ENSAE ParisTech

2016-2017
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 lundi 9 janvier. Les binômes doivent être formés avant le vendredi 25 novembre, et les chargés de TD doivent être avisés par email de la constitution des groupes à cette date (ainsi que le responsable du projet en question).

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
Vincent Cottet
Bureau **E04**
✉ vincent.cottet@ensae.fr

Sujet 1 : Edwin Grappin

Introduction à la régression pénalisée et à l'estimateur lasso

1. Introduction

L'objectif de ce tutoriel est d'introduire la notion de régression pénalisée, d'en comprendre son utilité notamment dans le cas d'une pénalité ℓ_1 pour un modèle parcimonieux.

Nous considérons le cas où nous observons n variables aléatoires $y_1, \dots, y_n \in \mathbb{R}$ et p covariables $\boldsymbol{x}^1, \dots, \boldsymbol{x}^p \in \mathbb{R}^n$. Nous supposons qu'il existe un vecteur $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ tel que les résidus $\xi_i = y_i - \beta_1^\star \boldsymbol{x}_i^1 - \dots - \beta_p^\star \boldsymbol{x}_i^p$ soient des variables aléatoires de lois normales indépendantes. Matriciellement, le modèle s'écrit :

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\xi},$$

où $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ est le vecteur des variables à expliquer, $\mathbf{X} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^p) \in \mathbb{R}^{n \times p}$ la matrice des données et $\boldsymbol{\xi}$ le vecteur de bruit que l'on suppose $\mathcal{N}(0, \sigma^2 \mathbf{I}_n)$.

2. Régression OLS

2.1. Question 1 : démontrer la formule de la régression des moindres carrés sous forme matricielle.

2.2. Question 2 : implémenter cette solution en R.

Pour cette question, implémentez la solution sur le jeu de données, mysmalldata.txt qui est au format csv. La première colonne est les outcome, les autres colonnes sont les données. Présentez vos résultats. Il est attendu que vous recodiez vous même la solution des moindres carrés.

2.3. Question 3 : Expliquer les limites de cette méthodes si le nombre de variables est plus grand que le nombre d'observations

3. Régression pénalisée : le lasso

L'estimateur lasso se définit par :

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

3.1. Question 4 : Preuve du seuillage doux dans le cas $p = 1$

Afin de mieux comprendre l'estimateur lasso, nous nous intéressons au cas simple où une seule variable est à estimer. L'objectif est de comprendre pourquoi l'estimateur lasso est un seuillage doux dans le cas où les variables sont orthogonales (ici, il n'y a qu'une seule variable,

la question de données non orthogonales n'est donc pas à prendre en compte). Montrez que pour tout λ, il existe t tel que l'estimateur $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ peut s'écrire sous la forme :

$$sign(\widehat{\boldsymbol{\beta}}) * (\widehat{\boldsymbol{\beta}} - t)_+,$$

où $\widehat{\boldsymbol{\beta}}$ est l'estimateur des moindres carrés ordinaires et $(.)_+$ la fonction partie positive. Expliquez donc l'impact du seuillage doux sur l'estimation d'un paramètre et expliquez pourquoi cela induit une estimation nulle plus souvent que l'estimateur des moindres carrés. Montrez graphiquement ce résultat. Enfin, expliquer l'impact du paramètre *lambda* sur l'estimateur lasso.

Dorénavant, et cela jusqu'à la fin de ce tutoriel, nous supposons que $p > n$ et qu'un nombre important d'éléments du vecteur $\boldsymbol{\beta}^\star$ sont nuls, on dit que $\boldsymbol{\beta}^\star$ est parcimonieux (ou sparse). Bien que $p > n$, nous cherchons à trouver une méthode qui permette de prédire correctement les valeurs de $\boldsymbol{y}$ en fonction des données $\mathbf{X}$.

3.2. Question 5 : Expliquer ce qu'est la régression pénalisée et pourquoi l'estimateur lasso est adapté à notre problème.

3.3. Question 6 : Quelles sont les avantages du lasso sur la pénalisation ℓ_0 ? Même question avec la pénalisation Ridge.

3.4. Question 7 : Implémenter un estimateur lasso à partir de R (vous pouvez utiliser des packages).

Pour cette question, implémentez la solution sur le jeu de données, mydata.txt qui est au format csv. La première colonne est les outcome $\boldsymbol{y}$, les autres colonnes sont les données $\mathbf{X}$.

4. Le choix du paramètre λ

L'ensemble de cette section se fait sur le jeu de données mydata.txt.

