UNIVERSITÉ PAUL SABATIER **UPSSITECH**

EXAMEN D'ESTIMATION STOCHASTIQUE – 2ASRI

1° session – Mardi 13 Novembre 2018

Durée 1h30 – Tous documents de Cours, TD, TP autorisés – Tablettes et objets communicants interdits

I. Questions de cours/ Répondre *en trois phrases maximum, claires et convenablement construites,* à chacune des questions suivantes. *Les formules mathématiques seront particulièrement soignées.*

1. Soient $X \in \mathbb{R}^{n_X}$ et $Y \in \mathbb{R}^{n_Y}$ deux vecteurs aléatoires continus, $\mathrm{p}_X(x)$, $\mathrm{p}_Y(y)$ leurs densités de probabilités respectives, et $\mathrm{p}_{X,Y}(x,y)$ leur densité de probabilité jointe. Rappeler le sens de $\mathrm{p}_X(x)$, $\mathrm{p}_Y(y)$, $\mathrm{p}_{X,Y}(x,y)$. Citer, s'il y en a, les relations qui unissent ces lois.
2. Comment s'énonce un problème d'estimation en contexte classique (Fisher) ?
3. Soit $p_{Z|\theta}(z|\theta)$ le modèle d'observation qui unit un vecteur de paramètres $\theta \in \mathbb{R}^M$ déterministe inconnu au vecteur aléatoire de mesure $Z \in \mathbb{R}^N$, et soit $I(\theta)$ la matrice d'information de Fisher associée. On désigne par $\mathcal{N}(b,P)$ désigne la loi gaussienne multidimensionnelle d'espérance b et de matrice de covariance P. Que signifie le résultat suivant ?

   (Sous certaines hypothèses de régularité,) lorsque $N \to +\infty$, l'estimateur du maximum de vraisemblance $\hat{\Theta}_{\mathsf{MLE}}$ de θ satisfait $\hat{\Theta}_{\mathsf{MLE}} \sim \mathcal{N}(\theta, I^{-1}(\theta))$.

II. Synthèse d'un estimateur par maximum de vraisemblance/
Étant donnée la séquence de scalaires réels $x_1,\dots,x_N$, on souhaite estimer deux constantes scalaires réelles inconnues a,b sur la base de N observations $z_1,\dots,z_N$, réalisations des variables aléatoires $Z_1,\dots,Z_N$ liées à a, b et $x_1,\dots,x_N$ par

$$\forall n = 1,\dots,N,\ Z_n = ax_n + b + W_n \tag{1}$$

où $W_1,\dots,W_N$ sont des variables aléatoires scalaires réelles i.i.d. selon la loi Gaussienne $\mathcal{N}(0,\sigma^2)$ de moyenne nulle et de variance σ^2 donnée. Un problème physique correspondant à cet énoncé mathématique consiste à établir une régression linéaire entre les N points de coordonnées $\{\left(\begin{smallmatrix} x_n \\ z_n \end{smallmatrix}\right)\}_{n=1,\dots,N}$.

Répondre aux questions suivantes.

4. Reformuler le problème de façon à exprimer le vecteur d'observations $z = (z_1,\dots,z_N)^T$ comme la réalisation d'un vecteur aléatoire $Z = (Z_1,\dots,Z_N)^T$ lié au vecteur $\theta = (a,b)^T$ des paramètres cachés par une équation linéaire avec bruits additifs de la forme

   $$Z = H\theta + V, \tag{2}$$

   en explicitant les expressions de la matrice déterministe H et du vecteur de bruits V.
5. Expliquer pourquoi V est Gaussien. Donner son espérance et sa matrice de covariance.
6. Écrire, sans la développer, l'expression de l'estimé du maximum de vraisemblance $\hat{\theta}$ du vecteur θ, obtenu pour une réalisation z du vecteur Z. *On réutilisera autant que possible des résultats développés en cours.*
7. Préciser à quelles conditions cet estimé peut être obtenu.
8. Développer ce calcul. On pourra éventuellement introduire les quantités suivantes :

$$\bar{x} = \sum_{n=1}^{N} x_n\,;\quad v_x = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 = \left(\frac{1}{N}\sum_{n=1}^{N} x_n^2\right) - \bar{x}^2\,; \tag{3}$$

$$\bar{z} = \sum_{n=1}^{N} z_n\,;\quad c_{xz} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(z_n - \bar{z}) = \left(\frac{1}{N}\sum_{n=1}^{N} x_n z_n\right) - \bar{x}\bar{z}. \tag{4}$$

9. Soient $\hat{a}, \hat{b}$ (composantes de l'estimé $\hat{\theta}$) les estimés de a, b (composantes de θ) obtenus pour un vecteur d'observations z (réalisation de Z). On souhaite déterminer des intervalles de confiance autour de $\hat{a}$ et $\hat{b}$ dans lesquels se trouvent les paramètres cachés a, b avec une probabilité de 99.7%. Expliciter les étapes suivantes permettant d'obtenir ce résultat.

   (a) Rappeler les statistiques de l'estimateur $\hat{\Theta}$:

      i. Expliquer pourquoi il est Gaussien.

      ii. Démontrer qu'est est centré sur θ.

      iii. Rappeler, *sans démonstration, sur la base de résultats développés en cours*, l'expression de sa matrice de covariance.

   (b) En déduire les statistiques de l'erreur d'estimation $\varepsilon = \theta - \hat{\Theta}$:

      i. Montrer que cette variable (2D) est Gaussienne, centrée sur 0.

      ii. Développer autant que possible l'expression mathématique de sa matrice de covariance.

      iii. Déduire les expressions des lois Gaussiennes des erreurs d'estimation sur a et sur b.

   (c) En déduire les intervalles de confiance recherchés.

Le paramètre b est désormais connu, de sorte que le vecteur des paramètres cachés s'écrit $\theta = a \in \mathbb{R}$.

10. Réécrire (2) sous la forme
$$Y = La + V, \tag{5}$$
où V est le vecteur aléatoire de bruit établi plus haut, et $Y = (Y_1, \ldots, Y_N)^T$ est un « nouveau » vecteur aléatoire de mesure dont on précisera le lien avec Z.

11. Écrire l'expression de l'estimé $\hat{a}$ du maximum de vraisemblance de a dans ce nouveau contexte.

12. L'intervalle de confiance centré sur l'estimé $\hat{a}$, et dans lequel le paramètre caché a se trouve avec une probabilité de 99.7%, est-il de même taille que dans le cas précédent ? plus grand ? plus petit ?