

# Tests semi-paramétriques d'indépendance

Bernard Colin et Ernest Monga  
Département de Mathématiques  
Université de Sherbrooke  
Sherbrooke J1K-2R1  
(Québec) Canada

Rapport de recherche N° 139

## Abstract

Let  $\mu$  be a probability measure on  $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$  absolutely continuous with respect to a reference measure. We show that the random variable  $Z = \otimes_{i=1}^n \mu_i(\cdot)$ , where the  $\mu_i$   $i = 1, 2, \dots, n$  are the marginal distributions of  $\mu$ , has a non parametric law and we use this fact to construct a test of independence between the components of the vector of law  $\mu$ . Thereafter, one extends, by the means of an equivalence relation between negative orthants, this test of independence to the case of random vectors. Some examples and simulations show how to conduct this test and illustrate its power. Finally, the natural link between this test and a brownian bridge allows an easy calculation of its asymptotic power.

**Keywords :** Independence, distribution-free statistic, copula, adequation test, negative orthant, brownian bridge, asymptotic power.

## Résumé

Étant donnée une mesure de probabilité  $\mu$  définie sur  $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ , absolument continue par rapport à une mesure de référence, on considère la variable aléatoire  $Z = \otimes_{i=1}^n \mu_i(\cdot)$  où les  $\mu_i$   $i = 1, 2, \dots, n$  désignent les mesures de probabilité marginales déduites de  $\mu$ . On montre alors que la statistique  $Z$  suit une loi non paramétrique et l'on utilise cette propriété afin de construire un test d'indépendance entre les composantes du vecteur aléatoire de loi  $\mu$ . Par la suite, on étend, à l'aide d'une relation d'équivalence entre orthants négatifs, ce test d'indépendance au cas de vecteurs aléatoires. Quelques exemples et simulations illustrent la procédure de test utilisé ainsi que sa puissance. Enfin, le fait que le test d'indépendance proposé puisse se présenter sous la forme d'un pont brownien, permet d'en déduire aisément le comportement asymptotique de sa puissance.

**Mots-clés :** Indépendance, statistique libre, copule, test d'ajustement, orthant négatif, pont brownien, puissance asymptotique.

## 1 Introduction

Soit  $X$  un vecteur aléatoire de  $\mathbb{R}^n$  de composantes  $X_1, X_2, \dots, X_n$ . La validité des applications d'un certain nombre de modèles issus de la théorie statistique classique, repose en grande partie sur l'hypothèse d'indépendance des composantes du vecteur aléatoire  $X$ . Dans ce contexte on dispose, dans la littérature actuelle, de tests permettant, par exemple, d'accepter ou de rejeter cette hypothèse dans les cas suivants :

- a) Cas de la loi normale :
  - i) Test sur le coefficient de corrélation.
  - ii) Test du rapport de vraisemblance.
- b) Cas d'une loi quelconque :
  - i) Test de permutation, test de rang et test du "tau" de Kendall.
  - ii) Test asymptotique, test de rang.

On notera que les tests ci-dessus mentionnés, se situent généralement dans un cadre paramétrique restreint.

Cette circonstance nous a donc incités à rechercher un test d'indépendance des composantes d'un vecteur aléatoire, ou plus généralement d'une famille de vecteurs aléatoires qui soit facile d'usage et qui, s'inspirant du cas unidimensionnel (pour lequel avec les notations habituelles on a :  $F_X(X) \sim \mathcal{U}_{[0,1]}$ ), soit libre.

## 2 Résultats généraux

### 2.1 Énoncé du problème

Soit  $X$  un vecteur aléatoire de  $\mathbb{R}^n$  et soit  $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \mathbb{P}, \mu)$  l'espace probabilisé associé où  $\mu$  désigne une mesure de référence (en général la mesure de Lebesgue) telle que  $\mathbb{P}$  soit absolument continue par rapport à  $\mu$ . On notera respectivement par  $F(x_1, x_2, \dots, x_n)$  (ou par  $F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ ) et par  $F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_n}(x_n)$  les fonctions de répartition associées à la mesure de probabilité  $\mathbb{P}$  et aux mesures de probabilité marginales  $\mathbb{P}_i$   $i = 1, 2, \dots, n$ .

On se propose alors de construire un test afin de confronter les hypothèses nulle  $\mathcal{H}_0$  et alternative  $\mathcal{H}_1$  suivantes :

$$\begin{aligned} \mathcal{H}_0 & : \mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i \\ \mathcal{H}_1 & : \mathbb{P} \neq \otimes_{i=1}^n \mathbb{P}_i \end{aligned}$$

ce qui peut également s'écrire sous la forme :

$$\begin{aligned} \mathcal{H}_0 & : F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i) \quad \forall (x_1, x_2, \dots, x_n) \in \mathbb{R}^n \\ \mathcal{H}_1 & : F(x_1, x_2, \dots, x_n) \neq \prod_{i=1}^n F_{X_i}(x_i) \end{aligned}$$

**Remarque 1 :** On notera que les expressions précédentes ne sont évidemment pas les seules que l'on puisse utiliser afin de caractériser l'indépendance des composantes  $X_1, X_2, \dots, X_n$  du vecteur aléatoire  $X$ . De fait, tout système de pavés mesurables engendrant  $\mathcal{B}_{\mathbb{R}^n}$ , peut convenir à cette fin et nous utiliserons cette propriété afin de préciser, dans le cadre des tests d'hypothèses, une alternative particulière à l'hypothèse d'indépendance.

### 2.2 Conditions nécessaires et suffisantes d'indépendance d'une famille finie de variables aléatoires

Les résultats qui suivent constituent une caractérisation intéressante de l'indépendance en probabilité d'une famille finie de variables aléatoires.

Etant donné une famille  $(X_1, X_2, \dots, X_n)$  de  $n$  variables aléatoires, on considère deux sous-familles disjointes à respectivement  $l$  et  $m$  éléments que l'on désignera respectivement par  $(X_{i_1}, \dots, X_{i_l})$  et par  $(X_{j_1}, \dots, X_{j_m})$  et l'on appelle "fonction de dépendance conditionnelle" de la sous-famille  $(X_{i_1}, \dots, X_{i_l})$  par rapport à la sous-famille  $(X_{j_1}, \dots, X_{j_m})$ , la fonction notée :

$$H_{X_{i_1}, \dots, X_{i_l} | X_{j_1}, \dots, X_{j_m}}$$

définie par :

$$\begin{aligned} & H_{X_{i_1}, \dots, X_{i_l} | X_{j_1}, \dots, X_{j_m}}(x_{i_1}, \dots, x_{i_l} | x_{j_1}, \dots, x_{j_m}) \\ & = \mathbb{P}(X_{i_1} \leq x_{i_1}, \dots, X_{i_l} \leq x_{i_l} | X_{j_1} \leq x_{j_1}, \dots, X_{j_m} \leq x_{j_m}) \end{aligned}$$

Ainsi, à l'aide de cette définition, la condition d'indépendance en probabilité peut s'écrire sous la forme :

$$\prod_{i=1}^{n-1} H_{X_i | X_{i+1}, \dots, X_n}(x_i | x_{i+1}, \dots, x_n) = \prod_{i=1}^{n-1} F_{X_i}(x_i)$$

À titre d'exemple, dans le cas où la famille considérée ne comprend que deux variables, notées  $X$  et  $Y$ , les deux fonctions de dépendance conditionnelle sont données par :

$$H_{X|Y}(x|y) = \mathbb{P}(X \leq x | Y \leq y) = \frac{\mathbb{P}(X \leq x, Y \leq y)}{\mathbb{P}(Y \leq y)}$$

et :

$$H_{Y|X}(y|x) = \frac{\mathbb{P}(X \leq x, Y \leq y)}{\mathbb{P}(X \leq x)}$$

Dans le cas de trois variables  $X, Y, Z$ , on a plusieurs fonctions de dépendance conditionnelle; l'une d'elles étant, par exemple :

$$H_{X|Y,Z}(x|y, z) = \mathbb{P}(X \leq x | Y \leq y, Z \leq z) = \frac{\mathbb{P}(X \leq x, Y \leq y, Z \leq z)}{\mathbb{P}(Y \leq y, Z \leq z)}$$

Les lemmes présentés ci-après concernent les fonctions de dépendance conditionnelle et utilisent, afin d'alléger l'écriture, les notations suivantes :

Étant données  $n$  variables aléatoires  $X_1, X_2, \dots, X_n$  de fonctions de répartition conjointe  $F(x_1, x_2, \dots, x_n)$  et de densité  $f(x_1, x_2, \dots, x_n)$ , on utilisera, par abus de notation, les mêmes symboles  $F$  et  $f$  pour désigner les fonctions de répartition et de densité conjointes (marginales) d'un sous-ensemble quelconque de variables parmi  $X_1, X_2, \dots, X_n$ . Ainsi,  $F(x_{i+1}, x_{i+2}, \dots, x_n)$  et  $f(x_{i+1}, x_{i+2}, \dots, x_n)$  désigneront respectivement la fonction de répartition et la fonction de densité du vecteur aléatoire  $(X_{i+1}, X_{i+2}, \dots, X_n)$ .

Nous noterons par :

$$F_{X_i|X_{i+1}, X_{i+2}, \dots, X_n}(x_i|x_{i+1}, x_{i+2}, \dots, x_n)$$

et par :

$$f_{X_i|X_{i+1}, X_{i+2}, \dots, X_n}(x_i|x_{i+1}, x_{i+2}, \dots, x_n)$$

la fonction de répartition conditionnelle et la fonction de densité conditionnelle de  $X_i$  étant donné que  $X_{i+1} = x_{i+1}, X_{i+2} = x_{i+2}, \dots, X_n = x_n$ . L'exception restera le cas unidimensionnel où nous continuerons d'écrire  $F_{X_i}(x_i)$  et  $f_{X_i}(x_i)$

### Lemme 1.

Les variables aléatoires  $X_1, X_2, \dots, X_n$  sont indépendantes en probabilité, si et seulement si l'on a, pour tout  $x_1, x_2, \dots, x_n$  :

$$\prod_{i=1}^{n-1} H_{X_i|X_{i+1}, \dots, X_n}(x_i|x_{i+1}, \dots, x_n) = \prod_{i=1}^{n-1} F_{X_i}(x_i) \quad .$$

Lorsque les variables sont indépendantes, les fonctions de dépendance :

$$H_{X_i|X_{i+1}, \dots, X_n}(x_i|x_{i+1}, \dots, x_n)$$

ne sont autres que les fonctions de répartition marginales  $F_{X_i}(x_i)$  et la précédente égalité est trivialement vérifiée.

Réciproquement, supposons que pour tout  $x_1, x_2, \dots, x_n$ , l'égalité :

$$\prod_{i=1}^{n-1} H_{X_i|X_{i+1}, \dots, X_n}(x_i|x_{i+1}, \dots, x_n) = \prod_{i=1}^{n-1} F_{X_i}(x_i) \quad ,$$

ait lieu. Il vient alors :

$$\begin{aligned} \prod_{i=1}^{n-1} H_{X_i|X_{i+1}, \dots, X_n}(x_i|x_{i+1}, \dots, x_n) &= \prod_{i=1}^{n-1} \frac{F(x_i, x_{i+1}, \dots, x_n)}{F(x_{i+1}, \dots, x_n)} \\ &= \frac{F(x_1, x_2, \dots, x_n)}{F(x_2, \dots, x_n)} \times \frac{F(x_2, \dots, x_n)}{F(x_3, \dots, x_n)} \times \dots \\ &\quad \dots \times \frac{F(x_{n-1}, x_n)}{F_{X_n}(x_n)} \\ &= \frac{F(x_1, x_2, \dots, x_n)}{F_{X_n}(x_n)} = \prod_{i=1}^{n-1} F_{X_i}(x_i) \end{aligned}$$

et donc :

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i) \quad .$$

ce qui entraîne l'indépendance des variables aléatoires  $X_1, X_2, \dots, X_n$ .

**Lemme 2.**

i) Les variables aléatoires  $X$  et  $Y$  sont indépendantes en probabilité si et seulement si l'on a, pour tous  $x, y$  :

$$\frac{\partial}{\partial x} H_{X|Y}(x|y) \frac{\partial}{\partial y} H_{Y|X}(y|x) = f_{X|Y}(x|y) f_{Y|X}(y|x)$$

En effet, supposons que  $X$  et  $Y$  soient indépendantes. On a alors :

$$f_{X|Y}(x|y) = f_X(x) \quad \text{et} \quad f_{Y|X}(y|x) = f_Y(y)$$

et de même :

$$\begin{aligned} \frac{\partial}{\partial x} H_{X|Y}(x|y) &= \frac{\partial}{\partial x} F_X(x) = f_X(x) \\ \frac{\partial}{\partial y} H_{Y|X}(y|x) &= \frac{\partial}{\partial y} F_Y(y) = f_Y(y) \end{aligned}$$

d'où l'égalité.

Inversement, si pour tous  $x, y$  l'égalité :

$$\frac{\partial}{\partial x} H_{X|Y}(x|y) \frac{\partial}{\partial y} H_{Y|X}(y|x) = f_{X|Y}(x|y) f_{Y|X}(y|x)$$

est vérifiée, on a :

$$\frac{\frac{\partial}{\partial x} F(x, y)}{F_Y(y)} \frac{\frac{\partial}{\partial y} F(x, y)}{F_X(x)} = \frac{f(x, y)}{f_Y(y)} \frac{f(x, y)}{f_X(x)}$$

et donc, utilisant le fait que  $\frac{\partial^2}{\partial x \partial y} F(x, y) = f(x, y)$ , il s'ensuit que :

$$\frac{f_Y(y) f_X(x)}{F_Y(y) F_X(x)} = \frac{\frac{\partial}{\partial y} \left[ \frac{\partial}{\partial x} F(x, y) \right]}{\frac{\partial}{\partial x} F(x, y)} \frac{\frac{\partial}{\partial x} \left[ \frac{\partial}{\partial y} F(x, y) \right]}{\frac{\partial}{\partial y} F(x, y)}$$

ou encore, de manière équivalente :

$$\frac{\partial}{\partial y} \text{Log}(F_Y(y)) \frac{\partial}{\partial x} \text{Log}(F_X(x)) = \frac{\partial}{\partial y} \text{Log} \left( \frac{\partial}{\partial x} F(x, y) \right) \frac{\partial}{\partial x} \text{Log} \left( \frac{\partial}{\partial y} F(x, y) \right)$$

expression qui peut également s'écrire sous la forme :

$$\begin{aligned} \frac{\partial}{\partial y} \text{Log}(F_Y(y)) \frac{\partial}{\partial x} \text{Log}(F_X(x)) &= \frac{\partial}{\partial y} \text{Log} \left( F_Y(y) \frac{\partial}{\partial x} H_{X|Y}(x|y) \right) \times \\ &\quad \frac{\partial}{\partial x} \text{Log} \left( F_X(x) \frac{\partial}{\partial y} H_{Y|X}(y|x) \right) \end{aligned}$$

et ce pour tous  $x, y$ . Or le terme de gauche de l'égalité ci-dessus, est le produit de deux fonctions à une variable et il doit donc en être ainsi pour le terme de droite. Ne pouvant isoler dans le second terme du membre de droite, l'expression  $\frac{\partial}{\partial y} \text{Log}(F_Y(y))$ , cela entraîne que  $\frac{\partial}{\partial x} H_{X|Y}(x|y)$  doit être une fonction positive  $a(x)$  de la variable  $x$  seulement. Pour les mêmes raisons, l'expression  $\frac{\partial}{\partial y} H_{Y|X}(y|x)$  doit être une fonction positive  $b(y)$  de la variable  $y$  seulement. Ainsi on doit avoir :

$$\frac{\partial}{\partial x} F(x, y) = F_Y(y) a(x) \quad \text{et} \quad \frac{\partial}{\partial y} F(x, y) = F_X(x) b(y)$$

ce qui, tenant compte du fait que  $F(x, y)$  est une fonction de répartition, implique que  $a(x) = f_X(x)$  et que  $b(y) = f_Y(y)$  et donc que  $f(x, y) = f_X(x)f_Y(y)$ .

ii) Dans le cas de trois variables aléatoires  $X, Y$  et  $Z$ , la condition nécessaire et suffisante d'indépendance s'exprime sous la forme suivante :

$$\begin{aligned} & \frac{\partial^2}{\partial x \partial y} H_{X,Y|Z}(x, y|z) \frac{\partial^2}{\partial x \partial z} H_{X,Z|Y}(x, z|y) \frac{\partial^2}{\partial y \partial z} H_{Y,Z|X}(y, z|x) \\ &= f_{X,Y|Z}(x, y|z) f_{X,Z|Y}(x, z|y) f_{Y,Z|X}(y, z|x) \quad \forall x, y, z \end{aligned}$$

La condition nécessaire étant trivialement vérifiée, il suffit de vérifier la condition suffisante. Supposons alors que l'égalité ci-dessus soit vérifiée pour tous  $x, y, z$ . Utilisant la même démarche que dans le cas de deux variables, on obtient aisément l'égalité qui suit :

$$\frac{f_X(x)}{F_X(x)} \frac{f_Y(y)}{F_Y(y)} \frac{f_Z(z)}{F_Z(z)} = \frac{f(x, y, z)}{\frac{\partial^2}{\partial y \partial z} F(x, y, z)} \frac{f(x, y, z)}{\frac{\partial^2}{\partial x \partial z} F(x, y, z)} \frac{f(x, y, z)}{\frac{\partial^2}{\partial x \partial y} F(x, y, z)}$$

Tenant compte du fait que :  $\frac{\partial^3}{\partial x \partial y \partial z} F(x, y, z) = f(x, y, z)$ , le terme de droite de cette dernière égalité peut s'écrire sous la forme :

$$\frac{\frac{\partial}{\partial x} \left[ \frac{\partial^2}{\partial y \partial z} F(x, y, z) \right]}{\frac{\partial^2}{\partial y \partial z} F(x, y, z)} \frac{\frac{\partial}{\partial y} \left[ \frac{\partial^2}{\partial x \partial z} F(x, y, z) \right]}{\frac{\partial^2}{\partial x \partial z} F(x, y, z)} \frac{\frac{\partial}{\partial z} \left[ \frac{\partial^2}{\partial x \partial y} F(x, y, z) \right]}{\frac{\partial^2}{\partial x \partial y} F(x, y, z)}$$

d'où il s'ensuit que :

$$\begin{aligned} & \frac{\partial}{\partial x} \text{Log}(F_X(x)) \frac{\partial}{\partial y} \text{Log}(F_Y(y)) \frac{\partial}{\partial z} \text{Log}(F_Z(z)) \\ &= \frac{\partial}{\partial x} \text{Log} \left( F_X(x) \frac{\partial^2}{\partial y \partial z} H_{Y,Z|X}(y, z|x) \right) \times \\ & \quad \frac{\partial}{\partial y} \text{Log} \left( F_Y(y) \frac{\partial^2}{\partial x \partial z} H_{X,Z|Y}(x, z|y) \right) \times \\ & \quad \frac{\partial}{\partial z} \text{Log} \left( F_Z(z) \frac{\partial^2}{\partial x \partial y} H_{X,Y|Z}(x, y|z) \right) \quad \forall x, y, z \end{aligned}$$

et utilisant les mêmes arguments que dans le cas de deux variables, il en découle que l'on doit avoir :

$$\begin{aligned} \frac{\partial^2}{\partial y \partial z} H_{Y,Z|X}(y, z|x) &= g(y, z) \\ \frac{\partial^2}{\partial x \partial z} H_{X,Z|Y}(x, z|y) &= h(x, z) \quad \text{et} \\ \frac{\partial^2}{\partial x \partial y} H_{X,Y|Z}(x, y|z) &= k(x, y) \end{aligned}$$

où  $g, h$  et  $k$  sont respectivement des fonctions positives des variables  $(y, z)$ ,  $(x, z)$  et  $(x, y)$ . On aura alors :

$$\frac{\partial}{\partial x} \left[ F_X(x) \frac{\partial^2}{\partial y \partial z} H_{Y,Z|X}(y, z|x) \right] = f(x, y, z) = f_X(x)g(y, z)$$

ce qui permet de conclure, en utilisant un argument de circularité, que :

$$f(x, y, z) = f_X(x)f_Y(y)f_Z(z)$$

et donc que les variables aléatoires  $X, Y$  et  $Z$  sont indépendantes.

iii) Enfin dans le cas de  $n$  variables aléatoires  $X_1, X_2, \dots, X_n$ , la condition nécessaire et suffisante d'indépendance s'exprimera sous la forme :

$$\begin{aligned} & \prod_{i=1}^n \frac{\partial^{n-1}}{\partial x_1 \dots \partial x_{i-1} \partial x_{i+1} \dots \partial x_n} H_{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n | X_i}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n | x_i) = \\ & \prod_{i=1}^n f_{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n | X_i}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n | x_i) \quad \forall x_1, x_2, \dots, x_n \end{aligned}$$

et en suivant la même démarche que dans les deux précédents cas, on pourra en conclure que :

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

et donc que les variables aléatoires sont indépendantes.