4.1. Question 8 : Expliquer le rôle du paramètre λ.

4.2. Question 9 : Expliquer ce qu'est la méthode de cross-validation et en quoi elle peut vous aider à choisir une valeur de λ.

4.3. Question 10 : Appliquer cette méthode pour choisir λ sur le jeu de données et appliquer le lasso à ce jeu de données.

Combien d'éléments de $\boldsymbol{\beta}$ estimez vous non-nuls ? Evaluez l'erreur de votre modèle. Trouvez de bonnes mesures pour évaluer votre méthode.

4.4. Consignes pour le tutoriel

En plus des consignes communes à tous les projets, merci d'inscrire les membres de votre groupe sur le lien : https://goo.gl/Tnzbt0 , cela me permet de vous contacter si des informations supplémentaires sont nécessaires pour vous aider à répondre au sujet. Par ailleurs, si vous

avez des questions, vous pouvez me contacter par mail : edwin.grappin@ensae.fr. Enfin, la qualité de rédaction du rapport et des codes sera évaluéé. Une partie des points sera attribuée à l'aisance pour reproduire votre code sur un autre ordinateur que le votre. Il faut donc que votre script fonctionne d'un bloc. N'hésitez par à commenter votre script pour expliquer la fonction et l'utilité de chaque partie de votre code. Pour votre rapport vous pouvez utiliser LaTeX ou Markdown.

Statistics 1 project

Alexander Buchholz
alexander.buchholz@ensae.fr

1 The Bayesian logistic regression and Laplace approximation of its posterior

1.1 The logistic regression

The aim of this project is to study an approximation technique in the Bayesian logistic regression. A logistic regression models the dependence between a binary outcome $Y_i \in \{0, 1\}$ and a vector of covariates $X_i \in \mathbb{R}^d$ given an unknown parameter vector $\beta \in \mathbb{R}^d$ and observations $i \in \{1, \cdots, N\}$. As a model we assume that $Y_i \sim \mathcal{B}er(\mu)$, that is to say our observations are sampled from a Bernoulli-model with unknown parameter $\mu \in [0, 1]$. Since the dot product $X_i^T\beta \in \mathbb{R}$, we reparametrize the coefficient via the sigmoid function $S(t) = \frac{1}{1+\exp(-t)}$, such that for all X_i, β we have $S(X_i^T\beta) \in [0, 1]$. Hence, the model becomes

$$Y_i \sim \mathcal{B}er(S(X_i^T\beta)). \tag{1}$$

1.2 The posterior distribution

Let $\mathcal{L}(Y, X|\beta) = \mathcal{L}(y_1, \cdots, y_N, x_1, \cdots, x_N|\beta)$ be the likelihood of the model. We will approach the question of parameter inference via a Bayesian point of view and denote $p(\beta)$ the prior distribution over the parameter β. In our example we will use a Gaussian distribution centered around zero and with a covariance $\Sigma = \sigma^2 I_d$ as prior distribution. The resulting unnormalized posterior reads as

$$p(\beta|Y, X) \propto \mathcal{L}(Y, X|\beta) \times p(\beta). \tag{2}$$

Our aim is to approximate this posterior distribution.

1.3 Laplace approximation of the posterior distribution

We approximate the true log posterior $\log p(\beta|Y, X)$ via a second order Taylor expansion around its mode β^*. This method is known as Laplace approximation. The approximated log posterior $\widehat{\log p}(\beta|Y, X)$ is then transformed back to a density via $\widehat{p}(\beta|Y, X)) \propto \exp\left(\widehat{\log p}(\beta|Y, X)\right)$.

1.4 Numerical example

The numerical application will be based on the dataset *Pima indians* available in the `R` package `EPGLM` via `data(Pima.tr)`. The variable that we want to predict is *type*. For the prior variance we will use $\sigma^2 = 0.1, 1, 5, 10$. Exclude the last 20 observations of the dataset for testing purposes. Furthermore, the covariates need to get demeaned and rescaled by dividing them via their standard deviation. Moreover, Y_i needs to be transformed to be a $\{0, 1\}$ random variable.

Question 1

Theory: Write down the likelihood of the model as well as the log-likelihood and simplify the resulting expressions. Write down an expression for the posterior distribution.

Question 2

Theory: Express the mode of the log posterior distribution via the Newton-Raphson algorithm. Specify the algorithm and the arising terms.

Question 3

Theory: Derive the Laplace approximation of the posterior. Which distribution do you identify? What is the normalization constant of this distribution?

