Sup Galilée
MACS1

Février 2022

Initiation au calcul numérique – TP Proba. 3

Lois à densité (simulation par inversion)

Les exercices marqués d'un astérisque sont à réaliser en priorité.

Avant de commencer ce TP, créer un sous-répertoire `TP_Proba_3`, et créer un script (vide) `TP.m` dans ce sous-répertoire.

L'objectif de ce TP est d'étudier plusieurs façons de **simuler des lois à densité** : méthode par inversion de la fonction de répartition, et méthode du rejet.

Afin de valider les fonctions écrites, on va d'abord étudier comment identifier la densité (si elle existe) de la loi suivie par des tirages aléatoires réels : on va pour cela utiliser des **histogrammes normalisés**. Au prochain TP, on verra une autre méthode, plus générale, pour estimer la loi de tirages aléatoires, basée sur la fonction de répartition.

Rappels de cours

Soit X une variable aléatoire à valeurs dans $\mathbb{R}$. On dit que X a une **densité** f si $f : \mathbb{R} \to \mathbb{R}$ est une fonction mesurable positive telle que

$$\text{pour tout } A \in \mathcal{B}(\mathbb{R}), \qquad P(X \in A) = \int_A f(x)dx.$$

Inversement, toute fonction $f : \mathbb{R} \to \mathbb{R}$ mesurable positive telle que $\int_{\mathbb{R}} f(x)dx = 1$ est la densité d'une loi de probabilité sur $\mathbb{R}$.

Dans tous les cas (que X ait une densité ou non), sa **fonction de répartition** est la fonction $F_X : t \mapsto F_X(t) = \mathbb{P}(X \le t)$.. Sa propriété principale est qu'elle caractérise la loi de X : si $F_X(t) = F_Y(t)$ pour tout $t \in \mathbb{R}$, alors X et Y ont même loi. Pour connaître la loi de X, il suffit donc en principe de calculer sa fonction de répartition.

Dans le cas où X a pour densité f continue par morceaux, on constate que $F_X(t) = \int_{-\infty}^t f(x)dx$ est la primitive de f qui a une limite nulle en $-\infty$.

1 Préliminaire : Représentation d'histogramme normalisé

On verra plus en détails dans le prochain TP comment estimer la densité f si elle existe, et plus généralement la loi de variables aléatoires $X_1, X_2, \dots$ indépendantes et de même loi qu'une variable aléatoire X, à partir d'une réalisation $X_1(\omega), \dots, X_N(\omega)$ (c'est-à-dire N tirages donnés). Pour le moment, on se contentera des **histogrammes**.

On suppose que f est continue. Alors, pour tout $x \in \mathbb{R}$,

$$f(x) = \lim_{h\to 0} \frac{1}{h} \int_{x-h/2}^{x+h/2} f(t)dt = \lim_{h\to 0} \frac{1}{h} P(X \in \Big[x - \frac{h}{2}, x + \frac{h}{2}\Big])$$

et, par la loi des grands nombres, pour tout $x \in \mathbb{R}$, pour tout $h > 0$, presque sûrement,

$$P(X \in \Big[x-\frac{h}{2}, x+\frac{h}{2}\Big]) = \lim_{N\to\infty} \frac{\mathbf{1}_{\{X_1 \in [x-\frac{h}{2}, x+\frac{h}{2}]\}} + \cdots + \mathbf{1}_{\{X_N \in [x-\frac{h}{2}, x+\frac{h}{2}]\}}}{N} = \lim_{N\to\infty} \frac{1}{N} \mathrm{Card}(\{1 \le i \le N \,|\, X_i \in \Big[x-\frac{h}{2}, x+\frac{h}{2}\Big]\}).$$

En combinant les deux, on en déduit que, pour tout $x \in \mathbb{R}$, à condition que "N soit assez grand et h assez petit",

$$f(x) \simeq \frac{1}{Nh} \mathrm{Card}(\{1 \le i \le N \,|\, X_i \in \Big[x - \frac{h}{2}, x + \frac{h}{2}\Big]\}).$$

On peut démontrer que $h = 1/\sqrt{N}$ est un choix qui assure la convergence ci-dessus quand $N \to \infty$ (en fait, au sens L^1).

Ceci justifie que l'on peut approcher f par un **histogramme normalisé** des réalisations de $X_1, \dots, X_N$: c'est une représentation graphique, pour une subdivision de $\mathbb{R}$ en intervalles appelés **classes** (ou corbeilles), du **nombre** de données dans chaque classe, **divisé par** Nh, où h est la largeur de la classe. Dans une représentation en barres de largeur h, cela assure que l'*aire* de chaque barre est égale à la *proportion* des données contenue dans la classe correspondante, et donc que la somme des aires des barres vaut 1 : cette représentation est une densité de probabilité.

Souvent, par facilité, on se donnera non pas la *largeur* h mais le *nombre* k de classes souhaité. Approximativement, on prendra $k \simeq \frac{1}{h} \simeq \sqrt{N}$ par exemple. Attention, cette approximation ne tient que si les données ne sont pas trop dispersées.

