Chapitre I

Simulation de variables aléatoires

La simulation informatique de variables aléatoires, aussi complexes soient elles, repose sur la simulation de variables aléatoires indépendantes très simples, auxquelles sont appliquées des transformations adéquates. La variable aléatoire de base est celle de loi uniforme sur $[0, 1]$.

I.1 Générateur de nombres pseudo-aléatoires

Dans un code de calcul, un générateur de nombre aléatoire est une suite de réels $(u_1, \dots, u_m, \dots)$ déterministes en double précision, compris entre 0 et 1. Selon le langage de programmation et les librairies utilisées, l'appel au générateur passe par la commande `drand48` en C, `random` en C++ et Java, `rand` en Matlab, `grand` en Scilab avec différents paramètres d'appel. Nous le noterons simplement `rand` dans la suite.

Au premier appel du générateur, on obtient la valeur u_1, puis u_2 au second appel. En effet, la graine (*seed*) du générateur – c'est-à-dire l'indice du premier nombre renvoyé – est par défaut égale à 1. Cela n'a rien de très aléatoire car la suite appelée $(u_1, \dots, u_m, \dots)$ est toujours la même : c'est très pratique dans une première phase d'écriture de programme de simulation, où générer des résultats vraiment aléatoires rend difficile la recherche de *bug*. Dans un second temps, il est recommandé de changer la graine de manière *aléatoire*, en l'initialisant par exemple sur l'horloge de la machine (a priori toujours différente).

Le générateur est qualifié d'aléatoire si la suite $(u_1, \dots, u_m, \dots)$ – bien que déterministe – a tout d'un comportement aléatoire, similaire à une suite de variables aléatoires indépendantes de loi uniforme : pour s'en assurer, il existe une multitude de

tests statistiques permettant de rejeter l'hypothèse d'indépendance ou l'adéquation à une distribution donnée. Le lecteur intéressé pourra se référer à [AG07, Chapitre 2].

En pratique, un générateur est cyclique et après L appels, il redonne la valeur initiale etc... Il est évidemment important de s'assurer que la *période* L du générateur utilisé est suffisamment grande devant le nombre d'appels de celui-ci : en pratique, l'essentiel des générateurs actuellement disponibles satisfont cette contrainte.

Un exemple classique de générateur est le *générateur linéaire congruentiel* : il s'écrit pour trois paramètres a, b et L

$$x_{m+1} = ax_m + b \text{ modulo } L, \quad u_m = \frac{x_m}{L}$$

pour atteindre des périodes maximales égales à L. Pendant longtemps, un choix populaire a été $a = 7^5$, $b = 0$ et $L = 2^{31} - 1 = 2147483647$.

Le générateur *Mersenne Twister*[1] est un générateur plus récent, robuste et rapide, avec une période égale $2^{19937} - 1$, très largement suffisante pour bien des applications.

I.2 Simulation de variable aléatoire unidimensionnelle

Le lecteur pourra se référer au travail encyclopédique de Devroye [Dev86] pour prendre connaissance de la myriade d'algorithmes de simulation d'une variable aléatoire donnée. Notre présentation suit plutôt celle de [BM01] définissant directement les variables aléatoires via leur algorithme de génération à partir de variables aléatoires de loi uniforme.

I.2.1 Inversion de la fonction de répartition

Une première approche repose sur la méthode d'inversion de la fonction de répartition, méthode proposée par Von Neumann en 1947 [Eck87]. Dans toute la suite, U désigne une variable aléatoire de loi uniforme sur $[0, 1]$.

Proposition I.2.1 *Soit X une variable aléatoire réelle de fonction de répartition $F(x) = \mathbb{P}(X \leq x)$, dont l'inverse généralisé (appelé* quantile*) est défini par $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Alors*

$$F^{-1}(U) \overset{\text{loi}}{=} X.$$

Inversement, si F est continue, alors $F(X) \overset{\text{loi}}{=} \mathcal{U}([0, 1])$.