### 2.3 Famille de variables aléatoires libre sous $\mathcal{H}_0$

Dans le contexte du cadre théorique introduit ci-dessus, considérons les variables aléatoires  $Z_i$   $i = 1, 2, \dots, n$  définies par :

$$\begin{aligned} Z_i &= F_{X_i, X_{i+1}, \dots, X_n}(X_i, X_{i+1}, \dots, X_n) \\ &= F_{X_1, X_2, \dots, X_n}(\infty, \dots, \infty, X_i, \dots, X_n) \end{aligned}$$

Notons par  $Z$  le vecteur aléatoire de composantes  $Z_1, Z_2, \dots, Z_n$  et par  $\mathcal{D}$  le support de sa loi donné par :

$$\mathcal{D} = \{(z_1, z_2, \dots, z_n) \in \mathbb{R}^n : 0 \leq z_1 \leq z_2 \leq \dots \leq z_n \leq 1\}$$

On se propose alors de trouver la loi de  $Z$  et pour ce faire, on considère le difféomorphisme  $T$  de  $\mathbb{R}^n$  dans  $[0, 1]^n$  défini par :

$$z = T(x) = \{z_i = T_i = F_{X_1, X_2, \dots, X_n}(\infty, \dots, \infty, x_i, \dots, x_n)\} \quad \forall i = 1, 2, \dots, n$$

Désignant par  $f(x) = f(x_1, x_2, \dots, x_n)$  la densité du vecteur aléatoire  $X$ , il s'ensuit que la densité  $g(z) = g(z_1, z_2, \dots, z_n)$  du vecteur aléatoire  $Z$  est donnée par :

$$g(z) = f(T^{-1}(z)) |J(T^{-1})| \mathbb{I}_{\mathcal{D}}(z_1, z_2, \dots, z_n)$$

où  $J(T^{-1})$  désigne le jacobien de la transformation inverse. Observant que :

$$z_i = z_{i+1} \mathbb{P}(X_i \leq x_i | X_{i+1} \leq x_{i+1}, \dots, X_n \leq x_n) \text{ pour } i = 1, 2, \dots, n$$

et que :  $z_n = \mathbb{P}(X_n \leq x_n)$ , posons pour des raisons de commodité d'écriture :

$$\begin{cases} \mathbb{P}(X_i \leq x_i | X_{i+1} \leq x_{i+1}, \dots, X_n \leq x_n) = h_{x_n, \dots, x_{i+1}}(x_i) & i \leq n-1 \\ \mathbb{P}(X_n \leq x_n) = h_{x_{n+1}}(x_n) \text{ et } z_{n+1} = 1 \end{cases}$$

Les égalités précédentes peuvent alors s'écrire sous la forme :

$$z_i = z_{i+1} h_{x_n, \dots, x_{i+1}}(x_i) \quad i = 1, 2, \dots, n-1 \text{ et } z_n = z_{n+1} h_{x_{n+1}}(x_n)$$

Afin de déterminer le jacobien de la transformation  $T$ , on remarque que  $\partial z_i / \partial x_j = 0 \quad \forall j \leq i-1$  et par conséquent seuls les termes diagonaux de la matrice des dérivées partielles sont utiles à cette fin. Pour ces derniers il vient :

$$\frac{\partial z_i}{\partial x_i} = z_{i+1} \frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(x_i) \quad i = 1, 2, \dots, n-1$$

et

$$\frac{\partial z_n}{\partial x_n} = z_{n+1} \frac{\partial}{\partial x_n} h_{x_{n+1}}(x_n)$$

d'où :

$$|J(T)| = \prod_{i=1}^n z_{i+1} \frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(x_i)$$

et donc :

$$\begin{aligned} |J(T^{-1})| &= (\prod_{i=2}^n z_i)^{-1} \left[ \prod_{i=1}^n \left[ \frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(\cdot) \Big|_{h_{x_n, \dots, x_{i+1}}^{-1}(z_i/z_{i+1})} \right] \right]^{-1} \\ &= (\prod_{i=2}^n z_i)^{-1} \left[ \prod_{i=1}^n \frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(T_i^{-1}(z_i)) \right]^{-1} \end{aligned}$$

Ainsi, la densité  $g(z_1, z_2, \dots, z_n)$  du vecteur aléatoire  $Z$  a pour expression :

$$\frac{f \left[ h_{x_n, \dots, x_2}^{-1}(z_1/z_2), \dots, h_{x_n, \dots, x_{i+1}}^{-1}(z_i/z_{i+1}), \dots, h_{x_{n+1}}^{-1}(z_n/z_{n+1}) \right]}{\prod_{i=2}^n z_i \left\{ \prod_{i=1}^n \left[ \frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(\cdot) \Big|_{h_{x_n, \dots, x_{i+1}}^{-1}(z_i/z_{i+1})} \right] \right\}} \times \mathbb{I}_{\mathcal{D}}(z_1, z_2, \dots, z_n) .$$

Considérons également le vecteur aléatoire  $Z$  de composantes  $Z_1, Z_2, \dots, Z_n$  où :

$$Z_i = \prod_{j=i}^n F_{X_j}(X_j) \quad i = 1, 2, \dots, n .$$

Le difféomorphisme  $T$  de  $\mathbb{R}^n$  dans  $[0, 1]^n$  est alors défini dans ce cas par :

$$z = T(x) = \{z_i = T_i = F_{X_i}(x_i)F_{X_{i+1}}(x_{i+1})\dots F_{X_n}(x_n)\} \quad \forall i = 1, 2, \dots, n .$$

En suivant la démarche précédente et en notant que :

$$z_i = z_{i+1}h_{x_n, \dots, x_{i+1}}(x_i) = z_{i+1}\mathbb{P}(X_i \leq x_i) \quad i = 1, 2, \dots, n ,$$

que :

$$\prod_{i=1}^n \left[ \frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(\cdot) \Big|_{h_{x_n, \dots, x_{i+1}}^{-1}(z_i/z_{i+1})} \right] = \prod_{i=1}^n f_{X_i}(F_{X_i}^{-1}(z_i/z_{i+1}))$$

et que :

$$|J(T^{-1})| = (\prod_{i=2}^n z_i)^{-1} \left[ \prod_{i=1}^n f_{X_i}(F_{X_i}^{-1}(z_i/z_{i+1})) \right]^{-1}$$

on obtient aisément pour la densité de  $Z$  l'expression suivante :

$$\frac{f \left[ F_{X_1}^{-1}(z_1/z_2), \dots, F_{X_i}^{-1}(z_i/z_{i+1}), \dots, F_{X_n}^{-1}(z_n/z_{n+1}) \right]}{\prod_{i=2}^n z_i \left[ \prod_{i=1}^n f_{X_i}(F_{X_i}^{-1}(z_i/z_{i+1})) \right]} \times \mathbb{I}_{\mathcal{D}}(z_1, z_2, \dots, z_n)$$

De l'hypothèse d'indépendance en probabilité des composantes du vecteur aléatoire  $X$ , découle un certain nombre de simplifications dans l'expression de la densité de  $Z$ . En effet, dans le cas du vecteur aléatoire  $Z = (Z_1, Z_2, \dots, Z_n)$  où :

$$Z_i = F_{X_i, X_{i+1}, \dots, X_n}(X_i, X_{i+1}, \dots, X_n) \quad i = 1, 2, \dots, n$$

il vient :

$$\begin{aligned} h_{x_n, \dots, x_{i+1}}(x_i) &= \mathbb{P}(X_i \leq x_i | X_{i+1} \leq x_{i+1}, \dots, X_n \leq x_n) \\ &= \mathbb{P}(X_i \leq x_i) = F_{X_i}(x_i) \quad i = 1, 2, \dots, n-1 \\ \text{et } h_{x_{n+1}}(x_n) &= \mathbb{P}(X_n \leq x_n) = F_{X_n}(x_n) \end{aligned}$$

Il s'ensuit donc que :

$$\frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(x_i) = \frac{\partial}{\partial x_i} F_{X_i}(x_i) = f_{X_i}(x_i) \quad \forall i = 1, 2, \dots, n$$

et que :

$$x_i = F_{X_i}^{-1}(z_i/z_{i+1}) \quad \forall i = 1, 2, \dots, n$$

Ainsi l'expression de la densité de  $Z$  devient :

$$g(z_1, z_2, \dots, z_n) = \prod_{i=1}^n f_{X_i}(T_i^{-1}(z)) \times \frac{1}{\prod_{i=2}^n z_i [\prod_{i=1}^n f_{X_i}(T_i^{-1}(z))]} \times \mathbb{I}_{\mathcal{D}}(z_1, z_2, \dots, z_n)$$

et donc, sous l'hypothèse  $\mathcal{H}_0$  d'indépendance des composantes du vecteur aléatoire  $X$ , la loi du vecteur  $Z$  a pour densité :

$$g(z_1, z_2, \dots, z_n) = \frac{1}{z_2 z_3 \dots z_n} \mathbb{I}_{\mathcal{D}}(z_1, z_2, \dots, z_n)$$

ce qui permet de conclure que la famille  $Z_1, Z_2, \dots, Z_n$  de variables aléatoires est libre puisque la loi conjointe ne dépend pas de la mesure de probabilité initiale  $\mathbb{P}$ . On obtient évidemment la même conclusion dans le cas du vecteur aléatoire  $Z = (Z_1, Z_2, \dots, Z_n)$  où :

$$Z_i = \prod_{j=i}^n F_{X_j}(X_j) \quad i = 1, 2, \dots, n$$

Par ailleurs il est facile de vérifier que la condition :

$$\frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(x_i) = f_{X_i | X_{i+1}, \dots, X_n}(x_i | x_{i+1}, \dots, x_n) \\ \forall x_i, x_{i+1}, \dots, x_n ; \forall i = 1, 2, \dots, n$$

est une condition nécessaire et suffisante d'indépendance des variables aléatoires  $X_1, X_2, \dots, X_n$ . On peut donc énoncer le dernier résultat sous la forme du théorème suivant :

**Théorème 1 :** Les variables aléatoires  $X_1, X_2, \dots, X_n$  définies sur la même structure statistique dominée  $(\Omega, \mathcal{F}, \mathbb{P}, \mu)$  sont indépendantes en probabilité si et seulement si la loi du vecteur aléatoire

$$Z = (Z_i = F(X_i, X_{i+1}, \dots, X_n))_{i=1, 2, \dots, n}$$

admet pour densité de probabilité  $g(z) = g(z_1, z_2, \dots, z_n)$ , l'expression donnée par :

$$g(z_1, z_2, \dots, z_n) = \frac{1}{z_2 z_3 \dots z_n} \mathbb{I}_{\mathcal{D}}(z_1, z_2, \dots, z_n)$$

où  $\mathcal{D} = \{(z_1, z_2, \dots, z_n) \in \mathbb{R}^n : 0 \leq z_1 \leq z_2 \leq \dots \leq z_n \leq 1\}$ .

**Remarque 2 :** Malgré l'apparente complexité de l'expression de la densité de probabilité du vecteur aléatoire  $Z$  dans le cas général, il est assez aisé d'en déduire par récurrence, que la densité de probabilité  $g(z_k, z_{k+1}, \dots, z_n)$  d'un vecteur quelconque de la forme  $(z_k, z_{k+1}, \dots, z_n)$  est donnée par :

$$\frac{f_{X_k, \dots, X_n} \left[ h_{x_n, \dots, x_{k+1}}^{-1}(z_k/z_{k+1}), \dots, h_{x_n}^{-1}(z_{n-1}/z_n), h_{x_{n+1}}^{-1}(z_n/z_{n+1}) \right]}{\prod_{i=k+1}^n z_i \left\{ \prod_{i=k}^n \left[ \frac{\partial}{\partial x_i} h_{x_n, \dots, x_{i+1}}(\cdot) \Big|_{h_{x_n, \dots, x_{i+1}}^{-1}(z_i/z_{i+1})} \right] \right\}} \\ \times \mathbb{I}_{[0 \leq z_k \leq \dots \leq z_n \leq 1]}(z_k, z_{k+1}, \dots, z_n)$$

qui est du même type que celle du vecteur  $Z$ . En particulier on peut vérifier que  $g(z_n) = 1 \times \mathbb{I}_{[0,1]}(z_n)$  ce qui est, bien entendu, le résultat attendu.

**Remarque 3 :** Dans le sens de la remarque 1, il existe, comme on peut le vérifier facilement, d'autres familles de variables aléatoires libres sous l'hypothèse  $\mathcal{H}_0$ . Il en est ainsi par exemple, dans le cas bidimensionnel, du couple  $(Z_1 = F_{X_1}(X_1)(1 - F_{X_2}(X_2)) ; Z_2 = 1 - F_{X_2}(X_2))$  ou encore du couple  $(Z_1 = (1 - F_{X_1}(X_1))(1 - F_{X_2}(X_2)) ; Z_2 = 1 - F_{X_2}(X_2))$ .



## 2.4 Lois de $F(X_1, X_2, \dots, X_n)$ et de $\prod_{i=1}^n F_{X_i}(X_i)$

### 2.4.1 Cas général

La détermination dans le cas général des fonctions de répartition des variables aléatoires  $F(X_1, X_2, \dots, X_n)$  et  $\prod_{i=1}^n F_i(X_i)$ , est relativement aisée si l'on utilise la représentation de la loi  $(X_1, X_2, \dots, X_n)$  sous la forme de copule car cette dernière permet de se ramener à un cadre unique quelle que soit la fonction de répartition  $F(X_1, X_2, \dots, X_n)$ .

En effet, on sait (Fréchet[9], Sklar[22], Kimeldorf et Sampson[12]) que l'on peut, et ce de manière unique, associer à toute mesure de probabilité  $\mathbb{P}$  définie sur  $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \nu)$  et absolument continue par rapport à la mesure de référence  $\nu$ , une mesure de probabilité  $\mathbb{C}$  définie sur  $([0, 1]^n, \mathcal{B}_{[0,1]^n}, \mu)$  absolument continue par rapport à  $\mu$  (mesure de Lebesgue) et telle que ses lois marginales soient des lois uniformes sur l'intervalle  $[0, 1]$ . On dit alors que  $\mathbb{C}$  est la copule associée à  $\mathbb{P}$  ou encore que  $\mathbb{C}$  constitue une "représentation uniforme" de  $\mathbb{P}$ . Adoptant les notations habituelles, la correspondance bijective entre les mesures  $\mathbb{P}$  et  $\mathbb{C}$  s'effectue à l'aide de la transformation :

$$U_i = F_{X_i}(X_i) \quad \text{pour tout } i = 1, 2, \dots, n$$

et l'on vérifie aisément que les fonctions de répartition  $C(u_1, u_2, \dots, u_n)$  (que l'on notera également  $C_{U_1, U_2, \dots, U_n}(u_1, u_2, \dots, u_n)$ ) et  $F(x_1, x_2, \dots, x_n)$  des vecteurs aléatoires  $U = {}^t(U_1, U_2, \dots, U_n)$  et  $X = {}^t(X_1, X_2, \dots, X_n)$  sont liées entre elles par la relation :

$$C(u_1, u_2, \dots, u_n) = F(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2), \dots, F_{X_n}^{-1}(u_n))$$

On observera par ailleurs que, dans le cas de l'indépendance, on a de façon évidente :

$$Z_1 = C(U_1, U_2, \dots, U_n) = \prod_{i=1}^n U_i$$

Désignons par  $\pi_{U_1, U_2, \dots, U_n}$  la mesure de probabilité associée à la copule  $\mathbb{C}$  et pour tout  $z \in [0, 1]$  notons par  $D^n(z)$  le domaine de  $[0, 1]^n$  défini par :

$$D^n(z) = \{(u_1, u_2, \dots, u_n) \in [0, 1]^n : C(u_1, u_2, \dots, u_n) \leq z\}$$

Il vient :

$$\begin{aligned} G(z) &= \mathbb{P}(Z_1 \leq z) = \mathbb{P}(C(U_1, U_2, \dots, U_n) \leq z) \\ &= \mathbb{E}^{\pi_{U_1, U_2, \dots, U_n}} [\mathbb{I}_{D^n(z)}(u_1, u_2, \dots, u_n)] \\ &= \int_{D^n(z)} dC_{U_1, U_2, \dots, U_n}(u_1, u_2, \dots, u_n) \\ &= \mathbb{E}^{\pi_{U_2, \dots, U_n}} [\mathbb{E}^{\pi_{U_1|U_2, \dots, U_n}} [\mathbb{I}_{D^n(z)}(u_1, u_2, \dots, u_n) | U_2 = u_2, \dots, U_n = u_n]] \end{aligned}$$

Or le domaine  $D^n(z)$  peut s'exprimer sous la forme de l'union des deux domaines disjoints suivants :

$$\begin{aligned} &[0, 1] \times \{(u_2, \dots, u_n) : C_{U_2, \dots, U_n}(u_2, \dots, u_n) \leq z\} \quad \text{et} \\ &[0, h_{u_n, \dots, u_2}^{-1}(z/C_{U_2, \dots, U_n}(u_2, \dots, u_n))] \times \\ &\{(u_2, \dots, u_n) : C_{U_2, \dots, U_n}(u_2, \dots, u_n) > z\} \end{aligned}$$

Soit encore, en notant par  $D^{n-1}(z)$  le domaine de  $[0, 1]^{n-1}$  défini par :

$$D^{n-1}(z) = \{(u_2, \dots, u_n) : C_{U_2, \dots, U_n}(u_2, \dots, u_n) \leq z\}$$

$$\begin{aligned} D^n(z) &= [0, 1] \times D^{n-1}(z) \cup \\ &[0, h_{u_n, \dots, u_2}^{-1}(z/C_{U_2, \dots, U_n}(u_2, \dots, u_n))] \times \overline{D^{n-1}(z)} \end{aligned}$$

Cette décomposition est une conséquence directe des propriétés suivantes :

$$\text{i) } C_{U_2, \dots, U_n}(u_2, \dots, u_n) \leq z \implies C_{U_1, U_2, \dots, U_n}(u_1, u_2, \dots, u_n) \leq z \\ \forall u_1 \in [0, 1]$$

$$\text{ii) } C_{U_1, U_2, \dots, U_n}(u_1, u_2, \dots, u_n) = h_{u_n, \dots, u_2}(u_1) C_{U_2, \dots, U_n}(u_2, \dots, u_n),$$

et donc les conditions :

$$\begin{cases} h_{u_n, \dots, u_2}(u_1) C_{U_2, \dots, U_n}(u_2, \dots, u_n) \leq z \\ 0 \leq h_{u_n, \dots, u_2}(u_1) \leq 1 \end{cases}$$

sont équivalentes aux conditions :

$$\begin{cases} h_{u_n, \dots, u_2}(u_1) \leq z / C_{U_2, \dots, U_n}(u_2, \dots, u_n) \\ C_{U_2, \dots, U_n}(u_2, \dots, u_n) > z \end{cases}$$

Il découle des observations précédentes que :

$$\begin{aligned} & \mathbb{E}^{\pi_{U_1|U_2, \dots, U_n}} [\mathbb{I}_{D^n}(z)(u_1, u_2, \dots, u_n) | U_2 = u_2, \dots, U_n = u_n] \\ = & \mathbb{E}^{\pi_{U_1|U_2, \dots, U_n}} [\mathbb{I}_{[0,1] \times D^{n-1}}(z)(u_1, u_2, \dots, u_n) | U_2 = u_2, \dots, U_n = u_n] + \\ & \mathbb{E}^{\pi_{U_1|U_2, \dots, U_n}} \left[ \frac{\mathbb{I}_{[0, h_{u_n, \dots, u_2}^{-1}(z/C_{U_2, \dots, U_n}(u_2, \dots, u_n))]} \times \overline{D}^{n-1}(z)(u_1, u_2, \dots, u_n)}{|U_2 = u_2, \dots, U_n = u_n}} \right] \\ = & \mathbb{I}_{D^{n-1}}(z)(u_2, \dots, u_n) + \\ & C_{U_1|U_2, \dots, U_n} [h_{u_n, \dots, u_2}^{-1}(z/C_{U_2, \dots, U_n}(u_2, \dots, u_n))] \times \mathbb{I}_{\overline{D}^{n-1}}(z)(u_2, \dots, u_n) \end{aligned}$$

D'où en intégrant par rapport à la mesure de probabilité  $\pi_{U_2, \dots, U_n}$  il vient :

$$\begin{aligned} G(z) &= \int_{D^{n-1}(z)} dC_{U_2, \dots, U_n}(u_2, \dots, u_n) + \\ & \int_{\overline{D}^{n-1}(z)} C_{U_1|U_2, \dots, U_n} [h_{u_n, \dots, u_2}^{-1}(z/C_{U_2, \dots, U_n}(u_2, \dots, u_n))] \times \\ & dC_{U_2, \dots, U_n}(u_2, \dots, u_n) \end{aligned}$$

En utilisant la même démarche que ci-dessus, on vérifie aisément que le premier terme du membre de droite de la dernière égalité se décompose à son tour, à l'aide de notations évidentes, sous la forme :

$$\begin{aligned} & \int_{D^{n-2}(z)} dC_{U_3, \dots, U_n}(u_3, \dots, u_n) + \\ & \int_{\overline{D}^{n-2}(z)} C_{U_2|U_3, \dots, U_n} [h_{u_n, \dots, u_3}^{-1}(z/C_{U_3, \dots, U_n}(u_3, \dots, u_n))] \times \\ & dC_{U_3, \dots, U_n}(u_3, \dots, u_n) \end{aligned}$$

et en procédant par récurrence on peut vérifier facilement qu'à la  $i^{\text{ième}}$  étape on obtient l'expression :

$$\begin{aligned} & \int_{D^{n-i}(z)} dC_{U_{i+1}, \dots, U_n}(u_{i+1}, \dots, u_n) + \\ & \int_{\overline{D}^{n-i}(z)} C_{U_i|U_{i+1}, \dots, U_n} [h_{u_n, \dots, u_{i+1}}^{-1}(z/C_{U_{i+1}, \dots, U_n}(u_{i+1}, \dots, u_n))] \times \\ & dC_{U_{i+1}, \dots, U_n}(u_{i+1}, \dots, u_n) \end{aligned}$$

et qu'à la dernière étape on a :

$$\int_0^z dC_{U_n}(u_n) + \int_{\overline{D}^1(z)} C_{U_{n-1}|U_n}(h_{u_n}^{-1}(z/C_{U_n}(u_n))) dC_{U_n}(u_n)$$

Ainsi, après simplifications, la fonction de répartition  $G(z)$  de  $Z_1$  s'exprime sous la forme :

$$z + \sum_{i=1}^{n-1} \left[ \int_{\overline{D}^{n-i}(z)} C_{U_i|U_{i+1},\dots,U_n} \left[ h_{u_n,\dots,u_{i+1}}^{-1} \left( z / C_{U_{i+1},\dots,U_n} (u_{i+1}, \dots, u_n) \right) \right] \times dC_{U_{i+1},\dots,U_n} (u_{i+1}, \dots, u_n) \right]$$

Cette expression, tout à fait générale de la loi de  $Z_1 = C(U_1, U_2, \dots, U_n)$ , permet de traiter aisément le cas particulier associé à l'hypothèse d'indépendance des composantes  $U_1, U_2, \dots, U_n$  de la copule  $\mathbb{C}$  puisque, dans ce cas, on a pour tout  $i = 1, 2, \dots, n$  :

$$h_{u_n,\dots,u_{i+1}}(u_i) = C_{U_i}(u_i) = u_i$$

Quant à la fonction de répartition de la variable aléatoire

$$\prod_{i=1}^n C_{U_i}(U_i) = \prod_{i=1}^n U_i$$

on montre, en suivant la même démarche que ci-dessus, que celle-ci est donnée par :

$$\begin{aligned} & z + \sum_{i=1}^{n-1} \left[ \int_{\overline{D}^{n-i}(z)} C_{U_i|U_{i+1},\dots,U_n} \left[ C_{U_i}^{-1} \left( z / \prod_{j=i+1}^n C_{U_j}(u_j) \right) \right] \right. \\ & \quad \left. \times dC_{U_{i+1},\dots,U_n}(u_{i+1}, \dots, u_n) \right] \\ = & z + \sum_{i=1}^{n-1} \left[ \int_{\overline{D}^{n-i}(z)} C_{U_i|U_{i+1},\dots,U_n} \left( z / \prod_{j=i+1}^n u_j \right) \right. \\ & \quad \left. \times dC_{U_{i+1},\dots,U_n}(u_{i+1}, \dots, u_n) \right] \end{aligned}$$

en vertu du fait que  $U_i \sim \mathcal{U}_{[0,1]} \forall i = 1, 2, \dots, n$ , que  $h_{u_n,\dots,u_{i+1}}(u_i) = u_i$  et que le domaine  $D^n(z)$  s'exprime sous la forme :

$$(u_1, u_2, \dots, u_n) \in [0, 1]^n : \prod_{i=1}^n u_i \leq z$$

**Remarque 4 :** Les fonctions de répartition des variables aléatoires

$$\prod_{i=1}^n C_{U_i}(U_i) \quad \text{et} \quad C(U_1, U_2, \dots, U_n)$$

diffèrent à la fois par rapport aux arguments des fonctions de répartition conditionnelles et par rapport aux domaines d'intégration. En effet, dans le premier cas l'argument et le domaine sont donnés respectivement, au rang  $i$ , par :

$$\begin{cases} C_{U_i}^{-1} \left( z / \prod_{j=i+1}^n C_{U_j}(u_j) \right) = \frac{z}{\prod_{j=i+1}^n u_j} \quad \text{et par :} \\ D^{n-i}(z) = \left\{ (u_{i+1}, \dots, u_n) \in \mathbb{R}^{n-i} : \prod_{j=i+1}^n u_j \leq z \right\} \end{cases}$$

alors que dans le second cas les mêmes quantités sont données, au même rang  $i$ , par :

$$\begin{cases} h_{u_n,\dots,u_{i+1}}^{-1} \left( z / C_{U_{i+1},\dots,U_n}(u_{i+1}, \dots, u_n) \right) \quad \text{et par :} \\ D^{n-i}(z) = \left\{ (u_{i+1}, \dots, u_n) \in \mathbb{R}^{n-i} : C_{U_{i+1},\dots,U_n}(u_{i+1}, \dots, u_n) \leq z \right\} \end{cases}$$

On notera cependant que dans le cas de l'indépendance tant les arguments que les domaines sont égaux.

