UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES

PAR
FRANÇOIS ÉTHIER

À PROPOS DE DIVERS TESTS STATISTIQUES POUR L'ÉGALITE DE LOIS

DÉCEMBRE 2011

REMERCIEMENTS

En préambule à ce mémoire, je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire. Je tiens d'abord à remercier mon directeur de recherche, M. Jean-François Quessy, professeur au département de mathématiques et d'informatique de l'Université du Québec à Trois-Rivières. En plus de m'avoir proposé ce sujet passionnant, il s'est toujours montré à l'écoute et très disponible afin de faire progresser la réalisation de cet ouvrage.

Je désire, bien sûr, remercier tous les membres de ma famille pour m'avoir encouragé et supporté tout au long de mes études. Je tiens également à souligner le support de tous mes amis, dont leur présence m'a permis de me divertir, même dans les moments plus difficiles. Je remercie plus particulièrement Tommy Mailhot pour son aide concernant l'installation de logiciels et ses explications qui m'ont permis de bien implémenter plusieurs programmes informatiques utilisés dans ce mémoire.

Je remercie aussi les professeurs Mhamed Mesfioui et Alain Chalifour pour avoir accepté d'évaluer mon mémoire. Leurs commentaires ont permis d'améliorer la qualité de ce travail.

Mes études à la maîtrise ont été financées en partie par des octrois individuels accordés à Jean-François Quessy par le *Conseil de Recherche en Sciences Naturelles et en Génie du Canada* et par le *Fonds québécois de la recherche sur la nature et les technologies*. Enfin, je remercie l'*Institut des Sciences Mathématiques* du Québec pour les bourses d'étude qu'il m'a accordées.

Table des matières

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1

INTRODUCTION

L'objectif d'un test statistique pour l'égalité de lois est de déterminer si les distributions d'où proviennent deux ou plusieurs échantillons sont identiques. À ce titre, plusieurs méthodologies existent et les domaines d'application sont nombreux. On peut citer la biostatistique, où l'on pourrait vouloir tester si les courbes de survie de personnes atteintes par un certain cancer sont les mêmes pour deux sortes de traitement. Également, en hydrologie, on s'intéresse parfois à vérifier si les distributions des débits moyens ou maximaux annuels sont identiques sur deux ou plusieurs cours d'eau.

Une classe populaire de tests est constituée des méthodes statistiques qui consistent à comparer les moyennes et/ou les variances empiriques; on conclut à des distributions hétérogènes si ces paramètres estimés diffèrent de manière trop marquée. Ces tests ont toutefois une limitation évidente, à savoir que des distributions différentes peuvent néanmoins avoir des moyennes et/ou des variances identiques. De plus, le comportement de ces statistiques de test dépend de la loi des observations, ce qui ajoute une restriction supplémentaire. Une catégorie de procédures qui ne nécessitent pas l'émission d'hypothèses

sur la forme de la loi comprend les tests qu'on qualifie de *non-paramétriques*. On peut répartir ces tests en deux sous-catégories, à savoir (i) les tests de rangs et (ii) les tests basés sur les fonctions de répartition empiriques.

La majorité des tests décrits brièvement dans le paragraphe précédent sont applicables dans le cas où les échantillons sont indépendants. Dans le cas d'échantillons *pairés*, la validité de ces procédures ne tient plus; cela est dû au fait que la structure de dépendance n'est pas prise en compte. La voie est donc ouverte pour le développement de méthodes statistiques pour l'égalité de lois en présence d'échantillons pairés.

La première contribution de ce mémoire est le développement de nouveaux tests pour la comparaison de deux ou plusieurs lois lorsque les échantillons sont constitués de données dépendantes. Une première catégorie de tests sera basée sur une distance de Cramér–von Mises entre les fonctions de répartition empiriques; une deuxième classe fera intervenir les fonctions caractéristiques empiriques. Le comportement asymptotique de ces statistiques de test sera obtenu à l'aide de la théorie des processus empiriques. Ensuite, puisqu'aucune hypothèse ne sera émise sur la forme des lois sous-jacentes, cela pose un problème complexe pour le calcul des valeurs critiques. En effet, celles-ci dépendent des lois marginales inconnue, de même que de la forme inconnue de la dépendance. Pour contrer ce problème, une adaptation judicieuse de la méthode du *multiplicateur* sera utilisée. Sa validité asymptotique pour le calcul de valeurs critiques sera démontrée rigoureusement.

La deuxième contribution de ce mémoire porte sur la construction de tests non-paramétriques pour l'égalité de lois standardisées selon leurs moyennes et

leurs variances. Ces méthodologies statistiques permettent ainsi de conclure à l'égalité de lois lorsque celles-ci ont la même forme, mais que leurs deux premiers moments diffèrent. Les développements asymptotiques nécessaires pour obtenir le comportement des statistiques de test utilisent des adaptations non-triviales de résultats sur les processus empiriques et sur la méthode delta fonctionnelle. Deux approches basées sur le multiplicateur pour le calcul des valeurs critiques sont proposées et justifiées formellement. À la fois les situations d'échantillons indépendants et pairés sont traitées. De nombreuses simulations permettent de conclure à l'efficacité des méthodes proposées. Enfin, des applications sur de vrais jeux de données sont présentées.

Au Chapitre 2, une revue de littérature sur les tests d'égalité de lois les plus utilisés est présentée. Au Chapitre 3, plusieurs notions fondamentales relatives à l'étude de la dépendance par les copules et aux processus empiriques sont introduites; ces outils seront d'une grande utilité dans les développements subséquents. Le Chapitre 4 concerne le développement et l'étude des propriétés asymptotiques et à tailles finies de nouveaux tests pour l'égalité de lois pour données pairées. Au Chapitre 5, une extension de cette méthodologie qui permet aux lois de différer selon leur moyenne et leur variance, est offerte; ce sujet est présenté sous la forme d'un article qui a été soumis à une revue savante. Le mémoire se termine par une brève conclusion.

À noter que le contenu du Chapitre 4 a fait l'objet d'un article paru dans la revue *Computational Statistics and Data Analysis* (2012, Vol. 56, pp. 2097–2111). De même, le Chapitre 5 constitue la reproduction intégrale d'un article soumis à la revue *Test* qui concerne des tests d'égalité de lois standardisées.

CHAPITRE 2

REVUE DE LITTÉRATURE SUR LES TESTS D'ÉGALITÉ DE LOIS

Dans ce chapitre, plusieurs tests populaires pour tester l'égalité en loi de deux variables aléatoires X et Y seront décrits. Dans ce cas, les hypothèses nulle et alternative sont $\mathcal{H}_0 : X \stackrel{d}{=} Y$ et $\mathcal{H}_1 : X \stackrel{d}{\neq} Y$. Les sections 2.1 à 2.3 considèrent des statistiques de test construites à partir de deux échantillons indépendants. Ainsi, on suppose que $X_1, \ldots, X_n$ sont des copies indépendantes de X, que $Y_1, \ldots, Y_m$ sont des copies indépendantes de Y, et que ces deux échantillons sont indépendants. À la section 2.4, la situation d'échantillons appariés est traitée; dans ce cas, on suppose que l'on observe des copies indépendantes $(X_1, Y_1), \ldots, (X_n, Y_n)$ de (X, Y).

2.1 Tests basés sur la différence de moyennes

2.1.1 Test de Student pour variances supposées égales

L'idée du test de Student consiste à comparer les moyennes des deux échantillons. Ainsi, les hypothèses qui sont testées dans ce contexte sont

$$\mathcal{H}_0 : \mu_X = \mu_Y \quad \text{et} \quad \mathcal{H}_1 : \mu_X \neq \mu_Y,$$

où $\mu_X = \mathrm{E}(X)$ et $\mu_Y = \mathrm{E}(Y)$. Ce test est qualifié de *paramétrique* car on doit émettre les hypothèses suivantes :

(i) Les variances sont égales, c'est-à-dire que $\sigma_X^2 = \sigma_Y^2$, où $\sigma_X^2 = \mathrm{var}(X)$ et $\sigma_Y^2 = \mathrm{var}(Y)$;

(ii) La loi de X et de Y est normale.

Soient $\bar{X}_n$ et S_n^2, la moyenne et la variance empiriques de $X_1, \ldots, X_n$. De même, on définit la moyenne et la variance empiriques de $Y_1, \ldots, Y_m$ par $\bar{Y}_m$ et S_m^2. La statistique du test est donnée par

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{S_{n,m}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}},$$

où

$$S_{n,m}^2 = \lambda_{n,m}\, S_n^2 + (1 - \lambda_{n,m})\, S_m^2, \quad \text{avec } \lambda_{n,m} = \frac{n-1}{n+m-2},$$

est la variance combinée. Sous l'hypothèse que les variances sont égales, $S_{n,m}^2$ est sans biais pour la variance commune.

Dans un article célèbre paru dans la revue *Biometrika* en 1908, William Sealy Gosset, qui signa alors sous le pseudonyme *Student*, obtena la densité du rapport entre une variable aléatoire normale et la racine carrée d'une variable aléatoire khi-carré; cette densité porte le nom de loi de Student.

On déduit directement du résultat de Gosset que sous l'hypothèse nulle d'égalité des deux moyennes, la statistique $T_{n,m}$ suit une loi de Student à $n+m-2$ degrés de liberté. Par conséquent, on rejettera $\mathcal{H}_0$, au niveau de confiance $1-\alpha$, lorsque

$$|T_{n,m}| > t_{n+m-2,\alpha/2},$$

où le nombre $t_{n+m-2,\alpha/2}$ est tel que

$$\mathrm{P}\left(T > t_{n+m-2,\alpha/2}\right) = \frac{\alpha}{2},$$

avec T qui suit une loi de Student à $n+m-2$ degrés de liberté.

2.1.2 Test de Welch pour variances différentes

Pour des variances inégales, une version de la statistique $T_{n,m}$ est donnée par

$$T'_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\xi_n + \xi_m}},$$

où $\xi_n = S_n^2/n$ et $\xi_m = S_m^2/m$. Dans ce cas, $T'_{n,m}$ suit une loi de Student dont le nombre de degrés de liberté est estimé par

$$\nu = \frac{(\xi_n + \xi_m)^2}{\dfrac{\xi_n^2}{n-1} + \dfrac{\xi_m^2}{m-1}}.$$

Cette dernière expression s'appelle l'équation de Welch–Satterthwaite. On rejette donc $\mathcal{H}_0$, avec un niveau de confiance approximatif $1 - \alpha$, si

$$|T'_{n,m}| > t_{\nu,\alpha/2}.$$

Voir Welch (1938), Yuen (1974) et Rasch et al. (2007) pour plus de détails.

2.2 Tests non-paramétriques

Les méthodologies non-paramétriques font peu ou pas de suppositions sur la loi des observations, contrairement aux tests statistiques paramétriques. Typiquement, les méthodes non-paramétriques sont basées sur les rangs des observations ou sur les fonctions de répartition empiriques. Les tests d'égalité de lois qui sont décrits dans la suite sont libérés de la contrainte que les données soient normales. Ils permettent de tester les hypothèses

$$\mathcal{H}_0 : F = G \quad \text{et} \quad \mathcal{H}_1 : F \neq G$$

pour deux populations de lois inconnues F et G, respectivement, où $F(x) = \mathrm{P}(X \leq x)$ et $G(y) = \mathrm{P}(Y \leq y)$. Dans la suite, on suppose que l'on a observé des échantillons $X_1, \ldots, X_n$ i.i.d. de loi F $Y_1, \ldots, Y_m$ i.i.d. de loi G.

2.2.1 Test de Wilcoxon

Ce test a été proposé initialement par Wilcoxon (1945). Pour le décrire, on crée d'abord le vecteur des observations combinées $\mathbf{Z} = (Z_1, \ldots, Z_{n+m})$, où

$$Z_i = \begin{cases} X_i, & i \in \{1, \ldots, n\}; \\ Y_{i-n}, & i \in \{n+1, \ldots, n+m\}. \end{cases}$$

Ensuite, on note R_{Z_i}, le rang de Z_i dans $\mathbf{Z}$. La statistique de Wilcoxon est la somme des rangs de $X_1, \ldots, X_n$ dans $\mathbf{Z}$, c'est-à-dire

$$W_{n,m} = \sum_{i=1}^{n} R_{Z_i}.$$

Pour n et m suffisamment grands, on a l'approximation

$$W_{n,m} \approx \mathcal{N}\left(\mu_X, \sigma_X\right),$$

où

$$\mu_X = \frac{n(n+m+1)}{2} \quad \text{et} \quad \sigma_X = \frac{nm(n+m+1)}{12}.$$

Dans ce cas, la p-valeur du test est donnée par

$$p = 2\left\{1 - \Phi\left(\frac{W_{n,m} - \mu_X}{\sigma_X}\right)\right\},$$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$. On rejette $\mathcal{H}_0$ lorsque $p < \alpha$, où α est le seuil nominal du test.

