_k} \|X_j \beta_j - X_l \beta_l\|^2 = \sum_{j,l \in C_k} [\beta_j^T X_j^T X_j \beta_j - \beta_j^T r_{jl} \beta_l - \beta_l^T r_{jl} \beta_j + \beta_l^T X_l^T X_l \beta_l],$$

[2]Coefficient des j-èmes et l-èmes variables au sein d'un groupe

En tenant compte de la normalisation des colonnes de la matrice X, nous avons pour tout $j, l \in C_k$, $X_l^T X_l = 1$ ainsi que $X_j^T X_j = 1$ on a :

$$\begin{aligned}\sum_{j,l \in C_k} \|X_j\beta_j - X_l\beta_l\|^2 &= \sum_{j,l \in C_k} [\beta_j^T\beta_j - \beta_j^T r_{jl}\beta_l - \beta_l^T r_{jl}\beta_j + \beta_l^T\beta_l], \\ &= \sum_{j,l \in C_k} [\beta_j^2 - \beta_j r_{jl}\beta_l - \beta_l r_{jl}\beta_j + \beta_l^2], \\ &= \sum_{j,l \in C_k} [\beta_j^2 - 2r_{jl}\beta_j\beta_l + \beta_l^2], \end{aligned} \tag{3.2.3}$$

En ajoutant et en retranchant les termes $r_{jl}\beta_j^2$ et $r_{jl}\beta_l^2$ dans (3.2.3), puis en faisant une simple factorisation $P(\beta)$ devient :

$$P(\beta) = \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l \in C_k}[(1 - r_{jl})(\beta_j^2 + \beta_l^2) + r_{jl}(\beta_j - \beta_l)^2] \tag{3.2.4}$$

Par conséquent, si les variables j-ièmes et l-ièmes sont dans le même groupe et que r_{jl} est grand d'une part, alors le terme $(\beta_j - \beta_l)^2$ dans (3.2.4) dominera et le cluster elastic net réduira β_j et β_l l'un vers l'autre. D'autre part, si r_{jl} est proche de zéro, alors $(\beta_j^2 + \beta_l^2)$ un terme équivaut à une pénalité de Ridge (sur un sous-ensemble des variables) dominera et le cluster elastic net réduira β_j et β_l vers zéro. Et si r_{jl} est négatif alors:

$$(1 - r_{jl})(\beta_j^2 + \beta_l^2) + r_{jl}(\beta_j - \beta_l)^2 = (1 - |r_{jl}|)(\beta_j^2 + \beta_l^2) + |r_{jl}|(\beta_j - \beta_l)^2$$

En d'autres termes, en fonction des corrélations entre les variables au sein d'un groupe, les variables seront soient rétrécies les unes vers les autres, soient elles vont tendre vers zéro (parcimonie), soient elles seront rétrécies les unes vers les autres en valeurs absolues mais avec des signes opposés.

Nous constatons aussi à partir de l'équation (3.1.2) que les variables (absolument) corrélées qui sont associées à la réponse c'est à dire les variables pour lesquelles $X_j\beta_j \approx X_l\beta_l, \forall j, l \in C_k$, appartiennent dans le même groupe, car cela entraîne une diminution de leurs coefficients et donc des valeurs faibles.
En outre, on constate que :

$$\frac{1}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l \in C_k}[\beta_j^2 - 2r_{jl}\beta_j\beta_l + \beta_l^2] = \beta^T M\beta \tag{3.2.5}$$

Où M est une matrice définie semi-positive de la forme :

$$M_{jl} = \begin{cases} (|C_k| - 1)/|C_k| & si\ \ j = l \in C_k \\ -r_{jl}/|C_k| & si\ \ j \neq l\ \ et\ \ j, l \in C_k \\ 0 & sinon \end{cases} \tag{3.2.6}$$

M_{jl} est la (j,l)-ième entrée de M.

Par suite, le problème d'optimisation pour le cluster elastic net avec les groupes connus peut être équivalent à :

$$\min_{\beta \in \mathbb{R}^{p+1}} \{\|y - X\beta\|^2 + \delta\|\beta\|_1 + \lambda\beta^T M\beta\} \tag{3.2.7}$$

La matrice semi-positive M définie dans (3.2.6), est une matrice qui réduit effectivement les p-degrés de pénalisation appliqués à des paires de variables corrélées au sein d'un groupe donné.

Si $M = I_p$[3], c'est le cas où la matrice des données comporte des colonnes orthogonales et que si $|C_1| = |C_2| = ... = |C_K|$, alors le problème (3.2.7) se réduit à Elastic Net. ■

Il en résulte la conséquence suivante:

Propriété 3.2.3. Dans le cas d'une matrice de données orthogonale $X^T X = I_p$, et de groupes de variables de tailles égales, c'est à dire $|C_1| = |C_2| = ... = |C_K|$, le cluster elastic net est équivalent à elastic net.

3.3 Chemin de régularisation

Le chemin de régularisation est un graphe qui produit le tracé des paramètres estimés du modèle statistique en fonction du paramètre de pénalité.

On considère une matrice $(n \times p) - X$ avec n = 50 et p = 30[30], les lignes de X sont indépendantes et identiquement distribuées tirées d'une distribution $\mathcal{N}(0, \Sigma)$, où Σ est une matrice diagonale de $p \times p$ blocs avec trois blocs de taille égale. Σ a des 1 sur la diagonale, des $0,8$ dans chaque bloc et des 0 ailleurs.
La figure (3.1) présente les chemins de régularisation de cinq techniques de régression à savoir : Elastic Net, le group lasso, Cluster Elastic Net, Cluster Elastic Net avec groupe connus et le PACS[4] sur X.

[3] La matrice identité d'ordre p
[4] Pairwise Absolute Clustering and Sparsity

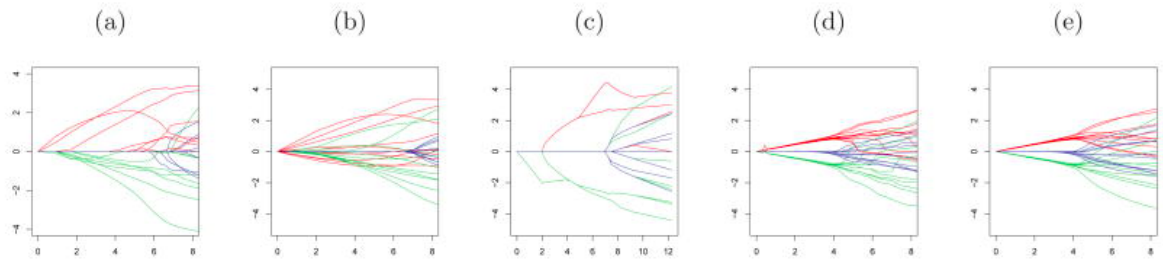


Figure 3.1: Chemins de régularisation pour (a) : $Elastic\ Net$, (b) : $lasso\ de\ groupe$, (c) : $PACS$, (d) : CEN, et (e) : CEN avec des groupes connus. La norme $\mathcal{L}^2$ du vecteur de coefficients estimé est affichée sur l'axe des abscisses x, et les estimations de coefficients sont sur l'axe des ordonnées y. Les couleurs indiquent le groupe auquel chaque ensemble de caractéristiques appartient ; le $rouge$, le $vert$ et le $bleu$ indiquent les valeurs réelles des coefficients de $1,\ -1\ et\ 0$, respectivement. Source : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4011669/figure/F2/

On voit bien que les coefficients correspondant aux variables d'un groupe donné sont regroupés de manière beaucoup plus étroite et plus claire par le Cluster Elastic Net que par Elastic Net. Aussi, le group lasso (qui ne peut s'appliquer que si les groupes sont connus a priori) donne des estimations de coefficients similaires pour les variables au sein d'un groupe, ce qui ne nous surprend en aucun cas. Quant au PACS, il ne parvient pas à regrouper les estimations de coefficients au sein de chaque groupe : seules les variables réellement pertinentes se voient attribuer des coefficients non nuls avant les variables réellement non pertinentes. Nous constatons également que dans ce contexte, les résultats du Cluster Elastic Net avec des groupes inconnus et du Cluster Elastic Net avec des groupes connus a priori sont pratiquement identiques (indiscernables).

Au regard de ce qui précède, on peut se permettre de dire que le CEN donne un regroupement des variables beaucoup plus satisfaisant en favorisant ainsi une bonne précision des chemins de régularisation par rapport à d'autres méthodes ci-dessus.

3.4 Contour Plots pour le CEN

Pour mieux comprendre, interpréter et comparer la pénalité du CEN par rapport aux pénalités existantes, nous considérons ses courbes de niveau qui sont représentées sur l'espace en deux dimensions donné par la figure(3.2).

Considérons une pénalité $\mathcal{L}^2$: $(\sum_i |\beta_i|^p)^{1/p}$ et $\|\beta\|_p = cste$ une isosurface[5].

Quand $p = 1$, on cherche à minimiser la fonction de perte $\|y - X\beta\|^2$ suivant la contrainte,

[5]Peut être définie comme l'analogue en 3D d'une courbe de niveau

$$\sum_{i=1}^{p} |\beta_i| \leqslant t, \; pour \; t > 0.$$

Dans ce cas, dans un espace en deux dimensions ($p = 2$) on a $|\beta_1| + |\beta_2| \leqslant t$, et donc les contours plots correspondant à la pénalité de $LASSO$ sont les courbes de niveau en forme des losanges représentés sur la figure 3.2(a).

Quand $p = 2$, on cherche à minimiser la fonction d'erreur $\|y - X\beta\|^2$ suivant la contrainte,

$$\sum_{i=1}^{p} \beta_i^2 \leqslant c, \; pour \; c > 0.$$

Dans ce cas ci, dans un espace en deux dimensions ($p = 2$) on a $\beta_1^2 + \beta_2^2 \leqslant c$, alors les contours plots correspondant à la pénalité de $Ridge$ sont les courbes de niveau en formes des cercles (centrés à l'origine) représentés sur la figure 3.2(b).

Lorsqu'on fait une combinaison convexe de la pénalité de $LASSO$ et celle de $Ridge$ on fait une pénalisation dite $Elastic \; Net$, on cherche à minimiser la fonction d'erreur $\|y - X\beta\|^2$ suivant les deux contraintes ci-dessus. Ces contraintes sont aussi équivalentes à la contrainte:

$$(1 - \alpha) \sum_{i=1}^{p} |\beta_i| + \alpha \sum_{i=1}^{p} \beta_i^2 \leqslant k, \; pour \; k > 0 \; avec \; \alpha = \frac{\lambda}{\delta + \lambda},$$

Les paramètres $\lambda > 0$ et $\delta > 0$ sont respectivement les coefficients de régularisation de $Ridge$ et de $LASSO$.

Dans un espace en deux dimensions ($p = 2$), les contours plots correspondant à la pénalité de $Elastic \; Net$ sont les courbes de niveau en formes des cercles représentés sur la figure 3.2(c). Ces courbes respectent : la singularité aux sommets (nécessaire pour la parcimonie) et les bords sont strictement convexes, la force de la convexité varie en fonction de $\alpha > 0$ (regroupement des variables).

La figure (3.2) montre les contours plots de la méthode de LASSO, Ridge, Elastic Net et le Cluster Elastic Net respectivement.