**Remarque 5 :** Il se peut, comme dans le cas semi-paramétrique, que la variable aléatoire  $F(X_1, X_2, \dots, X_n)$  appartienne à une certaine classe déterminée de fonctions de répartition dépendantes d'un certain nombre de paramètres inconnus. On peut alors estimer ces derniers et utiliser la variable aléatoire  $Z_1$  afin de "mesurer" l'écart à l'indépendance. On trouvera ci-dessous deux exemples traitant de ce cas.

### 2.4.2 Cas de l'indépendance

Puisque dans ce cas

$$Z_1 = F(X_1, X_2, \dots, X_n) = \prod_{i=1}^n F_{X_i}(X_i)$$

il vient, en intégrant la densité de probabilité de  $Z$  par rapport à  $z_n$  :

$$\begin{aligned} g(z_1, z_2, \dots, z_{n-1}) &= \frac{1}{z_2 \dots z_{n-1}} \int_{z_{n-1}}^1 \frac{dz_n}{z_n} \mathbb{I}_{[0 \leq z_1 \leq \dots \leq z_{n-1} \leq 1]}(z_1, z_2, \dots, z_{n-1}) \\ &= -\frac{\text{Log}(z_{n-1})}{z_2 \dots z_{n-1}} \mathbb{I}_{[0 \leq z_1 \leq \dots \leq z_{n-1} \leq 1]}(z_1, z_2, \dots, z_{n-1}) \end{aligned}$$

De même en intégrant par rapport à  $z_{n-1}$  on obtient :

$$\begin{aligned} g(z_1, z_2, \dots, z_{n-2}) &= -\frac{1}{z_2 \dots z_{n-1}} \int_{z_{n-2}}^1 \frac{\text{Log}(z_{n-1})}{z_{n-1}} dz_{n-1} \\ &\quad \times \mathbb{I}_{[0 \leq z_1 \leq \dots \leq z_{n-2} \leq 1]}(z_1, z_2, \dots, z_{n-2}) \\ &= \frac{1}{2} \frac{(\text{Log}(z_{n-2}))^2}{z_2 \dots z_{n-2}} \mathbb{I}_{[0 \leq z_1 \leq \dots \leq z_{n-2} \leq 1]}(z_1, z_2, \dots, z_{n-2}) \end{aligned}$$

et par récurrence, on vérifie sans peine que la densité de probabilité de la variable aléatoire  $Z_1 = \prod_{i=1}^n F_{X_i}(X_i)$  est donnée par :

$$g(z_1) = \frac{[\text{Log}(1/z_1)]^{n-1}}{(n-1)!} \mathbb{I}_{[0,1]}(z_1)$$

On désignera cette loi par le nom de loi  $\mathcal{CM}_n$  à  $n$  degrés de liberté, cette dernière n'étant autre, comme on devait s'y attendre, que la loi du produit de  $n$  variables aléatoires indépendantes de lois  $\mathcal{U}_{[0,1]}$ .

Un calcul élémentaire montre que la fonction de répartition  $G(z_1)$  de la variable aléatoire  $Z_1$  peut s'exprimer sous la forme :

$$G(z_1) = \mathbb{P}(Z_1 \leq z_1) = z_1 \sum_{k=0}^{n-1} \frac{[\text{Log}(1/z_1)]^k}{k!} \mathbb{I}_{[0,1]}(z_1) + \mathbb{I}_{]1,\infty[}(z_1).$$

On peut donc, en résumé, énoncer le théorème suivant ainsi que les corollaires évidents qui en découlent :

**Théorème 2 :** Soient  $X_1, X_2, \dots, X_n$ ,  $n$  variables aléatoires indépendantes définies sur le même espace probabilisé  $(\Omega, \mathcal{F}, \mathbb{P})$ , à valeurs dans  $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$  et de lois  $\mu_1, \mu_2, \dots, \mu_n$  absolument continues par rapport à une mesure de référence  $\mu$ . Si  $F_i$ ,  $i = 1, 2, \dots, n$  désigne la fonction de répartition de la mesure  $\mu_i$ , alors la variable aléatoire  $Z_1 = \prod_{i=1}^n F_i(X_i)$ , définie sur  $\otimes_{i=1}^n (\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mu_i)$  et à valeurs dans  $([0, 1], \mathcal{B}_{[0,1]})$ , suit une loi  $\mathcal{CM}_n$  à  $n$  degrés de liberté.

**Corollaire 1 :** Puisque  $G$  est justement, dans le cas de l'indépendance, la fonction de répartition de la variable aléatoire  $Z_1 = \prod_{i=1}^n F_i(X_i)$ , on peut encore de façon équivalente dire dans ce cas que,  $G(\prod_{i=1}^n F_i(X_i)) \sim \mathcal{U}_{[0,1]}$

**Corollaire 2 :** Soit  $(X_i)$ ,  $i = 1, 2, \dots, n$ , une suite de vecteurs aléatoires indépendants définis sur le même espace probabilisé  $(\Omega, \mathcal{F}, \mathbb{P})$ , à valeurs dans  $(\mathbb{R}^{k_i}, \mathcal{B}_{\mathbb{R}^{k_i}})$  et de lois  $\mu_i$  absolument continues par rapport à une mesure de référence  $\mu$ . Si l'on suppose que les composantes  $(X_{ij})$   $j = 1, 2, \dots, k_i$ , de chaque vecteur  $X_i$  sont indépendantes et si l'on note par  $(F_i)$  et  $(F_{ij})$  les fonctions de répartition des vecteurs  $X_i$  et des variables  $X_{ij}$ , alors la variable aléatoire

$$\mathcal{Z}_1 = \prod_{i=1}^n F_i(X_i) = \prod_{i=1}^n \prod_{j=1}^{k_i} F_{ij}(X_{ij}),$$

définie sur  $\otimes_{i=1}^n (\mathbb{R}^{k_i}, \mathcal{B}_{\mathbb{R}^{k_i}}, \mu_i)$  et à valeurs dans  $([0, 1], \mathcal{B}_{[0,1]})$ , suit une loi  $\mathcal{CM}_{\Sigma_i k_i}$  à  $k = \sum_{i=1}^n k_i$  degrés de liberté.

**Corollaire 3 :** Posant, dans le théorème 2 ci-dessus, le changement de variable  $Y = -\text{Log}(Z_1)$  on vérifie sans difficulté que  $Y \sim \Gamma(n, 1)$  et qu'en conséquence la variable aléatoire  $-2\text{Log}(Z_1)$  suit une loi du  $\chi_{2n}^2$  à  $2n$  degrés de liberté.

### 3 Test semi-paramétrique d'indépendance entre les composantes d'un vecteur aléatoire

#### 3.1 Procédure de test

Dans le cadre semi-paramétrique tel qu'évoqué à la remarque 5, on suppose, outre l'hypothèse d'indépendance des composantes du vecteur aléatoire  $X$ , que la fonction de répartition de  $X$  appartient à une classe  $F(X_1, X_2, \dots, X_n; \gamma)$  de fonctions de répartition dépendantes d'un paramètre "de forme"  $\gamma \in \mathbb{R}^p$ .

Désignant par  $\hat{\gamma}$  une estimation du paramètre  $\gamma$  obtenue à l'aide de  $k$  observations indépendantes, le test proposé consiste simplement à comparer la fonction de répartition empirique des valeurs observées  $Z_i = F(X_{1,i}, X_{2,i}, \dots, X_{n,i}; \hat{\gamma})$   $i = 1, 2, \dots, k$ , à la fonction de répartition d'une loi  $\mathcal{CM}_n$  à l'aide d'un test d'ajustement.

Dans le cas plus général, où l'on ne dispose d'aucune information sur la fonction de répartition  $F(X_1, X_2, \dots, X_n)$ , on ne peut utiliser la procédure décrite ci-dessus. Cependant en supposant que l'on puisse disposer d'un second échantillon de taille  $k'$  permettant d'obtenir une estimation  $\hat{F}(X_1, X_2, \dots, X_n)$  de  $F(X_1, X_2, \dots, X_n)$ , soit sous la forme de la fonction de répartition empirique soit encore sous la forme d'une estimation fonctionnelle (voir D. Bosq et J.P. Lecoutre[4]), on sait que les valeurs empiriques  $\hat{Z}_i = \hat{F}(X_{1,i}, X_{2,i}, \dots, X_{n,i})$   $i = 1, 2, \dots, k$  convergent presque sûrement, en vertu du théorème de *Glivenko-Cantelli*, vers les valeurs théoriques  $Z_i = F(X_{1,i}, X_{2,i}, \dots, X_{n,i})$ , ce qui permet de se ramener ainsi au précédent test d'ajustement à une loi  $\mathcal{CM}_n$  à l'aide de la fonction de répartition  $\hat{F}(X_1, X_2, \dots, X_n)$ .

**Remarque 6 :** Dans l'espace des fonctions de répartition multidimensionnelles, la "dépendance" représente, par rapport à l'hypothèse  $\mathcal{H}_0$ , une alternative excessivement complexe pouvant revêtir des formes très diverses et il est impossible, sauf cas très particulier, de la caractériser simplement à l'aide d'un nombre fini de paramètres. Par conséquent il est illusoire en pareille circonstance, de vouloir rechercher le test le plus puissant et donc a fortiori le test uniformément le plus puissant. Tout au plus pourra-t-on, en ce qui concerne la puissance du test, considérer par exemple le cadre très restreint des modèles semi-paramétriques pour lesquels des valeurs particulières des paramètres correspondent à l'indépendance entre les variables aléatoires et étudier la puissance du test aux voisinages de ces valeurs, afin de le comparer à d'autres tests d'indépendance.

Afin d'illustrer de façon détaillée la conduite du test, uniquement dans le cas semi-paramétrique, on présente ci-après, à l'aide d'un échantillon de taille 20, deux exemples portant sur la loi bi-exponentielle et sur la loi normale bidimensionnelle. Le test proposé, comme mentionné ci-dessus se ramène à un test d'ajustement de type *Kolmogorov-Smirnov* entre la fonction de répartition empirique de la variable aléatoire  $\hat{F}(X_1, X_2, \dots, X_n)$  et la fonction de répartition de la loi  $\mathcal{CM}_2$ .

Enfin, on présente brièvement ci-dessous l'algorithme utilisé dans le cadre de la simulation d'un échantillon bidimensionnel :

**Étape 1)** On considère les estimations  $\hat{\gamma} = 0, 0, 2$  et  $0, 5$  du paramètre  $\gamma$  de la loi  $\text{BiExp}(\gamma)$  et les estimations  $\hat{\rho} = 0, 0, 2$  et  $0, 5$  du paramètre  $\rho$  de la loi  $N(0, I + \rho J)$ .

**Étape 2)** Pour chacun de ces choix on forme le couple  $(x_1, x_2)$  à l'aide des lois de  $X_1$  et de  $X_2|X_1$  et l'on détermine suivant le cas  $F(x_1, x_2; \hat{\gamma})$  ou  $F(x_1, x_2; \hat{\rho})$ .

**Étape 3)** Si  $n$  désigne la taille de l'échantillon considéré, on itère la procédure précédente  $n$  fois (dans le cas présent,  $n = 20$ ).

**Étape 4)** Disposant de  $n = 20$  observations indépendantes de la variable aléatoire  $Z_1 = F(X_1, X_2; \hat{\gamma})$  ou de la variable aléatoire  $Z_1 = F(X_1, X_2; \hat{\rho})$ , on effectue un test d'ajustement de la fonction de répartition empirique de la variable aléatoire  $Z_1$  à celle de la loi  $\mathcal{CM}_2$ .

### 3.2 Loi bi-exponentielle $\mathcal{BiExp}(\gamma)$

Le vecteur aléatoire  $X = {}^t(X_1, X_2)$  suit une loi bi-exponentielle  $\mathcal{BiExp}(\gamma)$  de paramètre  $\gamma$  ( $|\gamma| \leq 1$ ) (voir Johnson [11]), si sa fonction de répartition a pour expression :

$$F(x_1, x_2) = (1 - e^{-x_1})(1 - e^{-x_2})(1 + \gamma e^{-x_1 - x_2})$$

La copule associée admet alors pour fonction de répartition :

$$C(u_1, u_2) = u_1 u_2 (1 + \gamma(1 - u_1)(1 - u_2))$$

et les densités correspondantes sont données respectivement par :

$$f(x_1, x_2) = e^{-x_1} e^{-x_2} (1 + \gamma(2e^{-x_1} - 1)(2e^{-x_2} - 1))$$

et par :

$$c(u_1, u_2) = (1 + \gamma(1 - 2u_1)(1 - 2u_2))\mathbb{I}_{[0,1]^2}(u_1, u_2)$$

dont on donne aux figures 1 et 2, dans le cas de la copule, les représentations graphiques pour  $\gamma = 0,5$  et pour  $\gamma = -0,5$ . Pour cette famille de lois, on peut vérifier aisément que les variables aléatoires  $X_1$  et  $X_2$  sont indépendantes, que si et seulement si le coefficient de corrélation linéaire  $\rho = \gamma/4$  est égal à 0.

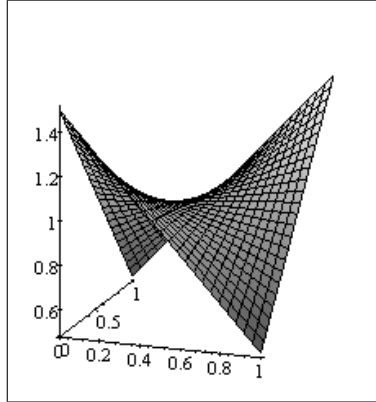


Fig. 1  $\gamma = 0,5$

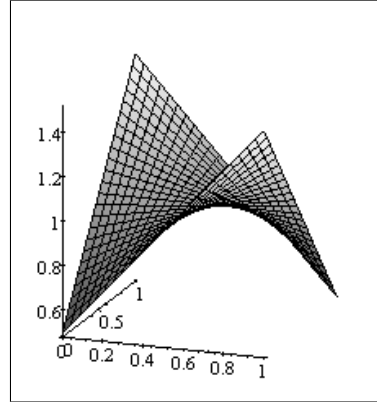


Fig 2  $\gamma = -0,5$

Il en est de même d'ailleurs pour les variables aléatoires  $U_1$  et  $U_2$  de la copule associée à la différence toutefois que dans ce cas on a  $\rho = \gamma/3$ . A l'aide des techniques usuelles telles que décrites, par exemple, dans Devroye[6], on simule à l'aide des lois de  $X_1$  et de  $X_2|X_1$  ou de  $U_1$  et de  $U_2|U_1$  un échantillon de taille  $n = 20$  pour les valeurs "estimées"  $\hat{\gamma} = 0$ ,  $\hat{\gamma} = 0,2$  et  $\hat{\gamma} = 0,5$  de  $\gamma$  à l'aide par exemple, dans le cas de la loi initiale, de l'estimateur :

$$\hat{\Gamma} = 4\hat{R}\mathbb{I}_{\{|\hat{R}| \leq 0,25\}} + \text{sign}\hat{R}\mathbb{I}_{\{|\hat{R}| > 0,25\}}$$

ou encore, dans le cas de la copule, de l'estimateur :

$$\hat{\Gamma} = 3\hat{R}\mathbb{I}_{\{|\hat{R}| \leq 1/3\}} + \text{sign}\hat{R}\mathbb{I}_{\{|\hat{R}| > 1/3\}}$$

(où  $\hat{R}$  représente l'estimateur usuel du coefficient de corrélation  $\rho$ ) et l'on conduit, dans chacun des cas, le test relatif à l'hypothèse d'indépendance  $\mathcal{H}_0$ , par le biais d'un test d'ajustement de *Kolmogorov-Smirnov* de la fonction de répartition empirique des "pseudo-observations"  $Z_i = F(X_{1,i}, X_{2,i}; \hat{\gamma})$  ou  $Z_i = C(U_{1,i}, U_{2,i}; \hat{\gamma})$   $i = 1, 2, \dots, 20$  à celle de la loi  $\mathcal{CM}_2$ , dont la densité  $g(z_1)$  et la fonction de répartition  $G(z_1)$  sont représentées respectivement aux figures 3 et 4.

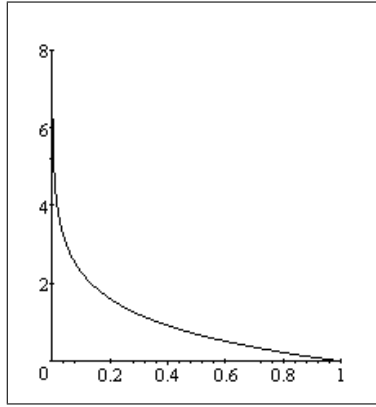


Fig. 3

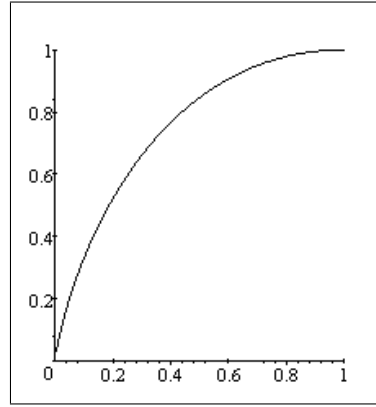


Fig. 4

### 3.2.1 *BiExp*(0)

Dans le cas particulier où les variables aléatoires  $X_1$  et  $X_2$  sont indépendantes, il est clair que le calcul effectué à l'étape 2 de l'algorithme présenté ci-haut donnera :

$$F(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2) = F_{X_1}(F_{X_1}^{-1}(u_1)) F_{X_2}(F_{X_2}^{-1}(u_2)) = u_1 u_2$$

Le test proposé se ramène donc, dans ce cas, au test d'indépendance de deux lois uniformes. Les résultats suivants, illustrent la conduite du test portant sur deux échantillons issus de deux lois uniformes indépendantes.

$u_1$	$u_2$	$u_1 u_2$	$G(z)$
0,92822	0,47147	0,43763	0,79928
0,83890	0,01670	0,01401	0,07380
0,73286	0,95785	0,70197	0,95037
0,73648	0,39511	0,29099	0,65021
0,36888	0,40199	0,14828	0,43131
0,58548	0,93053	0,54481	0,87568
0,80846	0,86955	0,70301	0,95074
0,15569	0,01543	0,00240	0,01689
0,99052	0,79783	0,79027	0,97628
0,78037	0,65491	0,51108	0,85413
0,87207	0,71530	0,62379	0,91818
0,00759	0,35110	0,00266	0,01846
0,85072	0,83041	0,70645	0,95194
0,39270	0,20239	0,07948	0,28074
0,77752	0,71966	0,55955	0,88444
0,04774	0,86494	0,04129	0,17289
0,51430	0,53761	0,27650	0,63194
0,09067	0,49977	0,04531	0,18552
0,18523	0,34174	0,06330	0,23800
0,68743	0,91418	0,62844	0,92035

Utilisant le test d'ajustement de *Kolmogorov-Smirnov* à la loi  $\mathcal{CM}_2$ , on trouve alors, en désignant par  $F_n(\cdot)$  la fonction de répartition empirique des valeurs observées de la variable  $Z_1$  :

$$D_n = \sup_z |F_n(z) - G(z)| = 0,25413$$

Pour  $n = 20$  et pour un risque de première espèce de  $0,1$ , on trouve dans la table de la loi de Kolmogorov-Smirnov (voir G.Saporta[23]) que la valeur critique  $d_n$  est donnée par :  $0,26473$  ce qui, en vertu du fait que  $D_n = 0,25413 < d_n$ , conduit à accepter l'hypothèse d'indépendance  $\mathcal{H}_0$ . La conclusion serait évidemment identique dans le cas d'un risque de première espèce de  $0,05$  dont la valeur critique associée  $d_n$  est égale à  $0,29408$ .