2.2.2 Test de Mann–Whitney

La statistique de test est

$$U_{n,m} = \frac{1}{n\,m} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}\left(X_i < Y_j\right).$$

Ainsi,

$$\mathrm{E}\left(U_{n,m}\right) = \mathrm{P}\left(X_i < Y_j\right) = \int_{\mathbb{R}} F(x)\,\mathrm{d}G(x).$$

Sous l'hypothèse nulle $\mathcal{H}_0 : F = G$, on a donc

$$\mathrm{E}\left(U_{n,m}\right) = \int_{\mathbb{R}} F(x)\,\mathrm{d}F(x) = \left.\frac{\{F(x)\}^2}{2}\right|_{x=-\infty}^{x=+\infty} = \frac{1}{2},$$

puisque $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$. On peut également démontrer que sous $\mathcal{H}_0$,

$$\mathrm{var}\left(U_{n,m}\right) = \frac{n+m+1}{12\,n\,m}.$$

En fait, on montre que pour n et m assez grands, on a l'approximation

$$U_{n,m} \approx \mathcal{N}\left(\frac{1}{2}, \frac{n+m+1}{12\,n\,m}\right).$$

Une expression simplifiée pour $U_{n,m}$ s'obtient en définissant les rangs moyens

$$\bar{R}_X = \frac{1}{n}\sum_{i=1}^{n} R_{Z_i} \quad \text{et} \quad \bar{R}_Y = \frac{1}{m}\sum_{i=n+1}^{m} R_{Z_i}. \tag{2.1}$$

On montre alors que

$$U_{n,m} = \frac{1}{2} + \frac{\bar{R}_X - \bar{R}_Y}{n+m}.$$

Une autre façon de comprendre le test est de partir du fait que

$$\int_{\mathbb{R}} F(x)\,\mathrm{d}G(x) = \frac{1}{2}$$

sous $\mathcal{H}_0$. Ainsi, on pourrait construire un test autour de la version empirique

$$\int_{\mathbb{R}} F_n(x)\,\mathrm{d}G_m(x),$$

où

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(X_i \leq x\right) \quad \text{et} \quad G_m(y) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(Y_i \leq y\right). \tag{2.2}$$

Puisque $\mathrm{d}\, G_m$ accorde un poids de $1/m$ à $Y_1, \ldots, Y_m$, on a

$$\begin{aligned}
\int_{\mathbb{R}} F_n(x)\, \mathrm{d}G_m(x) &= \frac{1}{m}\sum_{j=1}^{m} F_n\left(Y_j\right) \\
&= \frac{1}{m}\sum_{j=1}^{m}\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(X_i \leq Y_j\right)\right\} \\
&= \frac{1}{n\,m}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{I}\left(X_i < Y_j\right) \\
&= U_{n,m}.
\end{aligned}$$

2.2.3 Test des rangs de Welch

La statistique du test de Welch est

$$\tilde{W}_{n,m} = \frac{\bar{R}_X - \bar{R}_Y}{\sqrt{\dfrac{S^2_{R_X}}{m} + \dfrac{S^2_{R_Y}}{n}}},$$

où $\bar{R}_X$ et $\bar{R}_Y$ sont définies en (2.1). De plus, $S^2_{R_X}$ et $S^2_{R_Y}$ sont les variances des rangs. Tel que mentionné par Reiczigel et al. (2005), la loi de $\tilde{W}_{n,m}$ est approximativement Student à

$$\tilde{\nu} = \frac{\left(n\, S^2_{R_X} + m\, S^2_{R_Y}\right)^2}{\dfrac{\left(n\, S^2_{R_X}\right)^2}{m-1} + \dfrac{\left(m\, S^2_{R_Y}\right)^2}{n-1}}$$

2.3 Tests basés sur la fonction de répartition empirique

Soient F_n et G_m, les fonctions de répartition empiriques définies à (2.2). Ces estimateurs sont sans biais respectivement pour F et G car

$$\mathrm{E}\{F_n(x)\} = F(x) \quad \text{et} \quad \mathrm{E}\{G_m(y)\} = G(y).$$

De plus, comme on le verra plus formellement au chapitre suivant, le Théorème de Glivenko–Cantelli assure que

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \quad \text{et} \quad \sup_{y\in\mathbb{R}} |G_m(y) - G(y)|$$

convergent en probabilité vers 0. Comme F_n et G_m sont de bons estimateurs de F et G, on s'attend à ce qu'ils soient *proches* sous $\mathcal{H}_0 : F = G$. Une idée générale pour construire des tests est donc d'employer une mesure de distance fonctionnelle. Parmi les possibilités, on décrira des tests basés sur les distances de type Kolmogorov–Smirnov et Cramér–von Mises.

2.3.1 Test de Kolmogorov–Smirnov

Le test d'ajustement de Kolmogorov–Smirnov est l'un des plus utilisé pour tester l'adéquation à une loi de probabilité. Dans le contexte de l'égalité de lois de probabilité, la statistique de test est

$$K_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)| .$$

On peut montrer que $K_{n,m}$ converge en loi, lorsque $n, m \to \infty$, vers une variable aléatoire K dont la fonction de survie est donnée par

$$Q(x) = \mathrm{P}\left(K > x\right) = 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp\left(-2j^2x^2\right).$$

La p-valeur asymptotique d'un test basé sur $K_{n,m}$ est donc $p = Q(K_{n,m})$.

Puisque la représentation de Q sous forme de série converge très rapidement, une bonne approximation consiste à ne conserver que le premier terme; ainsi,

$$Q(x) \approx 2\exp\left(-2x^2\right).$$

Pour faciliter le calcul de $K_{n,m}$, soient $Z_1, \ldots, Z_{n+m}$, l'échantillon obtenu en combinant $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$, et en plaçant les observations en ordre croissant. On a alors

$$K_{n,m} = \max_{i\in\{1,\ldots,n+m\}} \left|F_n(Z_i) - G_m(Z_i)\right|.$$

Pour plus de détails, voir Bickel (1968), Schröer & Trenkler (1995), Præstgaard (1995), Büning (2002) et Greenwell & Finch (2004).

2.3.2 Test de Cramér–von Mises

La statistique de Cramér–von Mises est définie par

$$C_{n,m} = \frac{mn}{m+n}\int_{\mathbb{R}}\left\{F_n(x) - G_m(x)\right\}^2 \mathrm{d}H_{n,m}(x), \tag{2.3}$$

où

$$H_{n,m}(x) = \frac{1}{n+m}\left\{\sum_{i=1}^{n}\mathbb{I}\left(X_i \leq x\right) + \sum_{i=1}^{m}\mathbb{I}\left(Y_i \leq x\right)\right\}$$

est la fonction de répartition empirique combinée. À noter que

$$H_{n,m}(x) = \frac{n}{n+m} F_n(x) + \frac{m}{n+m} G_m(x). \tag{2.4}$$

Comme $H_{n,m}$ est la fonction de répartition de $Z_1, \ldots, Z_{n+m}$, on a

$$C_{n,m} = \frac{mn}{(n+m)^2} \sum_{i=1}^{n+m} \{F_n(Z_i) - G_m(Z_i)\}^2 .$$

Le lemme suivant donne une formule de calcul utile pour $C_{n,m}$.

Lemme 2.1. *Si R_{X_i} est le rang de X_i parmi $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ et R_{Y_i} est le rang de Y_i parmi $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, alors*

$$C_{n,m} = \frac{A_{n,m}}{n\,m(n+m)^2} - \frac{4\,n\,m-1}{6(m+n)},$$

où

$$A_{n,m} = n \sum_{i=1}^{n} (R_{X_i} - i)^2 + m \sum_{i=1}^{m} (R_{Y_i} - i)^2 .$$

L'hypothèse nulle $\mathcal{H}_0$ est rejetée lorsque $C_{n,m} > q_{1-\alpha}$, où les valeurs de $q_{1-\alpha}$ se retrouvent dans l'article de Büning (2002).

2.3.3 Test de Schmid & Trede (1995)

Schmid & Trede (1995) ont proposés la statistique de test

$$S_{n,m} = \sqrt{\frac{mn}{m+n}} \int_{\mathbb{R}} |F_n(x) - G_m(x)| \, \mathrm{d}H_{n,m}(x)$$

pour concurrencer les procédures basées sur les fonctionnelles de Kolmogorov–Smirnov et de Cramér–von Mises. On peut exprimer $S_{n,m}$ en terme des rangs

des observations. Pour ce faire, on définit R_{X_i} comme le rang de X_i, $i \in \{1, \ldots, n\}$, parmi $Z_1, \ldots, Z_{n+m}$ et R_{Y_j} comme le rang de Y_j, $j \in \{1, \ldots, m\}$, parmi $Z_1, \ldots, Z_{n+m}$. On peut montrer que

$$S_{n,m} = \frac{(mn)^{1/2}}{(m+n)^{3/2}} \left\{ \sum_{i=1}^{n} \left| \frac{R_{X_i}}{m} - i\left(\frac{1}{n} + \frac{1}{m}\right) \right| + \sum_{j=1}^{m} \left| \frac{R_{Y_j}}{n} - j\left(\frac{1}{n} + \frac{1}{m}\right) \right| \right\}.$$

2.4 Tests pour échantillons appariés

Soient des données pairées $(X_1, Y_1), \ldots, (X_n, Y_n)$. Dans ce cas, les échantillons ne sont pas indépendants dès lors qu'il y a de la dépendance entre X_i et Y_i. Par conséquent, les méthodes statistiques décrites depuis le début de ce chapitre ne sont pas applicables, à moins d'être absolument certains de l'indépendance entre les deux variables aléatoires.

2.4.1 Adaptation du test de Student

Ce test est entièrement basé sur les différences observées, c'est-à-dire sur $D_1, \ldots, D_n$, où $D_i = X_i - Y_i$, $i \in \{1, \ldots, n\}$. On suppose que $D_1, \ldots, D_n$ sont i.i.d. d'une loi Normale. Posons que $\bar{D}_n$ et S_D^2 sont la moyenne et la variance empiriques de $D_1, \ldots, D_n$. La statistique du test est donnée par

$$\tilde{T}_n = \frac{\sqrt{n}\,\bar{D}_n}{S_D}.$$

Comme T_n suit une loi de Student à $n-1$ degrés de liberté, on rejette l'hypothèse nulle d'égalité des moyennes lorsque $|\tilde{T}_n| > t_{n-1,\alpha/2}$.

2.4.2 Adaptation du test de Mann–Whitney

Ce test est une adaptation, au cas d'échantillons appariés, du test de Mann–Whitney–Wilcoxon. Considérons d'abord les statistiques d'ordre combinées des échantillons $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$, à savoir $Z_1, \ldots, Z_{2n}$. Ensuite, on note R_i le rang de X_i parmi $Z_1, \ldots, Z_{2n}$, et S_i le rang de Y_i parmi $Z_1, \ldots, Z_{2n}$. La statistique de test est donnée par

$$U_n = \frac{\sum_{i=1}^{n} R_i S_i - n\left(n + \frac{1}{2}\right)}{\sqrt{\frac{n^2(2n+1)(1-r_S)}{12}}},$$

où r_S est le coefficient de corrélation de Spearman de $(X_1, Y_1), \ldots, (X_n, Y_n)$. Lorsque $n \to \infty$, U_n converge en loi, sous $\mathcal{H}_0$, vers la distribution $\mathcal{N}(0,1)$; le critère de rejet de l'hypothèse nulle est donc $|U_n| > \Phi^{-1}(1-\alpha/2)$, où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

2.4.3 Adaptation du test des rangs pairés-signés de Wilcoxon

Soient $d_1, \ldots, d_n$, où pour $i \in \{1, \ldots, n\}$, on a $d_i = Y_i - X_i$. Considérons maintenant S_n, la somme des rangs parmi les observations $d_1, \ldots, d_n$ qui sont positives, c'est-à-dire

$$S_n = \sum_{i=1}^{n} R_{d_i} \mathbb{I}(d_i > 0).$$

La statistique de test est

$$Z_n = \frac{\left| S_n - \frac{n(n+1)}{4} \right| - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}.$$

Un résultat de type limite centrale assure que Z_n converge en loi vers $Z \sim \mathcal{N}(0,1)$ lorsque $n \to \infty$. Ainsi, la p-valeur du test est $p = 2\{1 - \Phi(|Z_n|)\}$. On rejette $\mathcal{H}_0$ si p est inférieure au seuil nominal α.

