2018-2019
2e année

– Statistique 1 – Énoncés des projets

Enseignant : M. Chopin

Consignes : Le projet de Statistique 1 est un travail à mener **en binôme.** Un binôme doit être constitué de **deux** étudiants du **même groupe de travaux dirigés**. La date limite de rendu du projet est le **vendredi 11 janvier**. Les chargés de TD doivent être avisés par email de la constitution **des binômes** avant **le lundi** 3 **décembre**. **Dans ce mail, chaque groupe doit hiérarchiser 3 choix de sujets**. Il y aura au maximum 2 binômes sur le même projet par groupe de TD.

Tout le travail de programmation doit être effectué sous le logiciel R. Le rapport doit être rédigé en LaTeXet une grande attention sera apportée à la rédaction, la présentation du rapport, la rigueur des réponses aux questions théoriques. Enfin, les codes R doivent être fournis en annexe du rapport, avec des commentaires précis dans le corps des programmes. Il est autorisé de rendre un rapport en anglais.

Correspondant
Christophe Gaillac
Bureau **3106**
✉ assistant-math@ensae.fr

Projet Statistique 1

Copules et modélisation de la dépendance entre marchés financiers

Alexis Derumigny, CREST-ENSAE, alexis.derumigny@ensae.fr

November 7, 2018

Toutes les variables aléatoires et tous les vecteurs aléatoires considérés ici seront toujours supposés continus, avec une densité par rapport à la mesure de Lebesgue. Si vous avez des questions, n'hésitez pas à me contacter par mail.

1 Introduction : autour du théorème de Sklar

Soient (X_1, X_2) un vecteur aléatoire de dimension 2. On notera $F_{1,2}$ leur fonction de répartition jointe, $f_{1,2}$ leur densité jointe, et respectivement F_1, F_2, f_1, f_2 leurs fonctions de répartitions et densités marginales.

1. Pour $u_1, u_2 \in [0,1]$, on pose $C_{1,2}(u_1, u_2) := F_{1,2}\big(F_1^{(-1)}(u_1), F_2^{(-1)}(u_2)\big)$, où $F^{(-1)}$ désigne l'inverse d'une fonction de répartition F, c'est-à-dire la fonction quantile associé à F. Montrer qu'on a alors $F_{1,2} = C_{1,2}\big(F_1, F_2\big)$.
2. On pose $U_1 := F_1(X_1)$, $U_2 := F_2(X_2)$.
   (a) Montrer que U_1 et U_2 suivent la même loi. Quelle est cette loi ?
   (b) Montrer que $C_{1,2}$ vérifie les propriétés d'une fonction de répartition à support sur $[0,1]^2$: croissance par rapport à chacun de ses arguments et limite de 0 en $(0,0)$ et de 1 en $(1,1)$.
   (c) Montrer que $C_{1,2}$ est la fonction de répartition du vecteur aléatoire (U_1, U_2).
   (d) On note $\mathcal{C}$ l'ensemble des fonctions de répartition sur $[0,1]^2$ à marges uniformes et dérivables. Dans toute la suite, on appellera copule tout élément de $\mathcal{C}$. Déduire de la question précédente que $C_{1,2}$ est une copule.
3. Montrer que l'application $\Lambda : F_{1,2} \mapsto (F_1, F_2, C_{1,2})$ est une bijection, dont la réciproque est donnée par $\Upsilon : (F_1, F_2, C_{1,2}) \mapsto C_{1,2}\big(F_1, F_2\big)$. Λ permet de décomposer une fonction de répartition bivariée en ses marges et sa copule (représentant toute la dépendance entre les deux variables), alors que Υ permet de faire la reconstruction inverse.

2 Estimation des copules Gaussiennes

Dans cette section uniquement, on supposera que $(X_1, X_2) \sim \mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$, avec $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$, $\rho \in]-1, 1[$.

1. Montrer que $C_{1,2}$ ne dépend pas du choix des paramètres $\mu_1, \mu_2, \sigma_1, \sigma_2$. On notera par la suite $C_{1,2} = C_\rho$.
2. On observe n couples $(Y_{i,1}, Y_{i,2})$, $i = 1, \dots, n$ i.i.d suivant la loi de marginales G_1, G_2 et de copule C_ρ, avec ρ inconnu à estimer. Autrement dit, la fonction de répartition de $(Y_{i,1}, Y_{i,2})$ est $\Upsilon(G_1, G_2, C_\rho) = \Big((t_1, t_2) \mapsto C_\rho\big(G_1(t_1), G_2(t_2)\big)\Big)$. Pour simplifier le problème, on suppose que l'on connaît G_1 et G_2. Dans cette question, on se donne un entier i fixé entre 1 et n et on pose $V_{i,1} = G_1(Y_{i,1})$ et $V_{i,2} = G_2(Y_{i,2})$.
   (a) Montrer que $\big(V_{i,1}, V_{i,2}\big) \overset{\text{loi}}{=} (U_1, U_2)$ et que la fonction de répartition de $\big(V_{i,1}, V_{i,2}\big)$ est C_ρ.
   (b) Soit Φ la fonction de répartition de la loi normale centrée réduite. Soient $X_{i,1} := \Phi^{-1}(V_{i,1})$ et $X_{i,2} := \Phi^{-1}(V_{i,2})$. Montrer que $\big(X_{i,1}, X_{i,2}\big) \overset{\text{loi}}{=} (X_1, X_2)$, avec $\mu_1 = \mu_2 = 0$ et $\sigma_1 = \sigma_2 = 1$.

3. On définit l'estimateur du maximum de vraisemblance de ρ par :

$$\hat{\rho} := \arg\max_{\rho} \sum_{i=1}^{n} \log c_\rho\big(G_1(Y_{i,1}), G_2(Y_{i,2})\big),$$

où c_ρ est la densité de la copule C_ρ, définie par $c_\rho := \partial^2 C_\rho(u_1, u_2)/\partial u_1 \partial u_2$. Dans cette question, on supposera qu'on a fixé $\mu_1 = \mu_2 = 0$ et $\sigma_1 = \sigma_2 = 1$.

   (a) Montrer que $\hat{\rho} = \arg\max_\rho \sum_{i=1}^{n} \log c_\rho\big(\Phi(X_{i,1}), \Phi(X_{i,2})\big)$.

   (b) Soit ϕ la densité de la loi normale centrée réduite. Montrer que $f_{1,2}(x_1, x_2) = \phi(x_1)\phi(x_2)c_\rho\big(\Phi(x_1), \Phi(x_2)\big)$ pour tout $x_1, x_2 \in \mathbb{R}$.

   (c) À l'aide des questions précédentes, montrer que $\hat{\rho}$ est l'estimateur du maximum de vraisemblance de ρ pour l'échantillon $(X_{i,1}, X_{i,2})_{i=1,\dots,n}$.

   (d) En déduire son comportement asymptotique.

4. (Question facultative) Dans la majorité des applications, on ne connaît pas les lois marginales G_1 et G_2, qui doivent donc être estimées. Comment modifierez-vous la définition de $\hat{\rho}$ pour en tenir compte ? Quelles en seraient les conséquences, selon vous ?

3 Application à des données simulées

Dans cette partie, on suit le cheminement inverse de la question 2.2. On commence par simuler $n = 500$ réalisations indépendantes $(X_{i,1}, X_{i,2})_{i=1,\dots,n}$ d'une loi normale centrée réduite avec une corrélation $\rho = 0.8$, puis on construit $V_{i,1} := \Phi(X_{i,1})$ et $V_{i,2} := \Phi(X_{i,2})$ pour chaque i. Finalement, on pose $Y_{i,1} := G_1^{-1}(V_{i,1})$ et $Y_{i,2} := G_2^{-1}(V_{i,2})$, où G_1 est la fonction de répartition d'une loi de Cauchy et G_2 celle d'une loi de Laplace de densité $g_2(x) = (1/2)\exp(-|x|)$.