### 3.2.2 *BiExp*(0, 2)

En suivant l'algorithme présenté ci-dessus et en utilisant la première composante du précédent échantillon  $(u_1, u_2)$  afin de générer à l'étape 2 l'observation  $x_1$ , on obtient :

$x_1$	$x_2$	$F(x_1, x_2)$	$G(z)$
2,63410	0,72208	0,48067	0,83035
1,82570	0,01944	0,01666	0,08281
1,32000	3,25230	0,70597	0,95127
1,33360	0,54149	0,31738	0,67064
0,46026	0,49351	0,15476	0,42246
0,88063	2,69830	0,54911	0,87644
1,65270	2,14010	0,71657	0,95431
0,16924	0,01367	0,00246	0,01513
4,65860	1,74890	0,81847	0,98237
1,51580	1,13740	0,53761	0,86658
2,05630	1,36130	0,65279	0,92938
0,00761	0,36967	0,00266	0,01654
1,90190	1,88680	0,72505	0,95711
0,49873	0,21771	0,08433	0,27398
1,50290	1,35080	0,58275	0,89381
0,04891	1,83700	0,04135	0,16919
0,72216	0,77442	0,28964	0,63288
0,09504	0,61454	0,04572	0,17396
0,20485	0,37774	0,06478	0,22392
1,16290	2,5222	0,63541	0,92211

Le test d'ajustement de *Kolmogorov-Smirnov* à la loi  $\mathcal{CM}_2$ , donne :

$$D_n = \sup_z |F_n(z) - G(z)| = 0,35326$$

Pour  $n = 20$ , la valeur expérimentale de  $D_n$  conduit à rejeter l'hypothèse d'indépendance  $\mathcal{H}_0$  que ce soit pour un risque de première espèce de  $0,05$  ( $d_n = 0,29408$ .) où de  $0,1$  ( $d_n = 0,26473$ )

### 3.2.3 *BiExp*(0, 5)

En procédant comme dans l'exemple précédent, on obtient :



$x_1$	$x_2$	$F(x_1, x_2)$	$G(z)$
2,63410	0,85813	0,54283	0,87448
1,82570	0,02527	0,02257	0,10816
1,32000	3,36930	0,71090	0,95348
1,33360	0,60463	0,35821	0,72596
0,46026	0,46423	0,16417	0,46080
0,88063	2,74380	0,55512	0,88185
1,65270	2,28120	0,73298	0,96067
0,16924	0,01157	0,00254	0,01772
4,65860	1,94980	0,85014	0,98816
1,51580	1,24600	0,57345	0,89234
2,05630	1,51090	0,68919	0,94573
0,00761	0,29714	0,00260	0,01849
1,90190	2,04070	0,74735	0,96499
0,49873	0,20596	0,09115	0,30950
1,50290	1,46520	0,61325	0,91312
0,04891	1,56010	0,04148	0,17350
0,72216	0,77905	0,30933	0,67228
0,09504	0,51319	0,04630	0,18857
0,20485	0,32641	0,06674	0,24742
1,16290	2,61570	0,64444	0,92759

Le test d'ajustement de *Kolmogorov-Smirnov* à la loi  $\mathcal{CM}_2$ , donne :

$$D_n = \underset{z}{\text{Sup}} |F_n(z) - G(z)| = 0,32448$$

On obtient donc les mêmes conclusions que dans le cas précédent.

### 3.3 Loi normale $N(\mu, \Sigma)$

On considère un vecteur  $X = {}^t(X_1, X_2)$  de loi normale  $N(0, \Sigma)$  où :

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = I + \rho J$$

Comme précédemment, on simule, pour les valeurs "estimées"  $\hat{\rho} = 0$  ;  $\hat{\rho} = 0,2$  et  $\hat{\rho} = 0,5$  de  $\rho$ , un échantillon de taille  $n = 20$  déduit des lois

$$X_1 \sim N(0, 1) \text{ et } X_2 | X_1 \sim N(\hat{\rho}x_1, 1 - \hat{\rho}^2)$$

On effectue alors, dans chacun des cas, le test relatif à l'hypothèse d'indépendance  $\mathcal{H}_0$ , à l'aide du test d'ajustement à la loi  $\mathcal{CM}_2$  identique au précédent.

#### 3.3.1 $N(0, \Sigma)$

$$\Sigma = I$$

Les valeurs simulées, d'un échantillon de taille 20 issu de cette loi, apparaissent ci-dessous :

$x_1$	$x_2$
-0,43968	-0,36384
+0,23599	+0,95087
+1,07125	-0,34858
-1,34640	+0,01228
-0,53556	+3,01160
-0,28834	+1,74062
-1,47427	+0,43855
+0,19950	+0,46316
+0,07054	-2,06789
-1,42025	-0,96442
+1,16016	-0,58917
-0,04919	+0,36419
+0,58645	+0,88677
+1,21662	+1,83403
-1,90441	+0,75046
-1,03414	-1,15172
+1,23791	+0,04986
+1,87534	+2,02903
+1,84138	-0,36170
-0,57816	-0,97534

On évalue alors les quantités  $\Phi(x_{1,i})$  et  $\Phi(x_{2,i})$  données par :

$$\Phi(x_{1,i}) = \int_{-\infty}^{x_{1,i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad \text{et} \quad \Phi(x_{2,i}) = \int_{-\infty}^{x_{2,i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

et l'on détermine les valeurs de la fonction de répartition de  $Z_1$  aux points  $z_i = \Phi(x_{1,i}) \Phi(x_{2,i})$   $i = 1, 2, \dots, 20$ .  
Les résultats des divers calculs apparaissent dans le tableau suivant :

$\Phi(x_1)$	$\Phi(x_2)$	$z$	$G(z)$
0,33008	0,35799	0,11816	0,37052
0,59328	0,82916	0,49192	0,84091
0,85797	0,36370	0,31204	0,67356
0,08908	0,50490	0,04498	0,18449
0,29613	0,99870	0,29574	0,65603
0,38654	0,95912	0,37073	0,73860
0,07020	0,66951	0,04700	0,19071
0,57906	0,67838	0,39282	0,75987
0,52812	0,01932	0,01020	0,05697
0,07776	0,16742	0,01301	0,06950
0,87701	0,27787	0,24369	0,58775
0,48038	0,64214	0,30847	0,67127
0,72121	0,81240	0,58591	0,89913
0,88813	0,96668	0,85853	0,98948
0,02842	0,77351	0,02198	0,10589
0,15054	0,12472	0,01877	0,09339
0,89213	0,51988	0,46380	0,82014
0,96963	0,97877	0,94904	0,99868
0,96722	0,35879	0,34702	0,71430
0,28158	0,16470	0,04637	0,18878

Utilisant à nouveau le test d'ajustement de *Kolmogorov-Smirnov* à la loi  $\mathcal{CM}_2$ , on trouve alors, en désignant par  $F_n(\cdot)$  la fonction de répartition empirique des valeurs observées de la variable  $Z_1$  :

$$D_n = \underset{z}{\text{Sup}} |F_n(z) - G(z)| = 0,15929$$

Pour  $n = 20$ , la valeur expérimentale de  $D_n$  conduit à accepter l'hypothèse d'indépendance  $\mathcal{H}_0$ , que le risque de première espèce soit de 0,05 ou de 0,1.

### 3.3.2 $N(0, \Sigma)$

$$\Sigma = \begin{bmatrix} 1 & 0,2 \\ 0,2 & 1 \end{bmatrix}$$

Les valeurs simulées  $x_{1,i}$  et  $x_{2,i}$  des composantes  $X_1$  et  $X_2$  ainsi que les valeurs de la fonction  $G(z_i)$ , pour  $i = 1, 2, \dots, 20$ , sont données dans le tableau suivant :

$x_1$	$x_2$	$G(z)$
-0,27933	+0,30869	0,64539
-1,38617	-0,31949	0,17851
+0,49829	-1,76889	0,14169
-0,22600	+1,04751	0,73542
-1,45794	-2,19496	0,01658
-0,17809	+0,73808	0,72171
-0,87554	+0,39967	0,42358
-1,67315	-0,32358	0,11857
+0,65949	-1,45906	0,23384
-0,11952	-0,56950	0,44565
+0,66622	-0,13210	0,72733
+0,16453	-1,03635	0,33675
+0,41950	-0,45069	0,58537
-0,52724	-0,50432	0,36792
+0,70302	-2,03962	0,09166
-1,50072	-0,94460	0,09449
+1,04970	-0,50977	0,63067
-1,71631	+0,69390	0,16085
-0,17169	-0,78138	0,36872
-0,56822	+0,04244	0,47899

où les valeurs  $z_i$  de la fonction de répartition  $F(X_1, X_2)$  sont obtenues à l'aide de la relation :

$$z_i = \int_{-\infty}^{x_{1,i}} \int_{-\infty}^{x_{2,i}} \frac{1}{2\pi\sqrt{0,96}} e^{-\frac{5}{48}(5x^2 - 2xy + 5y^2)} dx dy$$

On conduit alors un test d'ajustement de *Kolmogorov-Smirnov* à la loi  $\mathcal{CM}_2$  et l'on trouve pour valeur expérimentale de  $D_n$  :

$$D_n = \underset{z}{\text{Sup}} |F_n(z) - G(z)| = 0,28548$$

Pour  $n = 20$  et pour un risque de première espèce de 0,1, on rejeterait l'hypothèse d'indépendance  $\mathcal{H}_0$ , alors que pour un risque de première espèce de 0,05 on accepterait (tout juste cependant) cette dernière.

### 3.3.3 $N(0, \Sigma)$

$$\Sigma = \begin{bmatrix} 1 & 0,5 \\ 0,5 & 1 \end{bmatrix}$$

Les valeurs simulées  $x_{1,i}$  et  $x_{2,i}$  des composantes  $X_1$  et  $X_2$  ainsi que les valeurs de la fonction  $G(z_i)$ , pour  $i = 1, 2, \dots, 20$ , apparaissent dans le tableau qui suit :

$x_1$	$x_2$	$G(z)$
-0,21602	+0,24884	0,93471
-1,37514	-0,53185	0,55150
+0,18589	-1,60647	0,55196
-0,04416	+0,96264	0,98993
-1,75070	-2,33336	0,16992
-0,04911	+0,67519	0,98084
-0,77009	+0,23805	0,82326
-1,64967	-0,58274	0,46307
+0,39046	-1,28441	0,68002
-0,20773	-0,56304	0,81609
+0,61414	-0,01698	0,98100
-0,01266	-0,96205	0,74358
+0,32654	-0,36140	0,91897
-0,58572	-0,56760	0,74699
+0,3694	-1,83131	0,48025
-1,58680	-1,14715	0,39661
+0,91826	-0,31461	0,96118
-1,52426	+0,38118	0,59429
-0,29179	-0,77379	0,75449
-0,53530	-0,05253	0,84421

où les valeurs  $z_i$  de la fonction de répartition  $F(X_1, X_2)$  sont obtenues dans ce cas à l'aide de la relation :

$$z_i = \int_{-\infty}^{x_{1,i}} \int_{-\infty}^{x_{2,i}} \frac{1}{2\pi\sqrt{0,75}} e^{-\frac{1}{3}(x^2 - xy + y^2)} dx dy$$

A nouveau le test d'ajustement de *Kolmogorov-Smirnov* à une loi  $\mathcal{CM}_2$  donne pour valeur expérimentale de  $D_n$  : 0,31307 ce qui, avec un risque de première espèce de 0,05 où de 0,1, conduit à rejeter l'hypothèse d'indépendance  $\mathcal{H}_0$ . De fait on peut estimer dans ce cas que la "*p-value*" est d'environ 0,03.

## 4 Test semi-paramétrique d'indépendance entre vecteurs aléatoires

### 4.1 Généralités

Soient  $X_i \in \mathbb{R}^{k_i}$ ,  $i = 1, 2, \dots, n$ ,  $n$  vecteurs aléatoires définis sur les espaces probabilisés  $(\mathbb{R}^{k_i}, \mathcal{B}_{\mathbb{R}^{k_i}}, \mu_i, \nu_i)$ , où  $\mathcal{B}_{\mathbb{R}^{k_i}}$ ,  $\mu_i$  et  $\nu_i$  désignent respectivement la  $\sigma$ -algèbre de Borel de  $\mathbb{R}^{k_i}$ , la mesure de probabilité associée à  $X_i$  et une mesure de référence (en général la mesure de Lebesgue) telle que  $\mu_i$  soit absolument continue par rapport à  $\nu_i \forall i = 1, 2, \dots, n$ , et soit l'espace probabilisé produit :

$$\left( \times_{i=1}^{i=n} \mathbb{R}^{k_i}, \otimes_{i=1}^{i=n} \mathcal{B}_{\mathbb{R}^{k_i}}, \mu, \otimes_{i=1}^{i=n} \mu_i, \otimes_{i=1}^{i=n} \nu_i \right),$$

où  $\mu$  et  $\otimes_{i=1}^{i=n} \mu_i$  désignent la mesure de probabilité conjointe de

$$X = (X_1, X_2, \dots, X_n)$$

(avec  $\mu \ll \otimes_{i=1}^{i=n} \nu_i$ ) et la mesure de probabilité produit. On notera respectivement par  $F(x_1, x_2, \dots, x_n)$  et par  $F_i(x_i)$   $i = 1, 2, \dots, n$ , les fonctions de répartition associées aux mesures  $\mu$  et  $\mu_i$ ,  $i = 1, 2, \dots, n$ . Enfin, pour tout  $i = 1, 2, \dots, n$ ,  $O_i^-(s_i)$  désignera l'orthant négatif de  $\mathbb{R}^{k_i}$  de sommet  $s_i = (s_{i_1}, s_{i_2}, \dots, s_{i_{k_i}})$  c'est-à-dire, l'ensemble des points  $x_i = (x_{i_1}, x_{i_2}, \dots, x_{i_{k_i}})$  de  $\mathbb{R}^{k_i}$  tels que :

$$x_{i_1} \leq s_{i_1}, x_{i_2} \leq s_{i_2}, \dots, x_{i_{k_i}} \leq s_{i_{k_i}},$$

que l'on notera, par commodité d'écriture par :  $x_i \leq s_i$ . On remarquera, de plus, que pour tout  $i = 1, 2, \dots, n$  on a :

$$\mu_i(O_i^-(s_i)) = F_i(s_i)$$

et que pour tout  $\alpha_i \in [0, 1]$ , les ensembles suivants :

$$\{s_i \in \mathbb{R}^{k_i} : \mu_i(O_i^-(s_i)) = \alpha_i\},$$

ne sont autres que les surfaces de niveau,  $F_i^{-1}(\alpha_i)$ , des fonctions  $F_i(x_i)$ .

## 4.2 Relation d'équivalence

Dans le cadre de l'indépendance entre les vecteurs aléatoires  $X_1, X_2, \dots, X_n$ , les hypothèses nulle et alternative  $\mathcal{H}_0$  et  $\mathcal{H}_1$  s'écrivent, le plus souvent, sous la forme :

$$\mathcal{H}_0 : \mu = \otimes_{i=1}^{i=n} \mu_i \quad \text{et} \quad \mathcal{H}_1 : \mu \neq \otimes_{i=1}^{i=n} \mu_i,$$

ou encore, en posant  $k = \sum_{i=1}^{i=n} k_i$ , sous la forme :

$$\begin{aligned} \mathcal{H}_0 & : F(x) = \prod_{i=1}^{i=n} F_i(x_i) \quad \forall x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^k \quad \text{et} \\ \mathcal{H}_1 & : F(x) \neq \prod_{i=1}^{i=n} F_i(x_i). \end{aligned}$$

Notant par  $O^-(x)$  l'orthant négatif de  $\mathbb{R}^k$  de sommet  $x = (x_1, \dots, x_n)$ , on peut également écrire :

$$\begin{aligned} \mathcal{H}_0 & : \mu(O^-(x)) = \prod_{i=1}^{i=n} \mu_i(O_i^-(x_i)) \quad \forall x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^k \quad \text{et} \\ \mathcal{H}_1 & : \mu(O^-(x)) \neq \prod_{i=1}^{i=n} \mu_i(O_i^-(x_i)). \end{aligned}$$

Comme dans la pratique il est en général exclu de vérifier l'égalité entre deux mesures de probabilité en tout point  $x$  de leur support commun, cela suggère de restreindre la vérification, que l'on espère plus facile à mettre en oeuvre, à un sous-ensemble de points de ce dernier. Dans le cas présent, celui-ci est construit à l'aide de la relation d'équivalence définie de la façon suivante : soient, pour tout  $i$ , deux points  $s_i$  et  $t_i$  appartenant à  $\mathbb{R}^{k_i}$  et soient  $O_i^-(s_i)$  et  $O_i^-(t_i)$  les orthants négatifs de sommets respectifs  $s_i$  et  $t_i$ . On dira que  $s_i$  et  $t_i$  sont équivalents si  $\mu_i(O_i^-(s_i)) = \mu_i(O_i^-(t_i)) = \alpha_i$  où  $\alpha_i$  désigne la valeur commune de ces deux probabilités. Cette égalité définit une relation d'équivalence entre les points de l'espace  $\mathbb{R}^{k_i}$  et pour tout  $\alpha_i \in [0, 1]$ , les classes d'équivalences  $\mathcal{C}_{i,\alpha_i}$ , dont on désignera un représentant quelconque par  $s_{i,\alpha_i}$ , sont les surfaces de niveau,  $F_i^{-1}(\alpha_i)$ , de la fonction  $F_i(x_i)$ .

Ainsi, pour que la condition  $\mu(O^-(x)) = \prod_{i=1}^{i=n} \mu_i(O_i^-(x_i))$  soit vérifiée pour tout point  $x = (x_1, x_2, \dots, x_n)$  de  $\mathbb{R}^k$ , il faut et il suffit qu'elle le soit pour tout point  $x$  de la forme  $x = (s_{i,\alpha_i})_{i=1,2,\dots,n}$  et ce pour tout  $\alpha_i$  appartenant à l'intervalle  $[0, 1]$ . La condition relative à  $\mathcal{H}_0$  devient alors :

$$\begin{aligned} \mu(O^-(s_{1,\alpha_1}, s_{2,\alpha_2}, \dots, s_{n,\alpha_n})) & = \prod_{i=1}^{i=n} \mu_i(O_i^-(s_{i,\alpha_i})), \\ & = \prod_{i=1}^{i=n} \alpha_i \quad \forall (\alpha_1, \alpha_2, \dots, \alpha_n) \in [0, 1]^n. \end{aligned}$$

Reste alors à régler, afin de préciser la procédure du test, la question du choix d'un représentant de chacune des classes d'équivalence. Une façon commode de choisir un représentant  $s_{i,\alpha_i}$  par classe d'équivalence  $\mathcal{C}_{i,\alpha_i}$ , consiste à ne retenir, pour tout  $i$  et pour tout  $\alpha_i$ , que l'orthant négatif  $O_i^- (\tilde{s}_{i,\alpha_i})$  pour lequel le sommet  $\tilde{s}_{i,\alpha_i}$  possède toutes ses coordonnées égales, dont la valeur commune sera notée par  $s_{\alpha_i}$ .

Il s'ensuit que :

$$\begin{aligned} \mu_i (O_i^- (\tilde{s}_{i,\alpha_i})) &= \mu_i (x_i \leq \tilde{s}_{i,\alpha_i}) = \mathbb{P} (X_{i_1} \leq s_{\alpha_i}, X_{i_2} \leq s_{\alpha_i}, \dots, X_{i_{k_i}} \leq s_{\alpha_i}) \\ &= F_{V_i} (s_{\alpha_i}) \end{aligned}$$

où  $F_{V_i} (\cdot)$  désigne, pour tout  $i$ , la fonction de répartition de la variable aléatoire  $V_i = \sup (X_{i_1}, X_{i_2}, \dots, X_{i_{k_i}})$ . Ce choix s'avère particulièrement intéressant, puisque il permet de ramener le test d'indépendance d'une famille de vecteurs aléatoires (ou de blocs d'un vecteur aléatoire)  $(X_1, X_2, \dots, X_n)$ , à un test d'indépendance entre les composantes du vecteur aléatoire  $(V_1, V_2, \dots, V_n)$ . En effet, la condition portant sur l'hypothèse nulle  $\mathcal{H}_0$  s'écrit simplement de la façon suivante :

$$\mu (O^- (\tilde{s}_{1,\alpha_1}, \tilde{s}_{2,\alpha_2}, \dots, \tilde{s}_{n,\alpha_n})) = \prod_{i=1}^{i=n} \mu_i (O_i^- (\tilde{s}_{i,\alpha_i})) \quad \forall (\alpha_1, \alpha_2, \dots, \alpha_n) \in [0, 1]^n$$

c'est-à-dire :

$$\begin{aligned} F (s_{\alpha_1}, s_{\alpha_2}, \dots, s_{\alpha_n}) &= \mathbb{P} (X_1 \leq \tilde{s}_{1,\alpha_1}, X_2 \leq \tilde{s}_{2,\alpha_2}, \dots, X_n \leq \tilde{s}_{n,\alpha_n}) , \\ &= \mathbb{P} (V_1 \leq s_{\alpha_1}, V_2 \leq s_{\alpha_2}, \dots, V_n \leq s_{\alpha_n}) , \\ &= \prod_{i=1}^{i=n} F_{V_i} (s_{\alpha_i}) \quad \forall (s_{\alpha_1}, s_{\alpha_2}, \dots, s_{\alpha_n}) \in \mathbb{R}^n . \end{aligned}$$

La vérification de l'hypothèse nulle  $\mathcal{H}_0$  s'effectuera seulement sur ce sous-ensemble de points de  $\mathbb{R}^k$ , qui n'est autre que la trace, sur la surface  $F(x)$ , du premier hyperplan bissecteur de  $\mathbb{R}^{k+1}$ .

Ainsi, par le biais de la relation d'équivalence entre orthants négatifs, telle qu'introduite ci-dessus, le test d'indépendance entre vecteurs aléatoires se ramène à un test d'indépendance entre les composantes du vecteur

$$V = (V_1, V_2, \dots, V_n) ,$$

où  $V_i = \sup (X_{i_1}, X_{i_2}, \dots, X_{i_{k_i}})$  pour tout  $i = 1, 2, \dots, n$ . Bien entendu, tout autre choix d'un représentant par classe d'équivalence peut convenir, le critère du choix demeurant essentiellement, la facilité de la conduite des calculs.

