

REVUE DE STATISTIQUE APPLIQUÉE

D. DROUET MARI

La dépendance positive entre deux variables de durées : concepts de dépendance et mesures locales de liaison

Revue de statistique appliquée, tome 47, n° 4 (1999), p. 5-24

http://www.numdam.org/item?id=RSA_1999__47_4_5_0

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

LA DÉPENDANCE POSITIVE ENTRE DEUX VARIABLES DE DURÉES : CONCEPTS DE DÉPENDANCE ET MESURES LOCALES DE LIAISON.

D. Drouet Mari

Laboratoire SABRES,

Université de Bretagne Sud

Rue Yves Mainguy, Tohannic, 56000 Vannes

drouet@iu-vannes.fr

RÉSUMÉ

De nombreux concepts de dépendance positive entre deux variables aléatoires ont été introduits dans la littérature. Lehmann [17], par exemple, discute les concepts de dépendance positive par quadrant, de dépendance par la régression monotone, et de dépendance par le rapport de vraisemblance. Shaked [19] a défini d'autres notions adaptées au cas des variables de durées. On montre ici le lien entre ces différents concepts et divers indices de dépendance locale définis par Clayton *et al* [3], Anderson *et al* [1] et Dabrowska *et al* [5] dans le cadre de l'étude de variables de durées. Certains de ces indices permettent en outre d'étudier l'évolution de la dépendance au cours du temps.

Mots-clés : *dépendance positive, survie bivariée, mesure locale d'association, copule, positivité totale*

ABSTRACT

Various concepts of positive dependence between two random variables have been proposed in the literature. For example Lehmann [17] discussed three concepts of positive dependence : positively quadrant dependence, positively regression dependence and likelihood ratio dependence. Shaked [19] introduced dependence concepts in the field of reliability. We link here these concepts with various time-dependent association measures introduced by Clayton *et al* [3], Anderson *et al* [1] and Dabrowska *et al* [5] in the field of bivariate survival analysis. We point out that these local association measures may be used to come into sight the time of maximum association.

Keywords : *positive dependence, bivariate survival analysis, time-dependent association measure, copula, total positivity*

1. Introduction

De nombreux concepts de dépendance positive entre deux variables aléatoires ont été introduits dans la littérature. Lehmann [17], par exemple, pour proposer des tests d'indépendance qui soient sans biais discute trois concepts de dépendance. En se limitant aux variables de durées et en utilisant la théorie de la positivité totale [16], Shaked [19], a construit un cadre unificateur pour un grand nombre de concepts de dépendance.

Cependant les problèmes sont souvent plus pratiques, et on a parfois besoin de quantifier cette dépendance : par exemple les délais d'apparition d'une maladie génétique chez une paire de jumeaux peuvent être plus ou moins liés selon le degré de ressemblance de ces jumeaux.

Cette dépendance peut évoluer au cours du temps, certaines maladies génétiques par exemple ne s'expriment qu'à des ages élevés. De même si l'on s'intéresse aux délais d'apparition de deux événements successifs chez un même sujet (par exemple les délais début du traitement-rechute, rechute-décès dans le cas d'une maladie) on peut désirer identifier le moment d'association maximale entre ces durées. Des indices locaux d'association ont été proposés dans ce but par Oakes [18], Anderson *et al* [1] et Dabrowska *et al* [5].

Après avoir rappelé les différents concepts de dépendance et leurs emboîtements, on montre le lien entre certains d'entre eux et les différents indices. On montre aussi que l'évolution des indices au cours du temps est déterminée par la dépendance sous-jacente.

La dépendance précoce ou tardive peut être étudiée à travers les variations du rapport Θ de deux risques conditionnels. A partir de la présentation de trois distributions bivariées classiques, on illustre les notions de dépendance précoce, tardive et médiane et le calcul des différents indices.

Une application, avec le calcul de trois indices locaux sur des données réelles est aussi présentée.

2. Le risque bivarié et les risques conditionnels

Soient X et Y deux variables aléatoires de durées dont on étudie la liaison. Elles prennent des valeurs positives ou nulles, et on suppose que leur densité jointe $f(x, y)$ existe. On appelle $S(x, y) = P(X > x, Y > y)$, la fonction de survie du couple (X, Y) et $h(x, y) = f(x, y)/S(x, y)$, le risque bivarié.

La liaison entre (X, Y) peut s'étudier par l'intermédiaire du risque conditionnel $h(y/X = x) = \frac{f(y/X = x)}{S(y/X = x)}$, c'est-à-dire le risque de Y sachant que X est décédé en x , désigné par $h_{2/1}$, ou bien par $h(y/X > x)$, c'est-à-dire le risque de Y sachant que X a survécu au delà de x , ou est censuré en x , appelé h_2 (en inversant le rôle de x et y , on définit aussi $h_{1/2}$ et h_1).

Soit D_1 , le symbole $\partial/\partial x$, D_2 le symbole $\partial/\partial y$, et soit L la fonction $(x, y) \rightarrow -\text{Log}S(x, y)$ et $l = -D_1 D_2 L$. Alors, un calcul simple montre que :

$$\begin{aligned}
 h(x, y) &= \frac{D_1 D_2 S}{S} \\
 h_{2/1} &= \frac{-D_1 D_2 S}{D_1 S} \\
 h_2 &= \frac{-D_2 S}{S} = D_2 L \\
 D_2 D_1 L &= \frac{-D_1 D_2 S}{S} + \frac{D_1 S}{S} \cdot \frac{D_2 S}{S}
 \end{aligned} \tag{1}$$

D'où :

$$l(x, y) = h(x, y) - h_1(x, y).h_2(x, y) \tag{2}$$

Les deux risques $h_{2/1}$ et h_1 sont liés. En effet :

$$D_1 h_2 = D_2 h_1 = D_1 D_2 L$$

d'où :

$$h_{2/1} - h_2 = -\frac{D_1 D_2 L}{D_1 L} = -\frac{D_2 h_1}{h_1} \tag{3}$$

3. Les différentes notions de dépendance

Lehmann [17] introduit trois concepts de plus en plus forts de la dépendance positive : la dépendance positive par quadrant (DPQ), la dépendance par la régression monotone (DRM), la dépendance par le rapport de vraisemblance (DRV) :

$$DPQ : S(x, y) \geq S_1(x).S_2(y) \quad \forall x \quad \forall y$$

où $S_1(x) = P(X > x)$ et $S_2(y) = P(Y > y)$ désignent les fonctions de survies marginales.

$$DRM \text{ en } x : S(y/X = x) \nearrow x \quad \forall y$$

$$DRV : f(x, y)f(x', y') \geq f(x, y')f(x', y) \quad \forall x < x' \quad \forall y < y'$$

On a [17] :

$$DRV \rightarrow DRM \text{ en } x \text{ et en } y \rightarrow DPQ$$

Exemple : Si A, B, C , sont 3 variables aléatoires indépendantes et de même loi, par exemple exponentielle $(\gamma(1, 1))$, alors $X = A+B$ et $Y = A+C$ sont dépendantes, et de même loi $\gamma(1, 2)$. Le coefficient de corrélation linéaire $r(x, y) = 0,5$. Ce couple (X, Y) est DRV, parce que la densité exponentielle est une fonction Polya d'ordre 2 [16], et donc aussi DRM en x et y et DPQ.