1. Implémenter l'algorithme ci-dessus en R.

2. Représenter le nuage de points $(Y_{i,1}, Y_{i,2})_{i=1,\dots,n}$. Peut-on dire que les deux variables sont corrélées empiriquement ? Et d'un point de vue théorique ? Justifiez votre réponse.

4 Application à des données de marchés financiers

1. Télécharger les valeurs du CAC40 et du Dow Jones sur la période 2000 - 2018, en utilisant par exemple Yahoo Finance. On supprimera les jours où l'un des deux marchés n'est pas ouvert. Charger ces données dans un environnement R.

2. On pose $Y_{t,1} = (CAC40_t - CAC40_{t-1})/CAC40_{t-1}$ et $Y_{t,2} = (Dow_t - Dow_{t-1})/Dow_{t-1}$. On note $\hat{F}_1$ la fonction de répartition empirique de $(Y_{t,1})_t$ et $\hat{F}_2$ celle de $(Y_{t,2})_t$. On pourra utiliser la fonction `ecdf` de R. Pour chaque t, on pose $\hat{V}_{t,1} := \hat{F}_1(Y_{t,1})$ et $\hat{V}_{t,2} := \hat{F}_1(Y_{t,2})$. Représenter les nuages de points $(Y_{t,1}, Y_{t,2})_t$, puis $(V_{t,1}, V_{t,2})_t$. Que constatez-vous ?

3. En supposant que la copule des deux variables étudiées est Gaussienne, estimer ρ par maximum de vraisemblance à l'aide de la fonction `BiCopEst` du package R `VineCopula` et commenter la valeur obtenue.

4. Tracer les contours (lignes de niveau) de la densité de la copule gaussienne ainsi estimée en utilisant la fonction `plot.BiCop`. En utilisant la fonction `BiCopKDE`, tracer les contours d'un estimateur non-paramétrique de densité de la copule estimée sur l'échantillon $(V_{t,1}, V_{t,2})_t$. Comparer les deux graphiques ainsi obtenus et commenter. La copule Gaussienne est-elle une bonne approximation de la copule de (Y_1, Y_2), selon vous ? Justifier votre réponse.

Linear regression in finance: what is β?

Mehdi Tomas

Statistics 2A

November 20, 2018

1 Theory

In this section, we consider a zero-mean matrix $X \in \mathbb{R}^{n,m}$ and a zero-mean vector $Y \in \mathbb{R}^n$. We wish to solve the regression problem of the form:

$$y = X\beta,$$

For a certain vector $\beta \in \mathbb{R}^m$.

Question 1 *Compute the ordinary least square estimator.*

Question 2 *We wish to examine estimates of the regression problem with the sum of squares of coefficients capped. Consider therefore, for $c > 0$, the solution to the problem:*

$$\min_{\beta} \| y - X\beta \|_2,$$

$$\text{such that: } \|\beta\|_2 < c.$$

And show that there exists $\lambda > 0$ such that the solution to the above is of the form:

$$\beta_{ridge} = (XX^T + \lambda I)^{-1} X^T y.$$

We refer to this estimator as the ridge estimator.

Remark 1 *To do so, you may show that the problem is equivalent to minimizing the loss function*

$$\| y - X\beta \|_2 + \lambda \|\beta\|_2.$$

In further questions, we use the singular value decomposition of X and use the conventions $X = UDV^T$ with U and V orthogonal matrices and D a diagonal matrix ($U \in \mathbb{R}^{n,r}$, $D \in \mathbb{R}^{r,r}$, $V \in \mathbb{R}^{r,m}$).

Question 3 *Using the singular value decomposition, simplify both the expression of the ordinary least square estimator found in question 1 and of the ridge estimator found in question 2.*

Question 4 *Finally, express the predictions $\hat{y}_{OLS} = X\beta_{OLS}$ and $\hat{y}_{ridge} = X\beta_{ridge}$ using U, D and V.*

Question 5 *Comment on the influence of the parameter λ in the regression. What link is there between ridge regression and regression on the principal components of XX^T?*

2 Application to financial data

All further questions will be done using the data provided in the file *data.csv*. This data contains the closing prices of stocks, obtained from Yahoo Finance. There are 10 different companies from different sectors (Apple, Google, Microsoft, IBM, Amazon, Target, Walmart, JP-Morgan, Goldman Sachs, Morgan Stanley) and over 10 years (from January 2008 to January 2018).

Question 6 *Compute the daily log-returns of the price of each company and plot them as a function of time. Compute the mean of the log-returns of each company from January 2008 to January 2014. Compare them with the mean from January 2014 to January 2018.*

Question 7 *Fit an ordinary least square regression of the centered log-returns of all companies using the same day's log-returns of all other companies. This should yield an ordinary least square regression for each company. Comment on the coefficients of the linear regression.*

Question 8 *Apply the linear regression on out of sample on data from January 2014 to January 2018. Comment on your results.*

Question 9 *Repeat questions 7 and 8 and replace ordinary least squares by ridge regression. Comment on the difference between your results in light of the link between ordinary least squares and ridge regression.*

Question 10 *We now run the regression over the whole data. Plot the components of β_{ridge} as a function of the parameter λ for all companies. Comment on your results. Are there any common properties among different companies?*

3 Remarks

The report should consist in one pdf file and one code file. Code should be carefully commented. You can write your report with markdown to easily include formulas plots (Jupyter Notebooks and R notebooks are convenient). For ease of use and readability, Python is recommended.

Direct questions and send your report to: *mehdi.tomas17@imperial.ac.uk*.

Comparaison d'échantillons et tests multiples

Solenne Gaucher

Test simple

Les tests de comparaison d'échantillons sont utilisés pour détecter l'impact d'un traitement sur un ensemble de grandeurs d'intérêt. Par exemple, des scientifiques étudient en laboratoire l'influence de la consommation de maïs génétiquement modifié Mon863 sur le poids de rats. Dans ce cadre, ils nourrissent un groupe de rats avec un régime à base de maïs Mon863, et un groupe témoin de rats avec un régime à base de maïs non modifié génétiquement. Ils cherchent ensuite à tester l'hypothèse "les rats des deux groupes ont le même poids". Dans cette partie, on travaillera avec le jeu de données "Ratweight.csv".

1 - Variance connue

On observe des variables aléatoires $X_1^{(1)}, ..., X_{n_1}^{(1)}$ et $X_{n_2}^{(2)}, ..., X_1^{(2)}$ indépendantes, de loi respective $\mathcal{N}\left(\mu_1, \sigma^2\right)$ et $\mathcal{N}\left(\mu_2, \sigma^2\right)$. Dans un premier temps, on suppose la variance σ^2 connue. On souhaite tester l'hypothèse

$$\mathcal{H}_0 : \mu_1 = \mu_2$$
$$\text{contre } \mathcal{H}_1 : \mu_1 \neq \mu_2$$

1 - a) On note $\overline{X^{(1)}}$ la moyenne de l'échantillon $X^{(1)} \triangleq \left(X_1^{(1)}, ..., X_{n_1}^{(1)}\right)$ et $\overline{X^{(2)}}$ la moyenne de l'échantillon $X^{(2)} \triangleq \left(X_1^{(2)}, ..., X_{n_2}^{(2)}\right)$. Sous l'hypothèse $\mathcal{H}_0$, quelle est la loi de $\overline{X^{(1)}} - \overline{X^{(2)}}$?

1 - b) Déterminer un test de niveau α pour $\mathcal{H}_0$.