### 4.3 Principe et procédure du test

Le principe du test d'indépendance entre les vecteurs aléatoires  $X_1, X_2, \dots, X_n$ , repose sur la comparaison entre les mesures de probabilité conjointe et produit et consiste à rejeter l'hypothèse d'indépendance si "l'écart" entre les deux mesures dépasse un certain seuil (test unilatéral à droite). Il découle des précédents résultats que cette comparaison se ramène à l'étude de l'écart entre la fonction de répartition  $F_U(u)$  de la variable aléatoire  $U = G_n(Z_1)$  (qui, dans le cas présent, joue le rôle de "pseudo-observations"), avec celle d'une loi uniforme sur l'intervalle  $[0, 1]$ . Plus précisément, étant donné un risque  $\alpha$  de première espèce, on rejettera l'hypothèse  $\mathcal{H}_0$  si :

$$\sup_{0 < u < 1} | \hat{F}_{U,m}(u) - u | \geq d_\alpha$$

où  $\hat{F}_{U,m}(u)$  désigne la fonction de répartition empirique de  $U$ , obtenue à l'aide d'un échantillon de taille  $m$  des vecteurs aléatoires  $X_1, X_2, \dots, X_n$ , ce qui revient à utiliser un test de type *Kolmogorov-Smirnov*, bien que cela ne soit pas, dans le contexte des tests d'adéquation, l'unique test que l'on puisse utiliser. On reviendra brièvement sur cette question par la suite.

Le fait que le vecteur aléatoire  $V = (V_1, V_2, \dots, V_n)$  suive, sous l'hypothèse nulle  $\mathcal{H}_0$ , une loi  $\mathcal{CM}_n$  à  $n$  degrés de liberté permet alors de construire, dans un cadre semi-paramétrique, le test d'indépendance recherché, dont la procédure peut être décrite brièvement comme suit : on suppose que  $F(X) \in \Psi(X, \theta)$  où  $\Psi(X, \theta)$  désigne une famille de lois indicées par le paramètre  $\theta$ . On estime alors  $\theta$  par  $\hat{\theta}$  à l'aide d'un échantillon de taille  $m$  et l'on en déduit une estimation  $\hat{F}(X) = \Psi(X, \hat{\theta})$  de  $F(X)$ . À l'aide de cette dernière, on obtient une estimation  $\hat{F}_V$  du vecteur aléatoire  $V$  et l'on compare alors, à l'aide d'un test de type *Kolmogorov-Smirnov*, la fonction de répartition empirique de la variable aléatoire  $(G_n \circ \hat{F}_V)(V_1, V_2, \dots, V_n)$  à la fonction de répartition d'une loi uniforme sur  $[0, 1]$ .

## 4.4 Exemples et simulations de la puissance du test

### 4.4.1 Exemple 1

Dans le but de comparer le test décrit ci-dessus à des tests classiques, on se situera dans un cadre semi-paramétrique normal. Afin d'illustrer la procédure de test, on utilise dans l'exemple qui suit, un ensemble de données présenté à l'annexe 1 et tiré du manuel de K.V. Mardia, J.T. Kent et J.M. Bibby[16] (pages 3 et 4). Plus précisément, on dispose d'un ensemble de 88 étudiants pour lesquels on note les résultats (sur 100) obtenus lors d'examens, sur deux ensembles de matières  $X_1$  (examens sans documentation) et  $X_2$  (examens avec documentation), tels que décrits ci-dessous :

$$\begin{aligned} X_1 &= \begin{cases} X_{1,1} = \text{Mécanique} \\ X_{1,2} = \text{Calcul vectoriel} \end{cases} \\ X_2 &= \begin{cases} X_{2,1} = \text{Algèbre} \\ X_{2,2} = \text{Analyse} \\ X_{2,3} = \text{Statistique} \end{cases} \end{aligned}$$

Les auteurs supposent que les variables associées aux résultats des examens sur l'ensemble des 5 matières suivent une loi normale  $\mathcal{N}_5(\mu, \Sigma)$ , dont le vecteur moyen et la matrice de variance-covariance sont estimés à l'aide des estimateurs habituels.

Le test d'indépendance entre les vecteurs  $X_1$  et  $X_2$  a été effectué sur des données centrées et réduites et la matrice  $\hat{R}$  des corrélations empiriques apparaît à l'annexe 1. Le résultat numérique du test donne une "p-value" de  $2,3 \times 10^{-4}$  (voir l'extrait du programme en langage **R**[20] présenté également à l'annexe 1) et conduit donc au rejet de l'hypothèse nulle  $\mathcal{H}_0$  d'indépendance entre les vecteurs  $X_1$  et  $X_2$ . A titre indicatif, les auteurs utilisent ces données afin de déterminer en particulier les corrélations canoniques  $\rho_1$  et  $\rho_2$  entre les vecteurs  $X_1$  et  $X_2$  et trouvent que ces dernières ont pour valeurs respectives 0,6630 et 0,0412. A des fins de comparaison, on présente ci-dessous le test du rapport de vraisemblance généralisé habituellement utilisé dans le cadre de la loi normale et tel que présenté dans, entre autres, Mardia, Kent et Bibby[16], Kshirsagar[13] et Bilodeau et Brenner[3]. Suivant cette approche, on suppose que la loi conjointe des vecteurs aléatoires  $X_1 \in \mathbb{R}^{p_1}$  et  $X_2 \in \mathbb{R}^{p_2}$  est une loi normale  $\mathcal{N}_{p_1+p_2}(\mu, \Sigma)$  avec :

$$\Sigma = \left[ \begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right]$$

L'hypothèse nulle d'indépendance entre les vecteurs  $X_1$  et  $X_2$  s'exprime alors sous la forme :

$$\mathcal{H}_0 : \Sigma_{12} = 0$$

où  $\Sigma_{12}$  désigne la matrice des covariances entre les composantes de  $X_1$  et celles de  $X_2$ . Le rapport de vraisemblance généralisé étant désigné par  $\lambda$ , la statistique de test utilisée est donnée par :

$$\lambda^{2/n} = \prod_{i=1}^k (1 - \lambda_i)$$

où  $n$  désigne le nombre d'observations du vecteur  $X = (X_1, X_2)$ , où  $k = \min(p_1, p_2)$  et où  $\lambda_i$  désignent les valeurs propres non nulles de la matrice  $(\mathbb{I} - S_{22}^{-1} S_{21} S_{11}^{-1} S_{12})$ ,  $S_{ij}$   $i, j = 1, 2$  étant l'estimateur du maximum

de vraisemblance  $\hat{\Sigma}_{ij}$  de  $\Sigma_{ij}$ . Sous l'hypothèse  $\mathcal{H}_0$  et si  $n - 1 > \min(p_1, p_2)$  la variable aléatoire  $\lambda^{2/n}$  suit une loi de *Wilks*  $\Lambda(p_2, n - 1 - p_1, p_1)$  de paramètres  $p_1, p_2$  et  $n$ . Pour des valeurs suffisamment grandes de  $n$ , on peut utiliser en particulier l'approximation de *Bartlett* donnée dans le cas présent par :

$$- \left( n - \frac{1}{2}(p_1 + p_2 + 3) \right) \text{Log} | I - S_{22}^{-1} S_{21} S_{11}^{-1} S_{12} | \simeq \chi_{p_1 p_2}^2$$

On trouve alors après calculs (voir annexe 1) que la valeur du  $\chi_e^2$  empirique est de 48,7. Pour une loi du *chi-deux* à 6 degrés de liberté et pour un risque  $\alpha$  de première espèce de 0,1 on trouve dans la table  $\chi_6^2$ , un  $\chi_c^2$  critique de 10,6. On rejette donc également, avec le test du rapport de vraisemblance, l'hypothèse d'indépendance entre les vecteurs  $X_1$  et  $X_2$ . Il existe également dans ce contexte de normalité, un autre test connu sous le nom de test "d'union-intersection". Ce dernier fait intervenir la plus grande valeur propre  $\lambda_{\max}$  de la matrice  $S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}$  qui dans le cas présent a pour valeur 0,43958. A l'aide d'une table (voir Mardia, Kent et Bibby page 496[16]), l'on compare pour un risque  $\alpha$  de première espèce de 0,1, et pour des degrés de liberté donnés par :  $\nu_1 = n - p_1 - 1 = 85$  et  $\nu_2 = p_2 = 3$ , cette valeur empirique à la valeur critique 0,100, ce qui, de nouveau, entraîne le rejet de l'hypothèse nulle  $\mathcal{H}_0$ .

#### 4.4.2 Exemple 2

Les données traitées dans cet exemple sont présentées à l'annexe 2 et proviennent du manuel de W. Mendenhall et T. Sincich[19] (page 250 et suivantes). Elles se rapportent au nombre quotidien de sept types de transactions courantes, effectuées habituellement dans les grands magasins ou les grandes surfaces par le personnel de bureau, et regroupées en trois catégories distinctes  $X, Y$  et  $Z$  comprenant respectivement les variables  $(X_1, X_2)$ , les variables  $(X_3, X_4)$  et les variables  $(X_5, X_6, X_7)$ , dont les descriptions sont données ci-dessous :

- $X_1$  = Nombre de courriers de type courant traités.
- $X_2$  = Nombre de courriers divers traités suivant la disponibilité.
- $X_3$  = Nombre de mandats et de certificats cadeaux traités.
- $X_4$  = Nombre de modifications d'achats transigées.
- $X_5$  = Nombre de paiements effectués au guichet.
- $X_6$  = Nombre de chèques encaissés.
- $X_7$  = Nombre de traites bancaires.

On admet en général dans la pratique, en raison de la nature différente des transactions entre les catégories, qu'il y a indépendance entre les vecteurs aléatoires  $X = (X_1, X_2)$ ,  $Y = (X_3, X_4)$  et  $Z = (X_5, X_6, X_7)$ . Le test, présenté à l'annexe 2 et effectué sur des données centrées et réduites dans le cadre d'une famille de lois normales, conduit à accepter l'hypothèse d'indépendance entre les trois vecteurs identifiés ci-dessus. Il est par ailleurs intéressant de remarquer que le même test, conduit sur l'ensemble des sept variables entraîne, comme on pourra le vérifier également à l'annexe 2, le rejet de l'hypothèse d'indépendance entre les transactions prise dans leur ensemble. A des fins de comparaisons, on considère ci-après le test classique du rapport de vraisemblance tel que décrit, entre autres, dans Bilodeau[3], Kshirsagar[13], Anderson[1] et Giri[10]. Adoptant l'approche telle que préconisée par exemple dans Anderson[1], on considère dans le cas présent un vecteur aléatoire  $X$  suivant une loi normale  $\mathcal{N}_7(\mu, \Sigma)$  où la matrice  $\Sigma$  de variance-covariance s'écrit sous la forme :

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix}$$

L'hypothèse d'indépendance  $\mathcal{H}_0$  entre les blocs de variables associées aux matrices de variance-covariance  $\Sigma_{ii}$  pour  $i = 1, 2, 3$ , s'écrit :  $\Sigma_{ij} = 0 \forall i \neq j = 1, 2, 3$ . Si  $\hat{\Sigma}_{ij} \forall i, j = 1, 2, 3$ , désigne l'estimateur du maximum de vraisemblance de  $\Sigma_{ij}$ , alors le critère  $\lambda$  du rapport du maximum de vraisemblance s'exprime sous la forme :

$$\lambda = \frac{|\hat{\Sigma}|^{n/2}}{\prod_{i=1}^3 |\hat{\Sigma}_{ii}|^{n/2}},$$



où encore :

$$\lambda^{2/n} = \frac{|\hat{\Sigma}|}{\prod_{i=1}^3 |\hat{\Sigma}_{ii}|},$$

où  $n$  désigne la taille de l'échantillon considéré ( $n = 48$  dans le présent exemple). La région critique est alors de la forme :

$$\lambda \leq \lambda_\alpha$$

où  $\alpha$  désigne l'erreur de première espèce. Par ailleurs, on peut également écrire l'expression :

$$\frac{|\hat{\Sigma}|}{\prod_{i=1}^3 |\hat{\Sigma}_{ii}|}$$

sous la forme :

$$\frac{|\hat{R}|}{\prod_{i=1}^3 |\hat{R}_{ii}|}$$

où  $\hat{R}$  et  $\hat{R}_{ii}$   $i = 1, 2, 3$ , sont les matrices des corrélations empiriques. Notant par  $p_1 = p_2 = 2$  le nombre de variables des blocs  $X$  et  $Y$  et par  $p_3 = 3$  le nombre de variables du bloc  $Z$ , on montre que (Bilodeau[3], Kshirsagar[13]) dans le cas présent, la loi exacte de la statistique de test  $\lambda^{2/n}$  sous l'hypothèse nulle  $\mathcal{H}_0$  d'indépendance, est donnée par le résultat suivant :

La variable aléatoire  $\lambda^{2/n}$  a pour loi, celle de la variable aléatoire  $\prod_{i=2}^3 \prod_{j=1}^{p_i} Y_{ij}$  où  $Y_{ij}$  sont des variables aléatoires indépendantes et de lois *Bêta* données par :

$$Y_{ij} \sim \beta \left( \frac{1}{2}(n - \bar{p}_i + 1 - j), \frac{1}{2}\bar{p}_i \right)$$

avec  $\bar{p}_i = p_1 + p_2 + \dots + p_i$ . En d'autres termes on a :

$$\lambda^{2/n} \stackrel{d}{=} \prod_{i=2}^3 \prod_{j=1}^{p_i} \beta \left( \frac{1}{2}(n - \bar{p}_i + 1 - j), \frac{1}{2}\bar{p}_i \right)$$

La difficulté que l'on rencontre dans la manipulation de cette loi, fait que l'on a souvent recourt au résultat asymptotique suivant [voir Kshirsagar[13]] :  $-2\text{Log}\lambda$  suit asymptotiquement un  $\chi^2$  à  $\frac{1}{2}(p^2 - \sum_{i=1}^k p_i^2)$  avec dans le cas présent :  $p = 7; k = 3; p_1 = p_2 = 2$  et  $p_3 = 3$ . On aura donc :  $-2\text{Log}\lambda \simeq \chi_{16}^2$  dont la valeur du  $\chi_c^2$  critique pour un risque de première espèce de 0,1 est de 23,50. Les calculs présentés à l'annexe 2 montrent que  $-2\text{Log}\lambda$  est égal à 53,08 et comme l'on rejette l'hypothèse nulle si  $\lambda \leq \lambda_\alpha$  il semble donc que le test du rapport de vraisemblance rejette l'hypothèse d'indépendance. Malgré le comportement différent des deux tests, il convient de mettre en perspective les deux résultats. Dans le premier test, on utilise une loi exacte, à savoir la loi de la statistique de *Kolmogorov*, alors que dans le second cas on utilise une loi approximative pour un échantillon de taille 48 dans un espace euclidien à 7 dimensions et qui n'est pas nécessairement très précise dans ce cas. Il faudrait donc, afin d'avoir une idée plus réelle de la situation, déterminer la loi exacte de la variable aléatoire  $\lambda$  ce qui, a priori, ne semble pas être une tâche aisée. Enfin, comme on doit s'y attendre, les deux tests rejettent l'hypothèse d'indépendance entre les composantes du vecteur  $(X_1, X_2, \dots, X_7)$ . On dispose également d'une autre approximation telle que présentée dans Giri[10]. Posant dans le cas de  $k$  vecteurs aléatoires  $X_i \in R^{p_i}$   $i = 1, 2, \dots, k$  :

$$V = \frac{|\hat{\Sigma}|}{\prod_{i=1}^k |\hat{\Sigma}_{ii}|},$$

on montre que pour des tailles  $n$  d'échantillons suffisamment grandes on a :

$$\mathbb{P}(-a \log V \leq z) \simeq \mathbb{P}(\chi_f^2 \leq z)$$

avec :

$$\begin{aligned} a &= \rho n, \quad f = \frac{1}{2}(p^2 - \sum_{i=1}^k p_i^2), \quad \sum_{i=1}^k p_i = p \quad \text{et} \\ \rho &= 1 - \frac{2(p^3 - \sum_{i=1}^k p_i^3) + 18f}{12nf} \end{aligned}$$

On aura donc dans le présent exemple :  $n = 48$ ,  $f = 16$ ,  $\rho = 0,9037$  et  $a = 43,375$ . Après calculs, on trouve une valeur expérimentale pour la statistique  $-a \text{Log} V$  de 47,5 ce qui entraîne de nouveau le rejet de l'hypothèse nulle d'indépendance, mais avec cependant un écart moins prononcé que lors de l'approximation précédente.

**Remarque 7 :** Dans le même sens que la remarque 6, il convient de préciser à ce stade qu'il sera le plus souvent très difficile, voire impossible, de comparer le test proposé à des tests usuels, puisque ces derniers s'utilisent dans un contexte le plus souvent paramétrique, alors que celui que l'on présente ici, exploite la propriété selon laquelle, sous l'hypothèse nulle d'indépendance  $\mathcal{H}_0 : F(X_1, X_2, \dots, X_n) = \prod_{i=1}^n F_{X_i}(x_i)$ , la variable aléatoire

$$U = G_n(Z_1),$$

suit une loi uniforme sur l'intervalle  $[0, 1]$  et ce quelle que soit la fonction de répartition  $F$ . De plus l'extrême complexité de l'hypothèse alternative rend utopique le recherche d'un test non seulement qui soit uniformément le plus puissant, mais également qui soit sans biais et, plus souvent qu'autrement, on ne peut dans ce cas là que se restreindre à la considération d'alternatives locales suivant une "direction privilégiée" propre d'ailleurs en général à chacun des tests concurrents, ce qui rend vain tout espoir raisonnable de les comparer. Sauf dans des cas qui restent exceptionnels. On pourra consulter à ce sujet Lehmann and Romano[15].

**Remarque 8 :** Dans le cadre semi-paramétrique, la fonction de répartition  $F(X, \theta)$  dépend d'un paramètre  $\theta \in \mathbb{R}^p$  et si l'on estime ce dernier par  $\hat{\theta}$ , alors  $\hat{F} = F(X, \hat{\theta})$  peut être vue comme la réalisation d'une fonction aléatoire appartenant à l'espace fonctionnel  $C[0, 1]$  des fonctions continues sur l'intervalle  $[0, 1]$  muni le plus souvent de la topologie uniforme. Ainsi de façon naturelle, le test considéré peut être associé à un processus stochastique. En particulier, le fait que le test proposé puisse être considéré comme un pont brownien, permettra d'en déduire certaines propriétés asymptotiques concernant la puissance de ce dernier.

#### 4.4.3 Simulations 1

Les figures 1 et 2 ci-dessous illustrent une étude empirique de la puissance du test dans le cas de deux vecteurs normaux respectivement de  $\mathbb{R}^2$  et de  $\mathbb{R}^3$  et de deux vecteurs log-normaux, également de  $\mathbb{R}^2$  et de  $\mathbb{R}^3$  et ce pour un risque  $\alpha$  de première espèce de 0,1 et pour une structure de dépendance donnée, dans les deux cas, par la matrice de variance-covariance  $\Sigma$  définie par :

$$\Sigma = \left[ \begin{array}{cc|ccc} 1 & -0,5 & r & r & r \\ -0,5 & 1 & r & r & r \\ \hline r & r & 1 & 0,5 & 0,5 \\ r & r & 0,5 & 1 & 0,5 \\ r & r & 0,5 & 0,5 & 1 \end{array} \right] \text{ où } |r| \leq 0,4$$

et où les tailles  $m$  d'échantillons sont  $m = 20, 25, 30, 35, 40, 45, 50, 75, 100, 200$ . Dans les deux cas, l'hypothèse nulle correspond à  $r = 0$  et les hypothèses alternatives correspondent aux différentes valeurs de  $r \neq 0$  et variant de 5/100 en 5/100 de  $-0,4$  à  $0,4$ . Cette étude empirique permet de constater, tout au moins dans les cas considérés, que le test semble bien se comporter dans le sens où il est sans biais et qu'une puissance respectable est rapidement atteinte dès que  $|r| \geq 0,2$  pour autant que les échantillons soient de tailles  $m \geq 100$ .

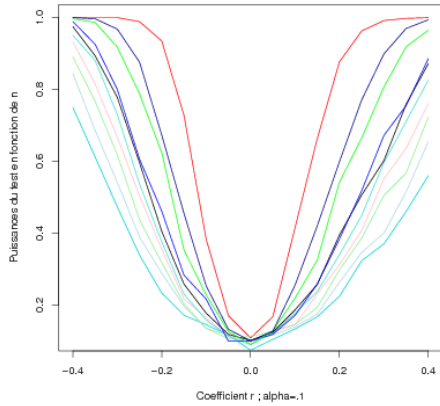


Fig.1 Vecteurs normaux

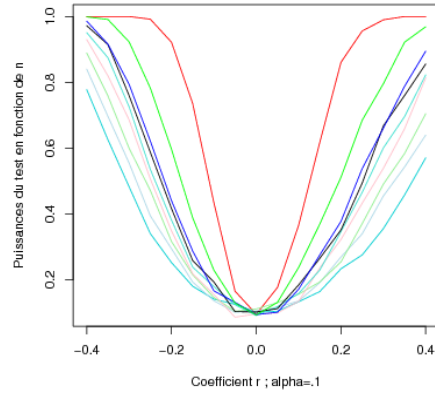


Fig.2 Vecteurs log-normaux

On trouvera, dans les deux cas, à l'annexe 3 un exemple de simulation (en langage  $\mathbf{R}$ [20]) pour une taille d'échantillon de 200.

#### 4.4.4 Simulation 2

La figure 3, quant à elle illustre, dans les mêmes conditions de simulation que précédemment, l'étude empirique de la puissance du test dans le cas de trois vecteurs normaux de  $\mathbb{R}^2$  et pour une structure de dépendance donnée par la matrice de variance-covariance  $\Sigma$  suivante :

$$\Sigma = \begin{bmatrix} 1 & -0,5 & 0 & r & r & 0 \\ -0,5 & 1 & r & 0 & 0 & r \\ \hline 0 & r & 1 & 0,5 & r & 0 \\ r & 0 & 0,5 & 1 & 0 & 0 \\ \hline r & 0 & r & 0 & 1 & 0 \\ 0 & r & 0 & 0 & 0 & 1 \end{bmatrix}$$

On constate alors que les précédentes conclusions sont également valables dans ce cas.

