Université de Montpellier — Année 2019-2020
HMMA304 - Modèles linéaires généralisés

# Examen final

## Durée 2h - Sans document - 4 novembre 2019

**Exercice 1 (4 pts)**

**1) (1 pt)** Donner un exemple de modèle log-linéaire hiérarchique qui n'est pas un modèle log-linéaire graphique.

**2) (1 pt)** Représenter le modèle log-linéaire graphique associé à l'équation suivante

$$x_1 * x_2 * x_3 + x_1 * x_4 + x_4 * x_5 * x_6 + x_7\,.$$

Donner deux relations d'indépendance conditionnelle entre les variables qui le composent.

**3) (2 pts)** Nous considérons 100 réalisations suivant deux variables qualitatives ayant chacune deux modalités. Donner la vraisemblance associée au modèle log-linéaire graphique poissonien d'indépendance. Montrer qu'il s'agit bien d'un modèle linéaire généralisé.

**Exercice 2 (8 pts)**

On considère la variable aléatoire $X$ admettant la densité de probabilité suivante :

$$f_X(x;k) = \frac{(2k+1)!\Phi(x)^k\Phi(-x)^k}{(k!)^2\sqrt{2\pi}\exp\left(x^2/2\right)} \tag{1}$$

où $k \geq 1$ est un entier et $\Phi$ est la fonction de répartition de la loi normale centrée réduite.

**1 (2 pts)** Montrer $\Phi(x)\Phi(-x) \leq 1/4$.

**2 (4 pts)** Proposer une méthode de simulation suivant (1). Discuter de l'efficacité de l'algorithme proposé.

**3 (2 pts)** On peut montrer que la loi de probabilité associée à la densité (1) correspond à la distribution de la médiane d'un échantillon de $n = 2k+1$ variables gaussiennes centrées réduites. On admet ce résultat. Donner alors un code `R` permettant de vérifier la validité de l'algorithme proposé dans la question précédente.

**Exercice 3 (4 pts)**

On souhaite estimer
$$I = \mathbb{E}_{\mathcal{N}(0,1)}\left(\sqrt{2+\sin(X)}\right) .$$

**1 (2 pts)** Proposer un estimer $\hat{I}_n$ de $I$ basé sur la méthode de Monte-Carlo standard.

**2 (2 pts)** En utilisant l'inégalité de Bienaymé-Chebyshev, déterminer le nombre de simulation nécessaire pour que l'erreur relative de $\hat{I}_n$ soit inférieure à 1% avec une probabilité supérieure ou égale à 99%.

**Exercice 4 (4 pts)**

On considère $n$ variables aléatoires indépendantes $y_1, \ldots, y_n$ telles que $y_i$ est distribué suivant la loi ci-dessous de paramètre $(\pi_i)$
$$f(y_i; \pi_i) = \pi(1-\pi)^{y_i} \mathbf{1}_{\mathbb{N}}(y_i) .$$

**1 (1 pt)** Montrer que cette loi appartient à la famille exponentielle.

**2 (1 pt)** Calculer $\mathbb{E}(y_i)$ et $\mathbb{V}(y_i)$.

Pour tout $i = 1, \ldots, n$, on suppose que $\log(\pi_i/(1-\pi_i)) = \beta_1 + \beta_2 x_i$.