2.5 Test des permutations

Les tests de permutation sont des méthodes générales qui permettent d'imiter le comportement d'une statistique de test sous une hypothèse nulle. Pour la décrire dans le cas d'un test d'égalité de deux moyennes, notons $E_1, \ldots, E_{n^2}$ l'ensemble de tous les échantillons distincts obtenus en permutant X_i et Y_i dans le couple (X_i, Y_i). Ainsi, on a

$$\begin{aligned} E_1 &= (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), \\ &\vdots \quad\quad \vdots \\ E_{n^2} &= (Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n). \end{aligned}$$

Soit maintenant une statistique de test $S_{n,1}$ calculée à partir des observations originales, à savoir E_1. Le critère de rejet de l'hypothèse nulle sera basé sur la distribution des statistiques de permutation $S_{n,1}, S_{n,2}, \ldots, S_{n,n^2}$, où $S_{n,j}$ est la statistique de test basée sur l'échantillon E_j. Ainsi, $\mathcal{H}_0$ sera rejetée si S_{n1} excède le α-percentile empirique basée sur $S_{n,1}, S_{n,2}, \ldots, S_{n,n^2}$. Pour plus de détails, voir Odiase & Ogbonmwan (2007).

CHAPITRE 3

NOTIONS IMPORTANTES

3.1 Problématique

En réalisant la revue de litérature du Chapitre 2, on constate qu'il n'existe pas beaucoup de tests concernant les échantillons dépendants. Ce constat est d'autant plus vrai à propos des tests non-paramétriques basés sur les fonctions de répartition empiriques, malgré qu'ils soient généralement puissants et simples à utiliser. Ainsi, les recherches sur les test d'égalité de lois se sont surtout attardés aux cas d'échantillons indépendants. Lorsqu'il y a de la dépendance entre deux échantillons, ces méthodologies ne sont plus valides, d'où l'intérêt d'en développer de nouvelles dans ce contexte.

Puisque plusieurs phénomènes observés font intervenir des variables aléatoires dépendantes, il serait utile de pouvoir déterminer si deux ou plusieurs distributions sont identiques ou non dans un tel contexte. Typiquement, le problème que l'on rencontre d'emblée concerne le calcul de p-valeurs. En effet, la structure de dépendance entre deux ou plusieurs variables aléatoires

est habituellement inconnue. Malheureusement, le comportement des statistiques de test est intimement lié à cette dépendance. Une solution serait d'utiliser la méthode des permutations, mais cette procédure est lourde du point de vue calculatoire; en plus, elle est valide uniquement si on peut supposer que les variables sont échangeables, c'est-à-dire que $(X, Y) \stackrel{d}{=} (Y, X)$.

Dans ce mémoire, le calcul des p-valeurs pour les différents tests qui seront développés sera basé sur des adaptations judicieuses de la méthode du multiplicateur. Cette technique est décrite, dans le cas *classique*, dans les ouvrages de van der Vaart & Wellner (1996) et (Kosorok, 2008).

Ce chapitre est consacré à une revue des résultats qui seront essentiels pour les développements et la validité des procédures de tests qui seront proposées. D'abord, une introduction aux copules est présentée. Ensuite, quelques résultats fondamentaux sur les processus empiriques, en particulier sur la méthode du multiplicateur et la méthode delta fonctionnelle sont décrits.

3.2 Les copules

L'étude des copules origine d'un théorème important de Sklar (1959). Ce résultat permet d'isoler la dépendance des comportements marginaux d'un couple de variables aléatoires (X, Y); il est donc d'une extrême importance pour la modélisation et la vérification d'hypothèses bivariées. Le résultat est maintenant formellement énoncé dans le cas à deux variables; l'extension multidimensionnelle est immédiate et sera présentée à la fin de cette section.

Théorème 3.1 (Sklar (1959)). *Soit H, une fonction de répartition bivariée de marges continues F et G. Alors il existe une unique fonction $C : [0,1]^2 \to [0,1]$ telle que pour tout $(x,y) \in \mathbb{R}^2$,*

$$H(x,y) = C\{F(x), G(y)\}. \tag{3.1}$$

La fonction C qui apparaît dans le résultat de Sklar s'appelle la *copule* de H. Si $(X,Y) \sim H$, alors C correspond à la loi conjointe de $(F(X), G(Y))$. Par conséquent, C est nécessairement une fonction de répartition définie sur le carré unitaire $[0,1]^2$ dont les marges sont uniformes. Inversement, une fonction $C : [0,1]^2 \to [0,1]$ est une copule si

(i) $C(u,0) = C(0,u)$ et $C(u,1) = C(1,u) = u$ pour tout $u \in [0,1]$;

(ii) $C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geq 0$ pour tout $u_1 \leq u_2$ et $v_1 \leq v_2$.

Le Théorème de Sklar permet de constater que la loi de (X,Y) est composée des comportements marginaux de X et de Y, représentés respectivement par F et G, et de la dépendance entre X et Y, représentée par la copule C. Ceci est d'une grande utilité pour la modélisation parce qu'on peut choisir des modèles univariés pour les marges F et G. De plus, on peut aussi choisir, indépendamment de l'étape précédente, un modèle adéquat pour la copule.

Une propriété fondamentale d'une copule est qu'elle invariante sous des transformations monotones croissantes. Autrement dit, la copule de (X,Y) est la même que celle de $(\tilde{X},\tilde{Y})$, où $\tilde{X} = A(X)$, $\tilde{Y} = B(Y)$, et A et B sont des fonctions monotones croissantes. En effet, soient H et $\tilde{H}$, les fonctions de

répartition conjointes de (X, Y) et de $(\tilde{X}, \tilde{Y})$, respectivement, et supposons que C est la copule de H. On a

$$\begin{aligned}
\tilde{H}(x,y) &= \mathrm{P}\left\{A(X) \leq x, B(Y) \leq y\right\} \\
&= \mathrm{P}\left\{X \leq A^{-1}(x), Y \leq B^{-1}(y)\right\} \\
&= H\left\{A^{-1}(x), B^{-1}(y)\right\} \\
&= C\left\{F \circ A^{-1}(x), G \circ B^{-1}(y)\right\} \\
&= C\left\{\tilde{F}(x), \tilde{G}(y)\right\},
\end{aligned}$$

où

$$\tilde{F}(x) = F \circ A^{-1}(x) \quad \text{et} \quad \tilde{G}(y) = G \circ B^{-1}(y)$$

sont les marges de $\tilde{X}$ et $\tilde{Y}$, respectivement. Par conséquent, du Théorème de Sklar, C est la copule associée à $\tilde{H}$.

Il est possible d'extraire la copule C associée à un couple (X, Y) de loi H et de marges F et G. Pour ce faire, il s'agit simplement de poser $u = F(x)$ et $v = G(y)$ dans l'équation (3.1), ce qui amène

$$C(u,v) = H\left\{F^{-1}(u), G^{-1}(v)\right\}. \tag{3.2}$$

Par exemple, les variables aléatoires X et Y sont indépendantes si et seulement si leur fonction de répartition conjointe est telle que

$$H(x,y) = \mathrm{P}(X \leq x, Y \leq y) = F(x)G(y),$$

où $F(x) = \mathrm{P}(X \leq x)$ et $G(y) = \mathrm{P}(Y \leq y)$. De l'équation (3.2), on déduit que la copule associée à l'indépendance est

$$\Pi(u,v) = F\left\{F^{-1}(u)\right\} G\left\{G^{-1}(v)\right\} = uv.$$

On peut également extraire la copule Normale du modèle normal bivarié classique. Ainsi, puisque la fonction de répartition standard (moyennes nulles, variances unitaires) de cette loi s'écrit

$$H_\rho(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} h_\rho(s,t)\,\mathrm{d}t\,\mathrm{d}s,$$

où h_ρ est la densité normale bivariée standard de corrrélation ρ, l'expression implicite de la copule Normale est

$$C_\rho(u,v) = \int_{-\infty}^{\phi^{-1}(u)}\int_{-\infty}^{\phi^{-1}(v)} h_\rho(s,t)\,\mathrm{d}t\,\mathrm{d}s.$$

On rappelle que

$$h_\rho(s,t) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left(x^2+y^2-2\rho xy\right)\right\}.$$

Un autre modèle de copules s'obtient de la famille des distributions de Farlie–Gumbel–Morgenstern, qui sont de la forme

$$H_\theta(x,y) = F(x)G(y) + \theta F(x)G(y)\left\{1-F(x)\right\}\left\{1-G(y)\right\}, \quad \theta\in[-1,1].$$

Ici, F et G correspondent aux marges de H_θ. On déduit facilement de (3.2) que la copule associée à H_θ est

$$C_\theta(u,v) = uv + \theta uv(1-u)(1-v), \quad \theta\in[-1,1].$$

On peut montrer que pour toute fonction de répartition conjointe H de marges F et G, on a pour n'importe quel $(x,y)\in\mathbb{R}^2$ que

$$\max\left\{F(x)+G(y)-1,\,0\right\} \le H(x,y) \le \min\left\{F(x),G(y)\right\}.$$

Par une application de l'équation (3.2), on déduit que toute copule C est comprise à l'intérieur des bornes dites de Fréchet–Hoeffding, à savoir que pour n'importe quel $(u, v) \in [0, 1]^2$,

$$W(u, v) \leq C(u, v) \leq M(u, v),$$

où

$$W(u, v) = \max(u + v - 1, 0) \quad \text{et} \quad M(u, v) = \min(u, v).$$

À noter que W et M sont elles-mêmes des copules; W correspond à la dépendance négative maximale, alors que M est associée à la dépendance positive parfaite.

Une version multivariée du Théorème de Sklar est maintenant offerte.

Théorème 3.2 (Sklar (1959)). *Soit une fonction de répartition k-dimensionnelle H de marges continues $F_1, \ldots, F_k$. Alors il existe une unique fonction $C : [0, 1]^k \to [0, 1]$, appelée copule, telle que*

$$H(x_1, \ldots, x_k) = C\left\{F_1(x_1), \ldots, F_d(x_k)\right\}.$$

Une fonction $C : [0, 1]^k \to [0, 1]$ est une copule si

(i) $C(u_1, \ldots, u_k) = 0$ lorsqu'au moins une des composantes du vecteur $(u_1, \ldots, u_k)$ est nulle;

(ii) La mesure induite par C de tout rectangle $A \subseteq [0, 1]^k$ est non-négative.

Tout comme dans le cas bivarié, la copule d'une loi H s'obtient de la relation

$$C(u_1, \ldots, u_k) = H\left\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_k)\right\}. \tag{3.3}$$

Par exemple, soit h_Σ, la densité de la loi Normale standard k-dimensionnelle de matrice de corrélation $\Sigma \in \mathbb{R}^{k\times k}$. Dans ce cas, la fonction de répartition associée s'exprime implicitement par

$$H_\Sigma(x_1,\ldots,x_k) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_k} h_\Sigma(s_1,\ldots,s_k)\,\mathrm{d}s_1\cdots\mathrm{d}s_k.$$

Finalement, une application de l'équation (3.3) permet de déduire que la copule Normale à k dimensions est

$$\begin{aligned} C_\Sigma(u_1,\ldots,u_k) &= H_\Sigma\left\{\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_k)\right\} && (3.4)\\ &= \int_{-\infty}^{\Phi^{-1}(u_1)}\cdots\int_{-\infty}^{\Phi^{-1}(u_k)} h_\Sigma(s_1,\ldots,s_k)\,\mathrm{d}s_1\cdots\mathrm{d}s_k. && (3.5)\end{aligned}$$

3.3 Processus empiriques

3.3.1 Échantillons univariés

Il est parfois utile de représenter un test ou une famille de tests statistiques en terme de fonctions de répartition empiriques. Ceci amène à étudier le comportement asymptotique des processus empiriques, ce qui permet l'obtention de la limite de statistiques dans un cadre très général.