1. `http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html`

PREUVE :

On vérifie de manière standard que la fonction F est croissante, continue à droite et limitée à gauche. De façon analogue, le quantile F^{-1} est croissant, continu à gauche, limité à droite. De plus, on a les propriétés générales suivantes (faciles à justifier) :

a) $F(F^{-1}(u)) \geq u$ pour tout $u \in]0,1[$.

b) $F^{-1}(u) \leq x \Longleftrightarrow u \leq F(x)$.

c) $F(F^{-1}(u)) = u$ si F continue en $F^{-1}(u)$.

Alors de b), on déduit $\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$ ce qui justifie la première assertion. Sous hypothèse de continuité de F, appliquer c) donne $F(X) \stackrel{\text{loi}}{=} F(F^{-1}(U)) = U$, ce qui prouve la seconde affirmation. □

Définition et proposition I.2.2 (loi exponentielle) *Soit $\lambda > 0$. Alors*

$$X = -\frac{1}{\lambda}\log(U)$$

a la loi d'une variable de loi exponentielle de paramètre λ (notée $\mathcal{E}xp(\lambda)$), dont la densité est $\lambda e^{-\lambda x}\mathbf{1}_{x\geq 0}$.

Définition et proposition I.2.3 (loi discrète) *Soit $(p_n)_{n\geq 0}$ une suite de réels strictement positifs satisfaisant $\sum_{n\geq 0} p_n = 1$ et $(x_n)_{n\geq 0}$ une suite de réels. Alors*

$$X = \sum_{n\geq 0} x_n \mathbf{1}_{p_0+\cdots+p_{n-1}\leq U<p_0+\cdots+p_n}$$

est une variable aléatoire discrète telle que $\mathbb{P}(X = x_n) = p_n$ pour $n \geq 0$.

Quelques exemples simples.

- La variable aléatoire de Bernoulli $\mathcal{B}(p)$ correspond au cas $(p_0, p_1) = (1-p, p)$ et $(x_0, x_1) = (0, 1)$.
- La variable aléatoire de loi binomiale $\mathcal{B}in(n,p)$ s'écrit comme une somme de n variables aléatoires de Bernoulli $\mathcal{B}(p)$ indépendantes : $\mathbb{P}(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}$ pour $1 \leq k \leq n$. La loi multinomiale est une généralisation.
- La loi de Poisson $\mathcal{P}(\theta)$ correspond à $x_n = n$ et $p_n = e^{-\theta}\frac{\theta^n}{n!}$ pour $n \geq 0$.

La loi géométrique est aussi une loi discrète mais elle peut se simuler plus simplement.

Définition et proposition I.2.4 (loi géométrique) *Soit $(X_m)_{m\geq 1}$ une suite i.i.d. de loi $\mathcal{B}(p)$. La variable aléatoire $X = \inf\{m \geq 1 : X_m = 1\}$ suit la loi géométrique de paramètre p (notée $\mathcal{G}(p)$) : $\mathbb{P}(X = n) = p(1-p)^{n-1}$ pour $n \geq 1$. On a aussi*

$$X \stackrel{\text{loi}}{=} 1 + \lfloor Y \rfloor \quad \textit{où} \quad Y \stackrel{\text{loi}}{=} \mathcal{E}xp(\lambda)$$

avec $\lambda = -\log(1-p)$*,et par conséquent*

$$1 + \Big\lfloor \frac{\log(U)}{\log(1-p)} \Big\rfloor \overset{\text{loi}}{=} \mathcal{G}(p).$$

Définition et proposition I.2.5 (loi de Cauchy) *Soit* $\sigma > 0$*. Alors*

$$X = \sigma \tan\big(\pi(U - \frac{1}{2})\big)$$

est une variable aléatoire de Cauchy de paramètre σ*, dont la densité est* $\frac{\sigma}{\pi(x^2+\sigma^2)}\mathbf{1}_{x\in\mathbb{R}}$.

Définition et proposition I.2.6 (loi de Rayleigh) *Soit* $\sigma > 0$*. Alors*

$$X = \sigma\sqrt{-\log U}$$

est une variable aléatoire de Rayleigh de paramètre σ*, dont la densité est* $\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}\mathbf{1}_{x\geq 0}$.

Définition et proposition I.2.7 (loi de Pareto) *Soit* $(a,b) \in \mathbb{R}\times]0,+\infty[$*. Alors*

$$X = \frac{b}{U^{\frac{1}{a}}}$$

est une variable aléatoire de Pareto de paramètres (a,b)*, dont la densité est* $\frac{ab^a}{x^{a+1}}\mathbf{1}_{x\geq b}$.