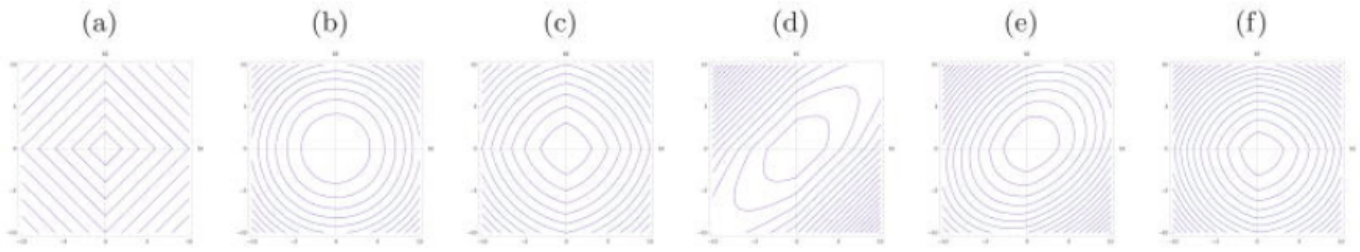


Figure 3.2: Les courbes de niveau sont présentées pour (a) le $lasso$; (b) la régression de Ridge ; (c) Elastic Net ; (d) le Cluster Elastic Net avec $p = 2$, avec corrélation positive entre les variables et $K = 1$; (e) le Cluster Elastic Net avec deux variables à corrélation positive dans le même groupe ; (f) le CEN avec deux variables non corrélées dans des groupes différents. Source : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4011669/figure/F3/

Quant aux contours plots du Cluster Elastic Net, les figure 3.2(d)-(f) décrivent les courbes de niveau correspondant à sa pénalité. La pénalité de CEN est donnée par l'expression:

$$P(\beta) = \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2$$

En particulier, les figures 3.2(d)-(f) montrent les contours de la pénalité CEN dans les trois cas suivants :

1er cas : $p = 2$, les variables sont positivement corrélées et $K = 1$. Figure 3.2(d)

$$P(\beta) = \delta\sum_{i=1}^{2}|\beta_i| + \frac{\lambda}{2|C_1|}\sum_{j,l\in C_1}\|X_j\beta_j - X_l\beta_l\|^2$$

Or $|C_1| = 1$ (le nombre de paires de variables en fonction de la corrélation) ici nous avons deux variables $(p = 2)$ positivement corrélées, donc une paire de variables, alors on a :

$$P(\beta) = \delta|\beta_1| + \delta|\beta_2| + \frac{\lambda}{2}\|X_1\beta_1 - X_2\beta_2\|^2$$

$$P(\beta) = \delta|\beta_1| + \delta|\beta_2| + \frac{\lambda}{2}(X_1\beta_1 - X_2\beta_2)^T(X_1\beta_1 - X_2\beta_2)$$

Après avoir développé, nous avons donc :

$$P(\beta) = \delta|\beta_1| + \delta|\beta_2| + \frac{\lambda}{2}[(X_1\beta_1)^TX_1\beta_1 - (X_1\beta_1)^TX_2\beta_2 - (X_2\beta_2)^TX_1\beta_1 + (X_2\beta_2)^TX_2\beta_2]$$

or nous savons que $(X_1\beta_1)^T = \beta_1^TX_1^T$,

$$P(\beta) = \delta|\beta_1| + \delta|\beta_2| + \frac{\lambda}{2}[\beta_1^TX_1^TX_1\beta_1 - \beta_1^TX_1^TX_2\beta_2 - \beta_2^TX_2^TX_1\beta_1 + \beta_2^TX_2^TX_2\beta_2]$$

On pose : $r_{12} = r_{21} = X_1^T X_2 = X_2^T X_1$, le coefficient attribué aux variables au sein d'un même groupe.

$$P(\beta) = \delta|\beta_1| + \delta|\beta_2| + \frac{\lambda}{2}[\beta_1^T X_1^T X_1 \beta_1 - \beta_1^T r_{12}\beta_2 - \beta_2^T r_{12}\beta_1 + \beta_2^T X_2^T X_2 \beta_2]$$

D'après les hypothèses, les colonnes de la matrice X ont été toutes normalisées, alors $X_1^T X_1 = X_2^T X_2 = 1$, on a :

$$P(\beta) = \delta|\beta_1| + \delta|\beta_2| + \frac{\lambda}{2}[\beta_1^T \beta_1 - r_{12}\beta_1^T \beta_2 - r_{12}\beta_2^T \beta_1 + \beta_2^T \beta_2]$$

$$P(\beta) = \delta|\beta_1| + \delta|\beta_2| + \frac{\lambda}{2}[\beta_1^2 - r_{12}\beta_1^T \beta_2 - r_{12}\beta_2^T \beta_1 + \beta_2^2]$$

Les coefficients β_1 et β_2 étant constants donc $(\beta_1^T = \beta_1 \; et \; \beta_2^T = \beta_2)$ alors;

$$P(\beta) = \frac{\lambda}{2}(\beta_1^2 + \beta_2^2 - 2r_{12}\beta_1\beta_2) + \delta|\beta_1| + \delta|\beta_2|$$

Par conséquent, les contours plot de $P(\beta)$ dans le cas d'une pénalité de CEN sont des ellipses (centrées à l'origine) d'après ce qui suit.

2em cas : Ici, il y'a $p = 4$ prédicteurs à corrélation positive appartenant à un seul groupe. β_1 se trouve sur l'axe des abscisses et β_2 sur l'axe des ordonnées, Figure 3.2(e). Nous supposerons que $\beta_3 = \beta_4 = 1$.

$$P(\beta) = \delta\sum_{i=1}^{4}|\beta_i| + \frac{\lambda}{2|C_1|}\sum_{j,l\in C_1}\|X_j\beta_j - X_l\beta_l\|^2$$

Or ici $|C_1| = 2$ car nous avons deux paires de prédicteurs (quatre prédicteurs) positivement corrélés, alors on a :

$P(\beta) = \frac{\lambda}{4}(\|X_1\beta_1 - X_2\beta_2\|^2 + \|X_1\beta_1 - X_3\beta_3\|^2 + \|X_1\beta_1 - X_4\beta_4\|^2 + \|X_2\beta_2 - X_3\beta_3\|^2 + \|X_2\beta_2 - X_4\beta_4\|^2)$

$+ \delta|\beta_1| + \delta|\beta_2| + \delta|\beta_3| + \delta|\beta_4|$

D'après l'hypothèse, $\beta_3 = \beta_4 = 1$ alors :

$P(\beta) = \frac{\lambda}{4}(\|X_1\beta_1 - X_2\beta_2\|^2 + \|X_1\beta_1 - X_3\|^2 + \|X_1\beta_1 - X_4\|^2 + \|X_2\beta_2 - X_3\|^2 + \|X_2\beta_2 - X_4\|^2)$

$+ \delta|\beta_1| + \delta|\beta_2| + 2\delta,$

$$P(\beta) = \frac{\lambda}{4}[\|X_1\beta_1 - X_2\beta_2\|^2 + \sum_{j=3}^{4}(\|X_1\beta_1 - X_j\|^2 + \|X_2\beta_2 - X_j\|^2)] + \delta|\beta_1| + \delta|\beta_2| + 2\delta,$$

En développant, on a :

$$P(\beta) = \frac{\lambda}{4}\|X_1\beta_1 - X_2\beta_2\|^2 + \delta|\beta_1| + \delta|\beta_2| + 2\delta + \frac{\lambda}{4}\sum_{j=3}^{4}(\|X_1\beta_1 - X_j\|^2 + \|X_2\beta_2 - X_j\|^2),$$

En se basant sur le premier cas, le terme $\frac{\lambda}{4}\|X_1\beta_1 - X_2\beta_2\|^2 + \delta|\beta_1| + \delta|\beta_2| + 2\delta$ est une ellipse qui dans ce cas est non centrée à l'origine à cause du dernier élément et le terme $\frac{\lambda}{4}\sum_{j=3}^{4}(\|X_1\beta_1 - X_j\|^2 + \|X_2\beta_2 - X_j\|^2)$ est un losange.

Par conséquent, les contours plots de $P(\beta)$ pour la pénalité du CEN sont la somme d'une ellipse (non centrée à l'origine) et d'un losange. Cela indique que β_1 et β_2 prendrons des valeurs similaires et positives.

3em cas : Ici, il y'a $p = 8$ prédicteurs et $K = 2$. β_1 est sur l'axe des abscisses et β_5 est sur l'axe des ordonnées. Les quatre premiers prédicteurs sont fortement corrélés et appartiennent à un groupe, tout comme les quatre autres prédictuers. Nous supposerons que $\beta_2 = \beta_3 = \beta_4 = 1$ et que $\beta_6 = \beta_7 = \beta_8 = -1$.Figure 3.2(f)

$$P(\beta) = \delta\sum_{i=1}^{8}|\beta_i| + \frac{\lambda}{2}\sum_{k=1}^{2}\frac{\lambda}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2,$$

$$P(\beta) = \frac{\lambda}{2}\times\frac{1}{4}\sum_{j,l\in C_1,C_2}\|X_j\beta_j - X_l\beta_l\|^2 + \delta(|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5| + |\beta_6| + |\beta_7| + |\beta_8|),$$

Car $|C_k|$ pour $k = 1, 2$ est égal à 4, parce que nous avons quatre paires de prédicteurs fortement corrélés disposées équitablement dans deux groupes. D'après l'hypothèse, $\beta_2 = \beta_3 = \beta_4 = 1$ et $\beta_6 = \beta_7 = \beta_8 = -1$, on a :

$$P(\beta) = \frac{\lambda}{8}\sum_{j,l\in C_1,C_2}\|X_j\beta_j - X_l\beta_l\|^2 + \delta|\beta_1| + \delta|\beta_5| + 6\delta,$$

En sommant le premier terme tout en respectant l'hypothèse ci-dessus, on a :
$P(\beta) = \frac{\lambda}{8}(\|X_1\beta_1 - X_2\beta_2\|^2 + \|X_1\beta_1 - X_3\|^2 + \|X_1\beta_1 - X_4\|^2 + \|X_5\beta_5 + X_6\|^2 + \|X_5\beta_5 + X_7\|^2 + \|X_5\beta_5 + X_8\|^2) + \delta|\beta_1| + \delta|\beta_5| + 6\delta,$

$$P(\beta) = \frac{\lambda}{8}(\sum_{j=2}^{4}\|X_1\beta_1 - X_j\|^2 + \sum_{j=6}^{8}\|X_5\beta_5 - X_j\|^2) + \delta|\beta_1| + \delta|\beta_5| + 6\delta,$$

$$P(\beta) = \frac{\lambda}{8}\sum_{j=2}^{4}(X_1\beta_1 - X_j)^T(X_1\beta_1 - X_j) + \frac{\lambda}{8}\sum_{j=6}^{8}(X_5\beta_5 - X_j)^T(X_5\beta_5 - X_j) + \delta|\beta_1| + \delta|\beta_5| + 6\delta,$$

$$P(\beta) = \frac{\lambda}{8}\sum_{j=6}^{8}[\beta_1^T X_1^T X_1\beta_1 - \beta_1^T X_1^T X_j - X_j^T X_1\beta_1 + X_j^T X_j] + \frac{\lambda}{8}\sum_{j=2}^{4}[\beta_5^T X_5^T X_5\beta_5 - \beta_5^T X_5^T X_j - X_j^T X_5\beta_5 + X_j^T X_j] + \delta|\beta_1| + \delta|\beta_5| + 6\delta,$$

$$P(\beta) = \frac{\lambda}{8}(\sum_{j=6}^{8}[-\beta_1^T X_1^T X_j - X_j^T X_1\beta_1 + X_j^T X_j] + \sum_{j=2}^{4}[-\beta_5^T X_5^T X_j - X_j^T X_5\beta_5 + X_j^T X_j]) + \frac{\lambda}{8}\beta_1^T X_1^T X_1\beta_1 + \frac{\lambda}{8}\beta_5^T X_5^T X_5\beta_5 + \delta|\beta_1| + \delta|\beta_5| + 6\delta,$$ les prédicteurs étant fortement corrélés et compte tenu de la normalisation des colonnes de la matrice X on a $X_1^T X_1 = X_5^T X_5 = 1$. On pose $r_{1j} = X_1^T X_j = X_j^T X_1$ et $r_{5j} = X_5^T X_j = X_j^T X_5$ respectivement les coefficients des prédicteurs dans les groupes C_1 constitués de quatre premiers prédicteurs et C_2 constitués des quatre autres, on a :

$$P(\beta) = \frac{\lambda}{8}(\sum_{j=6}^{8}[-\beta_1^T r_{1j} - r_{1j}\beta_1 + X_j^T X_j] + \sum_{j=2}^{4}[-\beta_5^T r_{5j} - r_{5j}\beta_5 + X_j^T X_j]) + \frac{\lambda}{8}\beta_1^T\beta_1 + \frac{\lambda}{8}\beta_5^T\beta_5 + \delta|\beta_1| + \delta|\beta_5| + 6\delta,$$

$$P(\beta) = \frac{\lambda}{8}(\sum_{j=6}^{8}[-2r_{1j}\beta_1 + X_j^T X_j] + \sum_{j=2}^{4}[-2r_{5j}\beta_5 + X_j^T X_j]) + \underbrace{\frac{\lambda}{8}\beta_1^2 + \frac{\lambda}{8}\beta_5^2 + \delta|\beta_1| + \delta|\beta_5| + 6\delta}_{\text{Cercle (non centré à l'origine)}}.$$

Alors, les contours plots de $P(\beta)$ pour la pénalité du CEN dans ce cas, c'est la somme d'un losange et d'un cercle (non centré à l'origine). Ensuite β_1 et β_5 sont obligés a prendre les valeurs positives et négatives respectivement.

Par conséquent, on peut dire que peu importe la structure de corrélation de la matrice de données X, on remarquera toujours que les contours plots pour la régression de Ridge et celle de LASSO ne changent pas de forme, les courbes de niveau pour ces deux méthodes seront toujours respectivement les cercles (centrés à l'origine) et les losanges. Pareil pour Elastic Net dont les courbes sont les cercles (centrés à l'origine) plus des losanges.