A partir de la notion de fonction réelle totalement positive d'ordre 2 en x et en y (TP2) [16], Shaked [19] propose un ensemble de définitions emboîtées de la

dépendance positive pour deux durées. Appliquée à la densité $f(x, y)$, la définition TP2¹ n'est autre que DRV, Shaked nomme cette dépendance DTP(0,0). Puis il définit :

– DTP(0, 1) (resp DTP(1, 0)) lorsque $-D_1S$ (resp $-D_2S$) est TP2,

– DTP(1, 1) si $S(x, y)$ est TP2.

– DTP(1, 2) si $E(Y - y/Y > y, X > x) = m(y/X > x)$ est TP2.

– DTP(m, n) si

$$\int_y^\infty \int_{y_{n-1}}^\infty \dots \int_{y_1}^\infty \int_x^\infty \int_{x_{m-1}}^\infty \dots \int_{x_1}^\infty f(x_0 y_0) dx_0 \dots dx_{m-1} dy_0 \dots dy_{n-1} \text{ est TP2.}$$

Shaked montre les implications suivantes :

$$DTP(0, 0) \rightarrow DTP(0, 1) \rightarrow DTP(1, 1) \rightarrow DTP(1, 2) \rightarrow \dots DTP(m, n) \quad \forall m > 1,$$

$$\forall n > 2$$

$$DTP(0, 1) \rightarrow DRM$$

Shaked démontre des propriétés caractéristiques de certaines de ces dépendances positives. En particulier :

– la dépendance DTP(0, 1) est équivalente à la décroissance du risque conditionnel $h_{2/1}(x, y) = h(y/X = x)$ selon x ,

– la dépendance DTP(1, 1) à la décroissance du risque $h_2(x, y) = h(y/X > x)$ selon x et aussi du risque $h_1(x, y) = h(x/Y > y)$ selon y ,

– la dépendance DTP(1, 2) est équivalente à la croissance en x de l'espérance de vie restante conditionnelle $m(y/X > x) = E(Y - y/Y > y, X > x)$.

4. Les indices locaux de dépendance

4.1 L'espérance de vie restante conditionnelle

A partir de la notion d'espérance de vie restante $m(y) = E(Y/Y > y) - y$ Anderson *et al* [1] considère un indice normalisé, mais non symétrique pour exprimer la dépendance :

$$\Phi(x, y) = \frac{m(Y/X > x)}{m(y)} \quad (4)$$

Cet indice est utilisé par les démographes, pour étudier, par exemple, le lien entre les durées de vie de deux personnes apparentées. Si X et Y sont indépendantes alors Φ vaut un. Plus les variables X et Y sont liées, plus $\Phi(x, y)$ est grand. A y constant Φ devrait croître avec x : cette dernière condition est équivalente à la dépendance DTP(1, 2).

¹ $f(x, y)$ est TP2 en x et y , si $f(x, y) \geq 0 \forall x, \forall y$ et f remplit la condition DRV

4.2 Le rapport des probabilités de survie

Anderson *et al.* utilise un deuxième indice :

$$\Psi(x, y) = \frac{S(x, y)}{S_1(x) \cdot S_2(y)} \quad (5)$$

S_1 et S_2 étant les survies marginales. Ψ vaut un si (X, Y) sont indépendantes. La condition $\Psi > 1$ correspond à la définition de *DPQ*. Or, on peut aussi écrire : $\Psi(x, y) = \frac{S(y/X > x)}{S_2(y)}$. Quand x vaut zéro, Ψ vaut un. Si la liaison est positive, $\Psi(x, y)$ ne peut donc décroître selon x : cette propriété (*RTI* : queue à droite croissante) est mentionnée par Lehmann comme intermédiaire entre *DPQ* et *DRM*.

4.3 Le rapport de covariance conditionnelle

Ψ ne traduit pas très finement l'évolution de la dépendance au cours du temps. A sa place Dabrowska *et al.* [5] propose d'utiliser $l(x, y)$:

$$l(x, y) = D_1 D_2 \log(\Psi(x, y)) = h(x, y) - h_1(x, y) \cdot h_2(x, y) \quad (6)$$

D'après l'équation 2, $l(x, y) = -D_1 D_2 L = -D_1 h_2 = -D_2 h_1$. Si X et Y sont indépendantes, le risque conditionnel h_2 ne dépend pas de x et donc sa dérivée par rapport à x est nulle. Donc l'indépendance entre X et Y se traduit par $l(x, y) \equiv 0$.

Si X et Y sont liées positivement alors $l(x, y)$ est positive, donc $-D_1 h_2 > 0$, donc h_2 décroît selon x , ce qui équivaut à la dépendance *DTP*(1, 1) (S est *TP2*).

4.4 Le rapport des risques conditionnels

Clayton [3], puis Oakes [18] pour étudier la dépendance entre les deux durées utilisent le rapport

$$\Theta = \frac{h_{2/1}}{h_2}$$

c'est-à-dire :

$$\Theta = \frac{D_1 D_2 S \cdot S}{D_1 S \cdot D_2 S} = \frac{h}{h_1 h_2} \quad (7)$$

Si X et Y sont indépendants, alors :

$$h_{2/1} = h_2 = h_0$$

h_0 étant le risque marginal de Y et leur rapport $\Theta \equiv 1$. Si X et Y sont en association positive alors Θ est plus grand que un. Contrairement aux deux risques conditionnels

Θ est symétrique en x et y . Remarquons que Θ est relié à 1, en effet, d'après les équations 6 et 7 :

$$\Theta - 1 = l/(h_1 h_2) \quad (8)$$

4.4.1 La condition « $\Theta > 1$ »

Cette condition est équivalente à S est TP2; en effet $\Theta > 1$ est équivalent à $h_{2/1} - h_2 > 0$, ce qui d'après l'équation (3) équivaut à $D_2 h_1 < 0$ puisque $h_1 > 0$ (de façon symétrique on a aussi $D_1 h_2 < 0$); ceci équivaut alors à S est TP2. D'après l'équation 8, « $\Theta > 1$ » équivaut aussi à « $l > 0$ ».