2 - Variance inconnue

Dans la pratique, la variance des grandeurs mesurées est inconnue et il faut l'estimer. On note $\widehat{\sigma_1^2}$ l'estimateur sans biais de la variance de l'échantillon $X^{(1)}$ et $\widehat{\sigma_2^2}$ l'estimateur sans biais de la variance de l'échantillon $X^{(2)}$.

2 - a) Rappeler l'expression de $\widehat{\sigma_1^2}$ et $\widehat{\sigma_2^2}$. Déterminer $\widehat{\sigma^2}$ le meilleur estimateur de la variance de l'échantillon $X \triangleq \left(X_1^{(1)}, ..., X_{n_1}^{(1)}, X_1^{(2)}, ..., X_{n_2}^{(2)}\right)$ parmi les estimateurs de la forme $\lambda\widehat{\sigma_1^2} + (1-\lambda)\widehat{\sigma_2^2}$ où $\lambda \in [0,1]$, et donner son expression en fonction de $X^{(1)}, X^{(2)}$. Montrer qu'il est sans biais.

2 - b) On rappelle le Théorème de Cochran :

Théorème (Cochran)**.** *Soit X un vecteur aléatoire de $\mathbb{R}^n$ de loi $\mathcal{N}(0_n, Id_n)$, F un sous espace de $\mathbb{R}^n$, F^T son orthogonal et P_F, P_{F^T} les matrices des projections orthogonales sur F et F^T. Alors $P_F X$, $P_{F^T} X$ sont indépendants et $\|P_F X\|^2$ suit une loi du χ^2 à dim(F) degrés de liberté.*

On note u le vecteur de $\mathbb{R}^{n_1+n_2}$ tel que $u_i = \frac{1}{n_1} 1_{1 \leq i \leq n_1}$, v le vecteur tel que $v_i = \frac{1}{n_2} 1_{n_1+1 \leq i \leq n_1+n_2}$ et $F = \text{Vect}(u, v)$. Montrer que le vecteur $P_F X$ a pour coordonnées $(\overline{X^{(1)}}, \overline{X^{(2)}})$ dans (u, v) une base orthonormée de F. Exprimer $P_{F^T} X$ dans la base canonique. En déduire que $(\overline{X^{(1)}}, \overline{X^{(2)}})$ est indépendant de $\widehat{\sigma^2}$. Quelle est la loi de $\widehat{\sigma^2}$?

2 - c) Sous l'hypothèse $\mathcal{H}_0$, quelle est la loi de $\frac{\overline{X^{(1)}} - \overline{X^{(2)}}}{\sqrt{\widehat{\sigma^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$? En déduire un test de niveau α pour $\mathcal{H}_0$.

2 - d) Charger le jeu de données "Ratweight.csv". En extraire un vecteur X correspondant aux poids à 14 semaines des rats nourris au Mon863, et un vecteur Y correspondant aux poids à 14 semaines des rats du groupe témoin. Implémenter le test décrit en question 2 - c) pour tester l'égalité des moyennes des échantillons X et Y. Conclure.

3 - Test non paramétrique

L'hypothèse $X \sim \mathcal{N}(\mu, \sigma)$ est justifiée dans la limite des grands échantillons par le théorème central limite. Cependant, dans le cadre de petits échantillons, cette hypothèse peut être problématique. On recourt donc à des tests non paramétriques, parmi lesquels se classe le test de Wilcoxon-Mann-Whitney. La logique de ce test est la suivante : dans un premier temps, on ordonne l'échantillon X. Si les espérances des deux populations sont différentes, la plupart des observations d'une population aura un faible rang, tandis que celle de l'autre aura un rang important. On considère comme statistique la somme des rangs des observations de l'échantillon $X^{(1)}$, notée R_1.

3 - a) À l'aide de méthodes de Monte-Carlo, estimer les quantiles à $0,025$ et à $0,0975$ de la statistique R_1 sous $\mathcal{H}_0$. Que peut-on en conclure pour nos données ?

3 - b) À l'aide de méthodes de Monte-Carlo, comparer les puissances du test du χ^2 et du test de Wilcoxon-Mann-Whitney dans le cas où les observations sont distribuées suivant une loi normale. Commenter.

Tests multiples

Dans le cadre d'études sur la leucémie, des chercheurs observent le niveau d'expression de différents gènes pour des patients atteints de leucémie de deux types. Ils tentent ensuite de déterminer les gènes liés au développement de ces maladies. Ils observent le jeu de données "leucemie.RData", c'est à dire une matrice X de taille 3051×38. Chaque ligne de cette matrice correspond au niveau d'expression d'un des $m = 3051$ gènes contrôlés, tandis que chaque colonne correspond à un patient. Les 27 premières colonnes correspondent aux patients atteints de leucémie de type "ALL", tandis que les 11 colonnes restantes correspondent aux patients atteints de leucémie de type "AML". Pour chaque gène i, on teste l'hypothèse $\mathcal{H}_0^i$: "l'espérance du niveau d'expression du gène i est la même dans les deux populations" contre l'hypothèse $\mathcal{H}_1^i$: "l'espérance du niveau d'expression du gène i est différentes selon les populations".

4 - Une approche naïve des tests multiples

4 - a) Pour un test d'hypothèse simple, on considère une statistique S et on note $T(s) = \mathbb{P}_{\mathcal{H}_0}(S \geq s)$. On rappelle que la p-valeur est définie comme $p = T(S)$. Montrer que sous $\mathcal{H}_0$, si la statistique T admet une densité, $p \sim \mathcal{U}([0,1])$. En conclure que le test consistant à rejeter $\mathcal{H}_0$ dès que $p < \alpha$ est de niveau α.

4 - b) Importer les données, et les séparer en deux matrices correspondant aux patients "ALL" et "AML". À l'aide de la fonction "t.test", calculer les p-valeurs p_i associées à ces hypothèses". Ordonner ces p-valeurs et les afficher. Commenter.

4 - c) Une approche naïve pour détecter les gènes liés au développement d'un cancer particulier consiste à calculer les p-valeurs liés aux différents niveaux d'expression des gènes, et à rejeter l'hypothèse "le niveau d'expression du gène n'a pas la même distribution dans les deux populations" dès que la p-valeur associée est inférieure à 0.05. En supposant que les niveaux d'expression des gènes sont indépendants, quel est le nombre moyen de faux positifs induit par cette procédure lorsque que pour chaque gène i, $\mathcal{H}_0^i$ est vraie ? Conclure.

5 - Contrôle de la "Family-Wise Error Rate"

La "Family-Wise Error Rate" (abrégée FWER) est définie comme la probabilité, lorsque pour chaque gène i $\mathcal{H}_0^i$ est vraie, de rejeter l'hypothèse $\mathcal{H}_0^i$ pour au moins un gène i. Pour

$\beta \in [0,1]$ et pour chaque gène i, on rejette l'hypothèse $\mathcal{H}_0^i$ si $p_i \leq \beta$, où p_i est la p-valeur associée à l'hypothèse $\mathcal{H}_0^i$.

5 - a) On suppose que les niveaux d'expression des différents gènes sont indépendants. Montrer que

$$FWER \leq 1 - (1-\beta)^m.$$

En déduire un choix de β tel que la FWER soit plus petite que 0.05. Appliquer cette procédure au jeu de données "leucemie.RData".

5 - b) On ne suppose plus l'indépendance des niveaux d'expression des différents gènes. Montrer que

$$FWER \leq m\beta.$$

En déduire un choix de β tel que la FWER soit plus petite que 0.05. Appliquer cette procédure au jeu de données "leucemie.RData".

6 - Contrôle du taux de faux positifs

Les méthodes proposées ci-dessus pour contrôler la FWER ont pour défaut de considérablement réduire la puissance des tests mis en place. Pour palier ce défaut, on peut préférer contrôler le taux de faux positifs $TFP = \mathbb{E}\left[\frac{FP}{VP+FP}\right]$, où FP et VP dénotent respectivement le nombre de faux positifs et de vrais positifs. Pour minimiser le nombre de faux positifs tout en maximisant le nombre de vrais positifs, on étudie une procédure rejetant $\mathcal{H}_0^i$ dès que la p-valeur associée p_i est suffisamment petite.