Mais nous constatons que les contours plots du CEN sont déterminés par les données : leur forme dépend de la structure de corrélation de la matrice de données X.

De plus, si la matrice de données est orthogonale, alors les contours seront simplement des cercles (centrés à l'origine) plus des losanges ; cela équivaut aux contours de Elastic Net. En revanche, en présence d'une forte corrélation entre les variables d'un groupe, les contours correspondant aux variables d'un groupe seront des ellipses (non centrées à l'origine) plus des losanges, et les contours correspondant à des paires de variables dans différents groupes seront des cercles (non centrés à l'origine) plus des losanges.

Si la matrice de données est non orthogonale, les contours de la fonction de pénalité ne sont pas centrés à l'origine car les variables corrélées sont encouragées à prendre des valeurs de coefficients similaires.

Enfin, contrairement à la régularisation de LASSO, Ridge et Elastic Net, dont les solutions

donnent toujours respectivement les losanges (conf. figure 3.2a), les cercles (conf. figure 3.2b) et les cercles plus les losanges (conf. figure 3.2c), le Cluster Elastic Net donne un contour plot beaucoup plus varié selon la structure de corrélation de la matrice de données. Cela est d'autant mieux car la structure des données n'est pas toujours simple et peut comporter diverses variantes. Les autres méthodes présentent une forme unique et cela peut être compromettant sur l'apprentissage d'un modèle statistique donné. Pour espérer obtenir un contour plot correspondant à un type de donnée beaucoup plus corrélées (en grande dimension) il est d'autant plus préférable et judicieux d'utiliser la méthode du Cluster Elastic Net car celle-ci garantit une forme sur n'importe quelle structure de corrélation de la matrice de conception (conf. figures 3.2d-f).

4. Considérations informatiques

4.1 Algorithme de résolution du problème CEN

Dans cette partie, nous considérons l'aspect de résolution du problème d'optimisation (3.1.2). Ce problème d'optimisation est non convexe, donc une résolution ordinaire ne pourrait marcher pour trouver un optimal global. Pour ce faire, lorsque p est très petit, on considère toutes les partitions possibles de $O(K^p)$ des p-variables en K-clusters, ensuite pour chaque partition, le problème (3.1.2) pourrait être résolu avec les groupes $C_1, C_2, ..., C_K$ maintenus fixes.

Mais lorsque p n'est pas très petit, cette approche n'est réalisable. Par conséquent, au lieu de chercher un optimum global au problème (3.1.2), nous allons chercher un optimum local.

Contrairement à d'autres méthodes dont la fonction à optimiser est convexe, ici nous adoptons une nouvelle approche qui est itérative, approche qui se déroule en deux étapes. Une première est celle de maintenir fixes les groupes $C_1, C_2, ..., C_K$, puis résoudre le problème par rapport à β; et une seconde étape celle de maintenir fixe le vecteur coefficient de régression β puis résoudre le problème par rapport aux groupes $C_1, C_2, ..., C_K$.

Dans cette seconde et dernière étape, nous allons trouver un optimum local du problème (3.1.2) avec le vecteur β maintenu fixe en appliquant l'algorithme de classification de $K-means\ clustering$ sur les variables $X_1\beta_1, X_2\beta_2, ..., X_p\beta_p$[10]. Cette approche de résolution du problème (3.1.2) est présentée dans **l'Algorithme 1** ci-dessous :

Algorithme 1 :

L'algorithme pour résoudre le problème d'optimisation (3.1.2)

1. Initialiser β comme solution du problème d'optimisation de $Elastic\ Net$,

$$\min_{\beta}\{\|y - X\beta\|^2 + \delta\|\beta\|_1 + \lambda\|\beta\|^2\}$$

.

Initialiser les groupes ou clusters C_k,

2. Itération jusqu'à la convergence:

a. On fixe β et on minimise (3.1.2) suivant $C_1, C_2, ..., C_K$, on résout le problème

$$\min_{C_1,...,C_K}\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2\}. \tag{4.1.1}$$

Un optimum local peut être trouvé en performant le $K-means\ clustering$ sur $X_1\beta_1, X_2\beta_2, ..., X_p\beta_p$ avec K-clusters.

b. On fixe $C_1, C_2, ..., C_K$ et on minimise (3.1.2) suivant β, on résout le problème

$$\min_{\beta}\{\|y - X\beta\|^2 + \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2\}. \tag{4.1.2}$$

■

Bien que **l'Algorithme 1** ne garantit pas la convergence à l'optimum global, néanmoins il nous donne une solution locale c'est à dire, les bons groupes ou clusters les mieux adaptés à la structure de données de notre problème (en tenant compte de leurs corrélations) en appliquant le $K-means\ clustering$ sur les variables $X_1\beta_1, X_2\beta_2, ..., X_p\beta_p$; contrairement aux méthodes dont les groupes sont connus a priori. Sur ce, la première étape de notre algorithme est résolue.

Pour résoudre la seconde étape, nous adoptons une nouvelle approche celle de descente en coordonnées[9], ce qui revient à effectuer de manière répétée une seule mise à jour du vecteur coefficient de régression β donnée dans la proposition (4.1.1) ci-après :

Proposition 4.1.1. Soit X_{-j} la sous-matrice $n \times (p-1)$ contenant tout sauf la j-ième colonne de X, β_{-j} est le $(p-1)$-vecteur contenant tout sauf le j-ièm élément de β, $\widetilde{y}_j = y - X_{-j}\beta_{-j}$. Supposons que $j \in C_k$, alors la mise à jour suivante de β_j minimise la fonction objectif dans (4.1.2) par rapport à β_j tout en maintenant fixes toutes les autres variables :

$$\beta_j \longleftarrow \frac{S(\widetilde{y}_j^T X_j + \frac{\lambda}{|C_k|}\sum_{l\in C_k, j\neq l}\beta_l r_{jl}, \delta/2)}{r_{jj}(1 + \lambda\frac{|C_k|-1}{|C_k|})} \tag{4.1.3}$$

Où $S(a,b)$ désigne l'opérateur de Soft-tresholding, défini par :

$$S(a,b) = sign(a)max(0,|a|-b) = \begin{cases} a-b & si\ \ a>0\ et\ b<|a| \\ a+b & si\ \ a<0\ \ et\ b<|a| \\ 0 & si\ \ b\geqslant|a| \end{cases}$$

L'approche de descente en coordonnée pour résoudre la seconde étape de **l'Algorithme 1** se présente comme suit :

Initialiser $(\widetilde{\beta}_0, \widetilde{\beta})$,

- Itérer de 2b) jusqu'à la convergence.
  Faire une descente par coordonnée pour j = 1, ..., p.
  i) Mettre à jour le β_j,
  ⋆ Calculer $\widetilde{y}_i = y_i(\widetilde{\beta}_0 - x_i^T\widetilde{\beta})$,

⋆ Calculer

$$\widehat{\beta}_j^{CEN} = \frac{S(\sum_{i=1}^{n} y_i(\widetilde{\beta}_0 - x_i^T\widetilde{\beta})x_{ij} + \frac{\lambda}{|C_k|}\sum_{l\in C_k, j\neq l}\beta_l r_{jl}, \delta/2)}{r_{jj}(1+\lambda\frac{|C_k|-1}{|C_k|})}.$$

⋆ poser $\widetilde{\beta}_j = \widehat{\beta}_j^{CEN}$,

ii) Mettre à jour l'ordonnée à l'origine:

Calculer $\widetilde{y}_i = y_i(\widetilde{\beta}_0 - x_i^T\widetilde{\beta})$,

⋆ Calculer $\widehat{\beta}_0 = \widetilde{\beta}_0 + \sum_{i=1}^{n} y_i(\widetilde{\beta}_0 - x_i^T\widetilde{\beta})$,

⋆ poser $\widetilde{\beta}_0 = \widehat{\beta}_0^{CEN}$...

Faire les itérations pour j = 1, ..., p jusqu'à atteindre la convergence.

Ce qui est intéressant dans cette approche de descente en coordonnée pour la régression du Cluster Elastic Net (3.1.2) c'est qu'à chaque itération nous allons juste répéter une seule mise à jour de β_j de la proposition (4.1.1), ce qui permet une convergence rapide vers un minimum global en β_j (j = 1, ... p), car la fonction à optimiser est convexe en β. Cet algorithme est assez puissant du fait que en plus de trouver l'optimum, il classifie simultanément chaque variable en appliquant le K-means. Contrairement aux méthodes qui disposent des groupes de variables au préalable, l'approche ci-dessus assure une convergence plus sûre et plus efficace du fait de la répétition de la même valeur de β_j de la proposition (4.1.1) dans chaque itérations, $j = 1, ..., p$.
.

4.2 Sélection des paramètres de régularisation

La plupart des méthodes de régularisation permettent non seulement de régler le problème de singularité, mais aussi celui lié aux bruits (variables non pertinentes). Les paramètres δ et λ du problème d'optimisation (3.1.2) encore appelés hyperparamètres du modèle, permettent de contrôler la complexité du modèle sur le jeu test. Ceux-ci permettent aussi de réduire le nombre de variables explicatives pour ne garder que celles qui sont significatives.
Les méthodes de régularisation introduisent un biais dans la recherche des paramètres du modèle, toutefois la variance est inférieure à celle du modèle non pénalisé. Il faut donc trouver le juste milieu entre le biais et la variance (conf. figure (4.1)) pour bien refléter les données et faire une bonne prédiction.

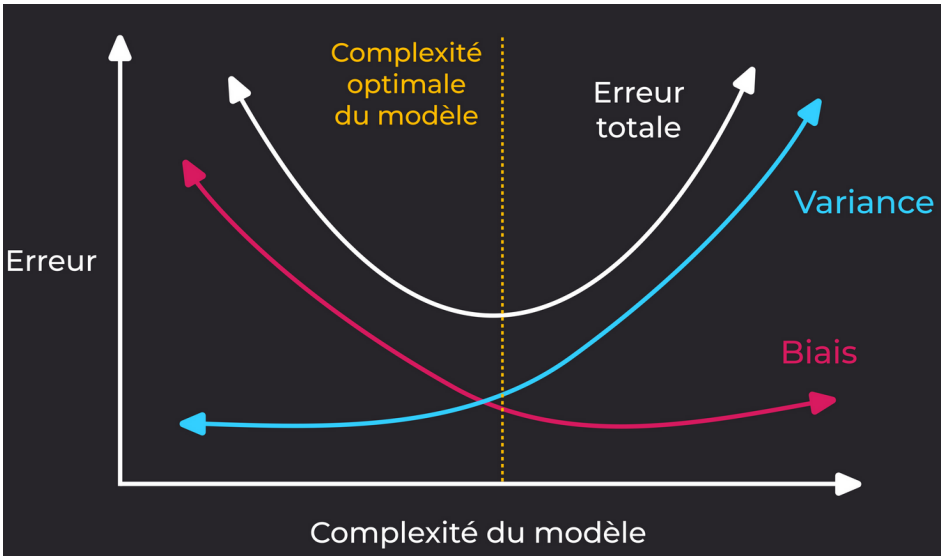


Figure 4.1: Courbe de compromis biais-variance. Source : cours "initiez-vous au machine learning", openclassroom.

Par conséquent, dans le problème du Cluster Elastic Net, la présence de la norme $\mathcal{L}^2$ permet de rétrécir les coefficients des variables qui sont fortement corrélées vers une même valeur, elle enlève la limitation sur le nombre de variables retenues et favorise le choix de groupes. Elle est donc très efficace lorsque les colonnes de la matrice de données sont corrélées tel est le cas dans notre étude, sans oublier que le Cluster Elastic Net dans son fonctionnement, ne sélectionne que les variables corrélées ayant une similarité avec la réponse.

La norme $\mathcal{L}^1$ quant à elle, a pour effet la création d'une matrice creuse (parcimonie) c'est à dire à la création d'une solution constituée des coefficients égaux à zéro exactement. Les variables ayant des coefficients égaux à zéro ne sont pas comptées dans le modèle, d'où l'effet de réduction de dimension[31].

Le paramètre K définit le nombre de variables et est très important sur la classification des variables en clusters ou groupes.

Plus les hyperparamètres δ et λ du problème d'optimisation (3.1.2) sont grands, plus le terme de pénalité est important. Plus ils sont petits, plus l'erreur est importante; s'ils sont tous les deux suffisamment faibles (en particulier égaux à zéro), on retrouve la solution de la régression non-régularisée. Si $\delta = 0$ on a le cluster ridge régression et si $\lambda = 0$ on a une régularisation de lasso d'après la remarque (1). Quelles valeurs donner donc à ces hyperparamètres ? En général, c'est une question que l'on règle en utilisant la technique de validation croisée.