4.4.2 La décroissance de Θ selon x

La notion de dépendance positive associée à « $\Theta > 1$ » est une notion assez sommaire. Une notion plus fine est « $\Theta \searrow 1$ quand $x \rightarrow \infty$ » (resp $y \rightarrow \infty$) [4]. En effet, au voisinage de $\{X = 0\}$, h_2 est égal au risque marginal h_0 : X n'apporte aucune information sur Y , $h_{2/1}$ est lui probablement très grand si la liaison est positive. Θ est donc grand aussi. Lorsque x augmente, les informations apportées par les événements $\{X = x\}$ et $\{X > x\}$ tendent à devenir équivalentes, et donc le rapport Θ tend vers un par valeurs supérieures.

Exemple : $X = A + B$ et $Y = A + C$, avec A, B, C indépendantes et de même loi $\gamma(1, 1)$. Dans ce cas :

$$\Theta = \frac{|y - x| + 2 - \exp(-\min(x, y))}{|y - x| + 1 - \exp(-\min(x, y))}. \quad (9)$$

Donc, Θ décroît de $\frac{|y| + 1}{|y|}$ vers un quand x tend vers l'infini quelque soit y .

A quelle notion de dépendance peut on associer la condition de décroissance vers un de Θ selon x ? Elle implique, puisque d'après 4.4.1, h_2 est une fonction décroissante de x , que $h_{2/1} - h_2$ décroît vers zéro par valeurs positives. Ceci implique d'après l'équation 3 que :

$$\frac{D_2 h_1(x, y)}{h_1(x, y)} \nearrow \text{selon } x \quad (10)$$

Ce qui correspond à h_1 est TP2, en utilisant une caractérisation des fonctions TP2 d'après Karlin [16]

Puisque $h_{2/1} - h_2$ décroît selon x , ainsi que h_2 , alors nécessairement $h_{2/1}$ décroît aussi, sinon Θ ne pourrait décroître. Or la condition « $h_{2/1} \searrow$ » selon x est équivalente à une dépendance $DTP(0, 1)$ pour le couple (X, Y) . Donc la condition Θ décroît selon x par valeurs positives est plus forte que la condition $DTP(0, 1)$, cependant on ne peut pas l'ordonner par rapport à $DTP(0, 0)$: nous la noterons $DTPL(0, 1)$.

5. Lien indices/dépendances

type de dépendance	Fonctions TP2	Condition sur les indices
<i>DPQ</i>		$\Psi > 1$
<i>RTI</i>		$\Psi \nearrow$ selon x
<i>DTP(1, 2)</i>	$m(y/X > x)$ est TP2	$\Phi \nearrow$ selon x
<i>DTP(1, 1)</i>	S est TP2	$h_2 \searrow$ selon x
		$l > 0$
		$\Theta > 1$
<i>DTP(0, 1)</i>	$-D_1S$ est TP2	$h_{2/1} \searrow$ selon x
<i>DTP(0, 0)</i>	f est TP2	$D_2f/f \nearrow$ selon x^*
<i>DTPL(0, 1)</i>	h_1^{**} est TP2	$\Theta \searrow 1+$ selon x

* : propriété caractéristique d'une fonction TP2.

** : cette condition est nécessaire, mais non suffisante pour que Θ décroisse selon x .

Les types de dépendance sont ordonnées de la première à la dernière ligne du tableau (sauf les deux dernières, et sauf *DTP(1, 2)* et *RTI*). La condition de décroissance sur Θ implique donc les conditions sur tous les autres indices. Donc, si l'on peut prouver la décroissance de Θ selon x (ou selon y) ou sa croissance selon S , alors il n'est pas nécessaire d'étudier les autres indices pour prouver la dépendance positive.

6. Exemples : trois distributions archimédiennes

6.1 Le type de dépendance

Les distributions archimédiennes désignent un ensemble de distributions bi-variées dont la fonction de survie S s'expriment en fonction des marges S_1 et S_2 sous la forme :

$$S(x, y) = p[q(S_1(x) + q(S_2(y))]$$

où p est une fonction positive ou nulle, décroissante, vérifiant $p(0) = 1$, et ayant une dérivée seconde positive ou nulle, q étant la fonction inverse de p . La dépendance s'exprime, en général, par un paramètre réel, dont les variations permettent d'obtenir tous les degrés de dépendance, de l'indépendance à la dépendance maximale (borne supérieure de Fréchet [8])

Oakes [18] a montré que l'on pouvait caractériser ces distributions par le fait que Θ s'exprime comme une fonction de $S(x, y)$.

$$\Theta(x, y) = \Theta^*(S(x, y))$$

où

$$\Theta^*(s) = -sq''(s)/q'(s)$$

q' et q'' étant les deux premières dérivées de q .

Étudions trois de ces distributions qui illustrent les notions de dépendance précoce, tardive, et médiane, et cherchons quels sont les indices qui rendent compte le mieux de ces dépendances.

1) Le modèle de Clayton :

Dans ce modèle (qui porte aussi le nom de «Cook et Johnson») :

$$q(v) = \left(\frac{1}{v}\right)^{\alpha-1} - 1$$

D'où :

$$S(x, y) = [(S_1(x))^{1-\alpha} + (S_2(y))^{1-\alpha} - 1]^{1/1-\alpha}, \quad \alpha > 1.$$

Le rapport Θ des deux fonctions de risques conditionnels est constant et égal à α . Cette propriété est caractéristique de ce modèle. Cette famille ne vérifie donc pas la propriété 4.4.2.