Question 4

Application: Implement the Newton-Raphson algorithm in `R` for the *Pima indians* dataset to identify the mode.

Question 5

Application: Calculate the covariance of the Laplace approximation. Plot the univariate marginal distributions of β for 4 different covariates of your choice. How do they change with different values of σ^2?

Question 6

Theory: Write down the posterior predictive distribution.
Application: Predict the outcomes via your model for the last 20 observations. What is your classification score? Note: Use Monte Carlo for the expectation since you cannot evaluate the integral exactly.

Question 7

Bonus: What other methods could you use to make parameter inference in a Bayesian model? Describe these methods.

Remarks

Comment your `R`–code carefully. Use the `R-Sweave` package to embed your code in a LaTeXfile. Write out and explain all statistical formulas you apply. Try to implement your procedure as general as possible by using functions. Send your final solution (one pdf and one code file) to `alexander.buchholz@ensae.fr`. Do not hesitate to ask for help, if you get stuck.

Sujet 3 : Clara Champagne

Estimation du taux d'infection à partir de données de séroprévalence

Clara Champagne

16 novembre 2016

Les études de séroprévalence sont très utiles en épidémiologie : il s'agit de tester dans une population la présence d'anticorps contre une certaine maladie à une date donnée. Ainsi, on peut savoir la part de la population qui est immunisée contre cette maladie et la part qui y reste susceptible. Cette information est décisive dans la mise en place de mesures de santé publique.

Afin d'étudier le contexte de la dengue au Vietnam, on dispose de données de séroprévalence stratifiées par âge, dans une population d'enfants de 7 à 14 ans. L'objectif est d'évaluer le taux annuel d'infection par la maladie, et également d'extrapoler ces résultats à l'ensemble des tranches d'âges de la population (enfants de moins de 7 ans et adultes, par exemple).

Age	Effectif	Nb. de séropositifs
7	132	70
8	184	106
9	195	121
10	202	132
11	182	146
12	44	38
13	17	15
14	5	4

Tableau 1 – Données. L'effectif correspond au nombre d'enfants testés, le nombre de séropositifs correspond au nombre d'enfants dont le test révèle qu'ils ont déjà eu la maladie.

On fait l'hypothèse que le taux d'infection λ est constant d'une année sur l'autre, et d'une tranche d'âge à l'autre. Sous cette hypothèse, la part de séropositifs augmente avec l'âge (les enfants ayant vécu plus d'années ont plus de chance d'avoir déjà eu la maladie).

On note $F(a_j)$ la proportion d'une cohorte qui, à l'âge a_j, a déjà connu l'infection (celle-ci a pu avoir lieu n'importe quand entre 0 et a_j ans). L'hypothèse du taux d'infection constant

λ se traduit par le choix de la fonction :

$$F_\lambda(a_j) = 1 - exp(-\lambda a_j)$$

On suppose ensuite que, dans chaque tranche d'âge a_j, le nombre R_j d'individus ayant eu la maladie suit une loi binomiale $B(N_j, p_j)$, avec $p_j = F_\lambda(a_j)$ et N_j l'effectif observé dans cette tranche d'âge.

L'objectif est d'estimer λ. Ainsi, on disposera d'un modèle qui permettra ensuite de calculer la séroprévalence par âge pour les enfants de 0 à 15 ans.

Question 1. Ecrire le modèle statistique, ainsi que la vraisemblance du modèle.

Question 2. Calculer l'estimateur du maximum de vraisemblance $\hat{\lambda}^{MV}$. Peut-il s'écrire de manière explicite ?

Question 3. Proposer une solution numérique au calcul du maximum de vraisemblance. Implémenter l'algorithme choisi à l'aide du logiciel R. Vous pouvez comparer les résultats de votre algorithme avec ceux obtenus avec la fonction "mle" du package "stats4".

Question 4. Calculer un intervalle de confiance asymptotique à 95% pour λ (expliciter la méthode utilisée, les hypothèses requises, et réaliser l'implémentation numérique).

Question 5. En utilisant la valeur estimée pour λ, reconstruire la proportion de séroprévalents par âge $\hat{p_j}$ prédite par le modèle.

Comparer les valeurs obtenues avec celles de l'estimateur ponctuel $\tilde{p}_j = R_j/N_j$. On pourra notamment faire un graphique.

Question 6. Conclure et répondre au problème posé.