4.3 Principe de fonctionnement de la validation croisée

La validation croisée présente une importance capitale dans l'application pratique de plusieurs modèles statistiques. C'est une technique empirique qui est utilisée pour estimer l'erreur du modèle, il s'agit en réalité de mesurer la capacité d'un modèle à s'adapter et à généraliser sur un ensemble de données indépendant de celui utilisé pour son estimation.
Le principe autrefois utilisé était de constituer une partition sur l'ensemble de données initial en deux ou trois sous-ensembles complémentaires, tels que:

✓ Un premier sous-ensemble dit d'apprentissage sur lequel est applique le modèle candidat, par exemple 50%,

✓ Un second sous-ensemble dit de validation sur lequel on fait la sélection du modèle, par exemple 30%,

✓ Un troisième sous-ensemble dit de test sur lequel on évalue le modèle final, par exemple 20%.

Ce principe est plusieurs fois contesté et révoqué compte tenu de son caractère heuristique. Il n'a pas de fondement théorique dans les problèmes de prédiction. Souvent très sensible à mettre en œuvre car l'utilisation d'un ou de deux sous-ensembles de validation ou de test réduit remarquablement la taille de données du modèle statistique.C'est là l'inconvénient de ce principe, car cela impacte sensiblement la qualité du modèle. Les paramètres trouvés dans ce cas peuvent être dépendants de la partition sélectionnée. Afin de corriger ce problème, il est d'autant plus souhaitable de répéter le principe ci-dessus en permutant les sous-ensembles d'apprentissage et de validation. Autrement dit, on met en place un second modèle qui comporte seulement les sous-ensembles de validation, puis en évaluer la performance sur le sous-ensemble d'apprentissage; faire une répétition de cette démarche en faisant un échantillonnage aléatoire de l'ensemble de données en plusieurs sous-ensembles (apprentissage et validation) indépendants.
Les mesures de performances ainsi calculées sont moyennées sur l'ensemble des partitions (ou sous-ensembles) utilisées.

L'objectif principal de la validation croisée est donc la sélection des paramètres du modèle et a fortiori la sélection de variables. En outre, elle est aussi utilisée comme technique de calibrage des méta-paramètres au sein d'un modèle statistique. Par contre, le principe le plus utilisé dans la technique de validation croisée est celui du k-fold[23], principe selon lequel (pour $k = 5$) l'ensemble de données d'apprentissage du modèle est divisé en cinq (05) sous-ensembles de tailles égales. Le modèle est alors construit sur l'ensemble constitué par les quatre (04) premiers sous ensembles, puis sa performance est évaluée sur le cinquième sous ensemble qui fait alors l'objet d'ensemble de validation. Puis cette étape est répétée pour les cinq (05) partitions (sous ensembles) apprentissages/validation possibles (conf. figure (5.1)).
Enfin, on estime l'erreur par validation croisée et le paramètre choisi est celui qui minimise cette erreur. La valeur du paramètre choisi est alors utilisée pour l'élaboration du modèle final sur l'ensemble de données.

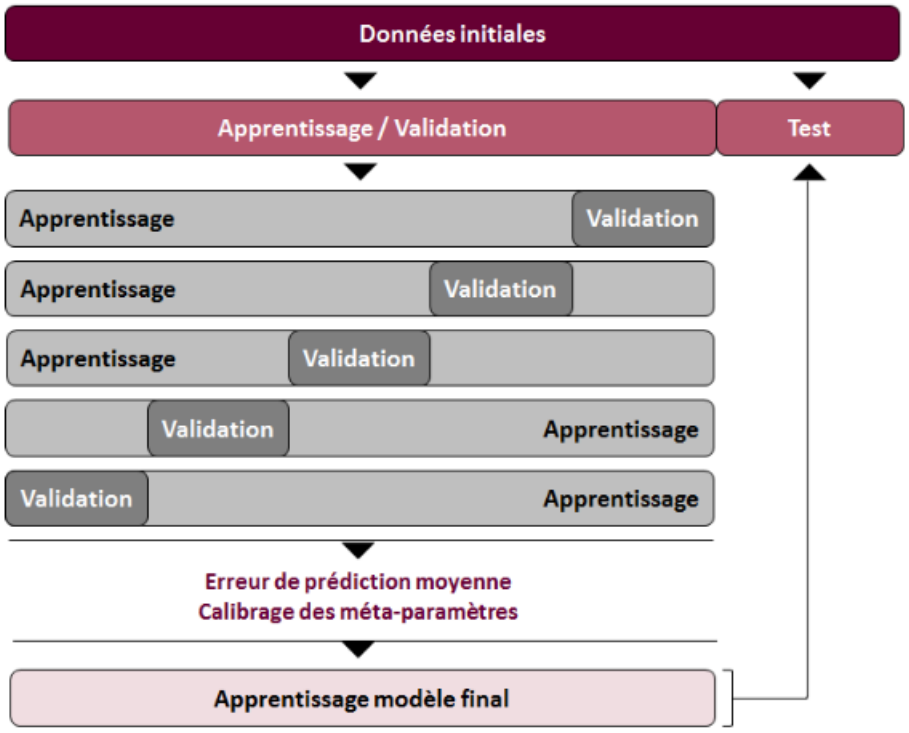


Figure 4.2: 5-fold, Validation croisée. Source : Apprentissage statistique en tarification non-vie, Antoine GUILLOT.

Pour la pénalisation du Cluster Elastic Net défini dans le problème d'optimisation (3.1.2), il existe trois paramètres de réglage qui interviennent dans la sélection de variables à savoir: δ le paramètre de pénalité de Lasso, λ celui de la pénalité de Ridge et K le paramètre qui est déterminant dans la constitution de groupes. Pour trouver ces paramètres on peut utiliser la validation croisée. Comme dans lasso, Ridge et Elastic Net le principe de la validation croisée reste le même, mais ces méthodes ne s'adaptent qu'à une structure particulière des variables. Souvent pour ces méthodes, lorsqu'il y'a une forte corrélation entre les variables, la technique de validation croisée peut donner des paramètres capable de biaiser le modèle statistique. En présence de cette situation, la méthode Cluster Elastic Net est celle qui est la plus optimale. On rappelle aux points (3.2.1) et (3.2.2) sur les propriétés du Cluster Elastic Net que, pour certaines valeurs de λ et de K, le cluster elastic net se réduit au lasso (pour K = p), ou bien à elastic net (pour K = 1 et pour autres valeurs de K : $1 < K < p$, aussi pour $\lambda < p$), ce qui ce qui revient à dire que la technique validation croisée s'effectuera sur une structure de données adéquate pour Ridge ou lasso. Par conséquent, pour un ensemble de données particulier, les hypothèses sous-jacentes du Cluster Elastic Net ne se vérifient pas aisément mais pour quelques valeurs de K, la validation croisée aboutirait éventuellement à une sélection des paramètres de lasso ou de elastic net.

Enfin, pour le Cluster Elastic Net certaines valeurs de K rendent la structure de la matrice de données similaire à celle de lasso ou elastic net, alors la validation croisée dans ce cas est susceptible à aboutir sur l'une des deux méthodes de Ridge, de lasso ou de Elastic Net et elle donnera des résultats satisfaisants pour une bonne sélection de variables.

5. Analyse du rétrécissement entre les groupes

Dans son principe de fonctionnement, Elastic Net règle les insuffisances de lasso[31]. En présence de groupes de variables, cette technique réduit toutes les estimations de coefficients vers l'origine tandis que la méthode du cluster elastic quant à elle, réduit les coefficients des variables corrélées qui appartiennent au même groupe les unes vers les autres plutôt que vers l'origine, et ne considère que les variables qui sont similaires à la variable réponse. On montre cette propriété du Cluster Elastic Net dans un cadre très simple, dans lequel on considère les groupes comme déjà connus et on établit les hypothèses suivantes :

(i) Il existe deux groupes connus, ayant chacun la taille $m = p/2$. Les variables sont ordonnées de telle sorte que celles du premier groupe précèdent celles du second;

(ii) Le véritable vecteur coefficient de régression β est $(\beta_1, ..., \beta_1, \beta_2, ..., \beta_2)^T$. Autrement dit, la valeur réelle de β est la même dans chaque groupe. On suppose également sans perdre de généralité que $\beta_1 > \beta_2$;

(iii) $y = X\beta + \epsilon$, ou $\epsilon \rightsquigarrow \mathcal{N}(0, \sigma^2 I_n)$;

(iv) $r_{jl} = r_1$ pour j et l appartenant au même groupe, et $r_{jl} = r_0$ pour j et l appartenant dans les groupes différents.

Ce qui signifie :

$$X^T X = (1 - r_1)I + \begin{bmatrix} r_1 & \cdots & r_1 & r_0 & \cdots & r_0 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ r_1 & \cdots & r_1 & r_0 & \cdots & r_0 \\ r_0 & \cdots & r_0 & r_1 & \cdots & r_1 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ r_0 & \cdots & r_0 & r_1 & \cdots & r_1 \end{bmatrix}.$$

On suppose en outre que $r_1 > r_0 > 0$.

Si les hypothèses **(i)**-**(iii)** sont vraies, alors l'hypothèse **(iv)** est assez simple. Notons qu'avec une matrice aléatoire de données, **(iv)** correspond à un simple modèle de corrélation en bloc et est donc raisonnable pour les données observées.

On va à présent comparer l'estimateur du Cluster Ridge Regression $\widehat{\beta}_{CRR}$ à l'estimateur de la Régression de Ridge $\widehat{\beta}_{RR}$. Il sied de noter que ces deux estimateurs sont normalement distribués. Ainsi, on annonce le théorème (5.1) ci-apres :

5.1 Théorème (comparaison entre CRR et RR)

Supposons que les hypothèses **(i)-(iv)** sont vraies et que $\lambda_{RR} = (1 - \frac{1-r_1}{m})\lambda_{CRR}$.

1. Si les j-ième et l-ième prédicteurs sont dans un même groupe alors,

$$\widehat{\beta}_{CRR,j} - \widehat{\beta}_{CRR,l} = \widehat{\beta}_{RR,j} - \widehat{\beta}_{RR,l} \rightsquigarrow \mathcal{N}(0, 2(1-r_1)\sigma^2(1+\lambda_{RR}-r_1)^{-2})$$

2. Si les j-ième et l-ième prédicteurs sont dans des groupes ou clusters différents alors,

$$E(\widehat{\beta}_{CRR,j} - \widehat{\beta}_{CRR,l}) = \frac{(\beta_j - \beta_l)(-1+m(r_0-r_1)+r_1)}{\lambda_{CRR}(-1+m)(-1+r_1)/m + (-1+m(r_0-r_1)+r_1)} \tag{5.1.1}$$

$$E(\widehat{\beta}_{RR,j} - \widehat{\beta}_{RR,l}) = \frac{(\beta_j - \beta_l)(-1+m(r_0-r_1)+r_1)}{-1-\lambda_{RR}+m(r_0-r_1)+r_1} \tag{5.1.2}$$

Pour démontrer le théorème (5.1), on définit quelques lemmes ci-contre :

5.2 Lemmes

Lemme 5.2.1. On définit une $p \times p$ matrice de la forme suivante :

$$D(a,b,c) = aI + \begin{bmatrix} b & \cdots & b & c & \cdots & c \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ b & \cdots & b & c & \cdots & c \\ c & \cdots & c & b & \cdots & b \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ c & \cdots & c & b & \cdots & b \end{bmatrix}.$$

où la deuxième matrice est composée de quatre blocs $m \times m$, $m = p/2$.

Pour toute matrice $N = D(a,b,c)$ et pour quelques valeurs de a, b et c, alors N admet comme inverse :

$$N^{-1} = D(a^{-1}, -a^{-2}(d_1^{-1} + a^{-1}m), d_2)$$

où $d_1 = \dfrac{ab + b^2m - c^2m}{a+bm}$ et $d_2 = -\dfrac{c}{a^2 + 2abm + (b^2-c^2)m^2}$.

Lemme 5.2.2. Soient $N = D(a,b,c)$ et $N_\star = D(a_\star, b_\star, c_\star)$ deux matrices et $a, b, c, a_\star, b_\star$ et $c_\star$ des réels.

On suppose que les j-èmes et l-èmes variables se trouvent dans un même groupe, on définit un vecteur v contenant les zéros, sauf pour $v_j = 1$ et $v_l = -1$. Alors,

$$v^T N N_\star N v = 2a^2 a_\star.$$

Lemme 5.2.3. Soient $N = D(a, b, c)$ et $N_\star = D(a_\star, b_\star, c_\star)$ deux matrices et $a, b, c, a_\star, b_\star$ et $c_\star$ les réels.