Dans Octave, on obtient un histogramme normalisé pour être une densité en utilisant `hist` puis `bar`. On récupère d'abord, par `hist`, le nombre d'éléments et le centre de chaque classe (vecteurs `nelem` et `centres` ci-dessous), puis on le normalise par Nh et on le représente graphiquement par `bar`. Dans le cas où se donne simplement le nombre de classes (`k`), on utilisera ainsi :

```
[nelem, centres]=hist(v,k);
N=numel(v);
h=centres(2)-centres(1);
bar(centres,nelem/(N*h),1);
```

Ou, si on précise directement la valeur de h (notée `h`) et l'intervalle $[\mathrm{a}, \mathrm{b}]$ où se trouvent les données :

```
centres=a:h:b;
nelem=hist(v,centres);
N=numel(v);
bar(centres,nelem/(N*h),1);
```

Exercice 1* : Exemples d'histogrammes

1. Écrire une fonction `hist_densite(vx,k)` qui représente un histogramme normalisé associé aux données contenues dans le vecteur `vx`, avec `k` classes. Vu le résultat précédent, on prendra en pratique $k \simeq \sqrt{N}$, sans quoi la convergence n'est pas assurée.
(Vu la double syntaxe de `hist`, la fonction obtenue pourra aussi s'utiliser sous la forme `hist_densite(vx,centres)` si on souhaite spécifier les centres des intervalles de l'histogramme plutôt que leur seul nombre)

2. (Dans `ex1.m`) Tester votre fonction en estimant, par un histogramme, la densité de $U+V$, où U et V sont indépendantes et de loi uniforme sur $[0,1]$. On utilisera $N = 10000$ tirages.

3. (Dans `ex1.m`) Essayer d'estimer la densité de la variable aléatoire $X = \max(U, 1/2)$, où U est une variable aléatoire de loi uniforme sur $[0,1]$. Expliquer le phénomène observé. *(Que vaut $P(X = 1/2)$?)*

2 Méthode de simulation par inversion de la fonction de répartition

Toute loi sur $\mathbb{R}$ peut, en principe, être simulée grâce à la proposition suivante :

Proposition 2.1. *Soit X une variable aléatoire réelle. On note F_X sa fonction de répartition, et F_X^{-1} sa réciproque généralisée :*

$$\text{pour tout } u \in]0,1[, \qquad F_X^{-1}(u) = \inf\{x \in \mathbb{R} \,|\, F_X(x) \geq u\}.$$

Si U suit la loi uniforme sur $[0,1]$, alors $F_X^{-1}(U)$ suit la même loi que X.

Exercice 2* : Justification de la méthode

Pour simplifier, on suppose que F_X est continue et strictement croissante sur l'intervalle ouvert $I =]\alpha, \beta[$, avec $\begin{cases} F_X(\alpha^+) = 0 \\ F_X(\beta^-) = 1. \end{cases}$
(et $\alpha, \beta \in \overline{\mathbb{R}}$). Par conséquent, $F_X :]\alpha, \beta[\to]0,1[$ est une bijection, et F_X^{-1} est sa fonction réciproque, au sens habituel.

1. Soit U une variable aléatoire de loi uniforme sur $[0,1]$. On pose $Y = F_X^{-1}(U)$. Calculer $F_Y(y) = P(Y \leq y)$ pour $y \in I$.

2. Conclure.

Exercice 3* : Exemples, pour des lois à densité

Dans chacun des cas suivants, on calculera la fonction réciproque de la fonction de répartition, pour utiliser la méthode précédente.

1. Écrire une fonction `x=simul_exponentielle(lambda)` qui simule la loi exponentielle de paramètre λ.
Rappel : densité $f(x) = \lambda e^{-\lambda x} \mathbf{1}_{[0,\infty[}(x)$, fonction de répartition $F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-\lambda x} & \text{si } x \geq 0. \end{cases}$

2. (Écrire dans `TP.m`) Produire un vecteur `tirage1` contenant $N = 10000$ réalisations de variables exponentielles de paramètre 2. Représenter un histogramme normalisé de `tirage1` et comparer à la fonction de densité attendue.

3. Écrire une fonction `x=simul_cauchy` (sans paramètre d'entrée) qui simule la loi de Cauchy.
Rappel : densité $f(x) = \frac{1}{\pi(1+x^2)}$, fonction de répartition $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$.

4. (Écrire dans `TP.m`) Produire un vecteur `tirage2` contenant $N = 10000$ réalisations de variables de loi de Cauchy. Représenter un histogramme normalisé de `tirage2` et comparer à la fonction de densité attendue. Il pourra être utile de préciser les centres (ou bornes) des classes attendues, par exemple `linspace(-10,10,100)` (paramètre `k` de `hist_densite`).

Exercice 4 : Un autre exemple de loi à densité

On considère la loi de densité $f : x \mapsto \frac{1}{2}\sin(x)\mathbf{1}_{[0,\pi]}(x)$.

1. Calculer sa fonction de répartition F.

2. En déduire une fonction `vx=simul_sin(N)` qui renvoit un vecteur contenant N réalisation de variables aléatoires de densité f. *On pourra utiliser la fonction* `acos` *d'Octave.*

3. Valider votre programme en comparant un histogramme normalisé de $N = 10000$ tirages avec la fonction f.