Définition et proposition I.2.8 (loi de Weibull) *Soit* $(a,b) \in]0,+\infty[^2$*. Alors*

$$X = b(-\log U)^{\frac{1}{a}}$$

est une variable aléatoire de Weibull de paramètres (a,b)*, dont la densité est* $\frac{a}{b^a}x^{a-1}e^{-(x/b)^a}\mathbf{1}_{x\geq 0}$.

Concernant la variable aléatoire gaussienne, il n'y a pas d'expression explicite pour sa fonction de répartition et son inverse ; néanmoins, d'excellentes approximations existent et permettent d'utiliser cette méthode d'inversion de manière approximative.

Définition et proposition I.2.9 (loi gaussienne, d'après [Mor95])
Définissons la fonction $u \in]0,1[\mapsto \mathcal{N}^{-1}_{\text{Moro}}(u)$ *par*

$$\mathcal{N}^{-1}_{\text{Moro}}(u) = \begin{cases} \dfrac{\sum_{n=0}^{3} a_n (u-0.5)^{2n+1}}{1+\sum_{n=0}^{3} b_n (u-0.5)^{2n}}, & 0.5 \leq u \leq 0.92, \\ \displaystyle\sum_{n=0}^{8} c_n (\log(-\log(1-u)))^n, & 0.92 \leq u < 1, \\ -\mathcal{N}^{-1}_{\text{Moro}}(1-u), & 0 < u \leq 0.5 \end{cases}$$

avec

$$\begin{aligned} a_0 &= \ \ 2.50662823884, & a_1 &= -18.61500062529, & a_2 &= \ \ 41.39119773534, \\ a_3 &= -25.44106049637, & b_0 &= \ -8.47351093090, & b_1 &= \ \ 23.08336743743, \\ b_2 &= -21.06224101826, & b_3 &= \ \ 3.13082909833, \end{aligned}$$

$c_0 = 0.3374754822726147$, $c_1 = 0.9761690190917186$, $c_2 = 0.1607979714918209$,
$c_3 = 0.0276438810333863$, $c_4 = 0.0038405729373609$, $c_5 = 0.0003951896511919$,
$c_6 = 0.0000321767881768$, $c_7 = 0.0000002888167364$, $c_8 = 0.0000003960315187$.

Alors $\mathcal{N}^{-1}_{\text{Moro}}$ *est une approximation*[2] *de l'inverse de* $\mathcal{N}$*, où* $\mathcal{N}$ *est la fonction de répartition de la loi gaussiennce centrée réduite :* $\mathcal{N}(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.
Ainsi, pour $\mu \in \mathbb{R}$ *et* $\sigma \geq 0$,

$$X = \mu + \sigma \mathcal{N}^{-1}_{\text{Moro}}(U)$$

suit approximativement la loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ *de moyenne* μ *et variance* σ^2.

I.2.2 Variable gaussienne

Il est possible de générer une variable aléatoire gaussienne directement, sans approximation, à l'aide de la transformation de Box-Muller. Elle s'appuie sur le résultat suivant.

Proposition I.2.10 *Soient* X *et* Y *deux variables aléatoires gaussiennes centrées réduites indépendantes. Définissons* (R, θ) *les coordonnées polaires de* (X, Y) *:*

$$X = R\cos(\theta), \quad Y = R\sin(\theta)$$

avec $R \geq 0$ *et* $\theta \in [0, 2\pi[$*. Alors* R^2 *et* θ *sont deux variables aléatoires indépendantes, la première est de loi* $\mathcal{E}xp(\frac{1}{2})$*, la seconde de loi uniforme sur* $[0, 2\pi]$.

```
r, theta : double ;
u, v : double ;
x, y : double ;
u ← rand ;
v ← rand ;
theta ← 2πu ;
r ← sqrt(-2 log(v)) ;
x ← r cos(theta) ;
y ← r sin(theta) ;
```

ALGORITHME I.1: simulation de deux variables aléatoires gaussiennes centrées réduites indépendantes par la méthode de Box-Muller

2. l'erreur est inférieure à 3×10^{-9} jusqu'à 7 écart-types (i.e. $\mathcal{N}(-7) \leq u \leq \mathcal{N}(7)$).

