Simulation d'un DAG à variables de Bernoulli

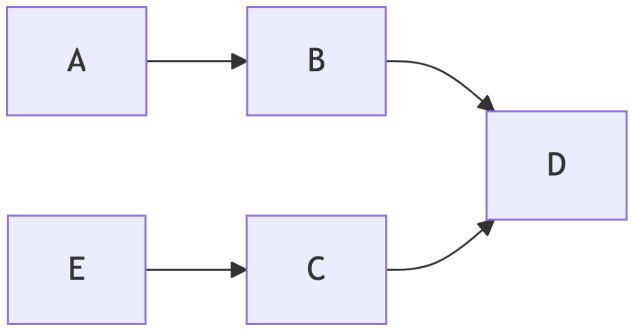
Table of contents

Modèle 1
Simulation 2
Contrôle des probabilités conditionnelles 3

Modèle

On considère le DAG $A \to B \to D \leftarrow C \leftarrow E$ sur cinq variables binaires.

```
flowchart LR
  A --> B --> D
  E --> C --> D
```



La loi jointe se factorise selon le graphe :

$$P(A, B, C, D, E) = P(A)\,P(E)\,P(B \mid A)\,P(C \mid E)\,P(D \mid B, C).$$

Le modèle est entièrement spécifié par $1 + 1 + 2 + 2 + 4 = 10$ probabilités conditionnelles $P(X = 1 \mid \mathrm{pa}(X))$:

Paramètre	Définition
`pA`	$P(A = 1)$
`pE`	$P(E = 1)$
`pB[a+1]`	$P(B = 1 \mid A = a)$, $a \in \{0, 1\}$
`pC[e+1]`	$P(C = 1 \mid E = e)$, $e \in \{0, 1\}$
`pD[b+1, c+1]`	$P(D = 1 \mid B = b, C = c)$, $b, c \in \{0, 1\}$

Simulation

Les variables sont tirées selon un ordre topologique ; chaque tirage conditionnel est vectorisé par indexation des tables.

```r
simulate_dag <- function(n, pA, pE, pB, pC, pD, seed = NULL) {
  stopifnot(length(pB) == 2, length(pC) == 2, all(dim(pD) == c(2, 2)))
  if (!is.null(seed)) set.seed(seed)
  A <- rbinom(n, 1, pA)
  E <- rbinom(n, 1, pE)
  B <- rbinom(n, 1, pB[A + 1])
  C <- rbinom(n, 1, pC[E + 1])
  D <- rbinom(n, 1, pD[cbind(B + 1, C + 1)])
  data.frame(A, B, C, D, E)
}
```

```r
pA <- 0.5
pE <- 0.3
pB <- c(0.2, 0.8)
pC <- c(0.3, 0.7)
pD <- matrix(c(0.1, 0.5,   # B=0 : C=0, C=1
               0.5, 0.9),  # B=1 : C=0, C=1
             nrow = 2, byrow = TRUE,
             dimnames = list(B = 0:1, C = 0:1))

df <- simulate_dag(n = 1e4, pA, pE, pB, pC, pD, seed = 1)
head(df)
```

```
  A B C D E
1 0 0 0 0 0
2 0 0 0 0 0
3 1 1 1 1 1
4 1 1 1 1 0
```

```
5 0 0 0 0 0
6 1 1 1 1 0
```

Contrôle des probabilités conditionnelles

Les fréquences empiriques doivent approcher les paramètres, à une erreur d'ordre $1/\sqrt{n}$ près.

```
c(A = mean(df$A), E = mean(df$E))
```

```
     A      E
0.4953 0.2985
```

```
tapply(df$B, df$A, mean)   # vs pB
```

```
        0         1
0.1963543 0.8079952
```

```
tapply(df$C, df$E, mean)   # vs pC
```

```
        0         1
0.3057733 0.7102178
```

```
with(df, tapply(D, list(B = B, C = C), mean))   # vs pD
```

```
   C
B           0         1
  0 0.1012570 0.5179655
  1 0.5074887 0.9033930
```