TP2 – Sur la Régression pénalisée

On utilisera le fichier de données **data_DNA.txt** disponible sur Campus associé à l'observation de 201 variables (en colonnes) sur 120 individus (en lignes). La première colonne est la réponse y à expliquer et qui correspond à l'expression d'une certaine protéine dans les tissus oculaires, protéine impliquée dans une maladie génétique. Les autres variables correspondent à l'expression de 200 gènes dans ces mêmes tissus. Ce TP aura pour but à la fois de comparer les méthodes de régression ridge et LASSO mais aussi d'identifier les variables (gènes) qui sont liées à la réponse (protéine). Ces données ont été obtenues grâce à des biopuces ou puces à ADN, voir wikipedia : https://en.wikipedia.org/wiki/DNA_microarray pour en savoir un peu plus.

PARTIE 1 : sur la méthode MCO lorsque p > n

1. Charger les données. Faire les transformations de centrage et réduction pour la matrice des prédicteurs et le centrage pour la réponse. On notera encore X la matrice des prédicteurs centrés et réduits et y la réponse centrée.

2. Faire la RLM (sans intercept β_0) de y contre tous les prédicteurs en utilisant la fonction **solve()** de R pour inverser le système des équations normales. Expliquer le message d'erreur. Quelle est la dimension de l'image de X ? Et la dimension de son noyau ?

3. Faire la même chose en utilisant cette fois la fonction usuelle **lm()** de R. Analyser les résultats obtenus. Tracer la réponse y contre la réponse prédite. Conclusion.

4. On va essayer de donner un sens précis à la solution par moindres carrés ordinaires (MCO) en reprenant l'idée de la régression pénalisée ridge. Pour cela, obtenir le graphique montrant les chemins de régularisation des coefficients $\hat{\beta}_j(\lambda)$ du modèle lorsque le paramètre de régularisation λ s'approche de la valeur 0 (on pourra utiliser une grille du type **lambda <- exp(- (1:20))**. Conclusion. De quel problème est solution le vecteur limite $\hat{\beta}(0)$ lorsque λ décroît vers 0 ? Tracer à nouveau la réponse y contre la réponse prédite.

PARTIE 2 : régression ridge

En utilisant le package **glmnet**, mettre en œuvre la régression ridge. Obtenir les coefficients ridge pour la valeur optimisée du paramètre λ par validation croisée. A quelle erreur de prédiction (MSE ou RMSE) cette valeur correspond ? Tracer encore la réponse y contre la réponse prédite pour le prédicteur ridge associé.

PARTIE 3 : régression LASSO pour identifier les « bons » gènes

Toujours en utilisant le package **glmnet**, mettre en œuvre la régression LASSO et faire la même analyse que précédemment. Combien de prédicteurs (gènes) sont ainsi « capturés » par le LASSO ?

PARTIE 4 : comparaison des différentes techniques pour la prédiction lorsque p > n

Pour valider et comparer au final différentes techniques de régression pour la prédiction, on découpera le jeu de données initial en un sous-ensemble de données pour l'apprentissage et un sous-ensemble de test pour évaluer la performance en prédiction ou l'erreur de généralisation des différentes techniques. Prendre par exemple 80% des données pour l'apprentissage et 20% pour l'évaluation.

Comparer ainsi les méthodes ridge, LASSO, PCR et PLS. On utilisera les packages **glmnet** et **pls** (pour les régressions PCR et PLS). Conclusions.