Pour obtenir une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$, il reste à multiplier x par l'écart-type σ et ajouter la moyenne μ.

L'algorithme de Marsiglia est une variante de simulation, qui évite le calcul de fonctions trigonométriques qui sont considérées coûteuses en temps calcul, voir [AG07].

I.3 Méthode de rejet

I.3.1 Simulation de loi conditionnelle

Pour obtenir une simulation de Z sachant un évènement A, il suffit de simuler de manière répétée et indépendante (Z, A), de rejeter les résultats tant que A n'est pas réalisé. Dans ce résultat, Z peut être une variable aléatoire multi-dimensionnelle.

Proposition I.3.1 *Soient Z une variable aléatoire et A un événement de probabilité non nul. Considérons $(Z_n, A_n)_{n\geq 1}$ la suite d'éléments aléatoires indépendants de même loi que (Z, A) Notons $\nu = \inf\{n \geq 1 : \ A_n$ est réalisé$\}$: alors, la v.a. Z_ν a la loi conditionnelle de Z sachant A.*

Preuve :

Pour tout borélien B, on a

$$\begin{aligned}\mathbb{P}(Z_\nu \in B) &= \sum_{n\geq 1} \mathbb{P}(Z_n \in B; A_1^c; \cdots; A_{n-1}^c; A_n) \\ &= \sum_{n\geq 1} (1-\mathbb{P}(A))^{n-1}\, \mathbb{P}(Z_n \in B; A_n) = \frac{\mathbb{P}(Z \in B; A)}{\mathbb{P}(A)} = \mathbb{P}(Z \in B|A).\end{aligned}$$

□

L'algorithme précédent est de durée aléatoire ν : cette dernière variable aléatoire est de loi $\mathcal{G}(\mathbb{P}(A))$. Ainsi, plus A est probable, plus la simulation est rapide (d'espérance $\frac{1}{\mathbb{P}(A)}$).

Donnons un exemple simple d'application de ce résultat : pour simuler une variable aléatoire X de loi uniforme sur un ensemble compact $D \subset \mathbb{R}^d$, il suffit de simuler des variables aléatoires Z de loi uniforme sur un cube contenant D (facile à faire), puis de retenir la première simulation qui tombe dans D. Elle aura en effet la loi de $Z|\{Z \in D\}$, dont la densité est $z \mapsto \frac{\mathbf{1}_{z\in \text{cube}}}{|\text{cube}|} \frac{\mathbf{1}_{z\in D}}{\mathbb{P}(Z\in D)} = \frac{\mathbf{1}_{z\in D}}{|\text{cube}|\mathbb{P}(Z\in D)} = \frac{\mathbf{1}_{z\in D}}{|D|}$, c'est-à-dire celle de X.

I.3.2 Simulation de loi (non conditionnelle) par méthode de rejet

Ici, nous supposons que la loi de la variable aléatoire X d'intérêt (éventuellement multidimensionnelle) possède une densité f connue, mais dont la simulation directe

n'est pas facile. Le principe de la méthode consiste à simuler une autre variable aléatoire Y de densité g et d'accepter le résultat de Y comme réalisation de X avec une probabilité proportionnelle au rapport $f(Y)/g(Y)$. Cette idée remonte à Von Neumann en 1947 [Eck87]. La propriété s'énonce précisément ainsi.

Proposition I.3.2 *Soient X et Y deux variables aléatoires à valeurs dans $\mathbb{R}^d$, dont les densités par rapport à une mesure de référence μ sont respectivement f et g. Supposons qu'il existe une constante $c(\geq 1)$ satisfaisant*

$$c\, g(x) \geq f(x) \quad \mu - p.p. \tag{I.3.1}$$

Soit U une variable aléatoire de loi uniforme sur $[0,1]$ indépendante de Y. Alors, la loi de Y sachant $\{c\, U\, g(Y) < f(Y)\}$ est la loi de X.