On suppose que les j-èmes et l-èmes variables se trouvent dans des groupes ou clusters différents et on définit un vecteur v contenant les zéros, sauf pour $v_j = 1$ et $v_l = -1$. En outre, on suppose que $\beta = (\beta_1, ..., \beta_1, \beta_2, ..., \beta_2)^T$. Alors,

$$\frac{v^T N N_\star \beta}{\beta_1 - \beta_2} = (a + (b - c)\frac{p}{2})(a_\star + (b_\star - c_\star)\frac{p}{2}).$$

Preuve du théorème (5.1): Nous présentons d'abord la preuve pour la régression de Ridge, on calcule les quantités suivantes :

$$E(v^T \widehat{\beta}_{RR}) = v^T (X^T X + \lambda_{RR} I)^{-1} X^T X \beta,$$

et

$$Var(v^T \widehat{\beta}_{RR}) = \sigma^2 v^T (X^T X + \lambda_{RR} I)^{-1} X^T X (X^T X + \lambda_{RR} I)^{-1} v.$$

Où v est un $p \times 1$-vecteur tel que $v_k = 1_{k=j} - 1_{k=l}$, j et l sont soient dans un même groupe, soient dans des groupes différents. On note que $D(a, b, c) = D(1 - r_1, r_1, r_0) = X^T X$ et on note que $X^T X + \lambda_{RR} I = D(1 + \lambda_{RR} - r_1, r_1, r_0)$.

Pour calculer l'espérance mathématique et la variance de $v^T \widehat{\beta}_{RR}$ pour j et l appartenant dans le même cluster, on appliquer les lemmes (5.2.1) et (5.2.2) ci-dessus.

Pour calculer l'espérance mathématique et la variance de $v^T \widehat{\beta}_{RR}$ pour j et l appartenant dans des groupes différents, on combine le lemme (5.2.3) et le lemme (**??**) ci-dessus.

De même, afin de calculer l'espérance mathématique et la variance de $v^T \widehat{\beta}_{CRR}$ dans le cas $\widehat{\beta}_{CRR}$ est l'estimateur du Cluster Ridge Regression; definies par :

$$E(v^T \widehat{\beta}_{CRR}) = v^T (X^T X + \lambda_{CRR} M)^{-1} X^T X \beta,$$

et

$$Var(v^T \widehat{\beta}_{CRR}) = \sigma^2 v^T (X^T X + \lambda_{CRR} M)^{-1} X^T X (X^T X + \lambda_{CRR} M)^{-1} v.$$

Pour les calculer, on procède de la même manière que pour Régression de Ridge. Notons que pour le CRR,

$$X^TX + \lambda_{CRR}M = D(1 + \lambda_{CRR}\frac{m-1}{m} - (1 - \frac{\lambda_{CRR}}{m})r_1, (1 - \frac{\lambda_{CRR}}{m})r_1, r_0).$$

En faisant ainsi, on aboutit à la démonstration du théorème (5.1).

Le théorème (5.1) indique qu'il existe une relation simple entre λ_{RR} et λ_{CRR} telle que $\widehat{\beta}_{CRR,j} - \widehat{\beta}_{CRR,l}$ et $\widehat{\beta}_{RR,j} - \widehat{\beta}_{RR,l}$ ont la même distribution quand j et l appartiennent dans le même groupe. Autrement dit, si on choisit les paramètres de réglages de cette manière, alors on verra que le Cluster Ridge Regression et la Régression de Ridge effectuent le même rétrécissement au sein d'un groupe.

Cependant, comment pouvons-nous comparer le rétrécissement effectué par le Cluster Ridge Regression et la Régression de Ridge pour les j-ièmes et l-ièmes caractéristiques appartenant dans différents groupes ?

Corollaire 1. Sous les hypothèses du théorème (5.1), si les j-ièmes et l-ièmes caractéristiques appartiennent dans les groupes ou clusters différents, alors

$$1 \geqslant \frac{E(\widehat{\beta}_{CRR,j} - \widehat{\beta}_{CRR,l})}{\beta_j - \beta_l} \geqslant \frac{E(\widehat{\beta}_{RR,j} - \widehat{\beta}_{RR,l})}{\beta_j - \beta_l} \geqslant 0$$

De plus, si $r_1 = 1$ alors :

$$\frac{E(\widehat{\beta}_{CRR,j} - \widehat{\beta}_{CRR,l})}{\beta_j - \beta_l} = 1$$

et,

$$\frac{E(\widehat{\beta}_{RR,j} - \widehat{\beta}_{RR,l})}{\beta_j - \beta_l} = \frac{m(1-r_0)}{\lambda_{RR} + m(1-r_0)} < 1.$$

Preuve : La preuve de ce corollaire est triviale. ■

Le corollaire (1) montre que le relation qui existe entre les paramètres de réglages et qui conduit à la même quantité de rétrécissement intra-groupes pour la régression de ridge et le Cluster Ridge Regression, par contre cette relation entraine un rétrécissement intergroupes très important en utilisant la régression de ridge que par Cluster Ridge Regression. Autrement dit, la régression de ridge et le Cluster Ridge Regression réduisent toutes les deux, les coefficients des caractéristiques au sein d'un même groupe les uns vers les autres. Quand les j-ièmes et l-ièmes caractéristiques sont dans des groupes différents et de plus si le coefficient des caractéristiques corrélées appartenant au sein d'un même groupe $r_1 = 1$, alors la régression de ridge a tendance à se contracter

entre les groupes, tandis que le Cluster Ridge Regression ne présente pas le même comportement, comme le montre la figure (5.1).

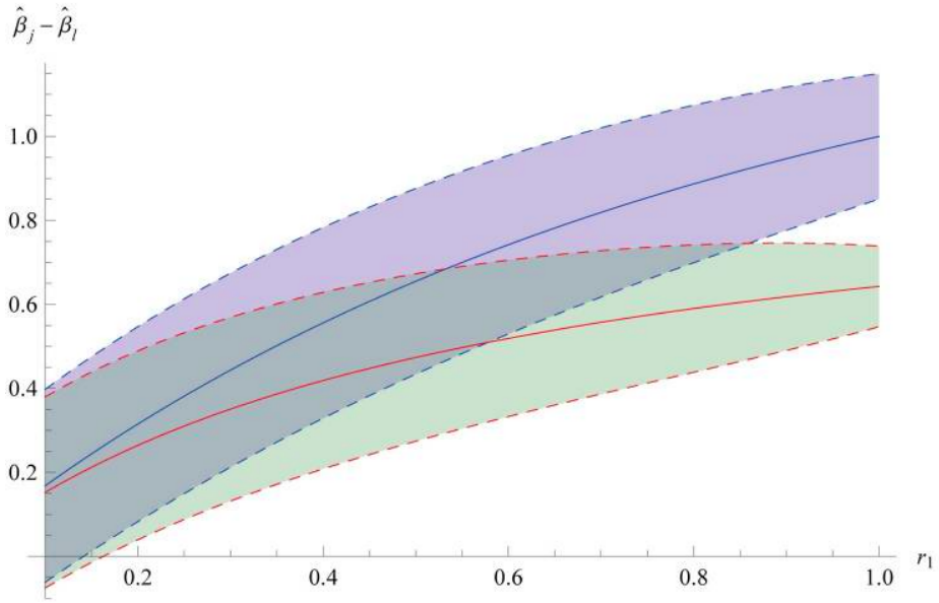


Figure 5.1: Effets du cluster elastic net et de la régression de ridge sur les variables quand $r_1 = 1$. Source : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4011669/figure/F4/

Toujours sous les hypothèses du théorème (5.1), dans la figure (5.1), la ligne bleue continue et rouge continue correspondent respectivement à $E(\widehat{\beta}_{CRR,j} - \widehat{\beta}_{CRR,l})$ et $E(\widehat{\beta}_{RR,j} - \widehat{\beta}_{RR,l})$ en fonction de b. Nous sommes dans le cas où les j-èmes et l-èmes caractéristiques appartiennent à des groupes différents. De plus, on considère $r_0 = 0.1$, $m = 10$, $\beta_j - \beta_l = 1$ *et* $\lambda_{RR} = 5$. L'écart-type plus ou moins attendu de $\widehat{\beta}_j - \widehat{\beta}_l$ est représenté par les lignes en pointillés.

Il sied de noter que le théorème (5.1) explique que si $r_1 > 0$ et même si $r_0 = 0$ (c'est à dire en l'absence de caractéristiques corrélées appartenant dans différents groupes), alors la régression de ridge réduit encore plus les coefficients des caractéristiques de différents groupes les uns par rapport aux autres, plus meilleur que le Cluster Ridge Regression. Dans ce cas, la régression de ridge réduit tous les coefficients vers zéro mieux que le Cluster Ridge Regression, tout en minimisant l'appartenance de caractéristiques à un groupe. C'est un comportement indésirable de la régression de ridge.

6. Relation du CEN avec d'autres approches

Dans la propriété (3.2.1) et (3.2.2) nous avons vu que pour certaines valeurs des paramètres K et λ, le problème (3.1.2) du cluster elastic net se réduit au lasso ou à elastic net. Nous abordons maintenant la relation du cluster elastic net avec d'autres approches.

6.1 Relation avec la régularisation graphique sous contrainte

Les graphes et les réseaux sont des moyens courants pour représenter l'information. On étudie la procédure de régularisation graphique sous contrainte[15] sur un graphe et ses propriétés théoriques pour l'analyse de régression, afin de prendre en compte les informations de voisinage des variables mesurées sur un graphe, où une pénalité de lissage sur les coefficients est définie comme une forme quadratique de la matrice de Laplacien associée au graphe.

On considère un graphe pondéré $G = (E, V, W)$ où $V = \{1, ..., p\}$ est un ensemble des sommets correspondant dans notre cas aux p-prédicteurs, $E = \{j \sim l\}$ est l'ensemble des arêtes entres sommets du graphe, et $W(j, l)$ correspond au poids (positif) de l'arête entre les j-ième et l-ième sommets. Soit $d_l = \Sigma_{j\sim l} w(j, l)$ le degré du l-ième sommet, supposons aussi que le graphe G est connu a priori. Alors (Li et Li, 2008)[15] et (Li et Li, 2010)[16] ont proposé un estimateur sous contrainte graphique (GRACE[1]) équivaut à la résolution (3.2.7), où la matrice $M = M^{grace}$ est donnée par :

$$M_{jl}^{grace} = \begin{cases} 1 - w(j,j)/d_j & \text{si } j = l \text{ et } d_j \neq 0 \\ -w(j,l)/\sqrt{d_j d_l} & \text{si } j \text{ et } l \text{ partagent une arête sur le graphe} \\ 0 & \text{sinon} \end{cases} \tag{6.1.1}$$

Nous examinons maintenant un cas particulier du graphe G afin de voir le lien existant entre le CEN et GRACE. On suppose que le graphe pondéré G est constitué de K composantes disjointes, et que $w(j, l) = 1$ si les j-ième et l-ième sommets sont dans une même composante. Ce qui signifie, dans une composante donnée, tous les sommets du graphe G sont réliés avec des poids égaux entre eux. Et que si les j-ième et l-ième sommets sont dans des composantes différentes, alors $w(j, l) = 0$.

On considère C_k comme l'ensemble contenant tous les indices des sommets dans la k-ième composante ($k = 1, ..., K$), il est clair de voir que si le j-ième sommet se trouve dans la k-ième composante alors $d_j = |C_k|$. Dans ce cas, l'expression (6.1.1) se réduit à :

$$M_{jl}^{grace} = \begin{cases} (|C_k| - 1)/|C_k| & \text{si } j = l \ \in C_k \\ -1/|C_k| & \text{si } j \neq l \text{ et } j, \ l \in C_k \\ 0 & \text{sinon} \end{cases} \tag{6.1.2}$$

[1] Graph-constrained estimator

En comparant l'expression (6.1.2) à (3.2.6), on constate que pour ce cas particulier de GRACE, les matrices M pour le CEN et M^{grace} sont presque similaires. En réalité, le cluster elatic net et ce cas spécial de GRACE seraient identiques si le cluster elastic net était réalisé avec des groupes connus et que si $r_{jl} = 1$ pour $j, l \in C_k$. En effet, dans la pratique, $|r_{jl}| < 1$ pour $j \neq l$.

Par conséquent, la pénalité appliquée par le CEN est moins stricte par rapport à celle appliquée par GRACE. En outre, GRACE exige que le graphe soit connu a priori, alors que dans le CEN les groupes ou clusters, ainsi que la structure du graphe sont déduits des données.