Question 7. L'hypothèse de taux d'infection constant est peut-être restrictive. On se propose de tester un modèle plus général, utilisant une distribution de Weibull de paramètres λ et p :

$$F_{\lambda,p}(a_j) = 1 - exp(-(\lambda a_j)^p)$$

Calculer numériquement l'estimateur du maximum de vraisemblance avec le nouveau modèle et comparer les résultats obtenus.

Sujet 4 : Léna Carel

Projet de Statistiques 1 — Léna CAREL
lena.carel@ensae.fr

Algorithme EM dans le cadre d'un mélange de gaussiennes

Dans une promotion de première année de licence composée de n étudiants, les moyennes du 1er semestre des étudiants dépendent de la filière de baccalauréat qu'ils ont obtenu l'année précédente. Au sein de chaque groupe d'étudiants ayant suivi la même filière au lycée, les moyennes suivent une distribution normale. Pour formaliser notre problème, X est la variable aléatoire réelle continue qui donne la moyenne obtenue et Z est la variable aléatoire discrète indiquant la filière suivie au lycée à valeur dans $\{1, 2\}$ tel que $\mathbb{P}(Z = 1) = \pi$ et $\mathbb{P}(Z = 2) = 1 - \pi$, où π est inconnu. $(X_1, \ldots, X_n)$ est l'échantillon observé et $(Z_1, \ldots, Z_n)$ est inobservé.

Partie théorique

Q1. Donnez $\mathbb{P}(X_i \leq k)$ et déduisez-en la fonction de densité que suit l'échantillon.

Q2. Donnez la vraisemblance $L(X, \theta)$ avec $\theta = (\theta_1, \theta_2)$ et $\theta_k = (\mu_k, \sigma_k^2)$

Q3. Dérivez le logarithme de la vraisemblance. Peut on estimer θ par la méthode du maximum de vraisemblance ? Si oui, donner l'EMV.

Q4. Déterminer la loi de $Z_i|X_i$.

Q5. Donnez la vraisemblance complétée $L(X, Z, \theta)$.

Q6. Calculer l'espérance de la vraisemblance complétée sous $Z|X, \pi$

Q7. Déduisez-en θ, en appliquant la méthode du maximum de vraisemblance à Q.

Partie implémentation sur R

Q8. Simuler un échantillon de $n = 100$, $\pi = 0.3$, $\mu_1 = 9$, $\sigma_1^2 = 5$, $\mu_2 = 14$, $\sigma_2^2 = 2$.

Q9. Implémenter l'algorithme EM correspondant à notre modèle.

Q10. Prenez comme point de départ de l'algorithme $\pi^{(0)} = \frac{1}{2}$, $\mu_1^{(0)} = \mu_2^{(0)} = 0$ et $\sigma_1^{2(0)} = \sigma_2^{2(0)} = 1$. Commentez les résultats obtenus et tracez l'évolution de la log-vraisemblance au cours des itérations. Réitérez les mêmes étapes pour différents points de départ de l'algorithme.

Q11. Faites varier π, l'algorithme arrive t-il bien à retrouver les groupes d'étudiants lorsque ceux-ci sont trop ? Tracez sur un même graphique l'évolution de la log-vraisemblance au cours des itérations pour chaque π.

Q12. Faites varier les variances, que pouvez-vous dire sur la capacité de l'algorithme à bien retrouver les groupes d'étudiants lorsque les variances sont grandes ? Tracez sur un même graphique l'évolution de la log-vraisemblance au cours des itérations pour chaque couple de variances (σ_1^2, σ_2^2) que vous testerez.

Q13. Concluez sur tous les tests que vous avez fait.

Sujet 5 : Clément Rousset

Tutoriel : Régression quantile et Bootstrap

Clément Rousset, clement.rousset@insee.fr

La régression linéaire "normale" se présente sous la forme

$$E(Y|X) = X'\beta,$$

les coefficients β permettent d'obtenir l'espérance de $Y|X$. La régression quantile consiste à remplacer l'espérance par le quantile τ de la distribution $Q_\tau(Y|X)$. Le modèle devient donc :

$$Q_\tau(Y|X) = X'\beta.$$

Pour $\tau = \frac{1}{2}$, on obtient un modèle sur la médiane de la distribution $Y|X$.

Dans la suite on ne considérera que la médiane et un seul régresseur (pas de constante) : $Q_{\frac{1}{2}}(Y|X) = \beta x$.

