Examen Final

Durée 2h. Les documents, la calculatrice, les téléphones portables, tablettes, ordinateurs ne sont pas autorisés. Les exercices sont indépendants. La qualité de la rédaction sera prise en compte.

Exercice 1. Régression Ridge

On considère le modèle de régression $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ où $\mathbf{Y}$ est un vecteur aléatoire de $\mathbb{R}^n$, $\mathbf{X}$ est une matrice de taille $n \times p$, $\boldsymbol{\beta}$ un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\epsilon}$ un vecteur de $\mathbb{R}^n$ de variables aléatoires supposées indépendantes et identiquement distribuées, centrées et de variance σ^2.

1. **Régression des moindres carrés**

   (a) Rappelez la définition de l'estimateur des moindres carrés (MC) $\hat{\boldsymbol{\beta}}$ pour la régression multiple. Retrouvez son expression, en détaillant les étapes de calcul. Sous quelles hypothèses cette expression est-elle valide ?
   [On pourra utiliser la formule suivante : pour toute matrice $\mathbf{A}$ de taille $n \times p$ et vecteur $\mathbf{b}$ de taille n, le gradient de $f : \mathbf{x} \mapsto \|\mathbf{A}\mathbf{x} + \mathbf{b}\|^2$ est donné par : $\nabla_{\mathbf{x}} f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} + \mathbf{b})$.]

   (b) Retrouvez, en justifiant, l'espérance et la variance de cet estimateur (à σ^2 fixé).

   (c) On suppose que $\mathbf{X}^T\mathbf{X}$ admet la décomposition spectrale : $\mathbf{X}^T\mathbf{X} = \mathbf{V}^T\mathbf{D}\mathbf{V}$, où $\mathbf{V}$ est une matrice orthogonale, et $\mathbf{D}$ une matrice diagonale de coefficients diagonaux $(d_1, \cdots, d_p)$. Montrez que la variance de l'estimateur MC s'écrit $\mathbb{V}[\hat{\boldsymbol{\beta}}] = \sigma^2\mathbf{V}^T\mathbf{D}^{-1}\mathbf{V}$.

   (d) On définit le risque quadratique de l'estimateur MC par la formule : $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2]$. Interprétez cette définition. Qualitativement, que peut on dire d'un estimateur dont le risque quadratique est faible ?

   (e) Montrez l'égalité : $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = \mathrm{tr}(\mathbb{V}[\hat{\boldsymbol{\beta}}])$, où on note tr la trace d'une matrice. En utilisant les notations de la question (1-c), en déduire que $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = \sigma^2 \sum_{k=1}^{d} \frac{1}{d_k}$.

   (f) Rappelez l'énoncé du théorème de Gauss-Markov pour une régression multiple (sans le démontrer). Que peut-on en déduire sur l'optimalité du risque quadratique de l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}$?

   (g) On suppose que $p > n$. Que dire de l'estimateur des moindres carrés dans ce cas ?

   (h) On se place dans le cas où $n > p$, $p = 2$ et $\mathbf{X} = (x\ z)$, avec deux prédicteurs x et z centrés, réduits, et corrélés: $\sum_{i=1}^n x_i = \sum_{i=1}^n z_i = 0$, $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n z_i^2 = 1$ et $\rho(x,z) = \sum_{i=1}^n x_i z_i = \sqrt{1-\delta}$, avec δ un réel positif ($0 < \delta < 1$). Explicitez $\mathbf{X}^T\mathbf{X}$ dans ce cas, et vérifiez que $1 + \rho(x,z)$ et $1 - \rho(x,z)$ sont les valeurs propres de $\mathbf{X}^T\mathbf{X}$. En déduire que $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = 2\sigma^2/\delta$. Comment se comporte le risque quadratique lorsque les deux variables sont très corrélées ?

2. **Régression ridge**

   (a) On appelle estimateur *ridge* de paramètre λ ($\lambda > 0$) l'estimateur de $\boldsymbol{\beta}$ suivant:
   $$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}' \in \mathbb{R}^p}{\mathrm{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}'\|^2 + \lambda \|\boldsymbol{\beta}'\|^2 \right\}.$$
   Dans toute la suite, on considère que $\lambda > 0$ est donné et fixé. Montrez l'égalité: $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}$, avec $\mathbf{I}_p$ la matrice identité de taille p.

   (b) Calculez l'espérance et la variance de l'estimateur ridge (à σ^2 fixé).

   (c) On suppose que $p > n$. Que dire de l'estimateur ridge dans ce cas ?