Preuve :

En effet, en posant $A = \{c\, U\, g(Y) < f(Y)\}$, pour tout B borélien de $\mathbb{R}^d$, on a

$$\begin{aligned}
\mathbb{P}(Y \in B \mid A) &= \frac{\mathbb{P}(Y \in B; A)}{\mathbb{P}(A)} \\
&= \frac{1}{\mathbb{P}(A)} \int_{\{(y,u): y\in B, c\ u\ g(y)<f(y)\}} g(y)\ \mathbf{1}_{g(y)>0}\ \mathbf{1}_{[0,1]}(u)\ \mu(dy)\ du \\
&= \frac{1}{\mathbb{P}(A)} \int_B g(y)\ \frac{f(y)}{c\, g(y)}\ \mathbf{1}_{g(y)>0}\ \mu(dy) \\
&= \frac{1}{c\, \mathbb{P}(A)} \int_B f(y)\ \mathbf{1}_{g(y)>0}\ \mu(dy) = \frac{1}{c\, \mathbb{P}(A)} \int_B f(y)\ \mu(dy)
\end{aligned}$$

car μ − p.p., si $g(y) = 0$ alors $f(y) = 0$. Le choix $B = \mathbb{R}^d$ conduit à $c\, \mathbb{P}(A) = 1$, et donc $\mathbb{P}(Y \in B \mid A) = \int_B f(y)\ \mu(dy)$ pour tout B. □

Ensuite, pour la simulation effective de la loi conditionnelle ci-dessus, on applique la Proposition I.3.1 et cela conduit à l'algorithme suivant :

```
c : double ;
y : double ;
u : double ;
c ← majorant de f/g ;
Repeat
    | y ← simulation de densité g ;
    | u ← rand ;
until (c u g(y) ≤ f(y))
return y ;    (la variable y en sortie a la loi de X de densité f)
```

Algorithme I.2: méthode de rejet

Lors de la mise en pratique de la méthode de rejet, il est relativement facile, pour une densité f donnée, de trouver une densité g et un nombre c vérifiant $c\,g(x) \geq f(x)$ pour tout x. Néanmoins, les choix de g sont satisfaisants si la constante c est petite, de sorte que le nombre moyen de rejets reste faible (en moyenne, $c = \frac{1}{\mathbb{P}(A)}$ appels) et que l'algorithme soit rapide.

Exemple I.3.3 (simulation de loi Beta) Une variable aléatoire de loi Beta $\mathcal{B}(\alpha,\beta)$ (avec $\alpha > 0$ et $\beta > 0$) a pour densité $\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbf{1}_{0<x<1}$. Supposons que $\alpha \geq 1$ et $\beta \geq 1$ de sorte que cette densité soit bornée. On peut alors utiliser la méthode de rejet avec $Y \overset{\text{loi}}{=} \mathcal{U}([0,1])$. La constante de rejet vaut

$$c_\alpha = \sup_{0<x<1} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} = \frac{1}{B(\alpha,\beta)} (x_{\alpha,\beta})^{\alpha-1}(1-x_\alpha,\beta)^{\beta-1}$$

où $x_{\alpha,\beta} = \frac{\alpha-1}{\alpha-1+\beta-1}$.

Exemple I.3.4 (simulation de loi Gamma) Une variable aléatoire de loi Gamma $\Gamma(\alpha,\theta)$ (avec $\alpha > 0$ et $\theta > 0$) a pour densité $\frac{1}{\Gamma(\alpha)} \theta^\alpha x^{\alpha-1} e^{-\theta x}\mathbf{1}_{x\geq 0}$.

- Si $\alpha = 1$, cela coïncide avec une variable aléatoire de loi $\mathcal{E}xp(\theta)$.
- Si α est un entier non nul, une variable aléatoire de loi $\Gamma(\alpha,\theta)$ s'écrit comme la somme de α variables aléatoires indépendantes de loi $\mathcal{E}xp(\theta)$: la simulation en découle immédiatement.
- Si α n'est pas entier, il est utile d'avoir recours à la méthode de rejet. Donnons-en une illustration, sans avoir le soucis d'optimalité. Pour simplifier, supposons $\theta = 1$ et $\alpha \in (n, n+1)$ avec $n \geq 1$. Prenons pour Y une loi $\Gamma(n, \frac{1}{2})$. On vérifie alors que la constante de rejet vaut

$$\begin{aligned} c_\alpha = \sup_{x>0} \frac{f(x)}{g(x)} &= \sup_{x>0} \frac{\frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}}{\frac{1}{\Gamma(n)} 2^{-n} x^{n-1} e^{-\frac{1}{2}x}} \\ &= \frac{\Gamma(n)}{\Gamma(\alpha)} 2^\alpha \sup_{y>0} y^{\alpha-n} e^{-y} = \frac{\Gamma(n)}{\Gamma(\alpha)} 2^\alpha (\alpha-n)^{\alpha-n} e^{-(\alpha-n)}. \end{aligned}$$