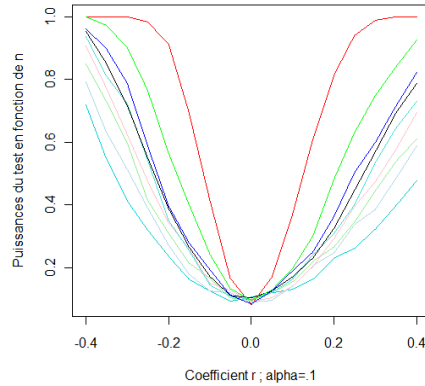


Fig.3 Cas de 3 vecteurs normaux

On trouvera également à l'annexe 3, un exemple de simulation pour une taille d'échantillon de 200.

## 5 Propriétés asymptotiques

### 5.1 Test et pont brownien

Puisque le test d'indépendance entre vecteurs aléatoires se réduit à celui d'un test d'indépendance entre les composantes d'un vecteur aléatoire, il suffit, aux fins de la présente étude, de se restreindre à ce dernier cas. On supposera pour la suite que l'on dispose de  $n$  observations indépendantes du vecteur aléatoire  $X = (X_1, X_2, \dots, X_k)$ . Désignant par  $F_U(u)$  la fonction de répartition inconnue de la variable aléatoire  $U = G_k(F(X_1, X_2, \dots, X_k))$  et par  $\hat{F}_{U,n}(u)$  sa version empirique, la procédure de test revient à considérer la quantité  $\Delta_n$  définie par :

$$\Delta_n = \sup_{0 \leq u \leq 1} |\hat{F}_{U,n}(u) - u|$$

et à rejeter l'hypothèse d'indépendance pour les valeurs de  $\Delta_n$  excédant un certain seuil  $\delta_\alpha$ . En d'autres termes, le test proposé peut, à l'aide de la variable aléatoire  $U = (G_k \circ F)(X)$  à valeurs dans  $[0, 1]$ , se formuler comme suit :

$$\mathcal{H}_0 : U \text{ suit une loi uniforme } \mathcal{U}_{[0,1]}$$

$$\mathcal{H}_1 : U \text{ ne suit pas une loi uniforme } \mathcal{U}_{[0,1]}$$

et se ramène donc à un test d'adéquation à la loi  $\mathcal{U}_{[0,1]}$  de type *Kolmogorov-Smirnov*. A ce sujet, puisque le problème de test tel que posé, se ramène à un test d'adéquation à une loi, rien ne s'opposerait à ce que l'on en choisisse un autre. En particulier dans le cas présent (la fonction de répartition  $F_0(u)$  sous  $\mathcal{H}_0$  n'est autre que  $u$ ), on peut penser à la famille de tests de *Cramer-von Mises* pour lesquels on rejette l'hypothèse nulle  $\mathcal{H}_0$ , si la quantité  $\Gamma_{n,\psi}$  donnée par ( $\psi$  : fonction positive définie sur  $[0, 1]$ ) :

$$\Gamma_{n,\psi} = \int_0^1 \left[ \hat{F}_{U,n}(u) - u \right]^2 \psi(u) du ,$$

dépasse, pour un risque  $\alpha$  de première espèce donné, une certaine valeur  $\gamma_\alpha$  ( $\psi(u) = 1$  correspond au test de *Cramer-von Mises*, alors que  $\psi(u) = [u(1-u)]^{-1}$  correspond au test d'*Anderson-Darling*). De fait, le

choix du test de *Kolmogorov-Smirnov* s'impose de manière naturelle dans la mesure où le test proposé est étroitement relié à la distance en variation entre deux mesures de probabilité  $\mathbb{P}$  et  $\mathbb{Q}$ . En effet, soient  $\mathbb{P}$  et  $\mathbb{Q}$  deux mesures de probabilité, de fonctions de répartition respectives  $F$  et  $G$ , définies sur le même espace probabilisable  $(\Omega, \mathcal{F})$ .

On appelle distance en variation entre les mesures de probabilité  $\mathbb{P}$  et  $\mathbb{Q}$ , la quantité, notée  $d_v(\mathbb{P}, \mathbb{Q})$ , définie par :

$$d_v(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$$

Dans le cas où  $\Omega = \mathbb{R}^k$ , on considère la famille  $\mathcal{F}_0$  des parties de  $\Omega$ , constituée de l'ensemble  $\{O^-(x) : \forall x \in \mathbb{R}^k\}$  des orthants négatifs (ensemble de parties de  $\mathbb{R}^k$  constituant un  $\pi$ -système qui engendre  $\mathcal{B}_{\mathbb{R}^k}$ ), la restriction de la distance en variation à cette dernière est alors donnée par :

$$\begin{aligned} d_v(\mathbb{P}, \mathbb{Q}) &= \sup_{O^-(x) \in \mathcal{F}_0} |\mathbb{P}(O^-(x)) - \mathbb{Q}(O^-(x))| , \\ &= \sup_{x \in \mathbb{R}^k} |F(x) - G(x)| , \end{aligned}$$

où  $\mathbb{P}(O^-(x)) = F(x)$  et  $\mathbb{Q}(O^-(x)) = G(x)$  représentent les probabilités du même orthant négatif, selon la mesure  $\mathbb{P}$  ou selon la mesure  $\mathbb{Q}$ . La mesure de leur distance correspond alors, tout au moins dans l'esprit, au test d'indépendance proposé, lequel, par le truchement de la relation d'équivalence entre orthants négatifs, conduit naturellement au test de *Kolmogorov-Smirnov* entre les fonctions de répartition  $F$  et  $G$ .

Posant alors, pour tout  $0 \leq u \leq 1$  :

$$Y_n(u) = \sqrt{n}(\hat{F}_{U,n}(u) - u) ,$$

il vient :

$$\sqrt{n}\Delta_n = \sqrt{n} \sup_{0 \leq u \leq 1} |\hat{F}_{U,n}(u) - u| = \sup_{0 \leq u \leq 1} |Y_n(u)| .$$

Le test proposé s'exprime donc naturellement sous la forme d'un "pont brownien"  $W^0(u)$  pour lequel, sous l'hypothèse nulle  $\mathcal{H}_0$ , on a (voir Billingsley [2], Le Cam[14]) :

$$\sqrt{n}\Delta_n \xrightarrow{\mathcal{L}} \sup_{0 \leq u \leq 1} |W^0(u)| ,$$

avec :

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(\sqrt{n}\Delta_n \leq d) &= \mathbb{P}\{x(\cdot) : \sup_{0 \leq u \leq 1} |x(u)| \leq d\} , \\ &= 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 d^2} \quad d > 0 . \end{aligned}$$

Il s'ensuit alors que le comportement asymptotique du test considéré dans le cas présent, découlera directement des propriétés similaires d'un pont brownien, dont on rappelle ici quelques propriétés élémentaires utiles pour l'étude de la puissance du test et telles qu'on les trouvera exposées dans Billingsley[2] ou encore dans Serfling[24].

On dit que le processus stochastique  $\{W(t) : t > 0\}$ , défini sur l'espace probabilisé  $(\Omega, \mathcal{F}, \mathbb{P})$ , est un mouvement *Brownien* ou encore un processus de *Wiener*, si les conditions suivantes sont remplies :

- i)  $W(0) = 0$ ,
- ii) Les accroissements  $W(t) - W(s)$  pour  $s \leq t$  sont indépendants et stationnaires,
- iii)  $W(t)$  suit une loi normale  $\mathcal{N}(0, c^2 t)$ . Lorsque  $c = 1$ , on dit que le mouvement brownien est "standard".

On montre alors que les trajectoires associées à ce mouvement brownien sont presque sûrement continues et presque sûrement non-différentiables. Par ailleurs, on montre également que le mouvement brownien

constitue un cas particulier d'un processus plus général, dit processus *Gaussien*  $\{X(t) : t \geq 0\}$ , caractérisé par la propriété suivant laquelle le vecteur de dimension finie  $(X(t_1), X(t_2), \dots, X(t_n))$  suit une loi normale pour tout  $t_1, t_2, \dots, t_n$ .

Soit  $\{W(t) : t > 0\}$  un mouvement brownien, pour lequel on considère les valeurs du processus sur l'intervalle  $[0, 1]$  sachant que  $W(1) = 0$ . En d'autres termes, on considère le processus défini par :

$$\{W(t) : 0 \leq t \leq 1 | W(1) = 0\}$$

On peut alors vérifier que ce processus conditionnel, appelé "pont brownien" (appellation découlant du fait que  $W(0) = W(1) = 0$ ) est également un processus *Gaussien* de moyenne nulle et de fonction de covariance donnée par :

$$Cov(W(s), W(t) | W(1) = 0) = s \wedge t - st$$

et pouvant s'exprimer sous la forme :

$$X(t) = W(t) - tW(1) \text{ où } 0 \leq t \leq 1.$$

Ce processus joue un rôle fondamental dans l'étude des fonctions de répartition empiriques puisqu'il est à l'origine du théorème de *Glivenko-Cantelli* qui permet d'affirmer que la fonction de répartition empirique  $F_n(t)$  converge presque sûrement et uniformément vers  $t$  lorsque  $n \rightarrow \infty$ . Autrement dit :

$$\sup_{0 < t < 1} |F_n(t) - t| \xrightarrow{p.s} 0$$

et qu'en vertu du théorème central limite, on a pour tout  $t \in [0, 1]$  :

$$\sqrt{n}(F_n(t) - t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, t(1-t))$$

De plus, si  $F$  est une fonction de répartition continue, mais par ailleurs quelconque, on montre que :

$$\alpha_n(s) = \sqrt{n}[F_n(y_s) - F(y_s)] ,$$

où  $y_s = F^{-1}(s)$ ,  $0 \leq s \leq 1$ , converge vers un pont brownien et que la loi limite de

$$\sqrt{n} \sup_x (F_n(x) - F(x)) ,$$

est celle du maximum d'un pont brownien  $X(t)$ . Il en découle que :

$$\lim_{n \rightarrow \infty} \mathbb{P}\{\sqrt{n} \sup_x |F_n(x) - F(x)| \leq d\} = \mathbb{P}\{\max_{0 \leq t \leq 1} |X(t)| \leq d\} .$$

Cette égalité permet ainsi d'étudier la puissance asymptotique du test et elle est une conséquence directe du résultat suivant :

Soit  $\{X_i\}_{i \geq 1}$  une suite de variables aléatoires indépendantes et de même loi, définies sur l'espace probabilisé  $(\Omega, \mathcal{F}, \mathbb{P})$ , de moyenne  $\mu$  et de variance  $\sigma^2$ . Pour chaque  $n = 1, 2, \dots$ , on considère alors le processus stochastique  $\{Y_n(t, \omega)\}_{n \geq 1}$  défini comme suit :

$$Y_n(t, \omega) = \frac{\sum_{i=1}^{[nt]} (X_i(\omega) - \mu) + (nt - [nt])(X_{[nt]+1}(\omega) - \mu)}{\sigma \sqrt{n}} \text{ où } : 0 \leq t \leq 1$$

Ce processus, dit "processus de sommes partielles", est un élément aléatoire de l'espace  $C[0, 1]$  des fonctions continues sur  $[0, 1]$ , muni de la métrique  $\rho$  usuelle donnée par :

$$\forall x, y \in C[0, 1] : \rho(x, y) = \sup_{0 \leq t \leq 1} |x(t) - y(t)| .$$

On désigne alors par  $\mathcal{B}_{C[0,1]}$  la  $\sigma$ -algèbre engendrée par les ensembles ouverts de  $C[0, 1]$  pour la métrique  $\rho$  et par  $\mathbb{Q}_n = \mathbb{P}^{Y_n}$  la mesure de probabilité image de  $\mathbb{P}$  par  $Y_n$ , sur  $(C[0, 1], \mathcal{B}_{C[0,1]})$ . Autrement dit on a :

$$\forall B \in \mathcal{B}_{C[0,1]} : \mathbb{Q}_n(B) = \mathbb{P}(\{\omega : Y_n(\cdot, \omega) \in B\}) .$$

Ainsi à chaque somme partielle  $Y_n(\cdot)$ , est associé l'espace probabilisé

$$(C[0, 1], \mathcal{B}_{C[0,1]}, \mathbb{Q}_n)$$

et afin que l'on puisse associer au processus  $\{Y_n(\cdot)\}_{n \geq 1}$  un processus limite  $Y(\cdot)$ , on doit disposer d'une mesure  $\mathbb{Q}$  sur  $(C[0, 1], \mathcal{B}_{C[0,1]})$ , de sorte que la suite  $\mathbb{Q}_n$  converge faiblement vers  $\mathbb{Q}$ . Le théorème de Donsker[7] montre que la mesure  $\mathbb{Q}$  n'est autre que la mesure de *Wiener*  $W$  sur  $(C[0, 1], \mathcal{B}_{C[0,1]})$  définie par :

i)  $W(\{x(\cdot) : x(0) = 0\}) = 1$  ,

ii) pour tout  $0 < t \leq 1$  et pour tout  $\alpha \in \mathbb{R}$ ,

$$W(\{x(\cdot) : x(t) \leq \alpha\}) = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\alpha} e^{-u^2/2t} du.,$$

iii) pour tout  $0 \leq t_0 \leq t_1 \leq \dots \leq t_k \leq 1$  et pour tout  $\alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{R}$ ,

$$W(\cap_{i=1}^k \{x(\cdot) : x(t_i) - x(t_{i-1}) \leq \alpha_i\}) = \prod_{i=1}^k W(\{x(\cdot) : x(t_i) - x(t_{i-1}) \leq \alpha_i\})$$

et que de surcroît l'on a bien :  $\mathbb{Q}_n \implies W$ . De ce dernier résultat découle alors le fait que :

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(\sqrt{n}\Delta_n \leq d) &= W^0\{x(\cdot) : \sup_{0 \leq u \leq 1} |x(u)| \leq d\} \\ &= 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 d^2} \quad d > 0 \end{aligned}$$

rendant ainsi possible le calcul de la puissance asymptotique du test.

## 5.2 Puissance du test

Comme les remarques 6 et 7 en font mention, l'étude de la puissance du test d'indépendance s'avère, dans un contexte où l'hypothèse alternative revêt une forme extrêmement complexe, excessivement ardue, voire utopique, dans la mesure où la "non-indépendance", qui caractérise l'hypothèse alternative, ne peut en aucun cas se décrire systématiquement, en dehors de quelques cas classiques particuliers déjà recensés dans la littérature et qui peuvent se résumer, dans l'essentiel, au trois cas suivants :

i) L'alternative  $\mathcal{H}_1$  est simple, ce qui revient à dire en d'autres termes, que la fonction de répartition  $F$  appartenant à l'alternative est fixée. C'est la démarche adoptée dans Raghavachari[21] dans le cas général où l'hypothèse nulle  $\mathcal{H}_0$  est réduite au seul élément  $F_0$ .

En dehors de ce cas particulier, la tendance veut que l'on s'intéresse au comportement de

$$\sqrt{n}\Delta_n = \sup_{0 \leq u \leq 1} |Y_n(u)|$$

sous une suite d'alternatives locales qui convergent faiblement vers  $F_0$  à une vitesse "convenable". Plus précisément (voir Lehmann[15], Chibisov[5]) on distingue les deux cas suivants :

ii) Si la vitesse de convergence de la suite d'alternatives vers l'hypothèse nulle n'est pas trop rapide, dans un sens qu'il conviendra de préciser, alors le test possède une puissance asymptotique qui tend vers 1 uniformément.

iii) Si la vitesse de convergence est trop rapide, la puissance du test est toujours inférieure au risque  $\alpha$  de première espèce. En d'autres termes, le test est systématiquement biaisé.

Les commentaires qui suivent traitent de chacun de ces cas et ne font, en réalité, que reprendre les propriétés asymptotiques du test de *Kolmogorov-Smirnov* puisque le test proposé est en fait de ce type.

### 5.2.1 Hypothèse alternative simple

Dans le cadre du présent test d'indépendance, l'hypothèse nulle  $\mathcal{H}_0$  s'écrit sous la forme :  $F_0(u) = u$ , tandis que l'hypothèse alternative  $\mathcal{H}_1$  correspond à un choix particulier d'une fonction de répartition  $F$ . Suivant Raghavachari[21] (voir également Lehmann[15]), on pose :

$$\Delta = \sup_{0 < u < 1} |F(u) - u|$$

et :

$$C_1 = \{u : F(u) - u = \Delta\}, \quad C_2 = \{u : F(u) - u = -\Delta\}$$

De plus, on note par  $\tilde{W}^0$  un élément aléatoire de l'espace  $(C[0, 1], \mathcal{B}_{C[0,1]})$ , que l'on suppose régir par l'unique mesure normale caractérisée par :

$$\mathbb{E}(\tilde{W}^0(u)) = 0$$

et par :

$$Cov(\tilde{W}^0(s), \tilde{W}^0(t)) = F(s)[1 - F(t)] \text{ pour } 0 \leq s \leq t \leq 1 .$$

On montre alors que :

$$\forall d \in \mathbb{R} : \lim_{n \rightarrow \infty} P(\sqrt{n}(\Delta_n - \Delta) \leq d) = \tilde{W}^0 \left( \left\{ x(\cdot) : \sup_{u \in C_1} x(u) \leq d; \sup_{u \in C_2} x(u) \geq -d \right\} \right) .$$

On notera que si  $C_1 = C_2 = [0, 1]$  alors  $\Delta = 0$  et la mesure  $\tilde{W}^0$  n'est autre que la mesure  $W^0$  introduite ci-dessus. Il est donc possible de déterminer à partir de cette expression et pour une alternative  $F$  fixée, la puissance du test d'indépendance.

### 5.2.2 Suite d'hypothèses alternatives simples

Étant donnée une alternative  $F$  appartenant à une suite d'alternatives données, trois cas peuvent se présenter suivant le comportement de la quantité

$$n^{1/2}d_K(F, u) = n^{1/2} \sup_{0 \leq u \leq 1} |F(u) - u|$$

lorsque  $n \rightarrow \infty$ . Les résultats qui suivent se trouvent exposés en détail dans, entre autres, Chibisov[5], Lehmann[15] et Massey[18] et concernent, dans chacun des cas, la puissance asymptotique du test de *Kolmogorov-Smirnov*, sous une famille d'alternatives particulières. On rappelle par ailleurs que le test de *Kolmogorov-Smirnov* est localement asymptotiquement convergent dans le sens où :

$$\lim_{n \rightarrow \infty} \mathbb{P}_F(n^{1/2}\Delta_n \geq d_{n,\alpha}) = 1 ,$$

où  $F$  désigne une alternative quelconque ( $F(u) \neq u, u \in [0, 1]$ ) et où  $\alpha$  désigne le risque de première espèce. Autrement dit, quelle que soit l'alternative  $F \neq I_d$ , la puissance du test, sous l'hypothèse  $F$  tend vers 1. Les résultats ci-dessous concernent donc la convergence uniforme de la puissance du test.

- **1<sup>er</sup> cas** :  $\lim_{n \rightarrow \infty} n^{1/2}d_K(F_n, F_0) = \infty$ .

Adoptant les notations de Lehmann[15], on considère une suite  $X_1, X_2, \dots, X_n$  de variables aléatoires indépendantes et de même loi, de fonction de répartition  $F$ . On montre alors que le test de *Kolmogorov-Smirnov* pour l'hypothèse nulle  $\mathcal{H}_0 : F = F_0$  contre l'hypothèse alternative  $\mathcal{H}_1 : F \neq F_0$ , possède une puissance qui tend uniformément vers 1 sur le sous-ensemble de toutes les alternatives  $F$  telles que :

$$n^{1/2}d_K(F, F_0) \geq \delta_n ,$$

où  $\delta_n$  tend vers l'infini avec  $n$ . En d'autres termes,

$$\inf\{\mathbb{P}_F(n^{1/2}\Delta_n > d_{n,\alpha}) : n^{1/2}d_K(F, F_0) \geq \delta_n\}$$



tend vers 1 si  $\delta_n \rightarrow \infty$ .

On notera que cette condition n'empêche nullement qu'une suite  $\{F_n\}_{n \geq 1}$  d'alternatives, puisse converger au sens de la métrique de la convergence uniforme, vers l'hypothèse nulle  $F_0$ , à condition toutefois que cette dernière se fasse à une vitesse contrôlée par  $\delta_n$ . Par exemple, si  $\delta_n$  est de la forme  $n^\epsilon$  avec  $\epsilon > 0$ , alors la distance entre  $F_n$  et  $F_0$  devra être, au moins, de l'ordre de  $n^{-1/2+\epsilon}$ .

Soit alors  $\{F_n\}_{n \geq 1}$  une suite quelconque d'hypothèses alternatives satisfaisant à la condition

$$n^{1/2}d_K(F_n, F_0) \geq \delta_n .$$

En vertu de l'inégalité du triangle, il vient :

$$d_K(F_n, F_0) \leq d_K(\hat{F}_n, F_0) + d_K(F_n, \hat{F}_n) ,$$

où  $\hat{F}_n$  désigne la fonction de répartition empirique de  $F_n$ . De cette dernière inégalité on en déduit que :

$$n^{1/2}\Delta_n \geq \delta_n - n^{1/2}d_K(F_n, \hat{F}_n) .$$

Il s'ensuit alors que :

$$1 \geq \mathbb{P}_{F_n}(n^{1/2}\Delta_n > d_{n,\alpha}) \geq \mathbb{P}_{F_n}(n^{1/2}d_K(F_n, \hat{F}_n) \leq \delta_n - d_{n,\alpha})$$

Or, sous l'hypothèse  $F_n$ , on montre que ( voir Lehmann[15]) la suite des variables aléatoires  $n^{1/2}d_K(F_n, \hat{F}_n)$  est tendue (on dit aussi bornée en probabilité, ce que l'on note par :  $n^{1/2}d_K(F_n, \hat{F}_n) = O_{F_n}(1)$ ). De plus :

$$\lim_{n \rightarrow \theta} d_{n,\alpha} = d_\alpha < \infty ,$$

où  $d_\alpha$  correspond à la frontière de la région critique pour la loi limite de  $n^{1/2}\Delta_n$  sous l'hypothèse nulle. Comme  $\delta_n$  tend vers l'infini avec  $n$ , alors  $\delta_n - d_{n,\alpha} \rightarrow \infty$  et

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_n}(n^{1/2}\Delta_n > d_{n,\alpha}) = 1 .$$

A distance finie, c'est-à-dire lorsque l'on fixe la valeur de  $n$ , on ne dispose pas de la valeur exacte de la puissance, mais il est possible d'obtenir une borne inférieure de celle-ci à l'aide de l'inégalité de *Dvoretzky-Kiefer-Wolfowitz* (voir par exemple Bosq et Lecoutre[4]) donnée par :

$$\forall d > 0 : \mathbb{P}_F\{d_K(\hat{F}_n, F) > d\} \leq C e^{-2nd^2} ,$$

où  $C$  est une constante universelle, que l'on peut choisir égale à 2 (Massart[17]). On montre alors que l'on obtient à l'aide de cette inégalité, une borne inférieure pour la puissance du test et pour l'alternatine  $F_n$  fixée donnée par :

$$\mathbb{P}_{F_n}(n^{1/2}\Delta_n > d_{n,\alpha}) \geq 1 - 2e^{-2(\delta_n - d_{n,\alpha})^2}$$

à condition que l'on ait :  $n^{1/2}d_K(F_n, F_0) \geq \delta_n$  et  $\delta_n > d_{n,\alpha}$ .