6.2 Relation avec PACS

On introduit une nouvelle pénalisation pour l'identification simultanée des groupes et la selection de variables. On utilise une nouvelle alternative, à la norme $\mathcal{L}^\infty$ de (Bondell et Reich 2008)[3] pour la fixation des coefficient des coefficients égaux en magnitude. Si l'on souhaite regrouper les coefficients de signes opposés en présence d'une corrélation négative élevée, cela est équivalent à un changement de signe des prédicteurs. Ainsi, l'égalité des coefficients est obtenue en pénalisant les différences entre paires et les sommes de coefficients entre paires. En particulier, (Sharma et al. (2013))[25] ont proposé un schéma de pénalisation avec des poids non négatifs w, dont les estimations sont des minimiseurs du problème :

$$\min_{\beta}\{\|y - X\beta\|^2 + \lambda(\sum_j w_j\beta_j + \sum_{j<k} w_{jk-}\,|\beta_k - \beta_j| + \sum_{j<k} w_{jk+}\,|\beta_j + \beta_k|)\} \tag{6.2.1}$$

Où $w_{jk+} = w_{jk-} = \alpha$ pour tout $1 \leqslant j < k \leqslant p$, et $w_j = 1$ pour tout $j = 1, ..., p$.

La pénalité prévue dans l'expression (6.2.1) consiste en une norme $\mathcal{L}^1$ pondérée des coefficients qui encourage leur annulation (modèle parcimonieux) et une pénalité sur les différences et les sommes de paires de coefficients qui encourage l'égalité de coefficients. La pénalité pondérée sur les différences de paires de coefficients encourage les coefficients de même signes à être fixés comme étant égaux, tandis que la pénalité pondérée sur les sommes de coefficients encourage les coefficients de signes opposés à être fixés à une valeur égale. Notons que les pondérations sont des nombres non négatifs prédéfinis, alors cette pénalité est appelée : la pénalité de regroupement absolu par paires et de dispersions, PACS (Sharma et al. (2013))[25].

(Sharma et al. (2013)) considèrent plutôt (6.2.1) avec $w_{jk+} = (1 + r_{jk})^{-1}$, $w_{jk-} = (1 - r_{jk})^{-1}$ ou bien $w_{jk+} = 1_{\{r_{jk} < -c\}}$, $w_{jk-} = 1_{\{r_{jk} > c\}}$, où $1_{\{A\}}$ est une variable indicatrice qui est égale à 1 si l'événement A se maintient, et 0 dans le cas contraire. Cette approche de regroupement absolu par paires et de dispersions (PACS) vise à encourager les caractéristiques corrélées à prendre des valeurs de coefficient similaires. Par contre, le cluster elastic net encourage les caractéristiques corrélées et qui ont une similarité avec la variable réponse y à prendre des valeurs de coefficient similaires. Cette différence est visible dans (3.2.4) qui rappelle que les groupes ou clusters $C_1, ..., C_k$ sont obtenus sur la base des variables $X_1\widehat{\beta}_1, ..., X_p\widehat{\beta}_p$. La corrélation entre les caractéristiques joue un rôle dans la détermination de la mesure dans laquelle les estimations des coefficients des caractéristiques sont mises en commun tant pour le cluster elastic net que pour le

PACS; toutefois, il faut noter que seul le cluster elastic net fait que l'association avec la variable réponse joue également un rôle.

Une autre différence concerne l'utilisation d'une pénalité $\mathcal{L}^1$ sur les paires de coefficients par le PACS, par opposition à une pénalité $\mathcal{L}^2$ par le cluster elastic net. Le cluster elastic net effectue une forme plus légère de rétrécissement, car il n'encourage pas les valeurs des coefficients à être exactement identiques.

7. Étude sur les données de simulation

7.1 Formulation

On fait une étude des données simulées selon le modèle de régression linéaire $y = X\beta + \epsilon$, on simule un jeu de donnée avec $n = 120$ observations et $p = 500$ variables. Les erreurs $\epsilon_1, ..., \epsilon_n$ sont i.i.d[1] avec une distribution $\mathcal{N}(0, 2.5^2)$. Les observations (lignes de X) sont identiques à celles d'une distribution $\mathcal{N}(0, \Sigma)$, où Σ est une $p \times p$ matrice diagonale par bloc définie par :

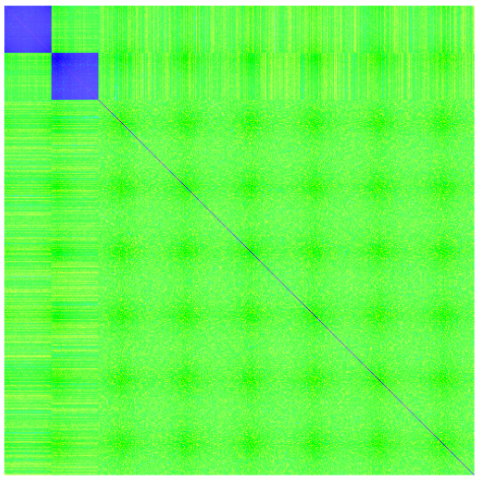


Figure 7.1: Heatmap de la matrice Σ, pour $\rho = 0.8$.

Comme suit :

$$\Sigma_{ij} = \begin{cases} 1 & \text{si } i = j \\ \rho & \text{si } i \leqslant 50\ ,\ j \leqslant 50\ ;\ i \neq j \\ \rho & \text{si } 51 \leqslant i \leqslant 100\ ;\ 51 \leqslant j \leqslant 100\ ;\ i \neq j \\ 0 & \text{sinon} \end{cases} \tag{7.1.1}$$

Nous allons explorer pour quelques valeurs de ρ comprises entre 0 et 0.8.

On simule également le vecteur β de la manière suivante :

$$\begin{cases} \beta_j \sim Unif[0.9, 1.1] & \text{si } 1 \leqslant j \leqslant 25 \\ \beta_j \sim Unif[-1.1, -0.9] & \text{si } 51 \leqslant j \leqslant 75 \\ \beta_j = 0 & \text{sinon} \end{cases}$$

En d'autres termes, nous allons générer trois groupes de prédicteurs C_1, C_2 et C_3 avec $K = 3$. Chaque groupe contient respectivement 50, 50 et 400 prédicteurs. Les prédicteurs des groupes C_1 et C_2 sont corrélés et il y'a 50% des prédicteurs qui sont associés à la variable réponse, c'est

[1] indépendantes et identiquement distribuées

à dire 25 prédicteurs dans chaque groupe. Les 400 prédicteurs du groupe C_3 ne sont pas corrélés entre eux et ne sont pas associés à la variable réponse.

En utilisant les critères de simulation dans (7.1.1), nous avons généré un ensemble d'entrainement de 120 observations, un ensemble de validation de 120 observations et un ensemble test de 300 observations.
Le vecteur coefficient de β prend la forme :

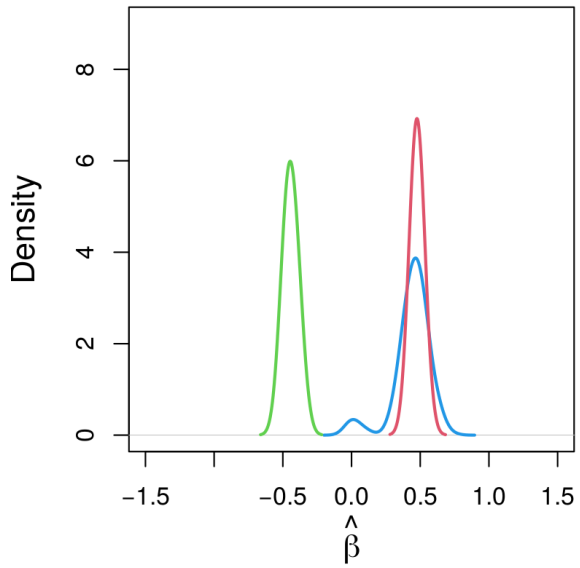


Figure 7.2: Graphique de densité de $\widehat{\beta}$ en utilisant le CEN avec $\delta = 0$, pour $\rho = 0.8$.

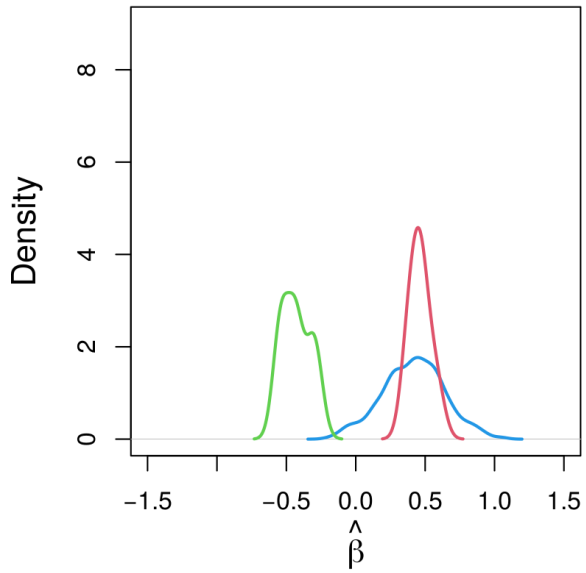


Figure 7.3: Graphique de densité de $\widehat{\beta}$ en utilisant la régression de ridge, pour $\rho = 0.8$.

Dans les figures (7.2) et (7.3), il existe trois ensembles de caractéristiques indiqués en vert, bleu et rouge. Dans les groupes vert et rouge, tous les coefficients sont respectivements égaux à -0.5 et 0.5. Dans le groupe bleu, la plupart des coefficients sont égaux à 0.5, aussi il y'a d'autres qui sont égaux à 0 et 1.

7.2 Résultats

Dans les tableaux ci-dessous, nous présentons les différents résultats de simulations pour la valeur de K = 3 groupes. Notons que, pour calculer ces statistiques nous avons pris une grille de 25 valeurs pour estimer les paramètres de réglage λ et δ, tous les modèles ont été entraînés avec la même grille de valeurs de λ et δ. Ainsi les resultats de toutes les méthodes sont affichés pour quelques valeurs de ρ comprises entre 0 et 0.8. Cependant, tous les modèles ont été ajustés sur un jeu d'entrainement en utilisant l'ensemble des paramètres de réglage qui a conduit à la plus petite valeur de $\|X\beta - X\widehat{\beta}\|^2$ sur un ensemble de validation (deuxième colonne); dans la troisième colonne nous avons l'erreur sur l'ensemble test; la quatrième colonne représente Num.non-zero qui désigne le nombre d'éléments non nuls dans $\widehat{\beta}$ (c'est à dire $\widehat{\beta}_j \neq 0$); enfin, dans la dernière colonne de chaque tableau on rapporte le Rand Index[27] qui mesure la concordance entre les groupes estimés et les vrais groupes. Tous ces modèles sont entraînés suivant 30 itérations et les résultats sont affichés comme suit : Moyenne(et SE : Standard Error).

Pour $\rho = 0.0$,

Méthodes	$\|X\beta - X\widehat{\beta}\|^2$	Correct Sparsity	Num. Non-Zeros	Rand Index
CEN	54.103(0.029)	0.159(0)	98.576(1.517)	0.814(0)
Ridge	49.906(0.001)	0.154(0)	500(0)	0.771(0.001)
Lasso	50.233(0)	0.154(0)	0(0)	0.815(0)
EN	50.22(0.003)	0.154(0)	20(3.922)	0.813(0)

Pour $\rho = 0.1$,

Méthodes	$\|X\beta - X\widehat{\beta}\|^2$	Correct Sparsity	Num. Non-Zeros	Rand Index
CEN	114.223(0.144)	0.314(0.001)	142.691(1.115)	0.826(0)
Ridge	172.719(0.013)	0.516(0)	500(0)	0.831(0)
Lasso	178.882(0)	0.535(0)	0(0)	0.815(0)
EN	178.635(0.048)	0.534(0)	20(3.922)	0.815(0)

Pour $\rho = 0.2$,

Méthodes	$\|X\beta - X\widehat{\beta}\|^2$	Correct Sparsity	Num. Non-Zeros	Rand Index
CEN	113.232(0.576)	0.313(0.001)	158.25(1.067)	0.853(0)
Ridge	283.259(0.049)	0.843(0)	500(0)	0.828(0)
Lasso	308.035(0)	0.919(0)	0(0)	0.815(0)
EN	307.044(0.195)	0.916(0.001)	20(3.922)	0.815(0)

Pour $\rho = 0.5$,

Méthodes	$\|X\beta - X\widehat{\beta}\|^2$	Correct Sparsity	Num. Non-Zeros	Rand Index
CEN	45.062(0.195)	0.142(0.001)	150.157(1.096)	0.845(0)
Ridge	514.419(0.292)	1.549(0.001)	500(0)	0.825(0)
Lasso	687.998(0)	2.08(0)	0(0)	0.815(0)
EN	680.963(1.362)	2.058(0.004)	20.424(3.92)	0.816(0)

Pour $\rho = 0.8$.