Cette constante augmente rapidement lorsque $\alpha \to +\infty$. Pour de meilleures procédures, voir [Dev86, Chapitre 9].

I.4 Simulation d'un vecteur aléatoire

Lorsque les composantes d'un vecteur aléatoire sont indépendantes, on est ramené au cas unidimensionnel sur chaque composante. Dans le cas de dépendance, une analyse plus poussée est nécessaire pour simuler le vecteur aléatoire.

I.4.1 Le cas de vecteur gaussien

Nous rappelons qu'un vecteur $X = (X_1, \dots, X_d)$ est dit gaussien si toute combinaison linéaire de ses composantes $\sum_{i=1}^d a_i X_i$ (avec $a_i \in \mathbb{R}$) a la loi gaussienne. Un vecteur gaussien X est caractérisé par sa moyenne m et sa matrice de covariance K, on note $X \overset{\text{loi}}{=} \mathcal{N}(m, K)$.

De manière générale, un vecteur gaussien se simule par transformation affine de variables aléatoires gaussiennes centrées réduites indépendantes (i.e. de loi $\mathcal{N}(0, \mathrm{Id})$).

Proposition I.4.1 *Soient d_0 et d deux entiers non nuls, X un vecteur gaussien d_0-dimensionnel de loi $\mathcal{N}(0, \mathrm{Id})$, $m \in \mathbb{R}^d$ et L une matrice de taille $d \times d_0$. Alors*

$$m + LX \overset{\text{loi}}{=} \mathcal{N}(m, LL^\perp),$$

i.e. $m + LX$ est un vecteur gaussien d-dimensionnel de moyenne m et de covariance $K = LL^\perp$.

Nous laissons la preuve au lecteur. Réciproquement, une matrice de covariance K – matrice symétrique positive de taille d – peut toujours se décomposer – de manière non unique – sous la forme

$$K = LL^\perp,$$

permettant ainsi de simuler tout vecteur gaussien en se ramenant au cas précédent. Pour calculer L, on peut utiliser l'algorithme de Choleski, qui fournit une matrice triangulaire inférieure (avec $d_0 = d$). Son coût calcul est d'ordre d^3 par rapport à la dimension.

Dans le cas de grande dimension, on peut chercher à accélérer la simulation. C'est possible pour certaines matrices. Par exemple, si

$$K = \begin{pmatrix} 1 & \rho & \cdots & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \rho & \cdots & \rho & 1 & \rho \\ \rho & \cdots & \cdots & \rho & 1 \end{pmatrix}$$

pour $\rho \in [0, 1]$, on peut prendre la matrice de taille $d \times d_0$ (avec $d_0 = d + 1$) suivante

$$L = \begin{pmatrix} \sqrt{\rho} & \sqrt{1-\rho} & 0 & \cdots & \cdots & 0 \\ \sqrt{\rho} & 0 & \sqrt{1-\rho} & 0 & \cdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \sqrt{1-\rho} & 0 \\ \sqrt{\rho} & 0 & \cdots & \cdots & 0 & \sqrt{1-\rho} \end{pmatrix}.$$

Dans ce cas, on utilise $d+1$ variables aléatoires gaussiennes centrées réduites indépendantes pour générer les d variables aléatoires gaussiennes avec la bonne covariance. Le coût calcul est de l'ordre de d au lieu de d^3 avec la méthode de Choleski usuelle, produisant une amélioration importante en grande dimension.