Procédure de Benjamini-Hochberg : Rejeter l'hypothèse $\mathcal{H}_0^i$ pour les gènes i tels que $p_i \leq \frac{\alpha\widehat{k}}{m}$, où $\widehat{k} \triangleq \max\{k : p_k \leq \frac{\alpha k}{m}\}$.

On peut démontrer la majoration suivante pour le TFP pour la procédure de Benjamini-Hochberg

$$TFP = \mathbb{E}\left[\frac{\text{card}\{i \in I_0 : p_i \leq \alpha\widehat{k}/m\}}{\widehat{k}} 1_{\widehat{k}\geq 1}\right] \leq \alpha.$$

Implémenter la procédure de Benjamini-Hochberg pour le jeu de données pour un taux de faux positifs de 0.05. Commenter.

Classification d'images de chats et de chiens.

Gautier Appert
gautier.appert.chess@gmail.com

Soit $\mathcal{D}_n = \{x_1, \dots, x_n\}$ une base de données d'images de chat et de chien où chaque image est représentée par un vecteur de pixels $x_i \in \mathbb{R}^p$, $i \in \{1, \dots, n\}$. Notons que le nombre de pixels p est potentiellement très largement supérieur à n. On note $Y_i \in \{0, 1\}$ la variable aléatoire correspondant au label chat ou chien associé à l'image $x_i \in \mathbb{R}^p$. En pratique la base données $\mathcal{D}_n$ stock les images en vecteur lignes

$$\mathcal{D}_n = \left[\; x_1 \; \middle| \; x_2 \; \middle| \; \dots \; \middle| \; x_n \; \right]^\top.$$

L'objet de ce tutoriel est de prédire le label chat ou chien à l'aide d'une analyse discriminante quadratique (QDA) pour une nouvelle image en dehors de la base d'apprentissage $x \notin \mathcal{D}_n$. L'implémentation doit être faite sous le langage `R`. Envoyer un mail au chargé de TD afin de pouvoir récupérer les données sur la dropbox.

1. Découverte de la base de données et réduction de la dimension

Deux bases de données intitulées X_{train}.RData et X_{test}.RData sont à disposition sur la dropbox. Ces bases contiennent des images de chiens et de chats stockées en vecteur ligne. En particulier on a $X_{\text{train}} \in \mathbb{R}^{315 \times 40000}$ et $X_{\text{test}} \in \mathbb{R}^{48 \times 40000}$. Deux autres bases de données Y_{train}.RData et Y_{test}.RData contiennent les labels associés au images X_{train} et X_{test}.

Question (1). Importer la base de données X_{train}.RData, X_{test}.RData, y_{train}.RData et y_{test}.RData à l'aide de la fonction `load`. Afficher deux à trois images des deux bases à l'aide de la fonction `image(...,col = grey(seq(0, 1, length = 256)))` en transformant préalablement les images en matrice de dimension 200×200 (fonction `matrix`). Enregistrer les images en format `pdf` ou `png` et les mettre dans votre rapport. A quoi correspond le label $y = 1$?

Question (2). On souhaite réduire la dimension des données $p = 40000$. Pour cela nous allons procéder à une analyse en composante principales (ACP) des images.

(a). Concatener X_{train} et X_{test} en utilisant la fonction `rbind` et centrer les vecteurs colonnes avec la fonction `scale`. On notera $X \in \mathbb{R}^{363 \times 40000}$ la matrice résultante. Construire une ACP en utilisant une décomposition en valeur singulière (SVD) de la matrice X à l'aide la fonction `svd`. On ne retiendra que les 15 premières composantes principales. On rappelle que la décomposition en valeur singulière permet de factoriser la matrice X de la manière suivante

$$X = UDV^\top,$$

où V est la matrice des vecteurs propres. Ainsi la matrice des composantes principales est donnée par $C = XV$.

(b). Quelle est la part de variance expliquée en ne retenant que 15 composantes principales ?

Désormais nous travaillerons sur les composantes principales $C \in \mathbb{R}^{363\times 15}$ au lieu des données d'origine X. Découper la base C en $C_{\text{train}} \in \mathbb{R}^{315\times 15}$ et $C_{\text{test}} \in \mathbb{R}^{48\times 15}$.

2. Analyse discriminante quadratique

On fait l'hypothèse du modèle suivant

- $Y_i \sim \mathcal{B}\big(\pi\big)$.
- $\mathbb{P}_{c_i|Y=1} = \mathcal{N}\big(\mu_1, \Sigma_1\big)$ et $\mathbb{P}_{c_i|Y=0} = \mathcal{N}\big(\mu_0, \Sigma_0\big)$ où c_i est le i-ième vecteur ligne de la matrice C.

Le paramètre inconnu est $\theta = (\pi, \mu_0, \mu_1, \Sigma_1, \Sigma_0)$ où $\pi \in]0,1[$, $(\mu_0, \mu_1) \in \mathbb{R}^{15} \times \mathbb{R}^{15}$ et $(\Sigma_0, \Sigma_1) \in \mathbb{R}^{15\times 15} \times \mathbb{R}^{15\times 15}$ sont des matrices définies positives. On definit $\mathbb{P}_\theta = \mathbb{P}_{c,Y}$ et on dispose d'un échantillon $(c_1, y_1), \ldots, (c_n, y_n) \overset{\text{i.i.d}}{\sim} \mathbb{P}_\theta$.

Question (3). Ecrire le modèle statistique associé aux observations $(c_1, y_1), \ldots, (c_n, y_n)$.

Question (4). On pose $N_1 = \sum_{i=1}^n y_i$ et $N_2 = n - N_1$. En utilisant le fait que $f_{c,Y}(c, y) = f_{c|Y=y}(c)\, f_Y(y)$, montrer que la log vraisemblance $\ell\big((c_1, y_1), \ldots, (c_n, y_n); \theta\big)$ s'écrit

$$\ell\big((c_1, y_1), \ldots, (c_n, y_n); \theta\big) = N_1 \log(\pi) + N_2 \log(1-\pi)$$
$$-\frac{N_1}{2}\log(\det(\Sigma_1)) - \frac{1}{2}\sum_{i:y_i=1}(c_i-\mu_1)^\top \Sigma_1^{-1}(c_i-\mu_1) - \frac{N_2}{2}\log(\det(\Sigma_0)) - \frac{1}{2}\sum_{i:y_i=0}(c_i-\mu_0)^\top \Sigma_0^{-1}(c_i-\mu_0).$$

Question (5). En utilisant les formules $\nabla_\Sigma \log(\det(\Sigma)) = \Sigma^{-1}$ et $\nabla_\Sigma\big(a^\top \Sigma^{-1} b\big) = -\Sigma^{-1} a b^\top \Sigma^{-1}$, écrire l'équation du premier ordre pour le maximum de vraisemblance et montrer que l'on obtient les estimateurs

$$\hat{\pi} = \frac{N_1}{n} \quad \hat{\mu}_1 = \frac{1}{N_1}\sum_{i:y_i=1} c_i \quad \hat{\mu}_0 = \frac{1}{N_0}\sum_{i:y_i=0} c_i$$

$$\widehat{\Sigma}_1 = \frac{1}{N_1}\sum_{i:y_i=1}(c_i-\hat{\mu}_1)(c_i-\hat{\mu}_1)^\top \quad \widehat{\Sigma}_0 = \frac{1}{N_0}\sum_{i:y_i=0}(c_i-\hat{\mu}_0)(c_i-\hat{\mu}_0)^\top.$$

Question (6). Montrer que la sous Hessienne $\nabla^2_{\pi,\mu_1,\mu_0}\ell(\theta)$ est bien définie négative. On ne regardera pas les conditions du second ordre avec Σ_1 et Σ_0.