Soit donc un échantillon $X_1,\ldots,X_n$ dans $\mathbb{R}$ tiré d'une loi dont la fonction de répartition est F. La fonction de répartition empirique associée à ces observations est

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(X_i \le x\right).$$

On montre facilement que

$$\mathrm{E}\{F_n(x)\} = F(x) \quad \text{et} \quad \mathrm{var}\{F_n(x)\} = \frac{F(x)\{1-F(x)\}}{n}.$$

Pour un $x \in \mathbb{R}$ fixé, $F_n(x)$ est simplement la proportion des observations qui sont inférieures à x. Le Théorème de la limite centrale de De Moivre (voir Casella & Berger, 1990) assure alors que

$$\mathbb{F}_n(x) = \sqrt{n}\{F_n(x) - F(x)\}$$

converge vers la loi Normale de moyenne zéro et de variance $F(x)\{1-F(x)\}$.

Le résultat suivant, appelé Théorème de Glivenko–Cantelli (voir Billingsley, 1999), est un résultat de convergence de type *uniforme*, c'est-à-dire qu'il s'applique à tous les $x \in \mathbb{R}$ simultanément.

Théorème 3.3 (Glivenko–Cantelli). *L'estimateur F_n est uniformément convergent pour F au sens où*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

converge en probabilité vers 0.

Dans la même lignée que le résultat de Glivenko–Cantelli, le Théorème de Donsker caractérise le comportement asymptotique en loi de $F_n(x)$, ou plus précisément de sa version standardisée $\mathbb{F}_n(x)$, pour tous les $x \in \mathbb{R}$.

Théorème 3.4 (Donsker). *Soit $D(\mathbb{R})$, l'espace des fonctions définies sur $\mathbb{R}$ qui sont continues à droites et qui possèdent une limite à gauche. Alors $\mathbb{F}_n$ converge en loi dans l'espace $D(\mathbb{R})$ vers un processus de représentation $\mathbb{F}(x) = \mathbb{B}\{F(x)\}$, où $\mathbb{B}$ est un pont Brownien.*

Notons qu'un pont Brownien est une fonction aléatoire définie sur $[0,1]$ et telle que $\mathbb{B}(s)$ est normale avec

$$\mathrm{E}\{\mathbb{B}(s)\} = 0 \quad \text{et} \quad \mathrm{var}\{\mathbb{B}(s)\} = s(1-s).$$

De plus, la fonction de covariance de $\mathbb{B}$ est donnée par

$$E\{\mathbb{B}(s)\,\mathbb{B}(t)\} = \min(s,t) - st.$$

Par conséquent, puisque $\mathbb{F}(x) = \mathbb{B}\{F(x)\}$, on a

$$\mathrm{E}\{\mathbb{F}(x)\} = 0 \quad \text{et} \quad \mathrm{var}\{\mathbb{F}(x)\} = F(x)\{1-F(x)\},$$

alors que la fonction de covariance est

$$E\{\mathbb{F}(x)\,\mathbb{F}(x')\} = \min\{F(x), F(x')\} - F(x)F(x').$$

3.3.2 Échantillons bivariés

Considérons maintenant un échantillon bivarié $(X_1, Y_1), \ldots, (X_n, Y_n)$, où pour $i \in \{1, \ldots, n\}$, $(X_i, Y_i) \sim H$. Alors une estimation de H s'obtient par la fonction de répartition empirique bivariée

$$H_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x, Y_i \leq y).$$

On peut montrer que pour $(x,y) \in \mathbb{R}^2$ fixé,

$$\mathrm{E}\{H_n(x,y)\} = H(x,y) \quad \text{et} \quad \mathrm{var}\{H_n(x,y)\} = H(x,y)\{1-H(x,y)\}.$$

De plus, le processus $\mathbb{H}_n(x,y) = \sqrt{n}\{H_n(x,y) - H(x,y)\}$ converge en loi vers la loi normale de moyenne zéro et de variance $H(x,y)\{1-H(x,y)\}$.

Maintenant, une version bivariée du Théorème de Donsker est présentée.

Théorème 3.5. *Le processus empirique* $\mathbb{H}_n$ *converge en loi dans l'espace* $D(\mathbb{R}^2)$ *vers un processus gaussien centré* $\mathbb{H}$ *de fonction de covariance*

$$\mathrm{E}\left\{\mathbb{H}(x,y)\,\mathbb{H}(x',y')\right\} = H\left\{\max(x,x'),\max(y,y')\right\} - H(x,y)H(x',y').$$

3.4 Méthode du multiplicateur

La méthode du multiplicateur est une méthode de ré-échantillonnage très utile qui permet de calculer la p-valeur de tests statistiques dans des situations où la loi n'est pas bien spécifiée sous l'hypothèse nulle. Elle sera d'abord décrite pour des variables et vecteurs aléatoires. L'extension aux processus empiriques, là où elle déploie toute sa puissance, sera ensuite présentée.

3.4.1 Dans $\mathbb{R}$

Soit un échantillon $X_1, \ldots, X_n$, où $X_i \in \mathbb{R}$, $\mu = \mathrm{E}(X_i)$ et $\sigma^2 = \mathrm{var}(X_i)$. Alors, le Théorème de la limite centrale stipule que

$$Z_n = \sqrt{n}\left(\bar{X}_n - \mu\right)$$

converge en loi vers la Normale de moyenne zéro et de variance σ^2. Ce résultat classique est énoncé, entres autres, dans Casella & Berger (1990).

Supposons que l'on désire *imiter* le comportement de Z_n. Afin d'y parvenir, définisssons d'abord les *multiplicateurs*.

Définition 3.1. *Les multiplicateurs sont des vecteurs indépendants*

$$\left(\xi_1^{(1)}, \ldots, \xi_n^{(1)}\right), \ldots, \left(\xi_1^{(M)}, \ldots, \xi_n^{(M)}\right)$$

de sorte que pour chaque $h \in \{1, \ldots, M\}$, $\xi_1^{(h)}, \ldots, \xi_n^{(h)}$ *sont des variables aléatoires indépendantes telles que pour chaque* $i \in \{1, \ldots, n\}$,

$$\mathrm{P}\left(\xi_i^{(h)} > 0\right) = 1 \quad \textit{et} \quad \mathrm{E}\left(\xi_i^{(h)}\right) = \mathrm{var}\left(\xi_i^{(h)}\right) = 1.$$

Ensuite on définit, pour tous les $h \in \{1, \ldots, M\}$,

$$Z_n^{(h)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\xi_i^{(h)}}{\bar{\xi}^{(h)}} \left(X_i - \bar{X}_n\right),$$

où

$$\bar{\xi}^{(h)} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(h)}.$$

Puisque $\sum_{i=1}^{n} \xi_i^{(h)} / \bar{\xi}^{(h)} = n$, on a la formule équivalente

$$Z_n^{(h)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\xi_i^{(h)}}{\bar{\xi}^{(h)}} - 1 \right) X_i.$$

Le résultat suivant établit formellement que la méthode fonctionne. En fait, il montre que les variables aléatoires $Z_n^{(1)}, \ldots, Z_n^{(M)}$ se comportent de la même façon que Z_n lorsque $n \to \infty$.

Proposition 3.1. *Le vecteur*

$$\left(Z_n, Z_n^{(1)}, \ldots, Z_n^{(M)}\right)$$

converge en loi vers

$$\left(Z, Z^{(1)}, \ldots, Z^{(M)}\right),$$

où $Z^{(1)}, \ldots, Z^{(M)}$ *sont des copies indépendantes de* $Z \sim \mathcal{N}(0, \sigma^2)$.

3.4.2 Dans $\mathbb{R}^d$

L'extension multidimensionnelle est immédiate. Soient $\mathbf{X}_1, \ldots, \mathbf{X}_n$ où $\mathbf{X}_i \in \mathbb{R}^d$, $\mu = \mathrm{E}(\mathbf{X}_i)$ et $\Sigma = \mathrm{var}(\mathbf{X}_i)$. En considérant que

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \quad \text{et} \quad \mathbf{Z}_n = \sqrt{n}\left(\bar{\mathbf{X}}_n - \mu\right),$$

on pose, pour $h \in \{1, \ldots, M\}$,

$$\mathbf{Z}_n^{(h)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\xi_i^{(h)}}{\bar{\xi}^{(h)}} - 1\right)\mathbf{X}_i.$$

3.4.3 Processus empirique univarié

Les tests qui seront développés dans les chapitres subséquents seront basés sur les fonctions de répartition empiriques. Il convient donc de présenter une version *processus* de la méthode du multiplicateur. Spécifiquement, les versions multiplicateurs de $\mathbb{F}_n = \sqrt{n}(F_n - F)$ sont

$$\mathbb{F}_n^{(h)}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\xi_i^{(h)}}{\bar{\xi}^{(h)}} - 1\right)\mathbb{I}\left(X_i \leq x\right).$$

Le résultat suivant indique que cette idée fonctionne.

Proposition 3.2. *Le vecteur*

$$\left(\mathbb{F}_n, \mathbb{F}_n^{(1)}, \ldots, \mathbb{F}_n^{(M)}\right)$$

converge en loi vers

$$\left(\mathbb{F}, \mathbb{F}^{(1)}, \ldots, \mathbb{F}^{(M)}\right),$$

où $\mathbb{F}^{(1)}, \ldots, \mathbb{F}^{(M)}$ sont des copies indépendantes de $\mathbb{F}(x) = \mathbb{B}\{F(x)\}$ et $\mathbb{B}$ est un pont Brownien.

3.4.4 Processus empirique bivarié

Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$, un échantillon dans $\mathbb{R}^2$. Alors, les versions multiplicateurs de $\mathbb{H}_n$ sont, pour $h \in \{1, \ldots, M\}$,

$$\mathbb{H}_n^{(h)}(x, y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\xi_i^{(h)}}{\bar{\xi}^{(h)}} - 1 \right) \mathbb{I}\left(X_i \leq x, Y_i \leq y\right).$$

Le résultat suivant est une version dans $\mathbb{R}^2$ de la Proposition 3.2.

Proposition 3.3. *Le vecteur*

$$\left(\mathbb{H}_n, \mathbb{H}_n^{(1)}, \ldots, \mathbb{H}_n^{(M)}\right)$$

converge en loi vers

$$\left(\mathbb{H}, \mathbb{H}^{(1)}, \ldots, \mathbb{H}^{(M)}\right)$$

où $\mathbb{H}^{(1)}, \ldots, \mathbb{H}^{(M)}$ *sont des copies indépendantes de* $\mathbb{H}$.

3.5 La méthode Delta fonctionnelle

3.5.1 Théorie

Plusieurs statistiques sont des fonctionnelles de processus empiriques. Pour étudier leur comportement asymptotique, on verra que la notion de dérivée d'Hadamard est centrale. Brièvement, l'objectif de la méthode Delta fonctionnelle est d'obtenir le comportement asymptotique de

$$\mathbb{L}_n = \sqrt{n} \left\{ \mathcal{L}(P_n) - \mathcal{L}(P) \right\},$$

où $\mathcal{L}$ est une fonctionnelle définie sur un espace de fonctions. À l'aide d'une notion de dérivée applicable aux fonctionnelles, cela découlera du comportement asymptotique de $\mathbb{P}_n = \sqrt{n}\,(P_n - P)$. En effet, en remarquant que

$$\mathbb{L}_n = \sqrt{n}\,\{\mathcal{L}(P_n) - \mathcal{L}(P)\} = \frac{\mathcal{L}(P + n^{-1/2}\mathbb{P}_n) - \mathcal{L}(P)}{n^{-1/2}},$$

on voit qu'asymptotiquement, $\mathbb{L}_n$ s'apparente à une sorte de dérivée de $\mathcal{L}$. Ainsi, l'idée d'une dérivée fonctionnelle consiste à trouver une fonctionnelle linéaire continue $\mathcal{L}'_P$ de telle sorte que

$$\mathbb{L}_n = \frac{\mathcal{L}'_P(n^{-1/2}\mathbb{P}_n)}{n^{-1/2}} + o_{\mathrm{P}}(1) = \mathcal{L}'_P(\mathbb{P}_n) + o_{\mathrm{P}}(1).$$

Donc, si $\mathbb{P}_n$ converge en loi vers $\mathbb{P}$, on aura que $\mathbb{L}_n \rightsquigarrow \mathcal{L}'_P(\mathbb{P})$, par une application du Théorème des fonctionnelles continues.