I.4.2 Modélisation de dépendance par les copules

Lorsque les variables sont de loi gaussienne, il est naturel de modéliser la dépendance par une matrice de covariance. Toutefois, les courbes de niveau de la densité gaussienne sont nécessairement des ellipses et peuvent ne pas bien rendre compte des dépendances dans les valeurs extrêmes. La modélisation de la dépendance est une question délicate, complexe et essentielle dans les applications. Cela ne résume pas à un coefficient de corrélation comme dans le cas de vecteur gaussien.

En fait, la dépendance peut être modélisée intrinsèquement sans prendre en compte les lois marginales, par la notion de *copule*. C'est une pure mesure de la dépendance, dont le fondement s'appuie sur le théorème de Sklar [Skl59].

Théorème I.4.2 *Considérons un vecteur aléatoire $X = (X_1, \cdots, X_d)$ d-dimensionnel de fonction de répartition jointe $F(x_1, \cdots, x_d) = \mathbb{P}(X_1 \leq x_1, \cdots, X_d \leq x_d)$. Alors il existe une fonction copule $C : [0,1]^d \mapsto [0,1]$ telle que :*

$$F(x_1; \cdots, x_d) = C(F_1(x_1); \cdots F_d(x_d)).$$

La copule C est unique lorsque les marginales sont continues.

On renvoie à [MFE05, Chapitre 5] pour des propriétés détaillées des fonctions copules. On peut facilement vérifier que la copule est invariante par transformation strictement croissante du vecteur aléatoire initial X, réaffirmant ainsi que c'est une mesure intrinsèque de la dépendance des composantes de X. Ce point de vue sépare la modélisation de la loi de X en la modélisation de chaque loi marginale d'une part, et la modélisation de leur dépendance d'autre part.

Donnons quelques exemples usuels de copule.

1. *Copule indépendante* : c'est la copule d'un vecteur avec composante indépendante, c'est-à-dire $C(u_1, \ldots, u_d) = u_1 \ldots u_d$.
2. *Copule co-monotone* : cela correspond au cas $X_i = \phi_i(Y)$ avec ϕ_i croissante, donnant la copule $C^+(u_1, \cdots, u_d) = \min(u_1, \cdots, u_d)$.
3. *Bornes de Fréchet-Hoeffding* : les copules sont universellement encadrées ainsi

$$(u_1 + \cdots + u_d - d + 1)_+ := C^-(u_1, \cdots, u_d) \leq C(u_1, \cdots, u_d) \leq C^+(u_1, \cdots, u_d).$$

4. *Copule gaussienne de matrice K inversible* : c'est la copule d'un vecteur gaussien $\mathcal{N}(0, K)$, c'est-à-dire

$$C(u_1, \cdots, u_d) = \int_{-\infty}^{\mathcal{N}^{-1}(u_1)} \cdots \int_{-\infty}^{\mathcal{N}^{-1}(u_d)} \frac{1}{(2\pi)^{d/2}\sqrt{\det(K)}} \exp\big(-\frac{x \cdot K^{-1}x}{2}\big) dx.$$

5. *Copule archimédienne* : elle prend la forme

$$C(u_1, \cdots, u_d) = \phi^{-1}(\phi(u_1) + \cdots + \phi(u_d))$$

où ϕ^{-1} est la transformée de Laplace d'une variable aléatoire Y positive non nulle, c'est-à-dire $\phi^{-1}(u) = \mathbb{E}(e^{-uY})$.

Simulations. Nous cherchons à générer un vecteur $(X_1, \cdots, X_d)$ de copule C et de marginales $F_1, \cdots, F_d$ données. Il suffit de

1. simuler des variables aléatoires $(U_1, \cdots, U_d)$ de marges uniformes et de copule C ;
2. puis calculer $X_i = F_i^{-1}(U_i)$.

Pour simuler $(U_1, \cdots, U_d)$, on peut

1. simuler des v.a. $(Y_1, \cdots, Y_d)$ de marges arbitraires, continues, et de copule C ;
2. puis calculer $U_i = F_{Y_i}(Y_i)$.