Question (7). Montrer que $\hat{\pi}$ est sans biais et montrer que $\hat{\mu}_1$ et $\hat{\mu}_0$ sont sans biais (conditionner par rapport à l'échantillon $\{y_1, \ldots, y_n\}$ via la loi des espérances itérées.)

Question (8). Montrer que les estimateurs issus de la méthode des moments coïncident avec les estimateurs du maximum de vraisemblance. (on pourra utiliser la définition de l'espérance conditionelle sachant un évenement $\mathbb{E}\big[C|Y=y\big] = \frac{\mathbb{E}[C\mathbb{1}(Y=y)]}{\mathbb{E}[\mathbb{1}(Y=y)]}$).

Question (9). Coder une fonction sous `R` intitulée `computeML(C,Y)` prenant en argument une matrice C et un vecteur Y, et qui renvoie sous forme de liste les estimateurs du maximum de vraisemblance $\hat{\pi}, \hat{\mu}_1, \hat{\mu}_0, \widehat{\Sigma}_1, \widehat{\Sigma}_0$. Lancer la fonction `computeML` sur `Ctrain,Ytrain`. Comparer les estimateurs obtenus avec la fonction `qda(Ctrain,Ytrain)` du package `MASS`. (La fonction `qda` ne fournit pas les estimateurs concernant les matrices de variances covariances mais fournit le log du determinant).

3. Prédiction des labels sur la base test

On souhaite dans cette partie prédire les labels correspondant au données C_{test} à l'aide de l'analyse discriminante quadratique dont les paramètres ont été estimés sur l'échantillon d'apprentissage (`Ctrain, Ytrain`). En effet, l'analyse discriminante quadratique permet de modéliser les probabilités $\mathbb{P}\big(Y=1|c\big)$ et $\mathbb{P}\big(Y=0|c\big)$. C'est pourquoi nous prendrons la règle de prédiction suivante $\hat{y} = \arg\max_{y\in\{0,1\}} \mathbb{P}\big(Y=y|c\big)$.

Question (10). A l'aide de la formule de Bayes, montrer que

$$\mathbb{P}\big(Y=1|c\big) = \frac{\pi\,\varphi(c;\mu_1,\Sigma_1)}{\pi\,\varphi(c;\mu_1,\Sigma_1) + (1-\pi)\,\varphi(c;\mu_0,\Sigma_0)}$$

où $\varphi(c;\mu,\Sigma)$ représente la densité de la Gaussienne multivariée $\mathcal{N}\big(\mu,\Sigma\big)$. Calculer de la même manière $\mathbb{P}\big(Y=0|c\big)$.

Question (11). En déduire que

$$\log\left(\frac{\mathbb{P}\big(Y=1|c\big)}{\mathbb{P}\big(Y=0|c\big)}\right) = -\frac{1}{2}\log(\det(\Sigma_1)) - \frac{1}{2}(c-\mu_1)^\top\Sigma_1^{-1}(c-\mu_1) + \log(\pi) \\ + \frac{1}{2}\log(\det(\Sigma_0)) + \frac{1}{2}(c-\mu_0)^\top\Sigma_0^{-1}(c-\mu_0) - \log(1-\pi).$$

Question (12). Montrer que

$$\mathbb{1}\left(\log\left(\frac{\mathbb{P}\big(Y=1|c\big)}{\mathbb{P}\big(Y=0|c\big)}\right) > 0\right) = \underset{y\in\{0,1\}}{\arg\max}\ \mathbb{P}\big(Y=y|c\big).$$

Ainsi, on utilisera la règle de prédiction suivante (méthode Plug-in): pour toutes lignes $c \in C_{\text{test}}$

$$\hat{y} = \mathbb{1}\bigg(-\frac{1}{2}\log(\det(\widehat{\Sigma}_1)) - \frac{1}{2}(c-\hat{\mu}_1)^\top\widehat{\Sigma}_1^{-1}(c-\hat{\mu}_1) + \log(\hat{\pi}) \\ + \frac{1}{2}\log(\det(\widehat{\Sigma}_0)) + \frac{1}{2}(c-\hat{\mu}_0)^\top\widehat{\Sigma}_0^{-1}(c-\hat{\mu}_0) - \log(1-\hat{\pi}) > 0\bigg).$$

Question (13). En utilisant le fait que $(y,\hat{y}) \in \{0,1\}^2$, montrer la double égalité

$$\mathbb{E}\big[(y-\hat{y})^2\big] = \mathbb{E}\big[|y-\hat{y}|\big] = \mathbb{P}\big(\hat{y}\neq y\big).$$

Empiriquement on prendra

$$\widehat{\mathbb{E}}\big[|y-\hat{y}|\big] = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|,$$

ce qui correspond à l'erreur de classification.

Question (14). Ecrire une fonction `R` intitulée `computeLogRatio(c,pi,mu1,mu0,Sigma1,Sigma0)` prenant en argument un vecteur $c \in C_{\text{test}}$ et le paramètre θ, et qui renvoie la quantité: $\log\left(\frac{\mathbb{P}\left(Y=1|c\right)}{\mathbb{P}\left(Y=0|c\right)}\right)$. Puis coder une fonction `computePred(C,pi,mu1,mu0,Sigma1,Sigma0)` prenant en argument une matrice C et le paramètre θ et qui renvoie la prédiction des labels pour chaque ligne de la matrice C.

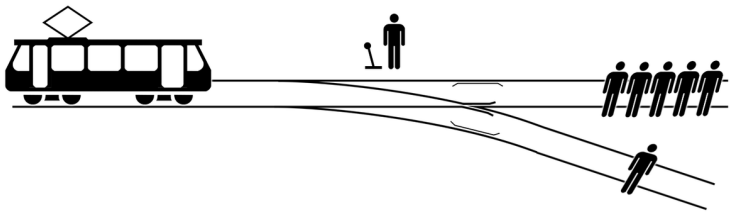
QUESTION (15). Prédire les labels de la base de données test C_{test} avec `computePred` et donner l'erreur de classification à l'aide de Y_{test}. Comparer avec la prédiction en utilisant l'estimation du modèle faite avec la fonction `qda` de `R`. La prédiction est-elle meilleur que le prédicteur aléatoire ?

RÉFÉRENCE

- ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction***, *Trevor Hastie, Robert Tibshirani, Jerome Friedman.*

Nombre de tramways et nombre d'estimateurs : EMV, Bayes, moments et découverte des GMM[1]

Lucas Girard - lucas.girard@ensae.fr



Histoire Vous êtes en voyage dans une ville étrangère pour rédiger un guide touristique. Une information cruciale est évidemment le nombre de lignes de tramway que comporte la ville en question. Malheureusement, la langue de ce pays vous est inconnue et vous ne savez lire que les nombres. Au hasard de vos visites, vous observez ainsi les numéros de quelques lignes : la ligne 1 à votre arrivée à la gare, la ligne 7 en allant au jardin botanique, la ligne 16 en vous rendant au musée des beaux-arts, etc. Vous savez par ailleurs que les lignes ont été numérotées successivement, c'est-à-dire qu'il n'y a pas de trous dans la numérotation[2]. Le but de ce projet est d'étudier différents estimateurs du nombre de lignes de tramway dans la ville.

Objectifs pédagogiques D'un point de vue pédagogique, ce projet a deux objectifs.