La notion de dérivée fonctionnelle en un point P la plus naturelle consiste à la *perturber* par un élément infinitésimal $P + t\Delta$. Ainsi, la dérivée de $\mathcal{L}$ au point P dans la direction Δ, appelée la dérivée de Gâteaux, est

$$\mathcal{L}'(\Delta) = \lim_{t \to 0} \frac{\mathcal{L}(P + t\Delta) - \mathcal{L}(P)}{t}.$$

Cette définition suggère que pour $t \approx 0$,

$$\frac{\mathcal{L}(P + t\Delta) - \mathcal{L}(P)}{t} \approx \mathcal{L}'(\Delta).$$

Cependant, $t = n^{-1/2}$ et $\Delta = \mathbb{P}_n$ dépendent de n et cela cause problème, car la dérivée de Gâteaux ne couvre pas cette situation particulière. La dérivée de Hadamard, plus générale, permet de régler ce problème.

Définition 3.2. *Une fonctionnelle $\mathcal{L} : A \to B$ est dérivable au sens de Hadamard au point $P \in A$ s'il existe une fonctionnelle linéaire continue*

$\mathcal{L}'_P : A \to B$ telle que pour toutes suites $\Delta_n \to \Delta$ et $t_n \to 0$,

$$\lim_{n\to\infty} \left| \frac{\mathcal{L}\left(P + t_n\,\Delta_n\right) - \mathcal{L}(P)}{t_n} - \mathcal{L}'_P(\Delta) \right|_B = 0,$$

en autant qu'il existe un $n_0 \in \mathbb{R}$ tel que $P + t_n\,\Delta_n \in A$ pour tout $n \geq n_0$.

Le résultat principale de la méthode Delta fonctionnelle est maintenant énoncé.

Théorème 3.6. *Si $\mathcal{L}$ est dérivable au sens de Hadamard, alors*

$$\mathbb{L}_n = \sqrt{n}\left\{\mathcal{L}(P_n) - \mathcal{L}(P)\right\} \rightsquigarrow \mathcal{L}'_P(\mathbb{P}),$$

où $\mathcal{L}'_P$ est la dérivée de $\mathcal{L}$ et $\mathbb{P}$ est la limite de $\mathbb{P}_n = \sqrt{n}(P_n - P)$.

On obtiendra l'expression de la dérivée fonctionnelle $\mathcal{L}'_P(\Delta)$ en calculant $h'(0)$, où $h(t) = \mathcal{L}(P + t\Delta)$. Quelques exemples sont décrits dans la suite.

3.5.2 Loi asymptotique de la moyenne empirique

La moyenne d'une loi de fonction de répartion F peut s'écrire

$$\mu = \int_{\mathbb{R}} x \,\mathrm{d}F(x).$$

Par conséquent, la moyenne empirique peut s'exprimer par

$$\bar{X}_n = \int_{\mathbb{R}} x \,\mathrm{d}F_n(x).$$

Pour obtenir la loi asymptotique de $\bar{X}_n$ par la méthode Delta fonctionnelle, on note d'abord que $\mu = \mathcal{L}(F)$, où

$$\mathcal{L}(P) = \int_{\mathbb{R}} x \,\mathrm{d}P(x)$$

Ainsi, on a

$$h(t) = \int_{\mathbb{R}} x \, \mathrm{d}\left\{P(x) + t\Delta(x)\right\},$$

ce qui fait que

$$\mathcal{L}'_P(\Delta) = h'(0) = \int_{\mathbb{R}} x \, \mathrm{d}\Delta(x) = \int_{\mathbb{R}} \Delta(x)\mathrm{d}x.$$

Ainsi, on déduit du Théorème 3.6 que

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \rightsquigarrow -\int_{\mathbb{R}} \mathbb{F}(x)\mathrm{d}x,$$

où $\mathbb{F}(x) = \mathbb{B}\{F(x)\}$ est la limite de $\mathbb{F}_n = \sqrt{n}(F_n - F)$.

3.5.3 Loi asymptotique des percentiles empiriques

Le percentile d'ordre α d'une loi dont la fonction de répartition est F est $\theta = F^{-1}(\alpha)$. Un estimateur naturel de θ est alors $\theta_n = F_n^{-1}(\alpha)$, où

$$F_n^{-1}(u) = \inf\left\{x \in \mathbb{R} : F_n(x) \geq u\right\}$$

est l'inverse généralisé de F_n. Dans ce cas, la fonctionnelle est $\mathcal{L}(P) = P^{-1}$, ce qui fait que $h(t) = (P + t\Delta)^{-1}(\alpha)$. Ainsi, $P\left\{h(t)\right\} + t\Delta\left\{h(t)\right\} = \alpha$. En dérivant, on obtient

$$\mathcal{L}'_P(\Delta) = h'(0) = \frac{-\Delta\{P^{-1}(\alpha)\}}{P'\{P^{-1}(\alpha)\}}.$$

Par conséquent, en autant que la densité f de F existe, le Théorème 3.6 assure que

$$\sqrt{n}\left(\theta_n - \theta\right) \rightsquigarrow \frac{-\mathbb{F}\{F^{-1}(\alpha)\}}{f\{F^{-1}(\alpha)\}} = \frac{-\mathbb{B}(\alpha)}{f\{F^{-1}(\alpha)\}}.$$

Enfin, comme $\mathbb{B}(\alpha) \sim \mathcal{N}(0, \alpha(1-\alpha))$, on déduit que

$$\sqrt{n}\left(\theta_n - \theta\right) \rightsquigarrow \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f^2\{F^{-1}(\alpha)\}}\right).$$

3.5.4 Loi asymptotique du tau de Kendall

Le tau de Kendall est une mesure de dépendance bivariée qui peut être utilisée comme alternative au coefficient de corrélation classique et au coefficient de corrélation de rangs de Spearman (1904). Si (X, Y) est un couple aléatoire de loi conjointe H, alors la mesure d'association de Kendall peut s'écrire

$$\tau = 4 \int_{\mathbb{R}^2} H(x, y) \, \mathrm{d}H(x, y) - 1.$$

Une version empirique est donc donnée par

$$\tau_n = 4 \int_{\mathbb{R}^2} H_n(x, y) \, \mathrm{d}H_n(x, y) - 1.$$

Cette définition n'est pas celle que l'on retrouve habituellement dans les ouvrages de statistique non-paramétrique; la version usuelle, telle que décrite dans Lee (1990), est basée sur une U-statistique d'ordre deux. Néanmoins, ces deux versions sont asymptotiquement équivalentes.

Pour obtenir la loi de $\sqrt{n}(\tau_n - \tau)$, notons que $\tau = \mathcal{L}(H)$, où

$$\mathcal{L}(P) = 4 \int_{\mathbb{R}^2} P(x, y) \, \mathrm{d}P(x, y) - 1.$$

Ainsi,

$$h(t) = 4 \int_{\mathbb{R}^2} \{P(x, y) + t\Delta(x, y)\} \, \mathrm{d} \{P(x, y) + t\Delta(x, y)\} - 1,$$

ce qui fait que

$$\mathcal{L}'_P(\Delta) = h'(0) = 4 \left\{ \int_{\mathbb{R}^2} \Delta(x, y) \mathrm{d}P(x, y) + \int_{\mathbb{R}^2} P(x, y) \mathrm{d}\Delta(x, y) \right\}.$$

Par une application du Théorème 3.6, on obtient

$$\sqrt{n}(\tau_n - \tau) \rightsquigarrow 4 \left\{ \int_{\mathbb{R}^2} \mathbb{H}(x, y) \, \mathrm{d}H(x, y) + \int_{\mathbb{R}^2} H(x, y) \, \mathrm{d}\mathbb{H}(x, y) \right\}.$$

CHAPITRE 4

NOUVEAUX TESTS D'ÉGALITÉ DE K LOIS POUR ÉCHANTILLONS DÉPENDANTS

Ce chapitre concerne le développement de nouveaux tests non-paramétriques pour l'égalité de deux ou plusieurs lois dans le cas d'échantillons dépendants. Le cas à $k = 2$ échantillons et à $k > 2$ échantillons sont traités séparément. Les statistiques de test sont de type Cramér–von Mises et fonction caractéristique. Leur efficacité sera étudiée à l'aide de simulations. De vrais jeux de données seront finalement analysés.

4.1 Cas à deux échantillons

4.1.1 Un processus empirique pour les tests

Soient $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$, des données pairées provenant d'une loi bivariée inconnue H ayant respectivement F_1 et F_2 comme lois marginales

continues. Par le Théorème 3.1 (Sklar (1959)), on sait qu'il existe une unique copule $C : [0,1]^2 \to [0,1]$ telle que pour tout $(x_1, x_2) \in \mathbb{R}^2$,

$$H(x_1, x_2) = C\{F_1(x_1), F_2(x_2)\}.$$

En vue de confronter les hypothèses $\mathcal{H}_0 : F_1 = F_2$ et $\mathcal{H}_1 : F_1 \neq F_2$, considérons la version empirique H_n de H définie par

$$H_n(x_1, x_2) = \frac{1}{n}\sum_{p=1}^{n} \mathbb{I}(X_{p1} \leq x_1, X_{p2} \leq x_2).$$

Comme $F_{n1}(x) = H_n(x, \infty)$ et $F_{n2}(x) = H_n(\infty, x)$ sont des estimateurs de F_1 et F_2, respectivement, un test basé sur une certaine distance entre F_{n1} et F_{n2} semble raisonnable. Ainsi, une statistique de test pour $\mathcal{H}_0$ et $\mathcal{H}_1$ pourrait se définir comme une fonctionnelle du processus empirique

$$\mathbb{A}_n(x) = \sqrt{n}\{F_{n1}(x) - F_{n2}(x)\}.$$

Le comportement asymptotique de ce processus est l'objet du prochain résultat dont la preuve est présentée à l'Annexe A. Dans la suite, $\rightsquigarrow$ signifie *convergence en loi* et $D(\mathbb{R})$ est l'espace des fonctions continues à droite ayant une limite à gauche définies sur $\mathbb{R}$. Enfin, $a \wedge b = \min(a, b)$.

Proposition 4.1. *Sous $\mathcal{H}_0$, $\mathbb{A}_n(x) \rightsquigarrow \mathbb{A}(x)$ dans l'espace $D(\mathbb{R})$, où $\mathbb{A}$ est un processus Gaussien centré et continue avec*

$$\mathrm{E}\{\mathbb{A}(x)\mathbb{A}(x')\} = 2F(x \wedge x') - C\{F(x), F(x')\} - C\{F(x'), F(x)\}$$

et F est la fonction de répartition sous l'hypothèse nulle.

Lorsque la copule C de H est symétrique, i.e. $C(u, v) = C(v, u)$ pour tout couple $(u, v) \in [0,1]^2$, la fonction de covariance dans la Proposition 4.1 est

de la forme

$$\mathrm{E}\left\{\mathbb{A}(x)\mathbb{A}(x')\right\} = 2\left\{F(x \wedge x') - C\left\{F(x), F(x')\right\}\right\}.$$

En particulier, $\mathrm{E}\left\{\mathbb{A}(x)\mathbb{A}(x')\right\} = 2\{F(x \wedge x') - F(x)F(x')\}$ sous l'indépendance, c'est-à-dire lorsque $C(u_1, u_2) = u_1 u_2$.

4.1.2 Versions *multiplicateur*

Le comportement de $\mathbb{A}_n(x)$ sous $\mathcal{H}_0$ dépend des lois marginales et de la structure de dépendance. Afin d'obtenir des p-valeurs valides, on doit être en mesure d'imiter $\mathbb{A}_n(x)$, du moins asymptotiquement. La stratégie qui sera développée est basée sur la méthode du multiplicateur. Cette idée a été utilisée par Scaillet (2005), Rémillard & Scaillet (2009) et Quessy (2011) pour tester des hypothèses à base de copules. Cette méthode a aussi été utilisée dans les travaux de Burke (2000) et de Gombay & Horváth (2002).

Les versions *multiplicateurs* de $\mathbb{A}_n(x)$ sont $\widehat{\mathbb{A}}_n^{(1)}, \ldots, \widehat{\mathbb{A}}_n^{(M)}$, où

$$\widehat{\mathbb{A}}_n^{(h)}(x) = \frac{1}{\sqrt{n}} \sum_{p=1}^{n} \left(\frac{\xi_p^{(h)}}{\bar{\xi}^{(h)}} - 1 \right) \left\{ \mathbb{I}\left(X_{p1} \leq x\right) - \mathbb{I}\left(X_{p2} \leq x\right) \right\}.$$

La prochaine proposition montre que $\widehat{\mathbb{A}}_n^{(1)}, \ldots, \widehat{\mathbb{A}}_n^{(M)}$ sont asymptotiquement des copies indépendantes de $\mathbb{A}_n$. La preuve se retrouve à l'Annexe A.