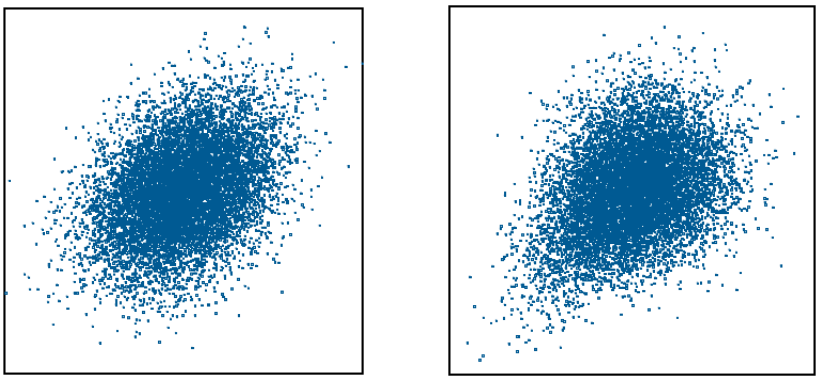

FIGURE I.1 – Deux échantillons de variables aléatoires en dimension 2, dont chaque marge suit la loi $\mathcal{N}(0, 1)$. A gauche : copule gaussienne (vecteur gaussien) avec corrélation 33%. A droite : copule de Clayton (copule archimédienne avec Y de loi $\mathcal{E}xp(1)$). Echantillon de taille 10000.

La séparation dépendance/marginale permet de générer des vecteurs aléatoires ayant une dépendance de type copule gaussienne avec une marginale de loi exponentielle,

une autre de loi de Cauchy, etc... La Figure I.1 montre deux échantillons de vecteur bidimensionnel de marginale gaussienne centrée réduite, avec une copule gaussienne d'un coté et une copule archimédienne de l'autre (Y de loi exponentielle) : la corrélation de la copule gaussienne est telle que les deux séries ont même corrélation empirique (montrant néanmoins des dépendances différentes).

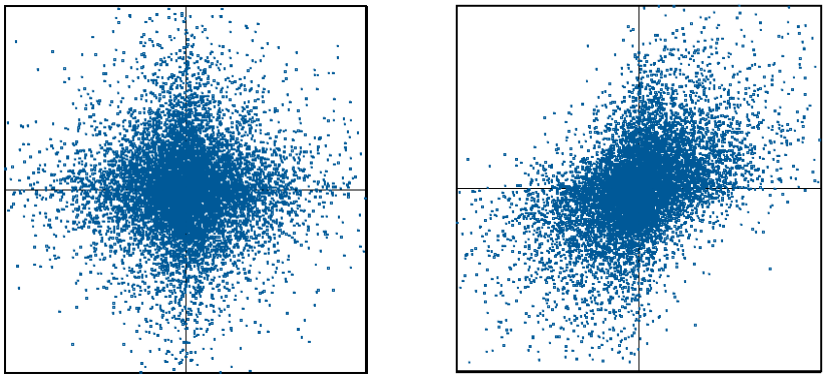

FIGURE I.2 – Deux échantillons de variables aléatoires en dimension 2, dont chaque marge suit la loi exponentielle avec signe randomisé. A gauche : composante indépendante. A droite : copule gaussienne avec corrélation 50%. Echantillon de taille 10000.

Sur la Figure I.2, chaque marge suit la loi exponentielle de paramètre 1 avec signe randomisé (c'est-à-dire obtenue par εX où $\varepsilon = \pm 1$ avec équiprobabilité et $X \overset{\text{loi}}{=} \mathcal{E}xp(1)$), avec soit des composantes indépendantes, soit une copule gaussienne avec corrélation 50%. Ces exemples montrent la variété de distributions possibles.

Mentionnons enfin que la dépendance archimédienne admet un algorithme de simulation ad hoc, voir [MO88].

Proposition I.4.3 (simulation avec copule archimédienne) *Soit C la copule archimédienne associée à la variable aléatoire Y (dont la transformée de Laplace est ϕ^{-1}), supposons $Y > 0$ p.s.. Soient $(X_i)_{1\leq i\leq d}$ des variables aléatoires indépendantes de loi uniforme sur $[0,1]$ et Y une variable aléatoire indépendante des $(X_i)_i$. Posons*

$$U_i = \phi^{-1}\Big(-\frac{1}{Y}\log(X_i)\Big).$$

Alors le vecteur $(U_1,\dots,U_d)$ a des marges uniformes et a C pour copule.

Pour en savoir plus, voir [MFE05].