Le principal est d'illustrer, sur un exemple simple, la multiplicité des estimateurs envisageables. Suite à plusieurs exercices de modèles réguliers classiques où l'estimateur du maximum de vraisemblance s'avère être également l'estimateur de la méthode des moments, on pourrait oublier qu'il y a de nombreux autres estimateurs disponibles. Au cours de ce projet, vous étudierez l'estimateur du maximum de vraisemblance (EMV), l'estimateur de Bayes pour la perte quadratique et la perte absolue, l'estimateur de la méthode des moments et vous définirez même un nombre dénombrable d'estimateurs différents en utilisant différents moments de la loi générant les données. En codant ces différents estimateurs, vous verrez sur simulations la multiplicité des estimations obtenues pour un même problème. On se demandera également comment comparer différents estimateurs.

Un objectif secondaire du projet est de vous faire découvrir la méthode des moments généralisés (ou GMM). Comme l'indique son nom, celle-ci généralise la méthode des moments. Elle fournit une nouvelle classe d'estimateurs, augmentant encore le nombre d'estimateurs disponibles. Il s'agira ici uniquement de se renseigner sur les GMM et de coder un estimateur GMM.

La notation sera attentive aux justifications des réponses et à la qualité de la rédaction.

[1] GMM: Generalized Method of Moments (méthode des moments généralisés).

[2] En observant un tramway de la ligne 3 par exemple, vous savez qu'il y a au minimum trois lignes : 1, 2, 3.

1 Formalisation du modèle statistique

Hypothèse On suppose observer les réalisations de variables aléatoires $X_1, \ldots, X_n$, avec $n \in \mathbb{N}^*$, indépendantes et identiquement distribuées (i.i.d.) selon une loi uniforme $\mathcal{U}([1, \theta])$, avec $\theta > 1$. Pour simplifier l'aspect technique, on suppose donc une loi uniforme continue, bien que le nombre véritable de lignes soit discret. θ s'interprète ainsi comme le nombre de lignes de tramway dans la ville.

(1) Écrivez le modèle statistique correspondant à ce problème. Ce modèle est-il identifiable ? Vérifie-t-il les conditions de régularité vues en cours (sections 2.3 et 4.3 des slides) ?

(2) En réalité, vous apprenez au cours de votre séjour que les lignes de tramway se décomposent en des lignes fonctionnant de jour (6h-22h) et des lignes fonctionnant de nuit (22h-6h).
(a) Écrivez un modèle statistique tenant compte de cette information (on demande une reparamétrisation du modèle de la Question 1).
(b) Sur votre carnet d'enquête, vous avez uniquement noté les lignes observées, $x_1 = 1$ en partant de la gare, $x_2 = 7$ en allant au jardin botanique, etc. Avec ces informations, pouvez-vous apprendre quelque chose sur le nombre de lignes de tramway diurnes et le nombre de lignes de tramway nocturnes ? Vous argumenterez votre réponse au moyen d'une notion statistique vue en cours.
(c) Concrètement, quelle autre information auriez-vous dû noter dans votre carnet afin de distinguer le nombre de lignes le jour et le nombre de lignes la nuit ? Supposons que vous avez cette information, détaillez le modèle statistique associé.

On oublie par la suite cette distinction entre lignes de jour et lignes de nuit et on se concentre désormais sur le modèle statistique de la Question 1.

2 Multiplicité des estimateurs

2.1 Estimateur du Maximum de Vraisemblance et simulations

(4) Déterminez $\hat{\theta}_n^{\mathrm{MV}}$, l'estimateur du maximum de vraisemblance de θ.

(5) On s'intéresse à différentes propriétés asymptotiques de $\hat{\theta}_n^{\mathrm{MV}}$. Montrez que $\hat{\theta}_n^{\mathrm{MV}}$ est un estimateur convergent. Déterminez la distribution asymptotique de $n(\theta - \hat{\theta}_n^{\mathrm{MV}})$.

(6) Montrez que pour tous $n \in \mathbb{N}^*$, $\theta > 1$, $\mathbb{E}_\theta[\hat{\theta}_n^{\mathrm{MV}}] = \frac{n}{n+1}\theta$. $\hat{\theta}_n^{\mathrm{MV}}$ est-il un estimateur sans biais ? Le cas échéant, proposez un estimateur sans biais $\tilde{\theta}_n^{\mathrm{MV}}$ à partir de $\hat{\theta}_n^{\mathrm{MV}}$. $\tilde{\theta}_n^{\mathrm{MV}}$ est-il convergent ? $\tilde{\theta}_n^{\mathrm{MV}}$ est-il asymptotiquement normal[3] ?

[3] On dit qu'un estimateur $\hat{\theta}_n$ de θ est asymptotiquement normal lorsqu'il existe une suite de réels positifs (r_n), tendant vers $+\infty$, et une distribution Z non dégénérée - i.e. $Z \neq \delta_0$, $Z \neq \delta_{+\infty}$, avec δ_a la notation d'une masse de Dirac en a - telles que : $\forall \theta, r_n(\hat{\theta} - \theta) \xrightarrow{d} Z$. La suite (r_n) est alors appelée vitesse de convergence de l'estimateur. *N.B.* malgré cette définition générale, on a souvent en tête $r_n = \sqrt{n}$ et $Z = \mathcal{N}(0, V)$ lorsqu'on parle de normalité asymptotique, dans ce cas V est appelée variance asymptotique.

(7) Cette question vise à vérifier, sur des simulations, le comportement asymptotique de $\hat{\theta}_n^{\mathrm{MV}}$. Choisissez un $\theta_0 > 1$ pour simuler des données $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}([1, \theta_0])$ et illustrer graphiquement les deux propriétés de $\hat{\theta}_n^{\mathrm{MV}}$ déterminées à la question 5. Vous êtes volontairement laissés assez libre quant à la façon de procéder.

2.2 Estimateur de Bayes

On suit désormais une approche bayésienne, ou bayésienne généralisée puisqu'on va considérer des lois a priori impropres. En effet, vous n'avez pas d'information a priori fiable sur le nombre de lignes de tramway et vous cherchez une loi a priori invariante par reparamétrisation.

(8) L'a priori de Jeffreys est-il utilisable ici ? Le cas échéant, déterminez la densité de l'a priori de Jeffreys.

On prend désormais pour loi a priori sur θ la loi définie par la densité suivante[4] (définie à une constante près) :

$$\pi(\theta) \propto \frac{1}{\theta}\,.$$

(9) Cette loi a priori est-elle une loi impropre ? Ce choix vous semble-t-il judicieux ?
Indice : on pourra se demander si cette loi a priori est invariante par changement d'échelle.

(10) L'estimateur de Bayes dépend de la fonction de perte considérée.
(a) Perte quadratique (ℓ^2) : $\mathcal{L}(\theta, d) = (\theta - d)^2$. Déterminez l'estimateur de Bayes associé à la perte quadratique, $\hat{\theta}_n^{\mathrm{B2}}$ (B2 pour Bayes et perte ℓ^2). Cet estimateur est-il défini pour tout $n \in \mathbb{N}^*$? L'estimateur $\hat{\theta}_n^{\mathrm{B2}}$ est-il sans biais ? L'estimateur $\hat{\theta}_n^{\mathrm{B2}}$ est-il convergent ?
(b) Perte en valeur absolue (ℓ^1) : $\mathcal{L}(\theta, d) = |\theta - d|$. Déterminez l'estimateur de Bayes associé à la perte ℓ^1, $\hat{\theta}_n^{\mathrm{B1}}$ (B1 pour Bayes et perte ℓ^1). Cet estimateur est-il défini pour tout $n \in \mathbb{N}^*$? L'estimateur $\hat{\theta}_n^{\mathrm{B1}}$ est-il sans biais ? L'estimateur $\hat{\theta}_n^{\mathrm{B1}}$ est-il convergent ?