Proposition 4.2. *Sous l'hypothèse nulle* $\mathcal{H}_0$,

$$\left(\mathbb{A}_n, \widehat{\mathbb{A}}_n^{(1)}, \ldots, \widehat{\mathbb{A}}_n^{(M)}\right) \rightsquigarrow \left(\mathbb{A}, \mathbb{A}^{(1)}, \ldots, \mathbb{A}^{(M)}\right)$$

dans $D(\mathbb{R})^{\otimes(M+1)}$, *où* $\mathbb{A}^{(1)}, \ldots, \mathbb{A}^{(M)}$ *sont des copies indépendantes de* $\mathbb{A}$.

Il est à noter que la méthode du multiplicateur présentée dans cette section est une alternative avantageuse à la technique des permutations. En effet, cette dernière suppose que toutes les permutations sont équiprobables, ce qui revient à imposer que la loi conjointe est symétrique au sens où $(X_1, X_2) \stackrel{d}{=} (X_2, X_1)$; cette supposition n'est pas vraie lorsque la copule n'est pas symétrique, c'est-à-dire qu'il existe $(u_1^\star, u_2^\star) \in [0,1]^2$ tel que $C(u_1^\star, u_2^\star) \neq C(u_2^\star, u_1^\star)$. De plus, cette méthode est lourde du point de vue computationnel, surtout pour des échantillons de grandes tailles.

4.1.3 Une statistique de Cramér–von Mises

Une statistique de test basée sur la fonctionnelle de Cramér–von Mises est

$$S_n = n \int_{\mathbb{R}} \{F_{n1}(x) - F_{n2}(x)\}^2 \, \mathrm{d}x = \int_{\mathbb{R}} \mathbb{A}_n^2(x) \, \mathrm{d}x.$$

Les versions multiplicateurs de S_n sont données par

$$\widehat{\mathbb{S}}_n^{(h)} = \int_{\mathbb{R}} \left\{ \widehat{\mathbb{A}}^{(h)}(x) \right\}^2 \mathrm{d}x, \quad h \in \{1, \ldots, M\}.$$

Comme la fonctionnelle de Cramér–von Mises est continue, une application du Théorème des fonctions continues, conjointement avec la conclusion de la Proposition 4.2, implique que

$$\left(\mathbb{S}_n, \widehat{\mathbb{S}}_n^{(1)}, \ldots, \widehat{\mathbb{S}}_n^{(M)} \right) \rightsquigarrow \left(\mathbb{S}, \mathbb{S}^{(1)}, \ldots, \mathbb{S}^{(M)} \right)$$

dans $(\mathbb{R}^+)^{\otimes(M+1)}$, où $\mathbb{S}^{(1)}, \ldots, \mathbb{S}^{(M)}$ sont des copies indépendantes de

$$\mathbb{S} = \int_{\mathbb{R}} \mathbb{A}^2(x) \, \mathrm{d}x.$$

Ainsi, une p-valeur asymptotiquement valide pour le test basé sur S_n est

$$p_S = \frac{1}{M} \sum_{h=1}^{M} \mathbb{I}\left(\widehat{\mathbb{S}}_n^{(h)} > S_n\right).$$

Le lemme suivant offre des formules pour S_n et ses versions multiplicateurs en termes de produits de matrices. La démonstration se retrouve à l'Annexe A.

Lemme 4.2. *Soit la matrice $L \in \mathbb{R}^{n \times n}$ dont les éléments sont*

$$L_{pq} = 2\max\left(X_{p1}, X_{q2}\right) - \max\left(X_{p1}, X_{q1}\right) - \max\left(X_{p2}, X_{q2}\right).$$

Alors

$$S_n = \frac{1}{n} \mathbf{1}\, L\, \mathbf{1}^\top \quad \textit{et} \quad \widehat{\mathbb{S}}_n^{(h)} = \gamma^{(h)}\, L \left(\gamma^{(h)}\right)^\top,$$

où $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^{1 \times n}$ et

$$\gamma^{(h)} = \left(\frac{\xi_1^{(h)}}{\bar{\xi}^{(h)}} - 1, \ldots, \frac{\xi_n^{(h)}}{\bar{\xi}^{(h)}} - 1\right).$$

4.1.4 Des statistiques *fonction caractéristique*

Puisque la fonction caractéristique d'une variable aléatoire caractérise son comportement stochastique, on pourrait construire un test d'égalité de lois basé sur l'écart entre les fonctions caractéristique marginales. Dans le cas d'échantillons indépendants, cette idée a été développée par Epps & Singleton (1986) pour ce qui concerne le cas de distributions univariées; elle a été généralisée au cas de lois multivariées par Alba Fernández et al. (2008).

Pour développer le test, soit la fonction caractéristique empirique bivariée

$$\phi_n(\mathbf{t}) = \frac{1}{n} \sum_{p=1}^{n} e^{i\mathbf{t}\,(X_{p1}, X_{p2})^\top},$$

où $\mathbf{t} = (t_1, t_2)$ et $i = \sqrt{-1}$. Il s'agit d'une estimation de la fonction caractéristique théorique, à savoir $\phi(\mathbf{t}) = \mathrm{E}\{e^{i\mathbf{t}\,(X_1,X_2)^\top}\}$. Une classe de statistiques de test sera basée sur le processus de fonctions caractéristiques

$$\begin{aligned}\Phi_n(t) &= \sqrt{n}\,\{\phi_n(t,0) - \phi_n(0,t)\} \\ &= \frac{1}{\sqrt{n}} \sum_{p=1}^{n} \left(e^{itX_{p1}} - e^{itX_{p2}}\right).\end{aligned}$$

Ainsi, Φ_n est la différence entre les fonctions caractéristiques marginales. Une statistique de test basée sur Φ_n est

$$T_n^{\Psi} = \int_{\mathbb{R}} |\Phi_n(t)|^2 \,\mathrm{d}\Psi(t),$$

où $|r| = \sqrt{a^2 + b^2}$ est le module du nombre complexe $r = a + b\,i$ et $\mathrm{d}\Psi$ est une fonction de poids. On doit supposer que $\mathrm{d}\Psi$ est intégrable. Si $\mathrm{d}\Psi(t) > 0$ presque partout, on peut montrer que $T_n^{\Psi} \to +\infty$ presque sûrement sous $\mathcal{H}_1$. Autrement dit, le test est convergent.

Pour le calcul de p-valeurs, on observe d'abord que

$$\Phi_n(t) = -\int_{\mathbb{R}} \mathbb{A}_n(x) \left(\frac{\mathrm{d}}{\mathrm{d}x}\, e^{itx}\right).$$

Ainsi, des versions multipicateur de Φ_n sont données par

$$\widehat{\Phi}_n^{(h)}(t) = -\int_{\mathbb{R}} \widehat{\mathbb{A}}_n^{(h)}(x) \left(\frac{\mathrm{d}}{\mathrm{d}x}\, e^{itx}\right), \quad h \in \{1, \ldots, M\}.$$

De là, une intégration par partie permet de montrer que

$$\widehat{\Phi}_n^{(h)}(t) = \frac{1}{\sqrt{n}} \sum_{p=1}^{n} \left(\frac{\xi_p^{(h)}}{\bar{\xi}^{(h)}} - 1\right) \left(e^{itX_{p1}} - e^{itX_{p2}}\right).$$

Ainsi, les versions multiplicateurs de T_n^{Ψ} sont

$$\widehat{\mathbb{T}}_n^{\Psi,(h)} = \int_{\mathbb{R}} \left|\widehat{\Phi}^{(h)}(t)\right|^2 \mathrm{d}\Psi(t), \quad h \in \{1, \ldots, M\},$$

Par conséquent,

$$\left(T_n^{\Psi}, \widehat{\mathbb{T}}_n^{\Psi,(1)}, \ldots, \widehat{\mathbb{T}}_n^{\Psi,(M)}\right) \rightsquigarrow \left(\mathbb{T}^{\Psi}, \mathbb{T}^{\Psi,(1)}, \ldots, \mathbb{T}^{\Psi,(M)}\right),$$

où $\mathbb{T}^{\Psi,(1)}, \ldots, \mathbb{T}^{\Psi,(M)}$ sont des copies indépendantes de la limite $\mathbb{T}^{\Psi}$ de T_n^{Ψ}.

Le prochain lemme présente des formules pour T_n^{Ψ} et $\widehat{\mathbb{T}}_n^{\Psi,(1)}, \ldots, \widehat{\mathbb{T}}_n^{\Psi,(M)}$ en terme d'opérations de matrices. La démonstration est dans l'Annexe A.

Lemme 4.3. *Soit la matrice $Q \in \mathbb{R}^{n\times n}$ dont les éléments sont*

$$Q_{pq}^{\Psi} = \beta^{\Psi}\left(X_{p1} - X_{q1}\right) - 2\beta^{\Psi}\left(X_{p1} - X_{q2}\right) + \beta^{\Psi}\left(X_{p2} - X_{q2}\right),$$

où

$$\beta^{\Psi}(a) = \int_{\mathbb{R}} \cos(ta)\, \mathrm{d}\Psi(t).$$

On a

$$T_n^{\Psi} = \frac{1}{n}\mathbf{1}\, Q\, \mathbf{1}^{\top} \quad \textit{et} \quad \widehat{\mathbb{T}}_n^{\Psi,(h)} = \gamma^{(h)}\, Q \left(\gamma^{(h)}\right)^{\top},$$

avec $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^{1\times n}$ et

$$\gamma^{(h)} = \left(\frac{\xi_1^{(h)}}{\bar{\xi}^{(h)}} - 1, \ldots, \frac{\xi_n^{(h)}}{\bar{\xi}^{(h)}} - 1\right).$$

4.2 Généralisation à k échantillons

Soit $\mathbf{X}_1, \ldots, \mathbf{X}_n$, où $\mathbf{X}_p = (X_{p1}, \ldots, X_{pk})$, $p \in \{1, \ldots n\}$, un échantillon aléatoire d'une population dont la fonction de répartition $H : \mathbb{R}^k \to [0,1]$ possède des marges continues $F_1, \ldots, F_k$. Le Théorème 3.2 de Sklar (1959)

assure qu'il existe une unique copule $C : [0,1]^k \to [0,1]$ telle que pour tout $(x_1, \ldots, x_k) \in \mathbb{R}^k$, on a

$$H(x_1, \ldots, x_k) = C\{F_1(x_1), \ldots, F_k(x_k)\}.$$

Comme extension au cas bivarié, on souhaite confronter les hypothèses

$$\mathcal{H}_0^k : F_1 = \cdots = F_k \quad \text{et} \quad \mathcal{H}_1^k : F_j \neq F_{j'} \text{ pour } j, j' \in \{1, \ldots, k\}.$$

Notons que l'hypothèse nulle est équivalente à $F_j - \bar{F} = 0$ *pour tout* $j \in \{1, \ldots, k\}$, où $\bar{F} = (F_1 + \cdots + F_k)/k$. Ainsi, comme

$$F_j - \bar{F} = \left(1 - \frac{1}{k}\right) F_j - \frac{1}{k} \sum_{\ell \neq j} F_\ell$$

et en posant $\mathbf{F} = (F_1, \ldots, F_k)^\top$, on peut écrire

$$\mathcal{H}_0^k : O\,\mathbf{F} = \mathbf{0}_k^\top \quad \text{et} \quad \mathcal{H}_1 : O\,\mathbf{F} \neq \mathbf{0}_k^\top,$$

où $\mathbf{0}_\ell \in \mathbb{R}^\ell$ est un vecteur de zéros et $O \in \mathbb{R}^{k \times k}$ est tel que $O_{jj} = 1 - 1/k$ et $L_{jj'} = -1/k$, $j \neq j'$. Les procédures de test développées dans la suite exploitent cette représentation.

4.2.1 Extension de la statistique de Cramér–von Mises

Soit la fonction de répartition empirique multivariée

$$H_n(\mathbf{x}) = \frac{1}{n} \sum_{p=1}^{n} \mathbb{I}\left(\mathbf{X}_p \leq \mathbf{x}\right),$$

où $\mathbf{x} = (x_1, \ldots, x_k)$. Les marges $F_1, \ldots, F_k$ sont estimées adéquatement par les marges de H_n, à savoir

$$F_{nj}(x) = H_n\left(\infty_{j-1}, x, \infty_{k-j}\right), \quad j \in \{1, \ldots, k\},$$

où les éléments de $\infty_\ell \in \mathbb{R}^\ell$ sont ∞. On définit ensuite $\mathbf{F}_n = (F_{n1}, \ldots, F_{nk})^\top$, de même que le vecteur de processus empiriques

$$\mathbb{D}_n(x) = \sqrt{n}\, O\, \mathbf{F}_n(x).$$

Le comportement asymptotique de $\mathbb{D}_n$ est décrit dans le résultat suivant.