1 Préliminaires mathématiques

1. Pour une v.a. X, quel est le réel α minimisant $E[(X-\alpha)^2]$?
2. Pour une v.a. X, quel est le réel α minimisant $E[|X-\alpha|]$?
3. On rappelle que l'estimateur de la régression linéaire est celui des moindres carrées ordinaires, donc qu'il vérifie

$$\hat{\beta}_{MCO} = \underset{\beta}{ArgMin} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i)^2.$$

Définir par analogie l'équation que vérifie l'estimateur $\hat{\beta}$ pour la régression sur la médiane.

2 Simulations sur R

4. Importer les séries y et x dans R puis programmer et exécuter l'algorithme qui donne l'estimateur $\hat{\beta}$ pour la régression sur la médiane (utilisation de la fonction *optimize*).
5. Contrairement à l'espérance, l'opérateur médiane n'est pas linéaire. Il en découle qu'il est plus difficile d'obtenir des formules d'écarts-type pour l'estimateur $\hat{\beta}$. Une des solutions alors est de passer par la méthode du Bootstrap.
   Mettre en œuvre cette méthode sous R (sans utiliser de packages spéciaux, la fonction *sample* suffit) et proposer un écart-type pour l'estimateur, bien justifier.
6. En utilisant la fonction *rq* du package "quantreg", obtenir un autre estimateur de β et de son écart-type. Comparer.

3 Ouverture

7. Chercher à refaire la question 3 avec un τ quelconque.

Sujet 6 : Gautier Appert

Régression linéaire robuste

Gautier Appert gautier.appert@ensae.fr

On considère le modèle de régression linéaire robuste suivant

$$y_i = x_i^\top \beta + \sigma \varepsilon_i, \qquad i \in \{1, \dots, n\},$$

avec $\varepsilon_i \overset{i.i.d.}{\sim} t_\nu$. Les vecteurs $x_i \in \mathbb{R}^p$ sont supposés déterministes, et le degré de liberté $\nu > 0$ est supposé connu. Le paramètre d'intérêt du modèle est $\theta = (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*$. Dans ce modèle les variables dépendantes $\{y_i\}_{i=1}^n$ suivent donc une loi de Student décentrée avec un paramètre d'échelle $\sigma > 0$, autrement dit pour tout $i \in \{1, \dots, n\}$, $y_i \sim t_\nu(x_i^\top \beta, \sigma^2)$. On obtient donc la densité suivante pour chaque y_i

$$f_\theta(y_i) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{(y_i - x_i^\top \beta)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

où $\Gamma(a) \overset{def}{=} \int_0^{+\infty} x^{a-1} \exp\{-x\} \, \mathrm{d}x$, pour tout $a > 0$.

Préliminaires

Question (1). La loi de Student correspond-elle à un modèle exponentiel ? (justifier). Soit $X \sim \mathcal{N}(0,1)$ et $Z \sim \chi^2(\nu)$ tel que $X \perp\!\!\!\perp Z_\nu$. La Student centrée T_ν est définie par $T_\nu = X/\sqrt{Z_\nu/\nu}$. A l'aide de l'inégalité de Tchebychev, montrer que $Z_\nu/\nu \xrightarrow[\nu\to+\infty]{\mathbb{P}} 1$. En déduire que $T_\nu \xrightarrow[\nu\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$.

Question (2). Ecrire la vraisemblance associée au modèle linéaire robuste et dériver les équations de vraisemblance. Peut-on obtenir des estimateurs du maximum de vraisemblance sous forme explicite ? *Bonus:* Montrer que la vraisemblance n'est pas concave globalement. *Notons que la méthode de Newton Raphson n'est pas appropriée lorsque la vraisemblance n'est pas globalement concave.*

Modèle de mélange et Algorithme EM

Il est possible de représenter une loi de Student comme un mélange d'une loi normale, avec comme loi mélangeante la distribution du $\chi^2(\nu)$

$$y|z \sim \mathcal{N}\left(x^\top \beta, \frac{\nu\sigma^2}{z}\right), \qquad z \sim \chi^2(\nu) \equiv \gamma\left(\frac{\nu}{2}, \frac{1}{2}\right).$$

La variable aléatoire z joue ici le rôle de la variable latente. L'objectif de cette partie est de déterminer un estimateur $\hat{\theta}$ à l'aide de l'algorithme d'Espérance Maximisation (EM) en se basant sur cette représentation d'une loi de Student vue comme un mélange. *On suppose jusqu'à la question (5) que l'on dispose seulement d'une seule variable aléatoire y associée à la variable latente z.*

Question (3). Justifier la représentation de la Student $t_\nu(x^\top \beta, \sigma^2)$ vu comme le mélange décrit précédement. De plus, montrer que $z|y \sim \gamma\left(\frac{\nu+1}{2}, \frac{(y-x^\top\beta)^2+\nu\sigma^2}{2\nu\sigma^2}\right)$.