**3 (1 pt)** Montrer qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ?

**4 (1 pt)** Donner la log-vraisemblance et les équations de vraisemblance.

# Correction examen final HMMA304 4/11/2019

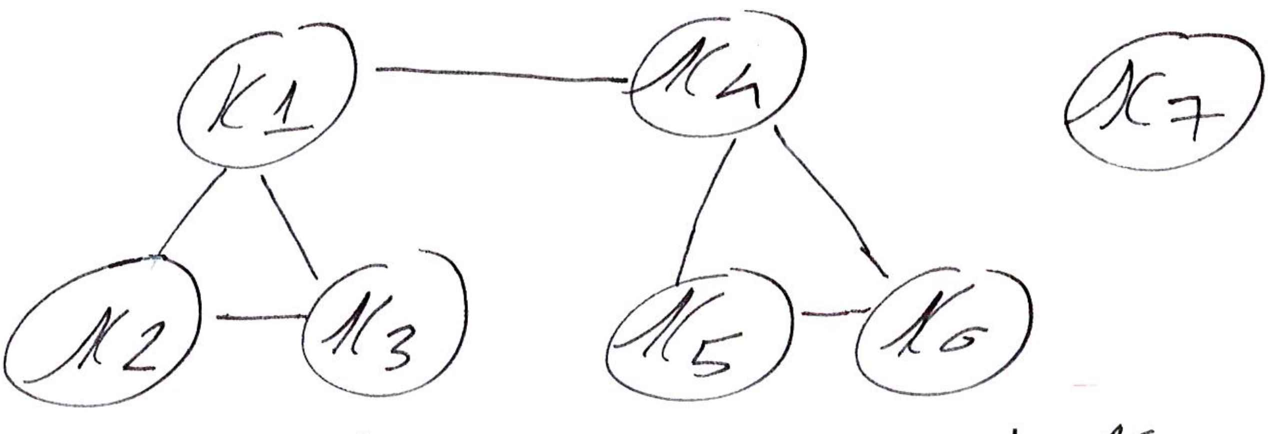
## Exercice 1

1| L'exemple typique est

$$x_1 * x_2 + x_2 * x_3 + x_1 * x_3$$

C'est un modèle linéaire hiérarchique mais pas un modèle linéaire graphique.

2|




$x_7 \perp x_4$ ; $x_1 \perp x_5 \mid x_4$

$x_3 \perp (x_4, x_5, x_6) \mid x_2, x_3$

3]

(2)

$$y_{ij} \sim \mathcal{P}(\mu_{ij}) \, , \quad \begin{matrix} i = 1,2 \\ j = 1,2 \end{matrix}$$

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j$$

modèle d'indépendance

contraintes identifiabilités :

$$\alpha_1 + \alpha_2 = 0 \text{ et } \beta_1 + \beta_2 = 0$$

$$V(\mu_{ij} ; y_{ij}) = \prod_{i=1}^{2} \prod_{j=1}^{2} \Big[ \exp(-\exp(\mu + \alpha_i + \beta_j))$$

$$\exp(y_{ij}(\mu + \alpha_i + \beta_j)) \frac{1}{y_{ij}!}$$

Exercice 2

1] $\Phi(x)\Phi(-x)$

$$= \Phi(x)(1 - \Phi(x))$$

Soit $h(u) = u(1-u)$

avec $u \in [0,1]$

$h'(u) = (1-u) - u$

$\Longrightarrow h'(u) = 0 \iff u = \frac{1}{2}$

$h''(u) = -2 < 0$ (3)

$h(\cdot)$ est une fonction strictement concave.

Donc $h(u) \leq h\left(\frac{1}{2}\right)$ $\forall u \in [0,1]$

$$\Rightarrow h(u) \leq \frac{1}{4}.$$

Et ainsi, $\Phi(x)\Phi(-x) \leq 1/4$

2) Soit $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

$$\frac{f_X(x;k)}{g(x)} = \frac{(2k+1)!}{(k!)^2}\left[\Phi(x)\Phi(-x)\right]^k$$

$$\frac{f_X(x;k)}{g(x)} \leq \frac{(2k+1)!}{(k!)^2}\left(\frac{1}{4^k}\right)$$

On peut utiliser une méthode acceptation-rejet avec comme loi instrumentale (loi de proposition) la loi normale centrée réduite.

la probabilité d'acceptation est égale à =

$$\alpha = \frac{4^k (k!)^2}{(2k+1)!}$$

3) On compare un histogramme de N simulation suivant l'algorithme acceptation - rejet précédent à N simulations consistant à simuler N fois n gaussiennes et en prendre la médiane.

```
> simu1 <- ar(N)
> simu2 <- rep(0,N)
> for (i in 1:N)
      simu2[i] <- median(rnorm(n))
> hist(simu1); hist(simu2, col="red",
                  add=TRUE)
```

## Exercice 3

$$I = \mathbb{E}_{\mathcal{N}(0,1)}\left(\sqrt{2+\sin(X)}\right)$$

1) $$\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N}\sqrt{2+\sin(X_i)}$$

où $X_1, \dots, X_N \overset{iid}{\sim} \mathcal{N}(0,1)$

2) $$ER(\hat{I}_N) = \frac{|\hat{I}_N - I|}{|I|}$$

On cherche le plus petit $N$ tel que

$$\mathbb{P}(ER(\hat{I}_N) \leq 0.01) \geq 0.99$$

$$\Leftrightarrow \mathbb{P}(ER(\hat{I}_N) \leq 0.01) \geq 0.99$$

$$\Leftrightarrow \mathbb{P}(|\hat{I}_N - I| \leq 0.01|I|) \geq 0.99$$

D'après l'inégalité de Bienaymé-Chebyshev, nous avons

$$\mathbb{P}(|\hat{I}_N - I| \leq 0.01|I|) \geq 1 - \frac{\mathbb{V}(\hat{I}_N)}{(0.01 I)^2}$$