Cependant ici :

$$h_1 = -\frac{D_1 S}{S} = -[S_1^{1-\alpha} + S_2^{1-\alpha} - 1]^{-1} S_1^{-\alpha} D_1 S_1$$

D'où :

$$D_2 h_1 = (1 - \alpha)[S_1^{1-\alpha} + S_2^{1-\alpha} - 1]^{-2} S_1^{-\alpha} D_1 S_1 S_2^{-\alpha} D_2 S_2$$

D'où :

$$\frac{D_2 h_1}{h_1} = (\alpha - 1)[S_1^{1-\alpha} + S_2^{1-\alpha} - 1]^{-1} S_2^{-\alpha} D_2 S_2 \quad (11)$$

Quand x augmente, S_1 diminue, l'expression entre crochet augmente, donc son inverse diminue, mais $D_2 S_2$ étant négatif, $\frac{D_2 h_1}{h_1}$ augmente, ce qui prouve que h_1 est TP2 [16].

Cette famille est aussi DRV (cf. annexe B). La figure 1-a montre que la densité est concentrée pour les valeurs élevées de x et y .

2) Le modèle de Hougaard [14] :

$$q(v) = (-\ln(v))^{\frac{1}{\alpha}}$$

$$S(x, y) = \exp(-[\{-\ln S_1\}^{1/\alpha} + \{-\ln S_2\}^{1/\alpha}]^\alpha) \quad 0 < \alpha < 1$$

On retrouve une famille de distributions aux valeurs extrêmes proposée par Gumbel [12]. Dans ce modèle

$$\Theta^*(s) = 1 + (1 - \alpha)/(-\alpha \ln(s))$$

Θ tend donc vers un, par valeurs supérieures quand S tend vers zéro. Ceci implique alors que h_1 est TP2. Comme pour la famille de Clayton, on montre (cf. annexe C) que cette famille est *DRV*. La figure 1-b illustre la dépendance précoce.

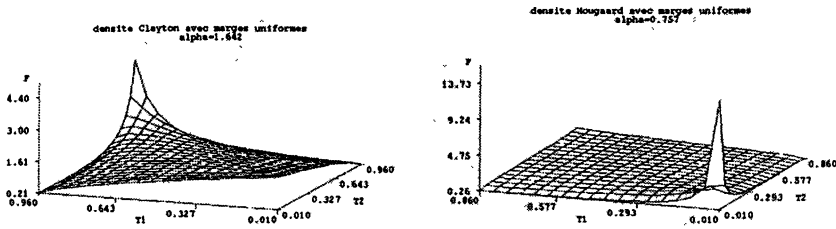


FIGURE 1
densités avec marges uniformes et $\tau = 0.243$,
a) modèle de Clayton, b) modèle de Hougaard

3) le modèle de Frank [7] :

$$q(v) = \ln\left(\frac{1-\gamma}{1-\gamma^v}\right) \quad 0 < \gamma < 1$$

$$S(x, y) = \log_\gamma\left(1 + \frac{(\gamma^{S_1} - 1)(\gamma^{S_2} - 1)}{\gamma - 1}\right) \quad 0 < \gamma < 1$$

Cette famille a été étudiée en détail par Genest [9], qui montre notamment qu'elle est *DRV*. Le rapport $\Theta^*(s) = -s \cdot \ln(\gamma) \cdot (1 - \gamma^s)^{-1}$ décroît de $\frac{\ln(\gamma)}{\gamma - 1}$ vers un S décroît de un à zéro. La densité est symétrique par rapport aux droites $x = y$ et $x = 1 - y$. On voit (figure 2) que la liaison entre X et Y est la plus forte pour les valeurs médianes de X et Y .

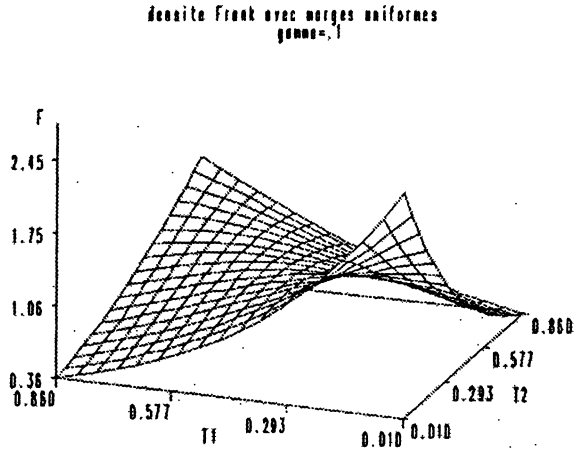


FIGURE 2

Densité modèle de Frank, $\tau = 0.243$, et marges uniformes

6.2 Evolution des indices

6.2.1 Le rapport Θ

La figure 3 représente l'évolution de Θ en fonction de S pour une même valeur du τ de Kendall ($\tau = 0.243$), pour les trois familles, avec les mêmes marges uniformes dans les trois cas.

- Pour les valeurs de S élevées ($S > 0.75$), c'est-à-dire au début de l'évolution selon le temps Θ est le plus grand pour la famille de Hougaard, exprimant la dépendance précoce.

- Pour les valeurs de S inférieures à 0.5, le rapport Θ est le plus élevé pour la famille de Clayton, exprimant le fait que la connaissance du décès de l'un des membres du couple (X, Y) , augmente encore beaucoup le risque de décès de l'autre membre : il y a donc une dépendance tardive.

- Pour les valeurs intermédiaires, c'est dans la famille de Frank que le rapport Θ est le plus élevé, on parlera donc de dépendance médiane.

- Pour les valeurs très faibles de S ($S < 0.1$), le rapport Θ est voisin de un pour les familles de Frank et Hougaard : la connaissance du décès de l'un des membres du couple (X, Y) n'apporte plus d'information sur le risque de décès de l'autre membre dans ces deux familles

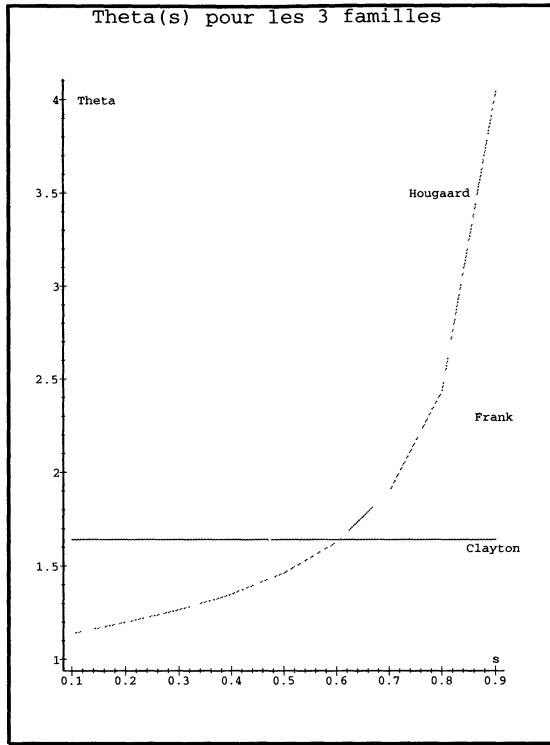


FIGURE 3

Evolution de Θ selon S pour les trois familles, $\tau = 0.243$ et marges uniformes