Proposition 4.3. *Sous $\mathcal{H}_0^k$, $\mathbb{D}_n \rightsquigarrow \mathbb{D}$ dans l'espace $D(\mathbb{R}^k)$, où*

$$\mathrm{E}\left\{\mathbb{D}(x)\,\mathbb{D}(x')^\top\right\} = O\,G(x,x')\,O,$$

où pour tout $(x,x') \in \mathbb{R}^2$, la matrice $G(x,x') \in \mathbb{R}^{k\times k}$ est définie par

$$G_{jj'}(x,x') = \begin{cases} C_{jj'}\left\{F(x), F(x')\right\} - F(x)F(x'), & j \neq j'; \\ F(x \wedge x') - F(x)F(x'), & j = j', \end{cases}$$

où F est la fonction de répartition sous l'hypothèse nulle et $C_{jj'}$, pour $j \neq j' \in \{1,\ldots,k\}$, sont les marges bivariées de C.

Si $\mathbb{D}_{nj}$ correspond à la j-ième composante de $\mathbb{D}_n$, alors une statistique de test raisonnable est donnée par

$$V_n = \sum_{j=1}^{k} \int_{\mathbb{R}} \left\{\mathbb{D}_{nj}(x)\right\}^2 \mathrm{d}x.$$

Maintenant, en vue d'obtenir des versions multiplicateurs de V_n, considérons $\widehat{\mathbb{F}}_n^{(h)} = (\widehat{\mathbb{F}}_{n1}^{(h)}, \ldots, \widehat{\mathbb{F}}_{nk}^{(h)})^\top$, où $\widehat{\mathbb{F}}_{nj}^{(h)}(x) = \widehat{\mathbb{H}}_n(\infty_{j-1}, x, \infty_{k-j})$ et

$$\widehat{\mathbb{H}}_n(\mathbf{x}) = \frac{1}{n}\sum_{p=1}^{n}\left(\frac{\xi_p^{(h)}}{\bar{\xi}^{(h)}} - 1\right)\mathbb{I}\left(\mathbf{X}_p \leq \mathbf{x}\right).$$

La version multiplicateur de $\mathbb{D}_n$ est alors $\widehat{\mathbb{D}}_n^{(h)}(x) = O\,\widehat{\mathbb{F}}_n^{(h)}(x)$. Une conséquence directe de la Proposition 4.3 est que

$$\left(\mathbb{D}_n, \widehat{\mathbb{D}}_n^{(1)}, \ldots, \widehat{\mathbb{D}}_n^{(M)}\right) \rightsquigarrow \left(\mathbb{D}, \mathbb{D}^{(1)}, \ldots, \mathbb{D}^{(M)}\right),$$

où $\mathbb{D}^{(1)}, \ldots, \mathbb{D}^{(M)}$ sont des copies indépendantes de $\mathbb{D}$. Ainsi,

$$\widehat{V}_n^{(h)} = \sum_{j=1}^{k} \int_{\mathbb{R}} \left\{ \widehat{\mathbb{D}}_{nj}^{(h)}(x) \right\}^2 \mathrm{d}x,$$

où $\widehat{\mathbb{D}}_{nj}^{(h)}$ est la j-ième composante de $\widehat{\mathbb{D}}_n^{(h)}$, sont des versions multiplicateurs asymptotiquement valides de la statistique V_n. Autrement dit,

$$\left(V_n, \widehat{V}_n^{(1)}, \ldots, \widehat{V}_n^{(M)} \right) \rightsquigarrow \left(\mathbb{V}, \mathbb{V}^{(1)}, \ldots, \mathbb{V}^{(M)} \right),$$

où $\mathbb{V}^{(1)}, \ldots, \mathbb{V}^{(M)}$ sont des copies indépendantes de

$$\mathbb{V} = \sum_{j=1}^{k} \int_{\mathbb{R}} \{\mathbb{D}(x)\}^2 \mathrm{d}x.$$

À l'instar du cas bivarié, des formules de calcul faciles à implanter sont disponibles pour V_n et $\widehat{V}_n^{(1)}, \ldots, \widehat{V}_n^{(M)}$. La démonstration est dans l'Annexe A.

Lemme 4.4. *Pour tout $(j, j') \in \{1, \ldots, k\}^2$, considérons la matrice $M^{jj'} \in \mathbb{R}^{n \times n}$ dont les éléments sont $M_{pq}^{jj'} = \max(X_{pj}, X_{qj'})$. Alors,*

$$V_n \;=\; \mathbf{1}\,\mathrm{diag}\,(O\,B\,O) \quad \textit{et} \quad \widehat{V}_n^{(h)} = \mathbf{1}\,\mathrm{diag}\left(O\,\widehat{B}^{(h)}\,O\right),$$

où $B, \widehat{B}^{(1)}, \ldots, \widehat{B}^{(M)} \in \mathbb{R}^{k \times k}$ sont tels que

$$B_{jj'} = -\frac{1}{n}\,\mathbf{1}\,M^{jj'}\,\mathbf{1}^\top \quad \textit{et} \quad \widehat{B}_{jj'}^{(h)} = -\gamma^{(h)}\,M^{jj'}\,\gamma^{(h),\top}.$$

4.2.2 Extension des statistiques *fonction caractéristique*

Considérons les fonctions caractéristique jointe théorique et empirique

$$\phi(\mathbf{t}) = \mathrm{E}\left(e^{i\mathbf{t}\mathbf{X}}\right) \quad \text{et} \quad \phi_n(\mathbf{t}) = \frac{1}{n}\sum_{p=1}^{n} e^{i\mathbf{t}\mathbf{X}_p},$$

où $\mathbf{t} = (t_1, \ldots, t_k)$. Soient ensuite les vecteurs $\underline{\phi} = (\phi_1, \ldots, \phi_k)^\top$ et $\underline{\phi}_n = (\phi_{n1}, \ldots, \phi_{nk})^\top$, où pour $j \in \{1, \ldots, k\}$,

$$\phi_j(t) = \phi\left(\mathbf{0}_{j-1}, t, \mathbf{0}_{k-j}\right) \quad \text{et} \quad \phi_{nj}(t) = \phi_n\left(\mathbf{0}_{j-1}, t, \mathbf{0}_{k-j}\right).$$

Les hypothèses nulle et alternative peuvent alors s'écrire

$$\mathcal{H}_0^k : O\,\underline{\phi} = \mathbf{0}_k^\top \quad \text{et} \quad \mathcal{H}_1^k : O\,\underline{\phi} \neq \mathbf{0}_k^\top.$$

Des statistiques de test seront construites en fonction du processus empirique

$$\mathbb{E}_n(t) = \sqrt{n}\, O\, \underline{\phi}_n(t).$$

Spécifiquement, si $\mathbb{E}_{nj}$ est la j-ième composante de $\mathbb{E}_n$, alors les statistiques de test proposées sont

$$W_n^\Psi = \sum_{j=1}^{k} \int_{\mathbb{R}} \left|\mathbb{E}_{nj}(t)\right|^2 \mathrm{d}\Psi(t),$$

où $\mathrm{d}\Psi$ est une fonction de poids intégrable. Pour obtenir des versions multiplicateurs de W_n^Ψ, on définit

$$\widehat{\phi}_n^{(h)}(\mathbf{t}) = \frac{1}{\sqrt{n}} \sum_{p=1}^{n} \left(\frac{\xi_p^{(h)}}{\bar{\xi}^{(h)}} - 1 \right) e^{i\mathbf{t}\mathbf{X}_p}.$$

Alors, $\widehat{\mathbb{E}}_n^{(h)} = O\,\widehat{\underline{\phi}}^{(h)}(t)$ sont les versions multiplicateurs de $\mathbb{E}_n$, où $\widehat{\underline{\phi}}^{(h)}(t) = (\widehat{\phi}_{n1}^{(h)}(t), \ldots, \widehat{\phi}_{nk}^{(h)}(t))^\top$ et

$$\widehat{\phi}_{nj}^{(h)}(t) = \widehat{\phi}_n^{(h)}\left(\mathbf{0}_{j-1}, t, \mathbf{0}_{k-j}\right).$$

Les versions multiplicateur de W_n^Ψ sont donc

$$\widehat{W}_n^{\Psi,(h)} = \sum_{j=1}^{k} \int_{\mathbb{R}} \left|\widehat{\mathbb{E}}_{nj}^{(h)}(t)\right|^2 \mathrm{d}\Psi(t).$$

Des formules pour W_n^Ψ et $\widehat{W}_n^{\Psi,(1)}, \ldots, \widehat{W}_n^{\Psi,(M)}$ sont décrites dans le lemme suivant. La preuve est dans l'Annexe A.

Lemme 4.5. *Pour tout* $(j, j') \in \{1, \ldots, k\}^2$, *considérons la matrice* $\underline{M}^{jj'} \in \mathbb{R}^{n \times n}$ *dont les éléments sont* $\underline{M}^{jj'}_{pq} = \beta^{\Psi}(X_{pj} - X_{qj'})$. *Alors,*

$$W_n^{\Psi} = \mathbf{1}\,\mathrm{diag}\,(O\,\underline{B}\,O) \quad et \quad \widehat{W}_n^{\Psi,(h)} = \mathbf{1}\,\mathrm{diag}\left(O\,\widehat{\underline{B}}^{(h)}\,O\right),$$

où $\underline{B}, \widehat{\underline{B}}^{(1)}, \ldots, \widehat{\underline{B}}^{(M)} \in \mathbb{R}^{k \times k}$ *sont tels que*

$$\underline{B}_{jj'} = \frac{1}{n}\,\mathbf{1}\,\underline{M}^{jj'}\,\mathbf{1}^{\top} \quad et \quad \widehat{\underline{B}}^{(h)}_{jj'} = \gamma^{(h)}\,\underline{M}^{jj'}\,\gamma^{(h),\top}.$$

4.3 Études de simulation

4.3.1 Puissance des tests dans le cas bivarié

Les procédures développées jusqu'ici pour le calcul de p-valeurs sont basées sur la méthode du multiplicateur. On a montré que cette technique est asymptotiquement valide dans tous les cas considérés. Cela n'indique cependant rien sur le comportement des tests pour de petites tailles d'échantillons.

Pour tester l'efficacité de la méthode du multiplicateur, un grand nombre d'échantillons aléatoires de loi $H(x_1, x_2) = C\{F(x_1), F(x_2)\}$ seront générés pour certains choix de F et de C. Parmi les choix possibles pour la structure de dépendance C, on retrouve la copule Normale

$$C_{\rho}^{\mathrm{N}}(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} h_{\rho}(x_1, x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1,$$

où h_{ρ} est la densité normale classique bivariée de corrélation ρ. On considérera aussi la copule de Clayton, à savoir

$$C_{\theta}^{\mathrm{CL}}(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}, \quad \theta > -1.$$

Ces copules sont symétriques, car $C_\rho^{\mathrm{N}}(u_1, u_2) = C_\rho^{\mathrm{N}}(u_2, u_1)$ et $C_\rho^{\mathrm{CL}}(u_1, u_2) = C_\rho^{\mathrm{CL}}(u_2, u_1)$. Pour étudier le comportement des tests sous des structures de dépendance asymétriques, on considéra des versions *asymétriques* de C_θ^{N} et C_θ^{CL}. Spécifiquement, suivant le critère de Khoudraji (1995),

$$K_{\delta,\rho}^{\mathrm{N}}(u_1, u_2) = u_1^{\delta}\, C_\rho^{\mathrm{N}}\left(u_1^{1-\delta}, u_2\right) \quad \text{et} \quad K_{\delta,\theta}^{\mathrm{CL}}(u_1, u_2) = u_1^{\delta}\, C_\theta^{\mathrm{CL}}\left(u_1^{1-\delta}, u_2\right),$$

sont des copules asymétriques, où $\delta \in (0, 1)$. Dans les résultats de simulation présentés ici, le paramètre d'asymétrie est fixé à $\delta = 1/2$.