Question (4). On définit la vraisemblance complète comme étant la vraisemblance augmentée de la variable latente $L(y, z; \theta)$. En utilisant le fait que $f_\theta(z|y) = \frac{L(y,z;\theta)}{L(y;\theta)}$, montrer que pour tout θ_0 fixé

$$\log L(y;\theta) = \mathbb{E}_{z\sim f_{\theta_0}(z|y)}\Big[\log L(y,z;\theta)\Big] - \mathbb{E}_{z\sim f_{\theta_0}(z|y)}\Big[\log f_\theta(z|y)\Big].$$

Question (5). On pose la fonction $Q(\theta;\theta_0,y) \overset{def}{=} \mathbb{E}_{z\sim f_{\theta_0}(z|y)}\Big[\log L(y,z;\theta)\Big]$. L'algorithme EM consiste à maximer cette fonction en θ, afin d'obtenir un premier estimateur $\hat{\theta}_1$. En remplaçant $\theta_0 \leftarrow \hat{\theta}_1$, et en maximisant à nouveau la fonction $Q(\theta;\theta_1,y)$, l'algorithme crée de manière itérative une séquence d'estimateurs $(\hat{\theta}_k)_{k\geq 1}$. Montrer que la vraisemblance augmente à chaque itération de l'algorithme EM, ie $L(y;\hat{\theta}_{k+1}) \geq L(y;\hat{\theta}_k)$.

Question (6). On dispose désormais d'un échantillon de variables aléatoires $\{y_i\}_{i=1}^n$ associées aux variables latentes $\{z_i\}_{i=1}^n$. Ecrire la log vraisemblance complète du modèle $\log L\left(\{(y_i,z_i)\}_{i=1}^n;\theta\right)$ et calculer la fonction $Q(\theta;\theta_0,\{y_i\}_{i=1}^n)$. On pourra utiliser le fait que si $X\sim\gamma(a,b)$ alors $\mathbb{E}[X]=a/b$ et $\mathbb{E}[\log(X)]=\Psi(a)-\log(b)$ où $\Psi(x)=\Gamma'(x)/\Gamma(x)$.

Question (7). Ecrire les conditions du premier ordre concernant la fonction $Q\left(\theta;\theta_0,\{y_i\}_{i=1}^n\right)$ et montrer que les équations de récurrence liées à la maximisation sont données pour tout $k\geq 0$ par

$$\hat{\beta}_{k+1} = \left(\sum_{i=1}^n \frac{x_i x_i^\top}{(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2}\right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2}$$
$$\hat{\sigma}_{k+1}^2 = \frac{1}{n}\sum_{i=1}^n \frac{(\nu+1)\hat{\sigma}_k^2(y_i - x_i^\top\hat{\beta}_{k+1})^2}{(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2}$$

Question (8). *Question Bonus:* On pose la matrice de poid $W_k = \mathbf{diag}\left(1/[(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2]\right)$ ainsi que la matrice $X = [x_1|\dots|x_n]^\top$. Réécrire les équations de récurrence précédentes en utilisant la matrice de poid W_k et la matrice X. A quelle type d'estimateur vu en cours d'Econométrie $\hat{\beta}_{k+1}$ correspond t'il ?

Programmation avec R

On souhaite dans cette section appliquer l'algoritme EM sur le jeu de données `election` issu du package `LearnBayes` sous `R`. Ce jeu de données contient en particulier le nombre de votes lors des élections présidentielles américaines en 2000 pour le condidat *Pat Buchanan* et le nombre de votes en 1996 pour le candidat *Ross Perot*, dans chacun des 67 comtés de Floride *. On veut estimer un modèle linéaire simple $y_i = \beta_0 + \beta_1 x_i + \sigma\varepsilon_i$, où y_i représente la racine carré du nombre de votes pour le candidat Pat Buchanan dans le comté numéro i et x_i la racine carré du nombre de votes pour le candidat Ross Perot.

Question (9). Faire un plot des données $\{y_i\}_{i=1}^n$ en fonction des $\{x_i\}_{i=1}^n$. Justifier l'utilisation d'un modèle linéaire simple robuste.