6.2.2 Evolution de $\Phi(x, x)$

L'espérance de vie restante peut se calculer par([6] *et al.*, page 24) :

$$m(y) = \frac{\int_y^{+\infty} S_2(u) du}{S_2(y)}$$

On obtient donc pour Φ :

$$\Phi(x, y) = \frac{m(y/X > x)}{m(y)} = \frac{S_2(y) \cdot \int_y^{+\infty} S(u/x) du}{\int_y^{+\infty} S_2(u) du \cdot S(y/x)}$$

On a représenté ici les variations de $\Phi(x, x)$, en supposant encore les marges uniformes pour les trois familles, et pour la même valeur de $\tau(0.243)$.

On a vu au paragraphe 4.1 que $\Phi(x, y)$ est croissant en x quelque soit y , dès que la dépendance est DTP(1,2). En revanche, rien ne prouve que $\Phi(x, x)$ soit croissant en x , parce que $\Phi(x, y)$ n'est pas symétrique en x et en y .

Dans la famille de Frank, cette fonction est, à cause des marges identiques et à cause des symétries de la densité (par rapport à $x = y$ et à $x = 1 - y$) symétrique par rapport à la droite $x = 1/2$. Donc pour cette famille $\Phi(x, x)$ décroît à partir de la date médiane : au delà de cette date, la connaissance de l'événement $\{X > x\}$, ne modifie plus beaucoup l'espérance de vie restante pour Y , et donc le numérateur et le dénominateur de Φ tendent à devenir égaux. Cette situation paraît assez naturelle.

Cependant, ce n'est pas ce qu'on observe pour les deux autres familles, où $\Phi(x, x)$ est toujours croissant. L'opposition précoce/tardif que l'on avait observé avec Θ pour ces deux familles n'apparaît pas ici, si ce n'est par la pente de Φ , beaucoup plus élevée dans le cas de la famille de Clayton.

6.2.3 Evolution de $\Psi(x, x)$

On a vu, dans le paragraphe 4.2 que $\Psi(x, y)$ est croissant en x quelque soit y , dès que la dépendance est RTI. Mais Ψ est symétrique lorsqu'on échange x et y . Donc $\Psi(x, x)$ est croissant en x . C'est bien ce qu'on observe pour les trois familles. On verra au paragraphe suivant, que, pour la famille de Clayton, $\Psi(x, x)$ tend vers l'infini lorsque x tend vers sa valeur supérieure.

Si les marges sont uniformes, la comparaison des trois $\Psi(x, x)$, revient à celles des trois fonctions de survie. La dépendance, tardive dans le cas de la famille de Clayton, précoce dans le cas de Hougaard, et médiane dans le cas de Frank implique que :

$$S_C(x, x) > S_F(x, x) > S_H(x, x)$$

(où S_C, S_F et S_H désignent les fonctions de survie de Clayton, Hougaard et Frank) à partir d'une certaine date x . C'est bien ce qu'on observe sur le graphique à partir de $x = 0.6$, mais l'examen seul de $\Psi(x, x)$, ne permettrait pas en retour d'en déduire le type de dépendance.

7. Illustration

Les données suivantes concernent 109 patientes soignées par deux chimiothérapies successives *AV* et *CMF* pour un cancer du sein [13]. Les malades reçoivent d'abord *AV* pendant huit cycles ou moins si l'intolérance survient ou si la maladie a progressé. Puis elles reçoivent *CMF* pour un total de six cycles, ou moins pour les mêmes raisons. Si le traitement est interrompu par la progression de la maladie ou pour des raisons non liées au traitement, les données sont censurées.

Le tableau (page 17) présente les effectifs pour chaque durée de tolérance aux deux produits. Il y a quatre nombres selon que les durées correspondent ou non à des censures : en haut à gauche pour les données non censurées, en haut à droite pour des données censurées en T_2 uniquement, en bas à gauche pour des données censurées en T_1 uniquement, et en bas à droite, pour les doubles censures.

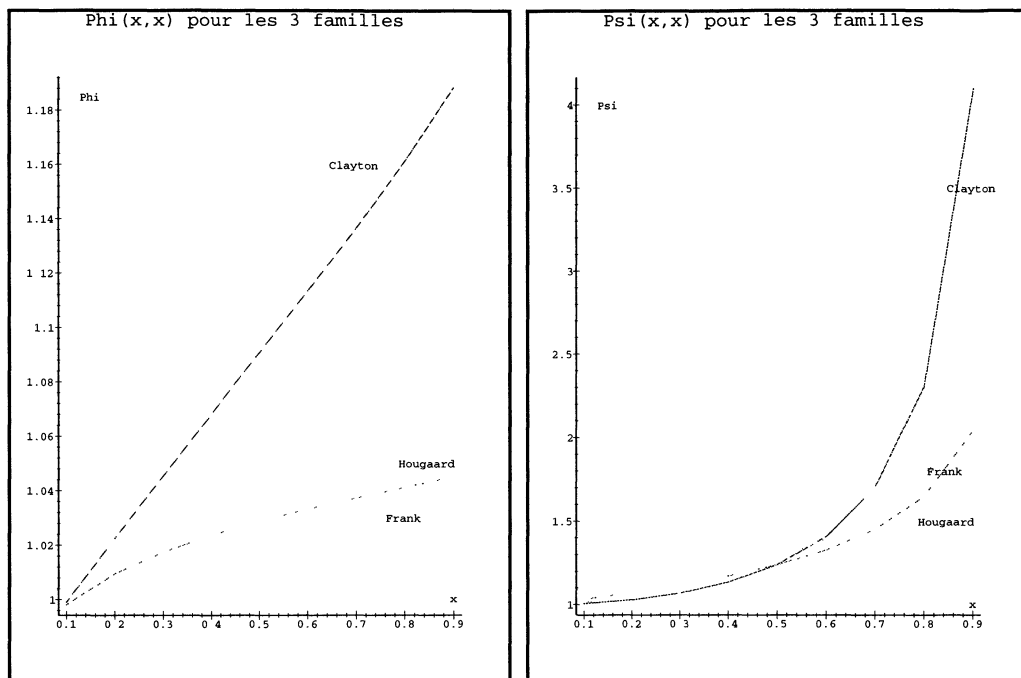


FIGURE 4

Evolution de $\Phi(x, x)$ et $\Psi(x, x)$ pour les trois familles, $\tau = 0.243$ et marges uniformes