2.3 Méthode des moments

(11) Déterminez $\hat{\theta}_n^{\mathrm{MM}}$, l'estimateur de la méthode des moments (EMM) de θ. Justifiez que $\hat{\theta}_n^{\mathrm{MM}}$ est un estimateur sans biais et convergent. $\hat{\theta}_n^{\mathrm{MM}}$ est-il un estimateur asymptotiquement normal ? Si oui, à quelle vitesse de convergence ?

2.4 Et les autres moments ?

La méthode des moments utilise le fait que, pour tout $\theta > 1$, $\mathbb{E}_\theta[X_1] = \theta/2$. Or, on connaît également le moment d'ordre 2 de X_1 : $\forall \theta > 1, \mathbb{E}_\theta[X_1^2] = \theta^2/3$.

[4] On utilise ici les notations du cours. Avec ces notations, on rappelle que dans l'écriture $\pi(\theta)$, θ a un double rôle : d'une part, il s'agit d'une variable muette en tant qu'argument de la densité et permettant de définir cette densité, d'autre part il s'agit d'une variable signifiante qui indique que la densité de probabilité considérée est la densité a priori sur θ. Une notation moins compacte pourrait être : notons π_θ la densité a priori de θ définie par, $\forall z \in \Theta, \pi_\theta(z) \propto 1/z$.

(12) *(a)* Vérifiez l'égalité précédente et en déduire un estimateur de θ reposant sur le moment d'ordre 2, qu'on notera $\hat{\theta}_n^{\mathrm{M2}}$. Cet estimateur est-il convergent ? A l'aide du théorème central limite et de la Delta-méthode, montrez que cet estimateur est asymptotiquement normal. En termes de variance asymptotique, $\hat{\theta}_n^{\mathrm{M2}}$ est-il préférable à $\hat{\theta}_n^{\mathrm{MM}}$?
(b) X_1 est une variable bornée et admet donc des moments de tout ordre. En vous inspirant de la question précédente, pour un entier quelconque $k \geq 2$, construisez un estimateur $\hat{\theta}_n^{\mathrm{M}k}$ reposant sur le moment d'ordre k de X_1. Montrez que cet estimateur est convergent et asymptotiquement normal avec une vitesse $\sqrt{n}$. $\hat{\theta}_n^{\mathrm{M}k}$ est-il sans biais ? Discutez du comportement de cet estimateur $\hat{\theta}_n^{\mathrm{M}k}$ lorsque k tend vers $+\infty$.
(c) Codez une fonction calculant l'estimateur $\hat{\theta}_n^{\mathrm{M}k}$ pour un entier $k \geq 2$ quelconque et un échantillon quelconque. Illustrez les propriétés de convergence et de normalité asymptotique de $\hat{\theta}_n^{\mathrm{M}k}$. Au moyen de simulations, déterminez le sens de variation du biais de $\hat{\theta}_n^{\mathrm{M}k}$ en fonction de k. Le biais évolue-t-il dans le même sens que la variance asymptotique ? Selon vos réponses aux deux questions précédentes, discutez d'un choix optimal de k en termes de Mean Square Error.

(13) Chaque estimateur $\hat{\theta}_n^{\mathrm{M}k}$ utilise un unique moment, le moment d'ordre k. Une extension naturelle serait de chercher à combiner ces différents moments afin d'estimer θ. Cette idée débouche sur les estimateurs GMM. On attend dans cette question uniquement un travail de code et de simulation afin de vous initier aux GMM. Au préalable, renseignez-vous sur la méthode des GMM - vous pouvez me contacter pour des références ou explications. Puis, codez une fonction implémentant un estimateur GMM de θ. Dans un premier temps, vous pouvez fixer le nombre de moments utilisés, disons les trois premiers moments. Puis, dans un second temps, définissez une fonction dont un des arguments est le nombre de moments utilisés.

3 Comparaison des estimateurs

(14) - cas particulier $n = 1$ Il s'avère que vous êtes seulement en transit dans la ville en question et n'apercevez ainsi qu'un seul numéro de tramway x_1 entre la gare routière et la gare ferroviaire. Autrement dit, $n = 1$. Comparez les estimateurs du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$, l'EMV débiaisé $\tilde{\theta}_n^{\mathrm{MV}}$, de la méthode des moments $\hat{\theta}_n^{\mathrm{MM}}$ et de Bayes (espérance a posteriori $\hat{\theta}_n^{\mathrm{B2}}$ et médiane a posteriori $\hat{\theta}_n^{\mathrm{B1}}$). Certains de ces estimateurs sont-ils identiques dans ce cas particulier ? Est-ce que certains de ces estimateurs vous semblent plus intuitifs que d'autres ? En particulier, choisiriez-vous l'EMV ?

(15) - cas général On cherche maintenant à comparer les estimateurs obtenus à la Section 2.
(a) En vous fondant sur la vitesse de convergence, répartissez ces estimateurs en deux grandes familles.
(b) Un critère possible pour arbitrer entre différents estimateurs est la comparaison de leur Mean Square Error (MSE) ou risque quadratique. Calculez la MSE de $\hat{\theta}_n^{\mathrm{MV}}$, $\tilde{\theta}_n^{\mathrm{MV}}$, $\hat{\theta}_n^{\mathrm{MM}}$ et $\hat{\theta}_n^{\mathrm{B2}}$. Quel estimateur préférez-vous en termes de risque quadratique ?

(16) En utilisant les différents estimateurs obtenus à la Section 2, illustrez la diversité des estimations obtenues, pour un même échantillon simulé. Comment cette diversité évolue-t-elle en fonction de la taille de l'échantillon n ?

Etude de la régression en grande dimension via l'estimateur LASSO

Badr-Eddine Cherief-abdellatif

Introduction

Nous étudions dans ce tutoriel le LASSO (Least Absolute Shrinkage and Selection Operator), une technique de régularisation et de sélection de variables particulièrement adaptée à la régression quand le nombre de variables p n'est pas négligeable devant le nombre d'observations n. D'un point de vue statistique, nous espérons trouver parmi les p variables qui expliquent une quantité Y les plus importantes d'entre elles. Cela est fondamental dans le domaine médical par exemple, pour prédire le volume d'une tumeur cancéreuse en fonction d'un grand nombre de mesures chez un individu.

Nous disposons d'un échantillon de données $(\mathbf{x}_i, Y_i)_{i=1,...,n}$, chaque couple étant à valeurs dans $\mathbb{R}^p \times \mathbb{R}$. Nous considérons un modèle de régression sur design fixe avec erreurs gaussiennes, ce qui s'écrit sous forme matricielle :

$$Y = f(\mathbf{X}) + \epsilon \tag{1}$$

avec $\epsilon = (\epsilon_1, ..., \epsilon_n)^T \in \mathbb{R}^n$, $Y = (Y_1, ..., Y_n)^T \in \mathbb{R}^n$ et $\mathbf{X} = (\mathbf{x}_1^T, ..., \mathbf{x}_p^T)^T \in \mathbb{R}^{n \times p}$ où les $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ sont indépendants et identiquement distribués (i.i.d.), les $\mathbf{x}_i = (x_{i1}, ..., x_{ip}) \in \mathbb{R}^p$ sont déterministes, et la variance du bruit $\sigma^2 > 0$ est connue. Notre but est d'estimer la vraie fonction f sous une hypothèse linéaire. Autrement dit, nous supposons qu'en réalité, $Y = \mathbf{X}\beta^0 + \epsilon$ où $\beta^0 \in \mathbf{R}^p$ est inconnu. Le but est donc d'estimer $\mathbf{X}\beta^0$ et non pas β^0. Toutefois, nous estimerons directement β^0 mais nous intéresserons à l'erreur quadratique moyenne d'un estimateur $\hat{\beta}$ comme mesure de sa performance :

$$R(\hat{\beta}) = \frac{\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^0\|_2^2}{n}.$$

1 Préliminaires

Dans un premier temps, nous étudions l'estimateur des moindres carrés ordinaires.