   (d) On suppose la même décomposition spectrale $\mathbf{X}^T\mathbf{X} = \mathbf{V}^T\mathbf{D}\mathbf{V}$ que précédemment. Montrez que la variance de l'estimateur ridge s'écrit: $\mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda] = \sigma^2\mathbf{V}^T\mathbf{F}\mathbf{V}$, avec $\mathbf{F}$ une matrice diagonale telle que, pour tout $1 \le k \le p$: $F_{kk} = d_k/[(d_k + \lambda)^2]$.

(e) Déduire des questions précédentes que, lorsque les deux estimateurs sont bien définis, $\mathbb{V}[\hat{\boldsymbol{\beta}}] - \mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda]$ est une matrice définie positive.

(f) Est-ce que ce résultat est en contradiction avec le théorème de Gauss-Markov ?

(g) On se place dans le cadre de la question (1-h), où $p = 2$ et $\mathbf{X} = (x\ z)$, avec deux prédicteurs x et z centrés, réduits, et corrélés: $\sum_{i=1}^n x_i = \sum_{i=1}^n z_i = 0$, $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n z_i^2 = 1$ et $\rho(x, z) = \sum_{i=1}^n x_i z_i = \sqrt{1-\delta}$, avec δ un réel positif ($\delta < 1$). Montrez que dans ce cas: $\mathrm{tr}(\mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda]) \leq \frac{2\sigma^2}{(1+\lambda)^2} + \frac{\sigma^2}{\lambda^2}$. En quoi est-ce que ce comportement est différent de celui de l'estimateur des moindres carrés ?

(h) Montrez l'égalité : $\mathbb{E}[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^T] = \mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda] + (\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta})(\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta})^T$.
[On pourra remarquer que : $(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}_\lambda - \mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda]) + (\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta})$, et développer.]

(i) En déduire que le risque quadratique de l'estimateur ridge s'écrit :

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|^2\right] = \mathrm{tr}(\mathbb{V}\left[\hat{\boldsymbol{\beta}}_\lambda\right]) + \left\|\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta}\right\|^2.$$

(j) Que peut on en déduire (de manière qualitative) sur les risques quadratiques de estimateurs des moindre carrés et ridge ? Est-ce que l'un des deux estimateurs est préférable en terme de risque ? On pourra regarder la différence des risques quadratiques, et discuter d'un compromis biais-variance.

(k) D'après ce qui précède, dans quel(s) cas particulier(s) conseilleriez vous l'utilisation de l'estimateur ridge ?

Exercice 2. Musique sur Spotify

On s'intéresse à la popularité des musiques sur Spotify. On dispose d'un jeu de données de 1466 pistes de musiques, toutes publiées avant l'an 2000, comportant les variables suivantes :

- `popularity` : popularité du morceau, entre 0 (non populaire) et 100 (très populaire).
- `danceability`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`: scores, compris entre 0 et 1, décrivant diverses caractéristiques du morceau.
- `tempo` : tempo en "beat par minutes".
- `duration` : durée du morceau en ms.
- `loudness` : volume sonore moyen du morceau en décibels.
- `genre` : style musical. On ne garde que les musiques "pop", "rap" et "r&b".
- `year` : année de publication.
- `artist` : interprète.

Table 1: Extrait de quelques lignes et colones du jeu de données.

	popularity	danceability	tempo	duration	loudness	genre	year	artist
Wannabe	79	0.768	110.008	173027	-6.145	pop	1996	Spice Girls
One Way Or Another	55	0.442	162.272	217364	-5.086	pop	1978	Blondie
Straight Outta Compton	70	0.834	102.848	258688	-9.484	rap	1988	N.W.A.
Still D.R.E.	75	0.816	93.431	270587	-3.323	rap	1999	Dr. Dre
All I Want for Christmas Is You	90	0.335	150.277	241107	-7.462	r&b	1994	Mariah Carey
Ain't No Sunshine	76	0.479	79.593	125093	-11.451	r&b	1971	Bill Withers