Durée	$T_2 = 1$ (CMF)	$T_2 = 2$	$T_2 = 3$	$T_2 = 4$	$T_2 = 5$	$T_2 \geq 6$
$T_1 = 1$ (AV)	10 3 01	2 2 1 1	2 0 0 1	1 0 0 0	0 0 0 0	1 0 0 0
$T_1 = 2$	9 2 3 0	2 2 2 0	1 0 0 0	2 0 0 0	0 0 0 0	2 0 1 0
$T_1 = 3$	3 1 2 1	0 0 0 0	0 0 0 1	2 1 0 0	1 0 0 0	1 0 0 0
$T_1 = 4$	1 2 0 0	0 1 1 1	1 0 1 0	1 0 0 1	0 1 0 0	1 1 0 0
$T_1 = 5$	1 1 0 0	0 0 0 0	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0
$T_1 \geq 6$	2 0 0 0	2 0 2 1	1 0 1 2	0 0 2 0	0 0 1 0	0 0 3 2

Afin de savoir si les durées de tolérance aux deux produits sont liées, on calcule les valeurs des trois indices Θ , Φ et Ψ dans les 25 cases du tableau où c'est possible. Parce que les données correspondent à deux événements successifs pour un même individu, avec des phénomènes de censure différents, la fonction de survie bivariable est difficile à estimer de façon cohérente. Hanley *et al.* [13] propose une estimation itérative en utilisant un algorithme EM. Le tableau suivant rappelle les probabilités qu'il a obtenues.

			\hat{p}_{ij}				\hat{p}_i
i/j	1	2	3	4	5	≥ 6	total
1	0,091	0,025	0,038	0,019	0,0	0,019	0,19
2	0,082	0,025	0,014	0,029	0,0	0,03	0,18
3	0,032	0,0	0,0	0,023	0,017	0,017	0,10
4	0,012	0,0	0,015	0,017	0,0	0,05	0,10
5	0,012	0,015	0,0	0,0	0,0	0,0	0,02
≥ 6	0,135	0,068	0,032	0,03	0,019	0,132	0,41
total	0,36	0,14	0,09	0,12	0,04	0,25	

A partir de ces probabilités, on estime les $S(i, j)$ par la somme des probabilités dans l'intervalle $\square i, 6 \square X \square j, 6 \square$, les $S(j)$ par $\sum_{k>j} \hat{p}_k$, on désigne par $D_1 \widehat{S}(i, j) = \sum_{k>j} p(\hat{i}, k)$, et de même pour $D_2 S(i, j)$. On estime alors Θ par :

$$\Theta(\widehat{i}, j) = \frac{\hat{p}_{ij} \hat{S}_{ij}}{D_1 \hat{S}_{ij} D_2 S_{ij}}$$

et Ψ par :

$$\Psi(\widehat{i}, j) = \frac{\hat{S}_{ij}}{\hat{S}_i \hat{S}_j}$$

On estime l'espérance de vie restante conditionnelle par :

$$\widehat{m}_{j/i} = \frac{\sum_{k \geq j} \hat{S}_{ik}}{\hat{S}_{ij}}$$

et de façon analogue \hat{m}_j . On estime alors Φ par :

$$\widehat{\Phi}(i, j) = \frac{\widehat{m}_{j/i}}{\widehat{m}_j}$$

Les trois paramètres ont dans chacune des vingt-cinq cases des estimations légèrement plus grandes que un montrant la dépendance entre les durées. Le rapport $\Psi(i, i)$ est bien croissant avec i . Mais il est difficile d'affirmer que Θ croît avec S , ou que $\Phi(i, i)$ croît avec i .

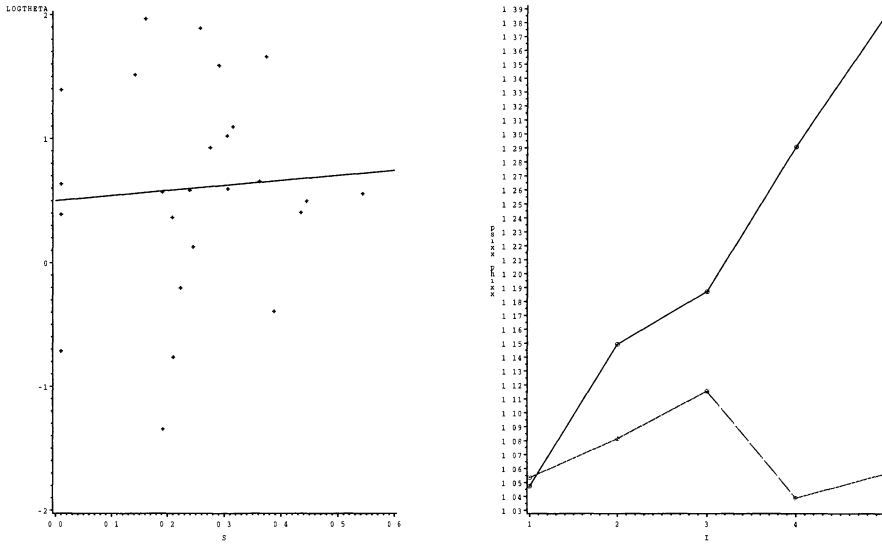


FIGURE 5

a : Evolution de $\log(\Theta)$ selon S , b : $\Psi(i, i)$ (ligne supérieure) et $\Phi(i, i)$ selon i

8. Discussion

Le rapport Θ a différentes interprétations, Oakes [18] montre notamment que $\frac{\Theta - 1}{\Theta + 1}$ est une version conditionnelle du τ de Kendall. Plus précisément si l'on dispose d'un échantillon de la distribution bivariee :

$$T_i = (X, Y)_i \quad i = 1, \dots, n$$

et si l'on définit la concordance (resp la discordance) d'une paire $(T_i, T_{i'})$ par : $(X_i - X_{i'})(Y_i - Y_{i'}) \geq 0$ (resp ≤ 0) alors :

$$\Theta(x, y) = \frac{Pr[(T_i, T_{i'}) \text{ concordantes} / X \geq x, Y \geq y]}{Pr[(T_i, T_{i'}) \text{ discordantes} / X \geq x, Y \geq y]}$$

Θ mesure donc la dépendance restante au delà de (x, y) . En cas de dépendance positive ce rapport est plus grand que un.