- **2<sup>ème</sup> cas** :  $\lim_{n \rightarrow \infty} n^{1/2}d_K(F_n, F_0) = 0$ .

Le résultat ci-dessus montre que la puissance asymptotique du test de *Kolmogorov* et, par voie de conséquence, la puissance asymptotique du test d'indépendance, est arbitrairement proche de 1 pour toute suite d'alternatives  $\{F_n\}_{n \geq 1}$  tendant vers  $F_0 = I_d$  ( $F_0(u) = I_d(u) = u : 0 \leq u \leq 1$ ) suffisamment lentement. A l'opposé, l'inégalité du triangle permet d'écrire que :

$$d_K(\hat{F}_n, F_0) \leq d_K(\hat{F}_n, F) + d_K(F, F_0)$$

d'où :

$$n^{1/2}\Delta_n \leq n^{1/2}d_K(\hat{F}_n, F) + n^{1/2}d_K(F, F_0)$$

soit encore :

$$\mathbb{P}_F(n^{1/2}\Delta_n > d_{n,\alpha}) \leq \mathbb{P}_F\{n^{1/2}d_K(\hat{F}_n, F) + n^{1/2}d_K(F, F_0) > d_{n,\alpha}\}$$

ce qui entraîne une puissance médiocre du test dans le cas d'une convergence trop rapide de la suite d'alternatives  $\{F_n\}_{n \geq 1}$  vers  $F_0$  ( $F_0(u) = u$ , dans le cas du test d'indépendance). Plus précisément on a le résultat suivant (voir Lehmann[15]) :

Le test de *Kolmogorov-Smirnov*, de niveau  $\alpha$ , pour l'hypothèse  $\mathcal{H}_0 : F = F_0$  contre l'hypothèse  $\mathcal{H}_1 : F \neq F_0$ , possède une puissance asymptotique inférieure ou égale à  $\alpha$  pour toute suite  $\{F_n\}_{n \geq 1}$  d'hypothèses alternatives telles que :

$$\lim_{n \rightarrow \infty} n^{1/2}d_K(F_n, F_0) = 0$$

ce qui, en d'autres termes, peut s'écrire :

$$\limsup_n \mathbb{P}_{F_n}\{n^{1/2}\Delta_n > d_{n,\alpha}\} \leq \alpha$$

Par conséquent le test d'indépendance ne peut distinguer des suites de contre-hypothèses qui se trouveraient à une distance de l'ordre de  $o(n^{-1/2})$  de  $F_0$ .

- **3<sup>ème</sup> cas :**  $\lim_{n \rightarrow \infty} n^{1/2}d_K(F_n, F_0) = \delta$ .

Les résultats qui suivent sont présentés en détail dans Chibisov[5] et également dans Lehmann[15]. On suppose alors que :

$$\lim_{n \rightarrow \infty} n^{1/2}d_K(F_n, F_0) = \delta \text{ avec } 0 < \delta < \infty$$

et qu'il existe par ailleurs une fonction  $d$  telle que :

$$\sup_u |d_n(u) - d(u)| \longrightarrow 0 ,$$

où  $d_n(u) = n^{1/2} [F_n(u) - F_0(u)]$ . Il vient :

$$n^{1/2} [\hat{F}_n(u) - F_0(u)] = n^{1/2} [\hat{F}_n(u) - F_n(u)] + d_n(u) .$$

Sous l'hypothèse alternative  $F = F_n$ ,  $n^{1/2} [\hat{F}_n(u) - F_n(u)]$  admet une moyenne nulle et une variance donnée par :

$$F_n(u) [1 - F_n(u)]$$

laquelle converge vers  $F_0(u) [1 - F_0(u)]$  lorsque  $n \longrightarrow \infty$ .

Par ailleurs, pour tout  $u$  fixé, le théorème central limite entraîne que, sous l'hypothèse  $F = F_n$  :

$$n^{1/2} [\hat{F}_n(u) - F_n(u)] \xrightarrow{\mathcal{L}} B(u) ,$$

où  $B(u)$  est un pont brownien de loi normale  $\mathcal{N}(0, F_0(u) [1 - F_0(u)])$ . Ainsi sous l'hypothèse alternative,

$$n^{1/2} [\hat{F}_n(u) - F_0(u)] \xrightarrow{\mathcal{L}} B(u) + d(u) \sim \mathcal{N}(d(u), F_0(u) [1 - F_0(u)]) .$$

On en déduit alors aisément que, sous l'hypothèse  $F = F_n$  et pour tout  $u_1, u_2, \dots, u_k$ , on a :

$$n^{1/2} [\hat{F}_n(u_1) - F_0(u_1), \dots, \hat{F}_n(u_k) - F_0(u_k)] \xrightarrow{\mathcal{L}} [B(u_1) + d(u_1), \dots, B(u_k) + d(u_k)] ,$$

d'où il découle que :

$$\max_{i=1,2,\dots,k} n^{1/2} |\hat{F}_n(u_i) - F_0(u_i)| \xrightarrow{\mathcal{L}} \max_{i=1,2,\dots,k} |B(u_i) + d(u_i)| ,$$

ce qui entraîne, toujours sous l'hypothèse  $F = F_n$ , que :

$$\sup_u n^{1/2} |\hat{F}_n(u) - F_0(u)| \xrightarrow{\mathcal{L}} \sup_u |B(u) + d(u)|$$

On peut donc en déduire que la puissance asymptotique du test d'indépendance peut s'exprimer sous la forme :

$$\mathbb{P}(\sup_u |B(u) + d(u)| > d_\alpha)$$

où dans ce cas on a :  $0 \leq u \leq 1$ ,  $F_0(u) = u$  et  $B(u)$  est un pont brownien de loi normale  $\mathcal{N}(0, u(1-u))$ . On montre alors que la puissance asymptotique du test d'indépendance, tend vers 1 lorsque  $n$  tend vers l'infini (voir Chibisov[5]).

### 5.3 Autres tests d'adéquation

Dans le but de proposer quelques tests concurrents, on traitera ici très brièvement, des tests d'adéquation de type *Cramer-von Mises* de la forme :

$$\Gamma_{n,\psi} = \int_0^1 [\hat{F}_{U,n}(u) - u]^2 \psi(u) du ,$$

dans le cas où la fonction de poids  $\psi(u)$  s'exprime sous la forme  $\psi(u) = 1$  (test de *Cramer-von Mises*), ou  $\psi(u) = [u(1-u)]^{-1}$  (test d'*Anderson-Darling*).

- Test de *Cramer-von Mises*.

Dans le cas particulier du test d'indépendance, pour lequel l'hypothèse nulle s'exprime par  $F_0(u) = u$  où  $0 \leq u \leq 1$ , la statistique  $C_n$  du test de *Cramer-von Mises* a pour expression :

$$C_n = n \int_0^1 [\hat{F}_n(u) - u]^2 du$$

où l'on a posé pour des raisons de commodité d'écriture,  $\hat{F}_{U,n}(u) = \hat{F}_n(u)$ . On considère alors la base orthonormée de l'espace de *Hilbert*  $L^2_{[0,1]}$  donnée par la suite de fonctions :  $\{\sqrt{2} \sin(\pi kx)\}$  où  $k = 1, 2, \dots$ . La quantité  $C_n$  étant, au coefficient  $n$  près, le carré de la norme  $L^2$  de la différence entre la fonction  $\hat{F}_n$  et la fonction identité  $I_d$ , il vient, en vertu de l'égalité de Parseval :

$$\begin{aligned} C_n &= n \int_0^1 [\hat{F}_n(u) - u]^2 du = n \| \hat{F}_n - I_d \|^2 \\ &= n \sum_{k=1}^{\infty} \langle (\hat{F}_n - I_d), \sqrt{2} \sin(\pi kx) \rangle^2 \\ &= n \sum_{k=1}^{\infty} \left[ \int_0^1 (\hat{F}_n(u) - u) \sqrt{2} \sin(\pi ku) du \right]^2 \end{aligned}$$

En intégrant par partie l'expression  $\int_0^1 (\hat{F}_n(u) - u) \sqrt{2} \sin(\pi ku) du$ , il vient (voir Billingsley[2]),

$$\begin{aligned} \int_0^1 (\hat{F}_n(u) - u) \sqrt{2} \sin(\pi ku) du &= \frac{-1}{k\pi} \int_0^1 \sqrt{2} \cos(\pi ku) d(\hat{F}_n(u) - u) , \\ &= \frac{-1}{k\pi} \int_0^1 \sqrt{2} \cos(\pi ku) d\hat{F}_n(u) , \\ &= \frac{-1}{nk\pi} \sum_{i=1}^n \sqrt{2} \cos(\pi kU_i) . \end{aligned}$$

Posant  $T_k(u) = \sqrt{2} \cos(\pi ku)$ , on obtient :

$$\int_0^1 (\hat{F}_n(u) - u) \sqrt{2} \sin(\pi ku) du = \frac{-1}{nk\pi} \sum_{i=1}^n T_k(U_i) = -\frac{V_{n,k}}{n^{1/2}k\pi} ,$$

où  $V_{n,k} = \frac{1}{n^{1/2}} \sum_{i=1}^n T_k(U_i)$ . D'où :

$$C_n = \sum_{k=1}^{\infty} \frac{1}{k^2 \pi^2} V_{n,k}^2 .$$

On montre alors que, sous l'hypothèse nulle d'indépendance on a :

$$C_n \xrightarrow{\mathcal{L}} \sum_{k=1}^{\infty} \frac{1}{k^2 \pi^2} V_k^2 ,$$

où  $Z_1, Z_2, \dots$  sont des variables aléatoires indépendantes et de même loi normale  $\mathcal{N}(0, 1)$ . Ainsi on pourra, en théorie, trouver la région critique du test de *Cramer-von Mises* pour un risque  $\alpha$  de première espèce, mais dans la pratique on n'utilise qu'un nombre fini de termes de la série de sorte que la région critique n'est pas en général exactement déterminée contrairement au test de *Kolmogorov-Smirnov* et il en est de même bien entendu en ce qui concerne la puissance du test pour une alternative donnée.

- Test de *Anderson-Darling*

La statistique  $A_n$  du test d'*Anderson-Darling* est donnée, toujours dans le cas particulier où  $F_0(u) = u$ , par :

$$A_n = n \int_0^1 \frac{[\hat{F}_n(u) - u]^2}{u(1-u)} du .$$

On montre (voir Lehmann[15]) que, si l'on désigne par  $T_k(u)$  le  $k^{\text{ème}}$  polynôme normalisé de *Legendre* sur  $[0, 1]$  et si l'on utilise des notations analogues à celles du précédent cas, on a sous l'hypothèse nulle d'indépendance :

$$A_n \xrightarrow{\mathcal{L}} \sum_{k=1}^{\infty} \frac{1}{k(k+1)} V_k^2 .$$

De même que dans le cas du précédent test, la région critique ne sera déterminée qu'approximativement. Il est bon de préciser toutefois que, dans certaines circonstances, ces deux tests peuvent avoir une puissance locale supérieure à celle du test de *Kolmogorov-Smirnov*.

## 5.4 Contiguïté

Bien que les tests d'adéquation puissent avoir, pour toute alternative fixée, une puissance asymptotique égale à 1 lorsque la taille de l'échantillon tend vers l'infini, il n'en demeure pas moins que cette puissance peut être relativement faible par rapport à une suite d'alternatives contiguës. Dans cette perspective, on énonce ci-dessous quelques principes suivant lesquels cette analyse pourrait s'effectuer ainsi que les résultats attendus de cette démarche.

Soit  $m_\nu$  une suite croissante d'entiers et notons par  $\mathcal{H}_{0\nu}$  et  $\mathcal{H}_{1\nu}$  les suites d'hypothèses nulles et alternatives associées aux  $m_\nu$  observations. Par exemple, dans le cas d'observations issues de lois normales, on pourrait supposer que sous les hypothèses  $\mathcal{H}_{0\nu}$  la loi des observations serait donnée par une loi  $\mathcal{N}_k(0, I)$ , alors que sous les hypothèses alternatives  $\mathcal{H}_{1\nu}$  elles seraient décrites par des lois de la forme  $\mathcal{N}_k(0, \Sigma_\nu)$ . En particulier, on examinera les cas où  $\lim_{\nu \rightarrow \infty} \Sigma_\nu = I$  qui est une expression de la notion de contiguïté entre les deux suites d'hypothèses. On considèrera alors les procédures de tests suivantes : la première, telle que présentée ci-dessus, pour laquelle on désignera la région critique par  $R_{1,m_\nu}$  et la seconde qui correspondra au test du rapport de vraisemblance, dont la région critique sera désignée par  $R_{2,m_\nu}$ . La conduite de l'étude des propriétés asymptotiques s'effectuera comme suit. Après avoir choisit deux régions critiques  $R_{1,m_\nu}$  et  $R_{2,m_\nu}$  telles que :

$$\mathbb{P}_{\mathcal{H}_{0\nu}}(R_{1,m_\nu}) = \mathbb{P}_{\mathcal{H}_{0\nu}}(R_{2,m_\nu}) = \alpha_\nu$$

on montrera que :

$$\lim_{\nu \rightarrow \infty} \mathbb{P}_{\mathcal{H}_{1\nu}}(R_{1,m_\nu}) = \lim_{\nu \rightarrow \infty} \mathbb{P}_{\mathcal{H}_{1\nu}}(R_{2,m_\nu}) = \beta$$

qui n'est autre que l'expression de l'équivalence asymptotique entre les deux procédures. A titre d'exemple, dans le cas de la normalité, ou plus généralement dans le cas des lois elliptiques, on considèrera des matrices de structure de dépendance  $\Sigma_\nu$  de la forme :  $\Sigma_\nu = I_k + \frac{1}{\sqrt{m_\nu}}\Delta_\nu$  où  $\lim_{\nu \rightarrow \infty} \Delta_\nu = \Delta$ . Les régions critiques  $R_{1,m_\nu}$  et  $R_{2,m_\nu}$  associées aux procédures décrites ci-dessus seront respectivement de la forme :

$$\sup_u |\tilde{F}_\nu(u) - G(u)| \geq c_{1,m_\nu} \quad \text{et} \quad \frac{L(\mathcal{H}_{0\nu})}{L(\mathcal{H}_{1\nu})} \leq c_{2,m_\nu}$$

où :  $\tilde{F}_\nu(u)$  désigne la fonction de répartition empirique de la variable aléatoire  $F_V(v)$ ,  $G(u)$  désigne la fonction de répartition de la loi  $\mathcal{CM}_n$  à  $n$  degrés de liberté et où  $\frac{L(\mathcal{H}_{0\nu})}{L(\mathcal{H}_{1\nu})}$  désigne le rapport de vraisemblance généralisé habituel.

## 6 Conclusions

Étant donné un vecteur aléatoire  $X = (X_1, X_2, \dots, X_n)$  de  $\mathbb{R}^n$ , de loi absolument continue par rapport à une mesure de référence, on considère le vecteur aléatoire  $Z = (Z_1, Z_2, \dots, Z_n)$  dont la composante  $Z_i$  est définie par :

$$\begin{aligned} Z_i &= F_{X_i, X_{i+1}, \dots, X_n}(X_i, X_{i+1}, \dots, X_n) \\ &= F_{X_1, X_2, \dots, X_n}(\infty, \dots, \infty, X_i, \dots, X_n) . \end{aligned}$$

On a alors montré que la famille des variables aléatoires  $\{Z_i\}$ ,  $i = 1, 2, \dots, n$  était libre et que parmi les éléments de cette famille, la composante  $Z_1$  possédait, dans le cas de l'indépendance entre les composantes de  $X$ , une loi particulièrement simple, dite loi  $\mathcal{CM}_n$  et qui n'est autre que la loi du produit de  $n$  variables aléatoires indépendantes et de même loi uniforme  $\mathcal{U}_{[0,1]}$ .

Ainsi, dans le cadre théorique retenu, le test proposé est, tout au moins conceptuellement parlant, particulièrement simple et naturel. En effet, partant du constat que la variable aléatoire

$$\prod_{i=1}^{i=n} F_i(x_i) ,$$

est, pour toute fonction de répartition  $F(X_1, X_2, \dots, X_n)$ , de loi  $\mathcal{CM}_n$  sous l'hypothèse d'indépendance des composantes, il semble naturel de mesurer l'écart à l'indépendance en considérant la distance

$$d(F, \prod_{i=1}^{i=n} F_i) = \sup_{x \in \mathbb{R}^n} |F(x) - (\prod_{i=1}^{i=n} F_i)(x)| ,$$

laquelle se ramène, à l'aide de la fonction de répartition  $G_n$  de la loi  $\mathcal{CM}_n$ , à la distance entre les fonctions de répartition de la variable aléatoire

$$(G_n \circ F)(x) ,$$

et celle de la variable aléatoire

$$(G_n \circ (\prod_{i=1}^{i=n} F_i))(x) ,$$

de loi uniforme  $\mathcal{U}_{[0,1]}$  puisque dans le cas de l'indépendance  $G_n$  est justement la fonction de répartition de  $F(\cdot) = \left[ \prod_{i=1}^{i=n} F_i \right](\cdot)$ .

On est ainsi conduit naturellement à considérer la distance de *Kolmogorov* (bien qu'à ce stade cela ne soit pas, comme il l'a été mentionné ci-dessus, la seule distance possible)

$$d_K(F_{U,n}, u) = \sup_{0 \leq u \leq 1} |F_{U,n}(u) - u| ,$$

entre les fonctions de répartition  $F_{U,n}(u)$  et  $u$  et à rejeter l'hypothèse d'indépen-

dance si cette distance dépasse un certain seuil. Cette approche s'inscrit alors, non moins naturellement, dans le cadre des processus empiriques et de ponts browniens et permet d'exploiter les propriétés asymptotiques de ces derniers afin de les adapter au test considéré.

Par ailleurs, d'un point de vue formel de la théorie de la décision, la fonction de répartition empirique  $\hat{F}_n(x)$ , utilisée dans le test de *Kolmogorov-Smirnov* (tout comme celui d'ailleurs de *Cramer-von Mises*), s'interprète naturellement en tant que règle de décision invariante par rapport à un groupe de transformations. Plus précisément, soient  $X_1, X_2, \dots, X_n$   $n$  observations indépendantes d'une variable aléatoire de fonction de répartition  $F(x)$  inconnue, mais que l'on supposera continue. On désigne alors par  $(\Theta, D, L)$  le problème décisionnel où  $\Theta$  est l'espace paramètre, constitué par l'ensemble des fonctions de répartition continues sur  $\mathbb{R}$ ,  $D$  est l'ensemble des décisions, identifié à l'ensemble des fonctions de répartition sur  $\mathbb{R}$  et  $L(\theta, d)$  est la fonction de perte donnée par :

$$L(\theta, d) = L(F, \hat{F}) = \sup_{x \in \mathbb{R}} |F(x) - \hat{F}(x)| \text{ où } F \in \Theta \text{ et } \hat{F} \in D .$$

Le théorème de *Glivenko-Cantelli* assure que si l'on choisit la décision  $\hat{F}(x) = \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{[X_i, \infty[}(x)$ ,  $L(F, \hat{F})$  converge en probabilité vers 0 lorsque  $n$  tend vers l'infini. Si l'on désigne par  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  les statistiques d'ordre associées à l'échantillon considéré, alors le problème décisionnel est invariant pour le groupe  $\mathcal{G}$  des transformations défini par :

$$\mathcal{G} = \{g_\varphi : g_\varphi(x_{(1)}, x_{(2)}, \dots, x_{(n)}) = (\varphi(x_{(1)}), \varphi(x_{(2)}), \dots, \varphi(x_{(n)}))\} ,$$

où  $\varphi$  est une fonction continue strictement croissante de  $\mathbb{R}$  dans  $\mathbb{R}$ . En effet, soit  $Y = (X_{(1)}, X_{(2)}, \dots, X_{(n)})$  le vecteur des statistiques d'ordre issues d'observations indépendantes  $X_1, X_2, \dots, X_n$  d'une loi de fonction de répartition  $F(x)$ , alors  $g_\varphi(Y)$  est un vecteur de statistiques d'ordre issues de la fonction de répartition

$$G(x) = P(\varphi(X) \leq x) = F(\varphi^{-1}(x)) .$$

Par conséquent les lois sont invariantes sous  $\mathcal{G}$  et  $\bar{g}_\varphi(F(x)) = F(\varphi^{-1}(x))$ . De même, la fonction de perte est également invariante sous  $\mathcal{G}$  avec  $\tilde{g}_\varphi(\hat{F}(x)) = \hat{F}(\varphi^{-1}(x))$  ce qui entraîne l'existence de règles de décision invariantes dont la forme générale est donnée par :

$$\hat{F}_Y(x) = \sum_{i=1}^{n-1} \nu_i \mathbb{I}_{[Y_i, Y_{i+1}[}(x) + \mathbb{I}_{[Y_n, \infty[}(x) ,$$

où  $0 \leq \nu_1 \leq \nu_2 \leq \dots \leq \nu_n \leq 1$ .