Méthodes	$\|X\beta - X\widehat{\beta}\|^2$	Correct Sparsity	Num. Non-Zeros	Rand Index
CEN	21.174(0.128)	0.07(0)	155.862(1.101)	0.831(0)
Ridge	609.217(0.608)	1.868(0.002)	500(0)	0.825(0)
Lasso	1054.634(0)	3.253(0)	0(0)	0.815(0)
EN	1031.06(3.635)	3.18(0.011)	24(3.968)	0.816(0)

Pour calculer ces statistiques nous avons eu à produire des groupes de prédicteurs. Pour la méthode du Cluster Elastic Net, nous avons déterminé les groupes de prédicteurs en procédant en deux étapes :

i) Nous avons calculé $\widehat{\beta}$ en utilisant **l'algorithme 1** discuté au chapitre (4);

ii) Nous avons fait une K-means basée sur $X_1\widehat{\beta}_1, X_2\widehat{\beta}_2, ..., X_p\widehat{\beta}_p$ pour estimer les groupes de prédicteurs.

La méthode du CEN estime les $\widehat{\beta}$ et regroupe les prédicteurs en même temps. Ainsi, les groupes de prédicteurs par cette méthode sont donnés directement par **l'algorithme 1** (voir chapitre 4) après convergence.

Comme nous l'avons dit au chapitre (1), la méthode (Cluster Elastic Net) tient compte de la structure des groupes de prédicteurs et l'association de ces groupes avec la variable réponse.

Pour ces différentes valeurs de ρ, les tableaux des statistiques ci-dessus montrent a fortiori que l'approche du CEN est meilleure en terme de regroupement des prédicteurs associés à la variable réponse que les autres méthodes. Il sied aussi de noter que dans le cadre des données simulées nous connaissons au départ le nombre de groupes K, mais cette information n'est pas disponible dans le cadre des données réelles. Cependant, pour mieux apprécier l'impact du nombre de groupes sur l'efficacité des résultats par la méthode du CEN, nous faisons une étude de simulation avec différentes valeurs de K (différents groupes), c'est à dire pour $K = \{2, 3, 5, 7\}$. Ainsi, pour la valeur de $\rho = 0.5$, nous obtenons les statistiques ci-dessous :

	$\|X\beta - X\widehat{\beta}\|^2$	Correct Sparsity	Num. Non-Zeros	Rand Index
K = 2	50.95(0.174)	0.16(0)	160.645(1.136)	0.84(0)
K = 3	45.062(0.195)	0.142(0.001)	150.157(1.096)	0.845(0)
K = 5	66.429(0.165)	0.193(0)	146.053(1.162)	0.837(0)
K = 7	73.726(0.161)	0.227(0)	140.494(0.937)	0.844(0)

Dans ce tableau, on constate que les meilleurs résultats des erreurs d'apprentissage sur les jeux de données de validation et de test sont donnés pour le nombre de groupes K = 3. Les résultats obtenus pour les autres valeurs de K sont également bons. On remarque aussi que la mauvaise spécification du nombre de groupes a priori n'influence pas la performance de la méthode du cluster elastic net.

À l'instar du Cluster Elastic Net Regression proposé par Witten et al. (2014), Bradley S. Price et Ben Sherwood (2018), ont proposé une méthode qui est similaire à celle de Witten et al. (2014) appelée Cluster Elastic Net for Multivariate Regression[22]. Contrairement à la méthode proposée par Witten et al. (2014) où la variable réponse **y** est de dimension $n \times 1$, la méthode du Cluster Elastic Net for Multivariate Regression dispose de plusieurs sorties (réponses multiples) autrement dit, la variable réponse **y** est de dimension $n \times r, (r > 1)$. Cependant, la spécificité de la méthode du Cluster Elastic Net for Multivariate Regression est que, en plus d'estimer simultanément les groupes de variables et les coefficients de régression, cette méthode permet de regrouper les variables réponse dans un modèle de regression multivariée[1] (modèles ayant plusieurs sorties ou réponses multiples) afin d'augmenter la précision des prévisions et de donner une indication de la relation entre les variables réponse. Notons que si $r = 1$, le Cluster Elastic Net for Multivariate Regression se réduit au Cluster Elastic Net.

Ainsi, suite à leurs travaux, Bradley S. Price et Ben Sherwood ont mis en place un package R nommé **'mcen'**[2][26]. Le package mcen est conçu pour la régression multivariée c'est à dire lorsque nous avons des réponses multiples. En d'autres termes, si la variable réponse est une matrice, dans ce cas l'utilisation du package R mcen est conseillée pour l'implémentation. Dans ce document, nous avons travaillé avec une seule réponse (ici la variable réponse est un vecteur), donc nous n'étions pas en mésure d'utiliser le package mcen de Bradley S. Price et Ben Sherwood (2018), ce qui a conduit à procéder au codage à la main en utilisant le logiciel R[3] dans l'implémentation et la simulation des statistiques obtenues dans les tableaux ci-dessus.

[2] multivariate cluster elastic net

[3] R version 4.0.0 (2020-04-24),
Platform: x86-64-pc-linux-gnu (64-bit),
Running under: Debian GNU/Linux 9 (stretch).

8. Conclusion et perspectives

Suite au succès des approches de Tibshirani (1996) et de Yang et Zou (2013) pour les calculs des solutions des coefficients de la régression de lasso et de elastic net respectivement, Daniela M. Witten et al. (2014) proposent une méthode de régularisation du cluster elastic net qui tient compte de la structure des groupes de prédicteurs et de leurs associations avec la variable réponse. Cette méthode s'inscrit dans le cadre de l'apprentissage supervisé. Autrement dit, ils avaient présenté une approche (pour les données en grande dimension) qui fait une selection des variables et qui trouve les coefficients de regression simultanément. Dans ce mémoire, il était question de comparer la méthode du cluster elastic net proposée par Daniela M. Witten et al. (2014), aux autres méthodes de régularisation qui disposent des groupes de variables a priori. En effet, dans ce document nous avons fait la comparaison du cluster elastic net avec quelques méthodes disposant des groupes de variables a priori à savoir : ridge, lasso et elastic net. Ainsi, les résultats sur les données simulées montrent une bonne performance sur la méthode du cluster elastic net comparée à celles disposant des groupes de variables a priori.

Comme perspectives, nous envisageons de comparer le cluster elastic net avec toutes les autres méthodes de régularisations disposant des groupes de variables a priori telles que : le group de lasso, le cluster group lasso, le cluster elastic net avec groupes connus... afin de savoir avec plus de précision et plus de généralité lequel des deux groupes de méthodes s'adapte mieux sur les données. Enfin, une autre perspective dans ce mémoire c'est d'entraîner la méthode du cluster elastic net et les autres méthodes disposant des groupes de variables a priori sur les données réelles afin d'évaluer et de comparer les performances. Notons que pour les données réelles, l'information du nombre de groupes K au départ n'est pas disponible.

Appendice

8.1 Preuve de la proposition 4.1.1

Preuve : L'estimateur de β par la méthode de CEN est défini par :

$$\widehat{\beta}^{CEN} = \arg\min_{\beta}\{\|y - X\beta\|^2 + \delta\|\beta\|_1 + \frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}\|X_j\beta_j - X_l\beta_l\|^2\} \tag{8.1.1}$$

Nous allons maintenant chercher la forme explicite de la solution de l'équation (8.1.1) pour chaque β_j, en gardant tous les autres paramètres fixes. L'équation (8.1.1) est équivalente à l'écriture matricielle suivante :

$$\begin{aligned}\widehat{\beta}^{CEN} &= \arg\min_{\beta}\{(y-X\beta)^T(y-X\beta)+\delta\sum_{j=1}^{p}|\beta_{-j}|+\frac{\lambda}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{j,l\in C_k}(X_{-j}\beta_{-j}-X_l\beta_l)^T(X_{-j}\beta_{-j}-X_l\beta_l)\}\\ &= \arg\min_{\beta}\{y^Ty - 2y^TX\beta + \beta^TX^TX\beta + \delta\sum_{j=1}^{p}|\beta_{-j}| + \lambda\sum_{j,l\in C_k}\beta_{-j}^TM_{jl}\beta_l\}.\end{aligned}$$

Nous rappelons de (3.2.6) que :

$$M_{jl} = \begin{cases}(|C_k|-1)/|C_k| & si\ \ j = l \in C_k\\ -r_{jl}/|C_k| & si\ \ j \neq l\ \ et\ \ j,l \in C_k\\ 0 & sinon\end{cases}$$

Si nous gardons seulement les quantités qui dépendent de β, alors la fonction à minimiser en β_j sera :

$$L(\beta_j) = -2y^Tx_j\beta_j + \sum_{i=1}^{n}x_{ij}^2\beta_j^2 + 2\sum_{i=1}^{n}\sum_{k\neq j}x_{ik}x_{ij}\beta_k\beta_j + \delta|\beta_{-j}| + \lambda\sum_{j,l\in C_k}\beta_{-j}^TM_{jl}\beta_l$$

D'une part, pour $j \neq l$, on a :

$$L(\beta_j) = -2y^Tx_j\beta_j + \sum_{i=1}^{n}x_{ij}^2\beta_j^2 + 2\sum_{i=1}^{n}\sum_{k\neq j}x_{ik}x_{ij}\beta_k\beta_j + \delta|\beta_{-j}| - \lambda\sum_{l\in C_k, j\neq l}\frac{r_{jl}}{|C_k|}\beta_{-j}^T\beta_l$$

C'est une équation quadratique en β_j dérivable pour $\beta_j \neq 0$ et qui a pour sous-gradient,

$$\frac{\partial L(\beta_j)}{\partial\beta_j} = -2y^Tx_j + 2\sum_{i=1}^{n}x_{ij}^2\beta_j + 2\sum_{k\neq j,i}x_{ik}x_{ij}\beta_k + \delta sign(\beta_{-j}) - \frac{\lambda}{|C_k|}\sum_{l\in C_k, j\neq l}r_{jl}\beta_l,$$

Alors,

$$-2y^T x_j + 2\sum_{i=1}^{n} x_{ij}^2\beta_j + 2\sum_{k\neq j,i} x_{ik}x_{ij}\beta_k + \delta sign(\beta_{-j}) - \frac{\lambda}{|C_k|}\sum_{l\in C_k, j\neq l} r_{jl}\beta_l = 0,$$

Nous savons que $\sum_{i=1}^{n} x_{ij}^2 = 1$, car les variables sont normalisées. Ainsi nous obtenons,

$$\widehat{\beta}_j^{(1)} = \widetilde{y}_j^T x_j + \frac{\lambda}{2|C_k|}\sum_{l\in C_k, j\neq l} r_{jl}\beta_l - \frac{\delta}{2}sign(\beta_{-j})$$

Avec $\widetilde{y}_j^T = y^T - \sum_{k\neq j,i} x_{ik}\beta_k = y - X_{-j}\beta_{-j}$.
C'est équivalent à :

$$\widehat{\beta}_j^{(1)} = sign\{\widetilde{y}_j^T x_j + \frac{\lambda}{2|C_k|}\sum_{l\in C_k, j\neq l} r_{jl}\beta_l\}\max(0, |\widetilde{y}_j^T x_j + \frac{\lambda}{2|C_k|}\sum_{l\in C_k, j\neq l} r_{jl}\beta_l| - \frac{\delta}{2}),$$

Alors,

$$\widehat{\beta}_j^{(1)} = S(\widetilde{y}_j^T x_j + \frac{\lambda}{2|C_k|}\sum_{l\in C_k, j\neq l} r_{jl}\beta_l, \frac{\delta}{2}).$$

D'autre part, pour $j = l$, on a :

$$L(\beta_j) = -2y^T x_j\beta_j + \sum_{i=1}^{n} x_{ij}^2\beta_j^2 + 2\sum_{i=1}^{n}\sum_{k\neq j} x_{ik}x_{ij}\beta_k\beta_j + \delta|\beta_{-j}| + \lambda\sum_{j\in C_k}\frac{|C_k|-1}{|C_k|}\beta_{-j}^2$$

En dérivant suivant β_j et en ne gardant que les quantités qui dépendent de β, on a :

$$\beta_j + \sum_{j\in C_k}\beta_j + \lambda\frac{|C_k|-1}{|C_k|}\sum_{j\in C_k}\beta_{-j} = 0,$$

Ceci est équivalent à :

$$\widehat{\beta}_j^{(2)} = r_{jj}(1 + \frac{|C_k|-1}{|C_k|}),$$

Avec, $r_{jj} = \sum_{j\in C_k}\beta_j$.