Tant que le moment de dépendance maximale n'a pas eu lieu, ce rapport doit rester pratiquement constant, puis baisser rapidement juste après. Ceci est illustré par le comportement des trois familles que l'on vient d'étudier. Dans la famille de Clayton, le rapport Θ est constant : le rapport des nombres de paires concordantes et discordantes ne varie pas au cours du temps, on est donc dans le cas d'une dépendance tardive. Le cas de la famille de Hougaard est inverse : le rapport Θ décroît de l'infini à un quand la fonction de survie S décroît de un à zéro : on est donc dans le cas d'une dépendance précoce. La famille de Frank est intermédiaire, avec un rapport Θ qui décroît d'une valeur finie $\frac{\ln(\gamma)}{\gamma-1}$ ($0 < \gamma < 1$) à un dans les mêmes conditions.

Le modèle de Clayton (à cause de sa définition comme modèle de fragilité gamma) est souvent choisi pour ajuster une distribution bivariable à cause de la plus grande «simplicité» des calculs, en particulier lorsqu'il s'agit de maximiser la vraisemblance. Il paraît plus judicieux de choisir la famille de distributions, c'est-à-dire le type de dépendance après un examen de l'évolution de Θ : Oakes [18] en propose une estimation après discrétisation; on peut aussi utiliser les estimations non paramétriques de h , h_1 et h_2 de Dabrowska *et al.* [5] pour estimer Θ .

8.1 Lien avec les notions de «upper tail dependence» et «lower tail dependence» de Joe [15]

Les familles de Clayton, Hougaard et Frank sont des cas particuliers de copules, c'est-à-dire ont la propriété que :

$$S(x, y) = C(S_1(x), S_2(y))$$

où la fonction $C : [0, 1]^2 \rightarrow [0, 1]$ et où S_1 et S_2 sont les fonctions de survie marginales.

Joe [15] définit la notion de «upper tail dependence» comme la limite δ si elle existe de $\frac{C(u, u)}{u}$ lorsque u tend vers zéro (par valeurs supérieures); si U et V sont des variables aléatoires uniformes sur $[0, 1]$, alors :

$$\delta = \lim_{u \rightarrow 0+} \frac{C(u, u)}{u} = \lim_{u \rightarrow 0+} Pr(U \leq u/V \leq u)$$

Si C' désigne la copule associée à la fonction de répartition correspondant à S , c'est-à-dire :

$$C'(u, v) = 1 - u - v + C(u, v)$$

alors, Joe définit la «lower tail dependence» par :

$$\gamma = \lim_{u \rightarrow 1-} \frac{C'(u, u)}{1 - u}$$

γ s'interprète alors comme la limite de $Pr(U > u/V > u)$ quand u tend vers un.

On voit que ces deux notions sont très proches des notions de dépendance tardive et précoce définies à partir de Θ . En particulier pour la famille de Clayton, à la dépendance tardive correspond une valeur de δ non nulle, $\delta = 2^{-1/(\alpha-1)}$, en revanche $\gamma = 0$. Pour la famille de Hougaard, la dépendance précoce se traduit par $\gamma = 2 - 2^\alpha$ et par $\delta = 0$. Pour la famille de Frank ces deux limites sont nulles, exprimant le fait qu'il n'y a ni dépendance précoce, ni dépendance tardive. Les indices Ψ et δ sont liés. Si le couple (X, Y) défini sur $[0, 1]^2$ a des marges uniformes $S_1(x) = 1 - x$, $S_2(y) = 1 - y$ alors :

$$\delta = \lim_{x \rightarrow 1^-} \Psi(x, x)(1 - x)$$

Dans le cas de la famille de Clayton, on voit que Ψ tend vers l'infini, quand x tend vers un, puisque δ est non nul.

Remerciements

Merci à Michel Bonneau et Evans Gouno pour leurs commentaires, ainsi qu'à Pierre Cazes et aux relecteurs de la R.S.A.

Références

- [1] ANDERSON J.E., LOUIS T.A., HOLM N.V., HARVALD B., *Time-dependent association measures for bivariate survival distributions*, JASA, Vol. **87** n°419, 641–650, 1992
- [2] CAPERAÀ P. GENEST C., *Concepts de dépendance et ordres stochastiques pour des lois bidimensionnelles*, The Canadian Journal of Statistics, **18**, 315–326, 1990.
- [3] CLAYTON D., CUZICK J., *Multivariate generalizations of the proportional hazards model*, J.R. Statist. Soc. A, **148**, Part 2, 82–117, 1985.
- [4] DROUET MARI D., PETIOT J.F., *Etude de la liaison entre deux durées par le modèle de Cox*, Bulletin de l'Institut International de la Statistique, 50^e session, **livre 1**, 294–295, 1995.
- [5] DABROWSKAA D.M., DUFY D.L., ZHANG D.Z., *Hazard and density estimation from bivariate censored data*, Nonparametric Statist. (in press), 1997.
- [6] DROESBEKE J.J., FICHET B., TASSI P., *Analyse statistique des durées de vie*, Economica, 1989
- [7] FRANK M.J., *On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$* , Aequationes Math., **19**, 194–226, 1979.
- [8] FRECHET M., *Sur les tableaux de corrélation dont les marges sont données*, Annales de l'Université de Lyon, Sec A, Ser 3, **14**, 53–77, 1951.
- [9] GENEST C., *Frank's family of bivariate distributions*, Biometrika, **74**, 3, 549–555, 1987.