Question (10). Estimer un modèle linéaire simple robuste à l'aide de l'algorithme EM en prenant plusieurs valeurs de ν, $\nu = 1; 10; 100$ (*on peut prendre un critère d'arrêt basé sur la différence des log vraisemblances prises entre deux itérations successives*). Prendre plusieurs initialisations différentes pour le paramètre $\theta = (\beta, \sigma^2)$. Faire un plot de l'évolution de la log vraisemblance à chaque itération.

Question (11). Estimer le modèle par Moindres carrées ordinaires (MCO), et afficher sur le graphique de la question (9) les droites de régressions issues de l'estimation par MCO et de l'algorithme EM. Conclure.

Question (12). Faire un Harlem Shake.

*Variables intitulées `Buchanan` et `Perot` dans le jeu de données `election`.

Sujet 7 : Lionel Riou-Durand

Régression de Poisson

Lionel Riou-Durand

La régression de Poisson permet de modéliser le lien entre un ensemble de cofacteurs ($X \in \mathbb{R}^p$) et une variable dépendante de comptage ($Y \in \mathbb{N}$). On associe souvent cette loi à l'idée d'un grand nombre d'évènements possibles, individuellement très peu probables, mais qui mènent, globalement, à l'observation d'un nombre d'évènements non négligeable (e.g. nb d'appels téléphonique pendant un certain laps de temps, du nb d'incendies dans une certaine zone géographique, nb d'homicides dans une certaine population). La régression de Poisson trouve beaucoup d'applications en actuariat, mais aussi en démographie, en épidémiologie, etc...

La régression de Poisson peut être généralisée pour permettre l'étude d'échantillons où chaque observation i est liée à une mesure d'exposition $N_i > 0$ qui lui est propre. On modélise en pratique un nombre d'évènements standardisé par unité d'exposition (e.g. nb d'appels par minute, nb d'incendies par km^2, nb de meurtres pour 100 000 habitants), pour s'affranchir d'un effet taille. Ainsi, on supposera que pour une observation i : $Y_i \sim \mathcal{P}oisson\big(N_i\ e^{X_i^T\beta}\big)$. On suppose dès lors que le logarithme du nb d'évènements moyen par unité d'exposition est une combinaison linéaire des régresseurs...

1 Partie théorique (12 points + 2 bonus)

1.1 Design fixe

On suppose que les covariables $(N_i, X_i)_{i=1,...,n} \in \big(]0, +\infty[\times\mathbb{R}^p\big)^n$ sont déterministes, avec $n > p$ fixé. On suppose de plus que les variables dépendantes $(Y_i)_{i=1,...,n} \in \mathbb{N}$ sont indépendamment distribuées tel que pour chaque $i = 1, ..., n$: $Y_i \sim \mathcal{P}oisson\big(N_i\ e^{X_i^T\beta}\big)$, où $\beta \in \mathbb{R}^p$ est un vecteur de paramètres inconnus.

1) Ecrire la vraisemblance du modèle, ainsi que la log-vraisemblance.

2) L'estimateur du maximum de vraisemblance n'admet pas de forme analytique. Mais il s'écrit comme le zéro du gradient, on peut donc le calculer numériquement via l'algorithme de Newton-Raphson. Ecrire le pseudo-code correspondant à notre problème d'optimisation.

Bonus : Supposons que la matrice des régresseurs soit de plein rang, i.e. que $\mathbb{X} := (X_1, ..., X_n)^T$ *soit de rang p. Montrer que cette hypothèse suffit à assurer l'existence et l'unicité du maximum de vraisemblance.*

3) Mathématiquement, le modèle de régression en design fixe ne suffit pas pour appliquer les propriétés asymptotiques de l'estimateur du maximum de vraisemblance. Ecrire le modèle statistique, quelle hypothèse fondamentale n'est pas vérifiée ?

1.2 Design aléatoire

Pour pallier cette difficulté technique, on va changer de modélisation. Désormais, on supposera $(N_i, X_i, Y_i)_{i=1,...,n} \overset{i.i.d.}{\sim} \mathbb{P}_\beta$ tel que $(N_1, X_1) \in (\mathcal{N}, \mathcal{X}) \subset]0, +\infty[\times\mathbb{R}^p$ admette une densité $g(n, x)$ ne dépendant pas de β, et tel que $(Y_1 \mid N_1, X_1) \sim \mathcal{P}oisson(N_1 e^{X_1^T \beta})$.

Remarque : l'hypothèse sur le caractère i.i.d. des triplets ou sur la loi de (X_1, N_1) est souvent assez raisonnable, c'est généralement l'hypothèse sur la loi de $(Y_1 \mid N_1, X_1)$ qui reste forte.