$$\Leftrightarrow \mathbb{P}(ER(\hat{I}_N) \leq 1\%) \geq 1 - \frac{\mathbb{V}(\hat{I}_N)}{I^2}(100)^2$$

$$\mathbb{V}(\widehat{I}_N) = \frac{1}{N} \mathbb{V}_{\mathcal{N}(0,1)}\left(\sqrt{2+\sin(x)}\right)$$

$$\mathbb{V}_{\mathcal{N}(0,1)}\left(\sqrt{2+\sin(x)}\right) = \mathbb{E}_{\mathcal{N}(0,1)}\left(2+\sin(x)\right) - I^2$$

$$\Rightarrow \quad \frac{\mathbb{V}(\widehat{I}_N)}{I^2} = \frac{1}{N}\left[\frac{\mathbb{E}_{\mathcal{N}(0,1)}\left(2+\sin(x)\right)}{I^2} - 1\right]$$

$$\left(2+\sin(x)\right) \leq 3$$

$$\mathbb{E}_{\mathcal{N}(0,1)}\left(2+\sin(x)\right) \leq \sqrt{3}$$

ainsi,

$$2+\sin(x) \geq 1$$

$$\sqrt{2+\sin(x)} \geq 1$$

$$I \geq 1$$

$$\frac{1}{I^2} \leq 1$$

ainsi, $$\frac{\mathbb{V}(\widehat{I}_N)}{I^2} \leq \frac{1}{N}\left[\sqrt{3} - 1\right]$$

donc,

$$1 - \frac{\mathbb{V}(\hat{I}_N)}{I^2}(100)^2 \geq 1 - \frac{1}{N}(100)^2(\sqrt{3}-1)$$

on cherche donc la plus valeur N telle que

$$1 - \frac{1}{N}(100)^2(\sqrt{3}-1) \geq 0.99$$

$$N \geq (100)^3(\sqrt{3}-1)$$

Exercice 4

1) $f(y;\pi) = \exp\{y \log(1-\pi) + \log(\pi)\}$

$$\mathbb{1}_{\mathbb{N}}(y)\,\delta(dy)$$

$\nu(dy) = \mathbb{1}_{\mathbb{N}}(y)\,\delta(dy)$

$\theta = \log(1-\pi) \Rightarrow \pi = 1 - e^{\theta}$

$b(\theta) = -\log(1 - e^{\theta})$

2) $\mathbb{E}(y) = b'(\theta) = \frac{e^{\theta}}{1-e^{\theta}} = \frac{1-\pi}{\pi}$

$\mathbb{V}(y) = b''(\theta) = \frac{e^{\theta}}{(1-e^{\theta})^2} = \frac{1-\pi}{\pi^2}$

3] $\mathbb{E}(y) = \frac{1-\pi}{\pi} = e^{-\beta_1-\beta_2 x}$

$= f(\beta_1+\beta_2 x)$

Il s'agit bien d'un modèle linéaire généralisé.

$\mathbb{E}(y) \neq \theta = \log(1-\pi)$

ce n'est pas le lien canonique qui a été utilisé

4] $LV(\beta_1,\beta_2) = \sum_{i=1}^{n} \log(1+e^{-\beta_1-\beta_2 x_i})\left[-y_i-1\right]$

$+ \sum_{i=1}^{n} y_i(-\beta_1-\beta_2 x_i)$

or $\frac{1-\pi}{\pi} = e^{-\beta_1-\beta_2 x}$

$\iff 1-\pi = \pi\, e^{-\beta_1-\beta_2 x}$

$\iff \pi = \frac{1}{1+e^{-\beta_1-\beta_2 x}} = (1+e^{-\beta_1-\beta_2 x})^{-1}$

et $1-\pi = 1 - \frac{1}{1+e^{-\beta_1-\beta_2 x}} = \frac{e^{-\beta_1-\beta_2 x}}{1+e^{-\beta_1-\beta_2 x}}$

$$\frac{dLV}{d\beta_1}(\beta_1,\beta_2) = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}(y_i+1)\left(\frac{e^{-\beta_1-\beta_2 x_i}}{1+e^{-\beta_1-\beta_2 x_i}}\right)$$

$$\frac{dLV}{d\beta_2}(\beta_1,\beta_2) = -\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n}(y_i+1)\,x_i\left(\frac{e^{-\beta_1-\beta_2 x_i}}{1+e^{-\beta_1-\beta_2 x_i}}\right)$$