Question 1 : Rappeler la formule définissant l'estimateur des moindres carrés, et donner son expression explicite sous forme matricielle dans le cas où $p < n$. Détailler les calculs. Expliquer en quoi le cas $p > n$ est problématique.

Question 2 : On reproche généralement à l'estimateur des moindres carrés de poser des problèmes de prédiction ainsi que d'interprétation lorsque le nombre de variables est élevé. Expliquer pourquoi.

Nous faisons dorénavant (et ce jusqu'à la question 6) l'hypothèse que $p > n$ et que β^0 est s-sparse, c'est-à-dire qu'au plus s composantes de β^0 sont non nulles ($s < p$).

2 Propriétés du LASSO

Nous allons désormais nous attarder sur l'estimateur LASSO. Celui-ci est défini par l'équation :

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|y - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{2}$$

où $\|.\|_1$ et $\|.\|_2$ sont des normes de $\mathbb{R}^p$, et $\lambda > 0$. Notons que ce problème d'optimisation a potentiellement une infinité de solutions, mais que toutes donnent la même valeur de $\mathbf{X}\beta$ et $\|\beta\|_1$.

Question 3 : L'estimateur LASSO est une variante de l'estimateur aux moindres carrés auquel on a ajouté une pénalité $\|.\|_1$. Expliquer brièvement le principe de la pénalisation en statistique.

Question 4 : En réalité, la vraie définition de l'estimateur LASSO est la suivante :

$$\hat{\beta}_\lambda = \underset{\beta\in\mathbb{R}^p, \|\beta\|_1 \leq t}{\arg\min} \|y - \mathbf{X}\beta\|_2^2 \tag{3}$$

où $t > 0$. Expliquer en quoi (3) est équivalent à (2) pour une certaine valeur de λ. Justifier.

Question 5 : Étant donné que le vecteur β^0 est s-sparse, il aurait pu sembler plus judicieux de définir l'estimateur LASSO par :

$$\underset{\beta\in\mathbb{R}^p, \|\beta\|_0 \leq s}{\arg\min} \|y - \mathbf{X}\beta\|_2^2$$

où $\|.\|_0$ désigne le nombre de composantes non nulles. Expliquer pourquoi ce choix n'est pas pertinent en pratique, et pourquoi la norme $\|.\|_0$ est remplacée par la norme $\|.\|_1$.

Question 6 : En pratique, l'estimateur LASSO a bien souvent un grand nombre de composantes nulles bien que ce ne soit pas la pénalisation $\|.\|_0$ qui soit utilisée. En s'aidant de la formulation (3), donner une explication géométrique simple de ce phénomène en dimension $p = 2$.

Question 7 : Dans cette question, nous relâchons l'hypothèse $p > n$. Sous l'hypothèse d'orthogonalité : $\mathbf{X}^T\mathbf{X}/n$ *est la matrice identité de* $\mathbb{R}^{p\times p}$, le LASSO a une solution explicite. La donner en fonction de l'estimateur des moindres carrés, en partant de la formulation (2). Indiquer pourquoi on parle de *seuillage doux*, et en quoi cela explique en partie la sparsité de la solution.

Question 8 : Montrer que l'estimateur LASSO peut s'écrire comme le maximum a posteriori d'une certaine distribution a posteriori. *Indication : on pourra considérer la loi a priori de Laplace ayant pour densité par rapport à la mesure de Lebesgue de dimension d* :

$$f(\beta) = \exp\left(-\frac{\lambda}{2\sigma^2}\|\beta\|_1\right).$$

Convergence du LASSO

L'espérance de l'erreur quadratique de l'estimateur des moindres carrés converge vers 0 *avec une vitesse* p/n *sous des hypothèses classiques lorsque* n *et* p *tendent vers l'infini avec* $p/n \to 0$. *Cette vitesse n'est par conséquent pas satisfaisante pour des valeurs de* $p > n$. *L'un des avantages de l'estimateur LASSO est de permettre d'obtenir la convergence de l'erreur quadratique même quand* p *n'est pas négligeable devant* n.

En réalité, sous l'hypothèse que la norme $\|.\|_2$ *de chaque colonne de* $\mathbf{X}$ *est égale à* $\sqrt{n}$ *(plus faible que l'hypothèse d'orthogonalité), il est possible de montrer qu'avec une probabilité supérieure ou égale à* $1 - \delta$:

$$R(\hat{\beta}_\lambda) \leq 4\|\beta^0\|_1\sqrt{\frac{2\sigma^2\log(\frac{ep}{\delta})}{n}}.$$

Ainsi, l'erreur quadratique est faible dès que p *est petit devant* e^n *et non pas devant* n, *au contraire de l'erreur de l'estimateur des moindres carrés. Nous obtenons donc une vitesse de convergence de l'ordre de* $\sqrt{\log(p)/n}$. *Notons qu'en ajoutant une hypothèse sur la matrice de design* $\mathbf{X}$, *il est possible d'obtenir une vitesse de l'ordre de* $s\log(p)/n$.

Précisons par ailleurs que la convergence de l'erreur quadratique vers 0 *n'implique pas la consistence de l'estimateur (i.e.* $\|\hat{\beta}_\lambda - \beta^0\|_2 \to 0$ *en probabilité) ni sa sparsistence (i.e.* $\mathbb{P}\{\texttt{supp}(\hat{\beta}_\lambda) = \texttt{supp}(\beta^0)\} \to 0$ *où le support* `supp` *désigne l'ensemble des composantes non nulles du vecteur considéré). Autrement dit, la prédiction, l'estimation et la sélection de variables sont trois problématiques bien différentes en régression.*

3 Implémentation du LASSO

Dans cette partie, nous allons implémenter l'estimateur LASSO sur un jeu de données à l'aide du logiciel R ou Python (au choix de l'étudiant). Le dataset (contenant $n = 20$ observations et $p = 8$ variables) est accessible sur *Pamplemousse*. Il est demandé de détailler au maximum le code qui est à joindre au rapport.

Question 9 : L'estimateur LASSO peut aisément être implémenté à l'aide d'une *coordinate descent*. Expliquer le principe d'un tel algorithme et en quoi il particulièrement adapté au LASSO. Détailler au maximum.

Question 10 : Calculer l'estimateur LASSO $\hat{\beta}_\lambda$ pour différentes valeurs de λ variant entre 0 et 1. Commenter. *Il est autorisé d'utiliser des packages de R ou Python. En revanche, un bonus sera attribué à l'étudiant s'il implémente lui-même l'algorithme décrit à la question précédente.*

Question 11 : Expliquer ce qu'est la méthode de validation croisée et en quoi elle peut aider à choisir une valeur de λ, puis sélectionner λ par validation croisée sur le jeu de données.

Question 12 : Representer graphiquement la fonction $\lambda \to \|\hat{\beta}_\lambda\|_0$. Commenter.

Question 13 : En réalité, les données ont été simulées selon (1) avec $\beta^0 = (3, 1.5, 0, 0, 2, 0, 0, 0)$ et $\sigma^2 = 1$. Représenter graphiquement la fonction $\lambda \to \|\mathbf{X}(\hat{\beta}_\lambda - \beta^0)\|_2^2$. Comparer avec les résultats des deux questions précédentes. Commenter.

4 Question bonus

Le LASSO a malgré tout certaines limites en très grande dimension ($p \gg n$). Une alternative bien connue est l'Elastic Net. Présenter le principe de l'Elastic Net, ses avantages sur le LASSO, et proposer une variante de l'algorithme présenté question 12 adaptée à l'Elastic Net. L'appliquer au jeu de données précédent. Commenter.

Envoyer le rapport accompagné du code R ou Python à l'adresse suivante :
badr.eddine.cherief.abdellatif@ensae.fr