5) Ecrire la vraisemblance du modèle en design aléatoire, montrer que la forme de $g(n, x)$ importe peu dans le problème d'estimation par maximum de vraisemblance.

6) Dans le modèle à design aléatoire, que vaut l'information de Fisher ?

Bonus : Supposons que $\mathcal{N} \times \mathcal{X}$ soit de la forme $[a, b] \times K$ où $b > a > 0$ et K est un compact de $\mathbb{R}^d$; et que la matrice de covariance $\mathbb{E}[X_1 X_1^T]$ soit définie positive. Montrer que pour tout $\beta \in \mathbb{R}^p$ la matrice d'information de Fisher existe bien et est définie positive.

7) On admettra l'existence et l'unicité de l'estimateur du maximum de vraisemblance noté $\widehat{\beta}^{(n)}_{MLE}$ à partir d'un certain rang. Donner la distribution asymptotique de $\sqrt{n}\big(\widehat{\beta}^{(n)}_{MLE} - \beta\big)$.

8) On définit l'approximation empirique de l'information de Fisher comme suit :
$\widehat{I}_1^{(n)}(\beta) := \frac{1}{n}\sum_{i=1}^n -\nabla^2_\beta \log f_\beta\big(N_i, X_i, Y_i\big)$ où f_β est la densité du vecteur (N_1, X_1, Y_1).

On admettra que : $\widehat{I}_1^{(n)}\big(\widehat{\beta}^{(n)}_{MLE}\big) \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} I_1(\beta)$, donner en les justifiant :

- un intervalle de confiance asymptotique à 95% pour une composante $j \in \{1, ..., p\}$ du vecteur β.
- une ellipse de confiance asymptotique à 95% pour $\beta \in \mathbb{R}^p$.

2 Mise en oeuvre (8 points)

En guise d'application pratique, nous allons étudier empiriquement le nombre d'homicides aux Etats-Unis pendant l'année 1976. On s'intéressera à l'effet de certaines caractéristiques socio-économiques, et géographiques de chaque état. Le nombre d'homicides par état sera standardisé par la population, afin de modéliser un taux d'homicides pour 100 000 habitants. L'objectif est de mettre en oeuvre l'estimation par maximum de vraisemblance sur ces données, en recodant entièrement une procédure d'estimation implémentable dans un logiciel.

Sur pamplemousse, vous trouverez un script PoissonReg.R qui importe les données.
Source : U.S. Department of Commerce, Bureau of the Census (1977)
Références : Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth and Brooks/Cole. McNeil, D. R. (1977) Interactive Data Analysis. New York: Wiley.

Notre modèle de régression de régression s'écrit $Y_i \sim \mathcal{P}oisson\big(N_i\ e^{X_i^T \beta}\big)$, $i = 1, ..., n$. Selon cette notation, les données composent ici un échantillon de $n = 50$ états, de plus :

- Y_i est le nombre d'homicides commis dans l'état i pendant l'année 1976.
- N_i représente la population estimée (1975) de l'état i, en centaines de milliers d'habitants.
- $X_i \in \mathbb{R}^{11}$ synthétise quelques caractéristiques socio-économiques et géographiques de l'état i (revenu par habitant, taux d'alphabétisation, pourcentage de diplômés, espérance de vie, densité de population, région géographique, nb de jours de gel par an, population urbaine en proportion). La modélisation inclura un régresseur constant ($X_i^1 = 1$).

1) Coder les fonctions *LogVraisemblance*, *Gradient*, *Hessienne*, permettant l'évaluation de la log-vraisemblance et de ses deux premières différentielles, en fonction des données et d'un vecteur paramètre $\beta \in \mathbb{R}^p$.

2) Coder une fonction, qui à partir des données et d'un vecteur initial $\beta_0 \in \mathbb{R}^p$, met en oeuvre l'algorithme de Newton-Raphson calculant une approximation numérique du maximum de vraisemblance. On s'arrangera pour que cette fonction renvoie à la fois : l'approximation du maximum, la Hessienne correspondante, l'évolution de la log-vraisemblance à chaque itération.

3) Estimer le modèle par maximum de vraisemblance. On présentera les résultats sous la forme d'un tableau récapitulant le vecteur maximum, et pour chacune de ses composantes : un écart-type approché, un intervalle de confiance à 95% approché. Représenter graphiquement l'évolution de la log-vraisemblance, on pourra partir du vecteur initial $\beta_0 = 0_{\mathbb{R}^p}$.