1. On cherche à prédire la popularité d'un morceau en fonction de ses caractéristiques. On exécute les commandes suivantes dans `R` :

```
fit1 <- lm(popularity ~
             danceability + speechiness + acousticness + tempo +
             loudness + instrumentalness + liveness + duration + year,
           data = spotify_songs)
summary(fit1)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + speechiness + acousticness +
##     tempo + loudness + instrumentalness + liveness + duration +
##     year, data = spotify_songs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.135 -11.783   1.112  12.322  45.047
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.517e+02  1.379e+02   5.452 5.85e-08 ***
## danceability      1.256e+00  3.785e+00   0.332 0.739948
## speechiness       2.626e-01  4.135e+00   0.064 0.949370
## acousticness      1.684e+00  2.347e+00   0.717 0.473252
## tempo            -6.799e-03  1.567e-02  -0.434 0.664376
## loudness          7.495e-01  1.435e-01   5.223 2.02e-07 ***
## instrumentalness -2.037e+00  3.195e+00  -0.638 0.523824
## liveness          1.790e+00  2.820e+00   0.635 0.525728
## duration         -2.713e-05  8.066e-06  -3.364 0.000789 ***
## year             -3.482e-01  6.939e-02  -5.019 5.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.52 on 1456 degrees of freedom
## Multiple R-squared:  0.04359, Adjusted R-squared:  0.03768
## F-statistic: 7.374 on 9 and 1456 DF,  p-value: 1.411e-10
```

(a) À quoi correspondent les colonnes `Std. Error` et `t value` ? Ecrivez de manière formelle les tests correspondants, avec les hypothèses, l'expression de la statistique, et sa loi sous H_0. Interprétez les différents coefficients.

(b) À quoi correspond le `Multiple R-squared` ? Donnez son expression. Interprétez.

(c) À quoi correspond la `F-statistic` ? Ecrivez de manière formelle le test correspondant, avec les hypothèses, l'expression de la statistique, et sa loi sous H_0. Interprétez.

2. On complète l'analyse avec les commandes suivantes :

```
fit2 <- lm(popularity ~ loudness + duration + year, data = spotify_songs)
AIC(fit1, fit2)

##      df      AIC
## fit1 11 12395.34
## fit2  5 12385.17
```

(a) Donnez la définition de l'AIC. Quelle est son utilité ?

(b) Quel modèle de régression préférez-vous ?

3. On cherche à savoir si le style musical a une influence sur la popularité, avec l'analyse :

```
summary(lm(popularity ~ genre, data = spotify_songs))

##
## Call:
## lm(formula = popularity ~ genre, data = spotify_songs)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -39.464 -12.626   1.374  13.374  46.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.464      1.124  44.904  < 2e-16 ***
## genrer&b      -7.313      1.288  -5.679 1.63e-08 ***
## genrerap      -5.837      1.332  -4.382 1.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.67 on 1463 degrees of freedom
## Multiple R-squared:  0.02163, Adjusted R-squared:  0.02029
## F-statistic: 16.17 on 2 and 1463 DF,  p-value: 1.133e-07
```

(a) Explicitez le modèle linéaire utilisé par `R`. À quoi correspondent les coefficients `genrer&b` et `genrerap` ?

(b) D'après cette analyse, quelles sont les popularités moyennes respectives des musiques de pop, r&b et de rap ? Peut-on dire qu'elles sont significativement différentes ?

(c) D'après cette analyse, peut-on rejeter l'hypothèse suivant laquelle tous les genres de musique ont la même popularité ?

4. On effectue l'anova suivante:

```
anova(lm(popularity ~ loudness + year + duration * genre, data = spotify_songs))

## Analysis of Variance Table
##
## Response: popularity
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## loudness          1   5052  5051.6 18.8211 1.534e-05 ***
## year              1   9275  9274.8 34.5556 5.127e-09 ***
## duration          1   3286  3286.5 12.2446 0.0004807 ***
## genre             2   5852  2926.2 10.9021 1.997e-05 ***
## duration:genre    2    683   341.6  1.2729 0.2803353
## Residuals      1458 391331   268.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Explicitez le modèle utilisé par `R` dans cette analyse.

(b) En utilisant la commande précédente, pouvez-vous répondre aux questions suivantes ? Si vous n'avez pas assez d'élément pour répondre à une question, indiquez pourquoi. Sinon, indiquez précisément quelle partie de la sortie vous permet de conclure.

  i. L'ajout du régresseur `loudness` à un modèle contenant uniquement l'intercept améliore-t-il significativement le modèle ?

  ii. L'ajout du régresseur `year` à un modèle contenant uniquement l'intercept améliore-t-il significativement le modèle ?

  iii. L'ajout du régresseur `duration` à un modèle contenant uniquement l'intercept améliore-t-il significativement le modèle ?

  iv. L'ajout du facteur `genre` à un modèle contenant déjà l'intercept et les trois regresseurs continus `loudness`, `year` et `duration` améliore-t-il significativement le modèle ?

  v. L'ajout d'une pente spécifique à chaque genre dans la regression contre la `duration` améliore-t-il significativement le modèle ?

5. J'aimerais composer un morceau qui aie beaucoup de succès. Au vue des analyses précédentes, quels conseils pourriez vous me donner ?