Ceci étant, de manière générale pour $j = l$ ou $j \neq l$ nous pouvons écrire :

$$\beta_j \longleftarrow \frac{S(\widetilde{y}_j^T X_j + \frac{\lambda}{|C_k|}\sum_{l\in C_k, j\neq l}\beta_l r_{jl}, \delta/2)}{r_{jj}(1+\lambda\frac{|C_k|-1}{|C_k|})}$$

■

8.2 Preuve du lemme 5.2.1

Preuve : On écrit la matrice N sous la forme,

$$N = \begin{bmatrix} A_\star & B_\star \\ B_\star & A_\star \end{bmatrix}$$

Où $A_\star = aI + b11^T$ et $B_\star = c11^T$. $\mathbf{1}$ est un vecteur colonne formé des 1 et de longueur m.

En utilisant la formule de l'inverse partitionnée[17], on a :

$$N^{-1} = \begin{bmatrix} (A_\star - B_\star A_\star^{-1} B_\star)^{-1} & -A_\star^{-1}B_\star(A_\star - B_\star A_\star^{-1}B_\star)^{-1} \\ -A_\star^{-1}B_\star(A_\star - B_\star A_\star^{-1}B_\star)^{-1} & (A_\star - B_\star A_\star^{-1}B_\star)^{-1} \end{bmatrix}$$

D'après la formule Sherman-Morrison-Woodbury[7],

$$A_\star^{-1} = a^{-1}I - a^{-2}(b^{-1} + a^{-1}m)^{-1}11^T,$$

Par conséquent,

$$\begin{aligned} A_\star - B_\star A_\star^{-1} B_\star &= aI + b11^T - c11^T(a^{-1}I - a^{-2}(b^{-1}+a^{-1}m)^{-1}11^T)c11^T, \\ &= aI + (b - a^{-1}mc^2 + a^{-2}c^2m^2(b^{-1}+a^{-1}m)^{-1})11^T, \end{aligned}$$

On pose:

$$d_1 = b - a^{-1}mc^2 + a^{-2}c^2m^2(b^{-1}+a^{-1}m)^{-1} = \frac{ab + b^2m - c^2m}{a+bm}$$

Ensuite en calculant l'inverse on trouve :

$$(A_\star - B_\star A_\star^{-1}B_\star)^{-1} = a^{-1}I - a^{-2}(d_1^{-1} + a^{-1}m)^{-1}11^T$$

Alors,

$$-A_\star^{-1}B_\star(A_\star - B_\star A_\star^{-1}B_\star)^{-1} = d_2 11^T,$$

Avec,

$$d_2 = -ca^{-2}(1-a^{-1}m(d_1^{-1}+a^{-1}m)^{-1}-a^{-1}m(b^{-1}+a^{-1}m)^{-1}+a^{-2}m^2(b^{-1}+a^{-1}m)^{-1}(d_1^{-1}+a^{-1}m)^{-1}),$$

Après calcul de l'expression ci-dessus, on a :

$$d_2 = -\frac{c}{a^2+2abm+(b^2-c^2)m^2}.$$

■

8.3 Preuve du lemme 5.2.2

Preuve : On peut écrire ces deux matrices sous la forme :

$$N = \begin{bmatrix} A_\star & B_\star \\ B_\star & A_\star \end{bmatrix}$$

et,

$$N_\star = \begin{bmatrix} A_{\star\star} & B_{\star\star} \\ B_{\star\star} & A_{\star\star} \end{bmatrix}$$

où $A_\star = aI + b11^T$, $B_\star == c11^T$, $A_{\star\star} = a_\star I + b_\star 11^T$ *et* $B_{\star\star} = c_\star 11^T$.

En faisant le produit de $NN_\star N$ on a :

$$NN_\star N = \begin{bmatrix} A_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) + B_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) & A_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) + B_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) \\ A_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) + B_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) & A_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) + B_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) \end{bmatrix}$$

Après avoir calculé les différentes quantités on trouve :

$$A_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) = a^2 a_\star I + K_1 11^T,$$

$$B_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) = L_1 11^T,$$
$$B_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) = K_2 11^T,$$
$$A_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) = L_2 11^T,$$

Par conséquent,

$$A_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) + B_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) = a^2 a_\star I + (K_1 + K_2)11^T,$$

et

$$A_\star(A_\star A_{\star\star} + B_\star B_{\star\star}) + B_\star(A_\star B_{\star\star} + B_\star A_{\star\star}) = (L_1 + L_2)11^T,$$

On pose $K = K_1 + K_2$ et $L = L_1 + L_2$ ou $K_1, K_2, L_1\, L_2$ sont les matrices composées de $a, b, c, a_\star, b_\star$ et $c_\star$, alors la matrice devient :

$$NN_\star N = \begin{bmatrix} a^2a_\star I + K11^T & L11^T \\ L11^T & a^2a_\star I + K11^T \end{bmatrix},$$

$$NN_\star Nv = \begin{bmatrix} a^2a_\star I \\ -a^2a_\star I \end{bmatrix},$$

Car v est un vecteur contenant les zéros, sauf pour $v_j = 1$ et $v_l = -1$. En multipliant par v^T, nous avons bien le résultat escompté.
D'où,

$$v^T NN_\star Nv = 2a^2a_\star$$

■

8.4 Preuve du lemme 5.2.3

Preuve : A partir de la preuve du lemme (5.2.2), nous déduisons que :

$$NN_\star = \begin{bmatrix} A_\star A_{\star\star} + B_\star B_{\star\star} & A_\star B_{\star\star} + B_\star A_{\star\star} \\ A_\star B_{\star\star} + B_\star A_{\star\star} & A_\star A_{\star\star} + B_\star B_{\star\star} \end{bmatrix}$$

où $A_\star A_{\star\star} + B_\star B_{\star\star} = aa_\star I + (cc_\star + ab_\star + ba_\star + bb_\star)11^T$ et $A_\star B_{\star\star} + B_\star A_{\star\star} = (ac_\star + bc_\star + ca_\star + cb_\star)11^T$,

On multiplie par le vecteur $\beta = (\beta_1, ..., \beta_1, \beta_2, ..., \beta_2)^T$ on a :

$$NN_\star\beta = \begin{bmatrix} (aa_\star I + (cc_\star + ab_\star + ba_\star + bb_\star)11^T)(\beta_1, ..., \beta_1) + (ac_\star + bc_\star + ca_\star + cb_\star)(\beta_2, ..., \beta_2)11^T \\ (ac_\star + bc_\star + ca_\star + cb_\star)(\beta_1, ..., \beta_1)11^T + (aa_\star I + (cc_\star + ab_\star + ba_\star + bb_\star)11^T)(\beta_2, ..., \beta_2) \end{bmatrix},$$

En multipliant par v^T et en sommant on a :

$$v^T NN_\star\beta = (aa_\star I + (cc_\star + ab_\star + ba_\star + bb_\star)11^T)(\beta_1 - \beta_2, ..., \beta_1 - \beta_2) - (ac_\star + bc_\star + ca_\star + cb_\star)(\beta_1 - \beta_2, ..., \beta_1 - \beta_2)11^T$$

Par consequent,

$$\frac{v^T NN_\star\beta}{\beta_1 - \beta_2} = aa_\star I + (cc_\star + ab_\star + ba_\star + bb_\star)(1, ..., 1)11^T - (ac_\star + bc_\star + ca_\star + cb_\star)(1, ..., 1)11^T,$$

D'après l'hypothèse **(i)**[1], il existe deux groupes connus ayant chacun la taille $m = p/2$. Ceci étant, on obtient après calcul :

$$\begin{aligned}\frac{v^T N N_\star \beta}{\beta_1 - \beta_2} &= aa_\star I + (ab_\star - ac_\star + ba_\star - ca_\star)\frac{p}{2} + (bb_\star - bc_\star + cc_\star - cb_\star)\frac{p}{2} \times \frac{p}{2} \\ &= (a + (b-c)\frac{p}{2})(a_\star + (b_\star - c_\star)\frac{p}{2}).\end{aligned}$$

■

[1] Hypothèse vue au chapitre (5)

Références

[1] Introduction to multivariate regression analysis. *Hippokratia*, pages 23–28, 12 2010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/.

[2] Yves Aragon. Régression linéaire par la méthode des moindres carrés. 01 2011. doi: 10.1007/978-2-8178-0208-4_3.

[3] Howard Bondell and Brian Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64:115–23, 04 2008. doi: 10.1111/j.1541-0420.2007.00843.x.

[4] David Borthwick. *Hilbert Spaces*, pages 5–33. 03 2020. ISBN 978-3-030-38001-4. doi: 10.1007/978-3-030-38002-1_2.

[5] Eric Dalissier, Charles Dapogny, Sofiane Hendili, Lise-Marie Imbert-Gérard, and Thomas Pradeau. Semaine d'etude mathématiques et entreprises 2 : Analyse de grands volumes de données en grande dimension. 12 2011.

[6] Frédéric de Gournay et Aude Rondepierre. pages 1–13. URL https://www.math.univ-toulouse.fr/~rondep/CoursTD/polyMIC3.pdf.

[7] Chunyuan Deng. A generalization of the sherman-morrison-woodbury formula. *Appl. Math. Lett.*, 24:1561–1564, 09 2011. doi: 10.1016/j.aml.2011.03.046.

[8] Marcel Dettling and Peter Bühlmann. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90:106–131, 07 2004. doi: 10.1016/j.jmva.2004.02.012.

[9] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1, 09 2007. doi: 10.1214/07-AOAS131.

[10] Trevor Hastie. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 01 2009. ISBN 9780387848570. doi: 10.1007/978-0-387-84858-7.

[11] Trevor Hastie, Robert Tibshirani, David Botstein, and Patrick Brown. Supervised harvesting of expression trees. *Genome biology*, 2, 01 2001.

[12] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 8:27–51, 01 1970.

[13] Peg Howland and Haesun Park. *Cluster-Preserving Dimension Reduction Methods for Efficient Classification of Text Data*, pages 3–23. 01 2004. doi: 10.1007/978-1-4757-4305-0_1.

[14] RACHID KHAROUBI. Une nouvelle approche pour la sélection des variables dans le cas de modèles de discrimination en grandes dimensions. pages 6–10, 06 2016.

[15] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics (Oxford, England)*, 24:1175–82, 06 2008. doi: 10.1093/bioinformatics/btn081.

[16] Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structure covariates with an application to genomics. *The annals of applied statistics*, 4:1498–1516, 09 2010. doi: 10.1214/10-AOAS332.

[17] TZON-TZER LU and SHENG-HUA SHIOU. pages 1–11, 10 2000. URL http://msvlab.hre.ntou.edu.tw/grades/now/inte/Inverse%20&%20Border/border-LuTT.pdf.

[18] Clustering methods. *Consulté*, 05 2020. URL https://www.google.com/search?q=clustering+families+methods&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiznIOBh57pAhUH-aQKHeT9DaUQ_AUoAXoECA4QAw&biw=1920&bih=938#imgrc=vj6BMMaej_t3aM&imgdii=Ird4Fsjn4upxrM.

[19] Henri Mineur. Technique de la méthode des moindres carres. 04 2020.

[20] Jan Novotny, Paul Bilokon, Aris Galiotos, and Frédéric Délèze. *Linear Regression with Regularisation*, pages 391–417. 11 2019. ISBN 9781119404729. doi: 10.1002/9781119404729.ch20.

[21] Mee Park, Trevor Hastie, and Robert Tibshirani. Averaged gene expressions for regression. *Biostatistics (Oxford, England)*, 8:212–27, 05 2007. doi: 10.1093/biostatistics/kxl002.

[22] Bradley Price and Ben Sherwood. A cluster elastic net for multivariate regression. *Journal of Machine Learning Research*, 18:1–39, 07 2017.

[23] Assaf Rabinowicz and Saharon Rosset. Cross-validation for correlated data, 04 2019.

[24] Shankar Rathinasamy. Unstructured data growth and challenges. page 7, 2015.

[25] Dhruv Sharma, Howard Bondell, and Hao Zhang. Consistent group identification and variable selection in regression with correlated predictors. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 22:319–340, 04 2013. doi: 10.1080/15533174.2012.707849.

[26] Ben Sherwood and Brad Price. *mcen: Multivariate Cluster Elastic Net*, 2018. URL https://CRAN.R-project.org/package=mcen. R package version 1.0.

[27] Douglas Steinley, Michael Brusco, and Lawrence Hubert. The variance of the adjusted rand index. *Psychological methods*, 21, 02 2016. doi: 10.1037/met0000049.

[28] Ewout Steyerberg. *Overfitting and Optimism in Prediction Models*, pages 95–112. 07 2019. ISBN 978-3-030-16398-3. doi: 10.1007/978-3-030-16399-0_5.

[29] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288, 01 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[30] Daniela Witten, Ali Shojaie, and Fan Zhang. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, 56:112–122, 02 2014. doi: 10.1080/00401706.2013.810174.

[31] Hui Zou and T. Hastie. Regularization and variable selection via the elastic nets. *J. Royal Stat. Soc. B*, 67:301–320, 01 2015.