- [10] GENEST C., GHOUDI K., RIVEST L.P., *A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions*, *Biometrika*, **82**, 3, 543–552, 1995.
- [11] GENEST C., MACKAY R.J., *Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données*, *the Canadian Journal of Statistics*, Vol. 14, **2**, 145–149, 1986.
- [12] GUMBEL E. J., *Bivariate exponential distributions*, *JASA*, **55**, 698–707, 1960.
- [13] HANLEY J. H., PARNES M.N., *Non parametric estimation of a multivariate distribution in the presence of censoring*, *Biometrics*, **39**, 129–139, 1983.
- [14] HOUGAARD P., *A class of multivariate failure time distributions*, *Biometrika*, **73**, 671–678, 1986.
- [15] JOE H., *Parametric Families of Mutivariate Distributions with given margins*, *Journal of Multivariate Analysis*, **46**, 262–282, 1993
- [16] KARLIN S., *Total positivity*, Stanford University Press, 1968
- [17] LEHMANN E.L., *Some concepts of dependence*, *Annals of Mathematical Statistics*, **37**, 1137–1153, 1966
- [18] OAKES D., *Bivariate survival models induced by frailties*, *JASA*, **84**, 406, 487–493, 1989.
- [19] SHAKED M., *A family of concepts of dependence for bivariate distributions*, *JASA*, **72**, 359, 642–650, 1977.

A. Annexe : copule archimédienne et dépendance DRV

Montrons qu'une copule archimédienne $S(x, y) = (p(q(S_1(x) + q(S_2(y))))$, q étant la fonction inverse de p et telle que p a une dérivée troisième, est DRV si et seulement si,

$$\frac{p'''(v)}{p''(v)} \text{ est une fonction croissante de } v. \quad (12)$$

En effet : Soit $v(x, y) = q(S_1(x) + q(S_2(y)))$

Alors,

$$S(x, y) = p[v(x, y)].$$

La fonction p étant décroissante, avec une dérivée seconde positive, il en est de même de q , donc : $q'(v) \leq 0$ et $q''(v) > 0$

a) v est une fonction croissante de x . En effet, quand x croît, $S_1(x)$ décroît, donc $q(S_1(x))$ croît, ainsi que v .

b) Si f est la densité associée à S , alors la propriété DRV est équivalente à D_2f/f est une fonction croissante de x [16] : cette dernière condition est équivalente

à (12). En effet :

$$\begin{aligned} D_1 S &= p'(v) \cdot q'(S_1) \cdot D_1 S_1 \\ f &= D_1 D_2 S = p''(v) \cdot q'(S_1) \cdot q'(S_2) \cdot D_1 S_1 \cdot D_2 S_2 \end{aligned} \quad (13)$$

$$D_2 f = q'(S_1) \cdot D_1 S_1 \cdot [p'''(v) \cdot q'^2(S_2) \cdot (D_2(S_2))^2 + p''(v) \cdot q''(S_2) (D_2 S_2)^2 + p''(v) \cdot q'((S_2)) \cdot D_2^2(S_2)]$$

$$D_2 f / f = \frac{p'''(v) \cdot q'(S_2) \cdot D_2(S_2)}{p''(v)} + \frac{q''(S_2) \cdot D_2(S_2)}{q'(S_2)} + \frac{D_2^2 S_2}{D_2 S_2} \quad (14)$$

Le deuxième et le troisième terme du membre de droite de (14) ne dépendent pas de x . Il suffit donc d'étudier le premier terme. Or, $q'(S_2)$ est négatif ainsi que $D_2 S_2$, leur produit est donc positif. La croissance de $D_2 f / f$ selon x équivaut donc à la croissance de $\frac{p'''(v)}{p''(v)}$ selon x , or v est une fonction croissante de x , d'où la condition annoncée.

B. Annexe : la famille de Clayton est DRV

La famille de Clayton vérifie la condition. En effet, pour cette famille :

$$p(v) = (v + 1)^{\frac{1}{1-\alpha}} \quad \alpha > 1$$

$$p'(v) = \frac{1}{1-\alpha} \cdot (v + 1)^{\frac{1}{1-\alpha}-1}$$

$$p''(v) = \frac{1}{1-\alpha} \cdot \frac{\alpha}{1-\alpha} \cdot (v + 1)^{\frac{1}{1-\alpha}-2}$$

$$p'''(v) = \frac{1}{1-\alpha} \cdot \frac{\alpha}{1-\alpha} \cdot \frac{2 \cdot \alpha - 1}{1-\alpha} (v + 1)^{\frac{1}{1-\alpha}-3}$$

D'où :

$$\frac{p'''(v)}{p''(v)} = \frac{(2\alpha - 1)}{(1 - \alpha)} \frac{1}{v + 1}$$

Quand v croît $1/(v + 1)$ décroît, $(1 - \alpha)$ est négatif car α est plus grand que un, d'où la croissance de $p'''(v)/p''(v)$ selon v , et donc la propriété DRV.

C. Annexe : la famille de Hougaard vérifie la condition *DRV*

La famille de Hougaard vérifie aussi la condition. En effet ici :

$$p(v) = \exp(-v^\alpha) \quad \alpha < 1$$

$$p'(v) = -\alpha \cdot p(v) \cdot v^{\alpha-1}$$

$$p''(v) = -\alpha \cdot v^{\alpha-2} \cdot p(v) \cdot [\alpha - 1 - \alpha \cdot v^\alpha]$$

$$p'''(v) = -\alpha \cdot v^{\alpha-3} \cdot p(v) \cdot [(\alpha - 2)(\alpha - 1 - \alpha \cdot v^\alpha) - \alpha \cdot v^\alpha (\alpha - 1 - \alpha \cdot v^\alpha) - \alpha^2 v^\alpha]$$

$$p'''(v) = -\alpha \cdot v^{\alpha-3} \cdot p(v) \cdot [(\alpha - 1 - \alpha \cdot v^\alpha)(\alpha - 2 - \alpha \cdot v^\alpha + \alpha) - \alpha(\alpha - 1)]$$

D'où :

$$\frac{p'''(v)}{p''(v)} = \frac{1}{v} \cdot (2\alpha - 2 - \alpha \cdot v^\alpha - \frac{\alpha(\alpha - 1)}{\alpha - 1 - \alpha \cdot v^\alpha})$$

$$\frac{p'''(v)}{p''(v)} = \frac{2(\alpha - 1)}{v} - \alpha \cdot v^{\alpha-1} - \frac{\alpha(\alpha - 1)}{(\alpha - 1)v - \alpha \cdot v^{\alpha+1}}$$

Puisque $(\alpha - 1)$ est négatif, les trois termes du second membre croissent avec v . Donc, la famille de Hougaard est *DRV*.