Dans le cadre du test d'hypothèse d'indépendance, la fonction  $\varphi$  introduite ci-dessus n'est autre que la fonction de répartition  $G_n$  de la loi  $\mathcal{CM}_n$  et les "observations" ne sont autres que les valeurs  $Z_{1i} = F(X_{1i}, X_{2i}, \dots, X_{ni})$   $i = 1, 2, \dots, k$ . Pour plus de détails sur le sujet, on pourra consulter Ferguson[8]. On a également la même propriété d'invariance pour les fonctions de perte données par :

$$L(F, \hat{F}) = \int \left( F(x) - \hat{F}(x) \right)^2 \psi(F(x)) dF(x) ,$$

où  $\psi$  est une fonction continue de  $[0, 1]$  dans  $\mathbb{R}^+$ .

Enfin, quelques exemples et simulations ont permis, dans un cadre semi-paramétrique, d'illustrer la procédure de test et de calculer la puissance de ce dernier. Concernant ce dernier aspect, le test proposé possède une puissance asymptotique égale à 1 pour toute alternative fixée  $F$ , alors qu'en revanche il est uniformément asymptotiquement convergent que dans le cas où les alternatives tendent vers l'hypothèse nulle d'indépendance à une vitesse suffisamment élevée.

## References

- [1] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd Edition, John Wiley, (1984).
- [2] P. Billingsley, *Convergence of Probability Measures*, John Wiley, (1968).
- [3] M. Bilodeau and D. Brenner, *Theory of Multivariate Statistics*, Springer-Verlag, (1999).
- [4] D. Bosq et J.P. Lecoutre, *Théorie de l'estimation fonctionnelle*, Economica, (1987).
- [5] D.M. Chibisov, *An investigation of the asymptotic power of the tests of fit*, Th. Prob. Applic.(1965), **10** 421-437.
- [6] L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, (1986).
- [7] M. Donsker, *An invariance principle for certain probability limit theorems*, Mem.Amer.Math.Soc.(1951), **6**:1-12.
- [8] T.S. Ferguson, *Mathematical Statistics, A Decision Theoretic Approach*, Academic Press, (1967).
- [9] M. Fréchet, *Sur les tableaux de corrélation dont les marges sont données*, Ann. Univ. Lyon, **Sér.3, 14** (1951), 53-77.
- [10] N.C. Giri, *Multivariate Statistical Analysis*, 2nd Edition, Marcel Dekker, (2004).
- [11] N.L. Johnson and S. Kotz, *Distributions in Statistics : Continuous Multivariate Distributions*, John Wiley, (1972).
- [12] G. Kimeldorf and A.R. Sampson, *Uniform representations of bivariate ditributions*, Com. Statist.(1975) **4** , 617-627.
- [13] A.M. Kshirsagar, *Multivariate Analysis*, Marcel Dekker, (1972).
- [14] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, (1986).
- [15] E.L. Lehmann and J.P. Romano, *Testing Statistical Hypotheses*, Third Edition, Springer, (2005).
- [16] K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis*, Academic Press, (1979).
- [17] P. Massart, *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*, Annals of Probability, (1990) **18**, 1269-1283.
- [18] F.J. Massey, *A note on the power of a non-parametric test*, Annals of Mathematical Statistics, (1950), **21**, 440-443.
- [19] W. Mendenhall and T. Sincich, *A Second Course in Statistics: Regression Analysis*, Fifth Edition. Prentice Hall, (1996).
- [20] R Development Core Team (2008). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [21] M. Raghavachari, *Limiting Distributions of Kolmogorov-Smirnov Type Statistics Under the Alternative*, The Annals of Statistics, (1973), **Vol1, No. 1**, pp.67-73.
- [22] A. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statist. Univ. Paris, **8** (1959), 229-231.
- [23] G. Saporta, *Probabilité, Analyse des données et Statistique*, Editions Technip, (1990).
- [24] R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, John Wiley, (1980).

# Annexe 1

$(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}, X_{2,3})$

77	82	67	67	81	63	78	80	70	81	75	73	71	66	81	55	72	63	70	68
63	63	65	70	63	53	61	72	64	73	51	67	65	65	68	59	70	68	62	56
62	60	58	62	70	64	72	60	62	45	52	64	60	63	54	55	67	59	62	44
50	50	64	55	63	65	63	58	56	37	31	55	60	57	73	60	64	56	54	40
44	69	53	53	53	42	69	61	55	45	62	46	61	57	45	31	49	62	63	62
44	61	52	62	46	49	41	61	49	64	12	58	61	63	67	49	53	49	62	47
54	49	56	47	53	54	53	46	59	44	44	56	55	61	36	18	44	50	57	81
46	52	65	50	35	32	45	49	57	64	30	69	50	52	45	46	49	53	59	37
40	27	54	61	61	31	42	48	54	68	36	59	51	45	51	56	40	56	54	35
46	56	57	49	32	45	42	55	56	40	42	60	54	49	33	40	63	53	54	25
23	55	59	53	44	48	48	49	51	37	41	63	49	46	34	46	52	53	41	40
46	61	46	38	41	40	57	51	52	31	49	49	45	48	39	22	58	53	56	41
35	60	47	54	33	48	56	49	42	32	31	57	50	54	34	17	53	57	43	51
49	57	47	39	26	59	50	47	15	46	37	56	49	28	45	40	43	48	21	61
35	35	41	51	50	38	44	54	47	24	43	43	38	34	49	39	46	46	32	43
62	44	36	22	42	48	38	41	44	33	34	42	50	47	29	18	51	40	56	30
35	36	46	48	29	59	53	37	22	19	41	41	43	30	33	31	52	37	27	40
17	51	52	35	31	34	30	50	47	36	46	40	47	29	17	10	46	36	47	39
46	37	45	15	30	30	34	43	46	18	13	51	50	25	31	49	50	38	23	09
18	32	31	45	40	08	42	48	26	40	23	38	36	48	15	30	24	43	33	25
03	09	51	47	40	07	51	43	17	22	15	40	43	23	18	15	38	39	28	17
05	30	44	36	18	12	30	32	35	21	05	26	15	20	20	00	40	21	09	14

**Matrice des corrélations  $\hat{R}$**

$$\left[ \begin{array}{cc|ccc} 1 & 0.553 & 0.547 & 0.410 & 0.389 \\ 0.553 & 1 & 0.610 & 0.485 & 0.437 \\ \hline 0.547 & 0.610 & 1 & 0.711 & 0.665 \\ 0.410 & 0.485 & 0.711 & 1 & 0.607 \\ 0.389 & 0.437 & 0.665 & 0.607 & 1 \end{array} \right]$$



## Extrait du programme en langage R

```

:
:
:
> m<-88
> supX<-numeric(m)
> for(i in 1:m){supX[i]<-max(x1[i],x2[i])}
> supY<-numeric(m)
> for(i in 1:m){supY[i]<-max(y1[i],y2[i],y3[i])}
> z<-numeric(m)
> u<-numeric(m)
> for(i in 1:m){u[i]<-pmvnorm(mu,sigma,lower=rep(-Inf,5),upper=c(supX[i],supX[i],
+ supY[i],supY[i],supY[i]))}
> z<-u*(1-log(u))
> ks.test(z,punif)
One-sample Kolmogorov-Smirnov test
data: z
D = 0.2236, p-value = 0.0002393

```

### Test du rapport de vraisemblance

$$S = \begin{bmatrix} 302.3 & 125.8 & 100.4 & 105.1 & 116.1 \\ 125.8 & 170.9 & 84.2 & 93.6 & 97.9 \\ 100.4 & 84.2 & 111.6 & 110.8 & 120.5 \\ 105.1 & 93.6 & 110.8 & 217.9 & 153.8 \\ 116.1 & 97.9 & 120.5 & 153.8 & 294.4 \end{bmatrix}$$

$$S_{22}^{-1}S_{21}S_{11}^{-1}S_{12} = \begin{bmatrix} 0.38042 & 0.41214 & 0.44127 \\ 3.7298 \times 10^{-2} & 4.2224 \times 10^{-2} & 4.3441 \times 10^{-2} \\ 1.6032 \times 10^{-2} & 1.7182 \times 10^{-2} & 1.8578 \times 10^{-2} \end{bmatrix},$$

Valeurs propres :  $\lambda_1 = 0,43958$ ,  $\lambda_2 = 1,6424 \times 10^{-3}$  et  $\lambda_3 \simeq 0$ .

$$\prod_{i=1}^2 (1 - \lambda_i) = (1 - 0.43958)(1 - 1.6424 \times 10^{-3}) = 0,55950.$$

$$-(n - \frac{1}{2}(p_1 + p_2 + 3)) \text{Log} | I - S_{22}^{-1}S_{21}S_{11}^{-1}S_{12} | = -84 \log 0.55950 = 48,780.$$

### Test “d’union-intersection”

Plus grande valeur propre de la matrice  $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$  :

$$\lambda_1 = 0,43958$$

$$\nu_1 = n - p_1 - 1 = 85 ; \nu_2 = p_2 = 3$$

La valeur critique pour un risque  $\alpha$  de première espèce de 0,1 est de : 0,100

## Annexe 2

$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$
7781	56	100	235	886	644	737	5149	59	92	124	389	405	236
7004	57	110	388	962	589	1029	5241	55	110	222	612	477	616
7267	59	61	398	1342	1081	830	2917	80	69	378	1057	970	1210
2129	57	102	457	1153	891	1468	4390	81	70	195	974	1027	1452
4878	49	45	577	803	537	335	4957	51	24	358	783	893	616
3999	64	144	345	1127	563	918	7099	62	130	374	1419	609	957
11777	60	123	326	627	402	335	7337	51	128	238	1137	461	968
5674	57	78	161	748	495	962	8301	74	115	191	946	771	719
7392	62	172	219	876	823	665	4889	69	86	214	750	513	489
8100	86	126	287	685	555	577	6308	49	81	132	461	430	341
4736	38	115	235	436	456	214	6908	57	145	164	864	549	902
4337	73	110	127	899	573	484	5345	48	116	127	604	360	126
3079	59	96	180	570	428	456	6994	53	59	107	714	473	726
7273	53	51	118	826	463	907	6781	74	78	171	917	805	1100
4091	76	116	206	1060	961	951	3142	70	106	335	809	702	1721
3390	77	70	284	957	745	1446	5738	52	27	126	546	455	502
6319	41	58	220	559	539	440	4931	71	174	129	891	481	737
7447	63	83	174	1050	553	1133	6501	47	69	129	643	334	473
7100	55	80	124	568	428	456	5678	52	94	107	828	384	1083
8035	78	115	174	709	498	968	4619	76	100	164	777	834	841
5579	79	83	223	568	683	660	1832	71	124	158	626	571	627
4338	84	78	115	900	556	555	5445	42	52	121	432	458	313
6895	41	18	118	442	479	203	4123	42	84	153	432	544	654
3629	57	133	155	644	505	781	5884	31	89	100	1061	391	280

### Matrice des corrélations $\hat{R}$

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
$x_1$	1.0000	-0.1380	0.0407	-0.0762	0.0094	-0.2329	-0.2406
$x_2$	-0.1380	1.0000	0.2855	0.1155	0.3236	0.5030	0.4837
$x_3$	0.0407	0.2855	1.0000	0.0130	0.2449	-0.0514	0.1331
$x_4$	-0.0762	0.1155	0.0130	1.0000	0.4600	0.4630	0.3194
$x_5$	0.0094	0.3236	0.2449	0.4600	1.0000	0.5465	0.5999
$x_6$	-0.2329	0.5030	-0.0514	0.4630	0.5465	1.0000	0.5683
$x_7$	-0.2406	0.4837	0.1331	0.3194	0.5999	0.5683	1.0000

### Extraits des programmes en langage R

### Calcul de $(G_3 \circ \hat{F})(v_1, v_2, v_3)$

```

:
:
:
> m<-48
> nu<-c(0,0,0,0,0,0)
> supX<-numeric(m)
> for(i in 1:m){supX[i]<-max(x1[i],x2[i])}
> supY<-numeric(m)
> for(i in 1:m){supY[i]<-max(y1[i],y2[i])}
> supZ<-numeric(m)
> for(i in 1:m){supZ[i]<-max(z1[i],z2[i],z3[i])}
> v<-numeric(m)
> u<-numeric(m)
> for(i in 1:m){u[i]<-pmvnorm(nu,sigma,lower=rep(-Inf,7),
+ upper=c(supX[i],supX[i],supY[i],supY[i],supZ[i],supZ[i]))}
> v<-u*(1-log(u)+.5*((log(u))^2))
> ks.test(v,punif)
One-sample Kolmogorov-Smirnov test
data: v
D = 0.1469, p-value = 0.2281

```

### Test d'indépendance des composantes d'un vecteur aléatoire à 7 composantes

```

> supT<-numeric(m)
> for(i in 1:m){supT[i]<-max(x1[i],x2[i],y1[i],y2[i],z1[i],z2[i],z3[i])}
> s<-numeric(m)
> t<-numeric(m)
> for(i in 1:m){t[i]<-pmvnorm(nu,sigma,lower=rep(-Inf,7),
+ upper=c(supT[i],supT[i],supT[i],supT[i],supT[i],supT[i],supT[i]))}
> s<-t*(1-log(t)+.5*((log(t))^2)-.16666*((log(t))^3)+.041666*((log(t))^4)
+ -.008333*((log(t))^5)+.0013889*((log(t))^6))
> ks.test(s,punif)
One-sample Kolmogorov-Smirnov test
data: s
D = 0.8946, p-value = 3.331e-16

```

### Test du rapport de vraisemblance

$$\det(\hat{R}) =$$

$$\det \begin{bmatrix} 1.0000 & -0.1380 & 0.0407 & -0.0762 & 0.0094 & -0.2329 & -0.2406 \\ -0.1380 & 1.0000 & 0.2855 & 0.1155 & 0.3236 & 0.5030 & 0.4837 \\ 0.0407 & 0.2855 & 1.0000 & 0.0130 & 0.2449 & -0.0514 & 0.1331 \\ -0.0762 & 0.1155 & 0.0130 & 1.0000 & 0.4600 & 0.4630 & 0.3194 \\ 0.0094 & 0.3236 & 0.2449 & 0.4600 & 1.0000 & 0.5465 & 0.5999 \\ -0.2329 & 0.5030 & -0.0514 & 0.4630 & 0.5465 & 1.0000 & 0.5683 \\ -0.2406 & 0.4837 & 0.1331 & 0.3194 & 0.5999 & 0.5683 & 1.0000 \end{bmatrix}$$

$$= 0.12694.$$

$$\det(\hat{R}_{11}) = \det \begin{bmatrix} 1.0000 & -0.1380 \\ -0.1380 & 1.0000 \end{bmatrix} = 0.98096.$$

$$\det(\hat{R}_{22}) = \det \begin{bmatrix} 1.0000 & 0.0130 \\ 0.0130 & 1.0000 \end{bmatrix} = 0.99983.$$

$$\det(\hat{R}_{33}) = \det \begin{bmatrix} 1.0000 & 0.5465 & 0.5999 \\ 0.5465 & 1.0000 & 0.5683 \\ 0.5999 & 0.5683 & 1.0000 \end{bmatrix} = 0.39112$$

$$\left( \frac{0.12694}{0.98096 \times 0.99983 \times 0.39112} \right)^{24} = 2.9721 \times 10^{-12},$$

$$-2 \log(2.9721 \times 10^{-12}) = 53.084.$$

## Test du $\chi^2$

$$\det \begin{bmatrix} 1.0000 & -0.1380 \\ -0.1380 & 1.0000 \end{bmatrix} = 0.98096.$$

$$\det \begin{bmatrix} 1.0000 & 0.0130 \\ 0.0130 & 1.0000 \end{bmatrix} = 0.99983.$$

$$\det \begin{bmatrix} 1.0000 & 0.5465 & 0.5999 \\ 0.5465 & 1.0000 & 0.5683 \\ 0.5999 & 0.5683 & 1.0000 \end{bmatrix} = 0.39112.$$

$$\log(V) = \log \left[ \frac{0.12694}{0.98096 \times 0.99983 \times 0.39112} \right] = \log(0.33091) = -1.1059, \text{ ce qui donne, avec } a = 43,375 : 1.1059 \times 43,375 = 47.554.$$

## Annexe 3

**Fig.1 : Simulation de la puissance dans le cas de deux vecteurs normaux**

```
> r<-c(-.4,-.35,-.30,-.25,-.20,-.15,-.10,-.05,.00,.05,.10,.15,.20,.25,.30,.35,.40)
> n<-200 ; t<-1000 ; m<-17
> mu<-c(0,0,0,0)
> phi<-numeric(m)
> X<-array(0,c(n,5,t,m))
> V<-array(0,c(n,2,t,m))
> u<-array(0,c(n,t,m))
> z<-array(0,c(n,t,m))
> p<-matrix(0,t,m)
> for(j in 1:m) {sigma<-matrix(c(1,-.5,r[j],r[j],r[j],-.5,1,r[j],r[j],r[j],
+ r[j],r[j],1,.5,.5,r[j],r[j],.5,1,.5,r[j],r[j],.5,.5,1),5,5)
+ for(k in 1:t){X[,k,j]<-mvrnorm(n,mu,sigma)
+ V[,k,j]<-c(pmax(X[,1,k,j,drop=F],X[,2,k,j,drop=F]),
+ pmax(X[,3,k,j,drop=F],X[,4,k,j,drop=F],X[,5,k,j,drop=F]))
+ for(i in 1:n) {u[i,k,j]<-pmvnorm(mu,sigma,lower=rep(-Inf,5),
+ upper=c(V[i,1,k,j],V[i,1,k,j],V[i,2,k,j],V[i,2,k,j],V[i,2,k,j]))}
+ z[,k,j]<-(u[,k,j])*(1-log(u[,k,j]))
+ out<-ks.test(z[,k,j], "punif")
+ p[k,j]<-out$p.value
+ phi[j]<-sum(1-1*(p[,j]>.1))/t}
+ plot(r,phi)}
> phi
[1] 1.000 1.000 1.000 0.989 0.935 0.728 0.383 0.170 0.108 0.167 0.415 0.663
[13] 0.877 0.963 0.992 0.998 1.000
```

**Fig.2 : Simulation de la puissance dans le cas de deux vecteurs log-normaux**

```
> r<-c(-.4,-.35,-.30,-.25,-.20,-.15,-.10,-.05,.00,.05,.10,.15,.20,.25,.30,.35,.40)
> n<-200 ; t<-1000 ; m<-17
> mu<-c(0,0,0,0)
> phi<-numeric(m)
> X<-array(0,c(n,5,t,m))
> V<-array(0,c(n,2,t,m))
> u<-array(0,c(n,t,m))
> z<-array(0,c(n,t,m))
> p<-matrix(0,t,m)
> for(j in 1:m) {sigma<-matrix(c(1,-.5,r[j],r[j],r[j],-.5,1,r[j],r[j],r[j],
+ r[j],r[j],1,.5,.5,r[j],r[j],.5,1,.5,r[j],r[j],.5,.5,1),5,5)
+ for(k in 1:t){X[,k,j]<-exp(mvrnorm(n,mu,sigma))
+ V[,k,j]<-c(pmax(X[,1,k,j,drop=F],X[,2,k,j,drop=F]),
+ pmax(X[,3,k,j,drop=F],X[,4,k,j,drop=F],X[,5,k,j,drop=F]))
+ for(i in 1:n) {u[i,k,j]<-pmvnorm(mu,sigma,lower=rep(-Inf,5),
+ upper=c(log(V[i,1,k,j]),log(V[i,1,k,j]),log(V[i,2,k,j]),log(V[i,2,k,j]),log(V[i,2,k,j])))}
+ z[,k,j]<-(u[,k,j])*(1-log(u[,k,j]))
+ out<-ks.test(z[,k,j], "punif")
+ p[k,j]<-out$p.value
+ phi[j]<-sum(1-1*(p[,j]>.1))/t}
+ plot(r,phi)}
> phi
```

```
[1] 1.000 1.000 1.000 0.992 0.922 0.736 0.435 0.166 0.096 0.177 0.366 0.618
[13] 0.861 0.956 0.990 1.000 1.000
```

**Fig.3 Simulation de la puissance dans le cas de trois vecteurs normaux**

```
> r<-c(-.4,-.35,-.30,-.25,-.20,-.15,-.10,-.05,.00,.05,.10,.15,.20,.25,.30,.35,.40)
> n<-200 ; t<-1000 ; m<-17
> mu<-c(0,0,0,0,0,0)
> phi<-numeric(m)
> X<-array(0,c(n,6,t,m))
> V<-array(0,c(n,3,t,m))
> u<-array(0,c(n,t,m))
> z<-array(0,c(n,t,m))
> p<-matrix(0,t,m)
> for(j in 1:m) {sigma<-matrix(c(1,-.5,0,r[j],r[j],0,-.5,1,r[j],0,0,r[j],0,r[j],1,.5,r[j],0,
+ r[j],0,.5,1,0,0,r[j],0,r[j],0,1,0,0,r[j],0,0,0,1),6,6)
+ for(k in 1:t){X[,,k,j]<-mvrnorm(n,mu,sigma)
+ V[,k,j]<-c(pmax(X[,1,k,j,drop=F],X[,2,k,j,drop=F]),
+ pmax(X[,3,k,j,drop=F],X[,4,k,j,drop=F]),pmax(X[,5,k,j,drop=F],X[,6,k,j,drop=F]))
+ for(i in 1:n) {u[i,k,j]<-pmvnorm(mu,sigma,lower=rep(-Inf,6),
+ upper=c(V[i,1,k,j],V[i,1,k,j],V[i,2,k,j],V[i,2,k,j],V[i,3,k,j],V[i,3,k,j]))}
+ z[,k,j]<-(u[,k,j])*(1-log(u[,k,j]))+.5*(log(u[,k,j])^2)
+ out<-ks.test(z[,k,j],"punif")
+ p[k,j]<-out$p.value
+ phi[j]<-sum(1-1*(p[,j]>.1))/t}
+ plot(r,phi)}
> phi
[1] 1.000 1.000 1.000 0.984 0.912 0.695 0.414 0.167 0.082 0.169 0.373 0.612
[13] 0.814 0.939 0.988 0.998 1.000
```