Au Tableau 4.1, la probabilité de rejeter l'hypothèse nulle pour les tests basés sur S_n et T_n^{Ψ}, avec $\mathrm{d}\Psi(t) = e^{-t^2}$, est estimée à l'aide de 10 000 échantillons de taille $n = 100$. Le nombre d'échantillons multiplicateurs a été fixé à $M = 1\ 000$. Pour chacune des quatre structures de dépendance considérées, on a $\tau(C) \in \{-2/3, -1/3, 1/3, 2/3\}$, où

$$\tau(C) = 4 \int_0^1 \int_0^1 C(u_1, u_2) \mathrm{d}C(u_1, u_2) - 1$$

est la mesure de dépendance de Kendall. Les modèles considérés pour les lois marginales F sont

(i) la distribution exponnentielle de moyenne λ, notée $\mathrm{Exp}(\lambda)$;

(ii) la distribution gamma, notée $\mathrm{G}(\alpha, \beta)$;

(iii) la distribution normale, notée $\mathcal{N}(\mu, \sigma^2)$.

En examinant les entrées du Tableau 4.1, on constate d'abord que les tests tiennent très bien leur seuil de 5%; c'est vrai pour toutes les copules considérées. En particulier, le fait que la structure de dépendance soit symétrique

ou asymétrique ne pose pas de problème aux tests sous $\mathcal{H}_0$. Également, la force de la dépendance, mesurée par $\tau(C)$, de même que le choix des marges, n'affectent pas vraiment la capacité des tests à conserver leur seuil.

Dans un deuxième temps, la puissance des tests, c'est-à-dire leur capacité à rejeter $\mathcal{H}_0$ lorsqu'elle est fausse, a été estimée. Les résultats se retrouvent au Tableau 4.2, pour $n = 100$, et au Tableau 4.3, pour $n = 250$. Ici, les mêmes structures de dépendance que celles considérées au Tableau 4.1 ont été utilisées, c'est-à-dire C_ρ^{N}, C_θ^{CL}, $K_{\delta,\rho}^{\mathrm{N}}$ et $K_{\delta,\theta}^{\mathrm{CL}}$. Pour se placer dans des situations où $\mathcal{H}_0$ est fausse, on a toutefois considéré des situations où les marges sont différentes. On constate que

(i) les deux tests sont très puissants, même quand $n = 100$;

(ii) le test basé sur S_n est toujours plus puissant que celui basé sur T_n^Ψ; la différence est plus marquée sous des marges Normales;

(iii) plus la valeur de $\tau(C)$ augmente, plus la puissance augmente;

(iv) pour une copule donnée, la puissance des tests est plus élevée pour sa version asymétrique que pour sa version symétrique.

4.3.2 Puissance des tests dans le cas à k échantillons

Au Tableau 4.4, on présente les résultats de la puissance des tests basés sur V_n et W_n^Ψ pour la comparaison de $k = 3$ et $k = 4$ lois. Pour ce faire, on a généré 10 000 échantillons de taille $n = 100$ de distributions multivariées dont les $k - 1$ premières marges sont F et la k-ième loi est $\tilde{F}$. Ainsi, on a

généré des observations à partir du modèle

$$H(x_1, \ldots, x_k) = C\left\{F(x_1), \ldots, F(x_{k-1}), \tilde{F}(x_k)\right\},$$

pour certains choix de F, $\tilde{F}$ et C. Pour les marges, on a utilisé les mêmes lois que pour le cas à deux échantillons; pour C, on a considéré la copule Normale C_Σ^{N} décrite à l'équation (3.4). Aussi, afin d'étudier le comportement des tests sous de la dépendance asymétrique, on a considéré la version *asymétrique* de C_Σ^{N} obtenue à l'aide de la méthode de Khoudraji (1995), à savoir

$$K_{\delta,\Sigma}^{\mathrm{N}}(u_1, \ldots, u_k) = u_1^{\delta}\, C_\Sigma^{\mathrm{N}}\left(u_1^{1-\delta}, u_2, \ldots, u_k\right),$$

où $\delta \in [0, 1]$. Dans les résultats de simulations présentés ici, le paramètre d'asymétrie a été fixé à $\delta = 1/2$. Le nombre d'échantillons *multiplicateurs* pour le calcul des p-valeurs a quant à lui été fixé à $M = 1\ 000$.

Les valeurs du Tableau 4.4 démontrent que les tests tiennent bien leur seuil de 5% sous l'hypothèse nulle. Lorsque la structure de dépendance est symétrique, les tests ont un peu plus de facilité à conserver leur seuil quand $\rho = 1/3$, donc pour de la faible dépendance, que lorsque $\rho = 2/3$. Enfin, on constate que la statistique W_n a tendance a conserver mieux son seuil que V_n dans presque tous les cas, bien que la différence soit minime.

Ensuite, la bonne capacité des tests à rejeter l'hypothèse nulle quand elle est fausse est démontrée par les entrées du Tableau 4.4. La statistique V_n est légèrement plus puissante que W_n dans tous les cas considérés. On remarque aussi que plus le paramètre de dépendance est élevé, plus les tests sont puissants. Finalement, les puissances sont plus élevées sous les versions symétriques des structures de dépendance que sous les versions asymétriques.

4.4 Analyse de vrais jeux de données

4.4.1 Consommation d'éléments nutritifs

En 1985, le département d'agriculture des États-Unis a conduit une étude auprès de 747 femmes âgées entre 25 et 50 ans. On a considéré la consommation quotidienne de cinq éléments nutritifs, à savoir

- calcium, mesurée en milligrammes;
- fer, mesurée en milligrammes;
- protéines, mesurée en grammes;
- vitamine A, mesurée en milligrammes;
- vitamine C, mesurée en milligrammes.

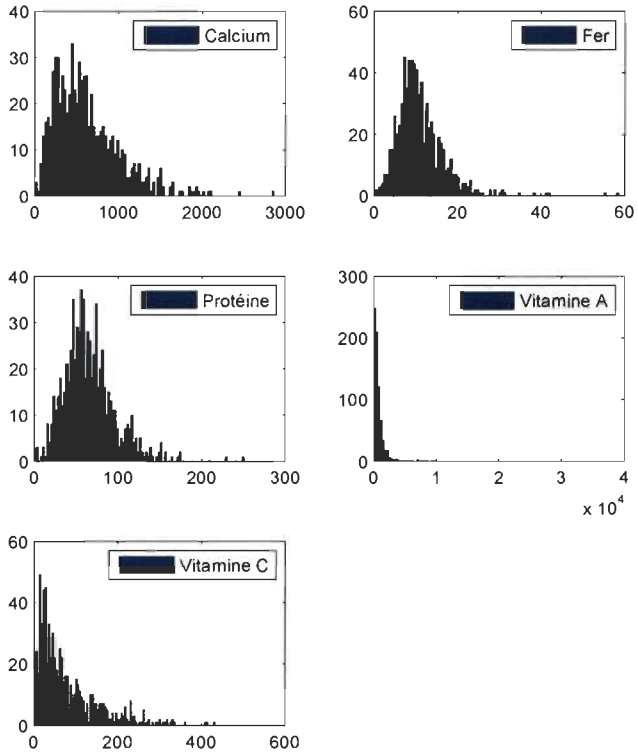
Dans la suite, ces variables seront notées Ca, Fe, Pr, vA et vB. Leurs histogrammes sont présentés à la figure 4.1.

Tout d'abord, les tests d'égalité simultanée de ces cinq échantillons rejettent clairement l'hypothèse nulle. En effet, on a

$$V_n = 310\ 090,0 \quad \text{et} \quad W_n = 261,4.$$

Les p-valeurs correspondantes sont substantiellement inférieures à $\alpha = 0,01$. Ensuite, pour vérifier si certaines paires de variables ont la même loi, les tests basés sur S_n et sur T_n^{Ψ}, avec $\Psi(t) = e^{-t^2}$, ont été effectués sur chaque paire possible. Les résultats se retrouvent dans le Tableau 4.5. On voit que les tests rejettent l'hypothèse nulle d'égalité de lois dans tous les cas.

Figure 4.1: Histogrammes de la consommation quotidienne en calcium, fer, protéines, vitamine A et vitamine B chez les femmes américaines



4.4.2 Concentration d'éléments chimiques dans l'eau

Cook & Johnson (1986) ont considéré un jeu de données comprenant 655 analyses chimiques provenant d'échantillons d'eau au Colorado. Les concentrations en uranium (U), lithium (Li), cobalt (Co), potassium (K), caesium (Ca), scandium (Sc) et titanium (Ti) ont été mesurées. Les histogrammes

de ces sept variables se retrouvent à la figure 4.2.

Dans un premier temps, les tests d'égalité simultanée des sept échantillons rejettent clairement l'hypothèse nulle, puisque $V_n = 2\,618,8$ et $W_n = 2\,914,5$. Ensuite, les tests basés sur les statistiques S_n et T_n^{Ψ} ont été effectués pour toutes les paires de variables possibles. Les résultats sont dans le Tableau 4.6. On voit que l'hypothèse nulle est rejetée dans tous les cas, sauf pour la paire (Co,Sc) testée avec T_n^{Ψ}; néanmoins, la p-valeur est à peine supérieure à 0,05.

Figure 4.2: Histogrammes pour les concentrations en uranium, lithium, cobalt, potassium, caesium, scandium et titanium

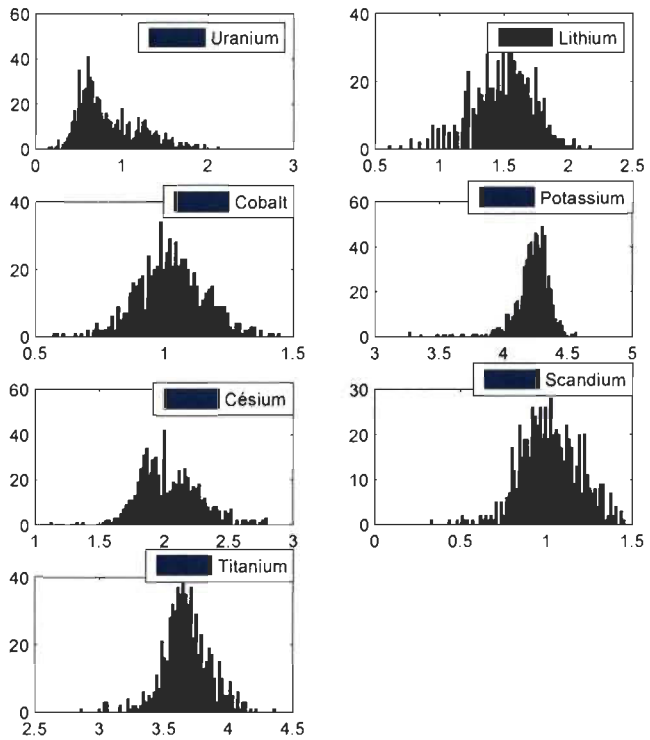


Table 4.1: Estimation, basée sur 10 000 itérations, de la probabilité de rejeter $\mathcal{H}_0$ sous différentes structures de dépendance et des marges identiques quand $n = 100$ et $M = 1\ 000$

		$\tau = -\frac{2}{3}$		$\tau = -\frac{1}{3}$		$\tau = \frac{1}{3}$		$\tau = \frac{2}{3}$	
C	F	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}
C^{CL}	Exp(1)	5.05	5.70	5.07	5.11	4.97	4.95	4.18	4.50
	G(2,1)	5.79	5.60	5.20	5.41	4.82	5.23	4.17	4.49
	$\mathcal{N}(0,1)$	5.55	5.45	5.48	5.36	5.23	5.32	4.35	5.29
C^{N}	Exp(1)	4.92	5.50	5.73	5.52	4.75	4.67	3.84	4.07
	G(2,1)	5.57	5.62	5.30	5.30	4.69	5.03	3.64	4.61
	$\mathcal{N}(0,1)$	5.41	5.61	5.26	5.59	4.87	5.23	3.89	5.16
K^{CL}	Exp(1)	5.43	5.63	5.35	5.26	4.55	5.42	4.96	5.12
	G(2,1)	5.06	5.55	5.26	5.33	4.97	4.97	4.92	4.92
	$\mathcal{N}(0,1)$	5.25	5.67	5.24	5.39	5.04	5.28	5.28	5.07
K^{N}	Exp(1)	5.61	5.55	5.00	5.07	5.02	5.03	5.26	5.01
	G(2,1)	5.61	5.37	5.17	5.17	5.51	5.20	4.79	5.00
	$\mathcal{N}(0,1)$	4.83	5.27	5.64	5.57	5.11	5.10	5.28	4.89

4.5 Preuves des résultats théoriques

4.5.1 Proposition 4.1

Selon la théorie classique, $\mathbb{H}_n(x_1, x_2) = \sqrt{n}\{H_n(x_1, x_2) - H(x_1, x_2)\}$ converge en loi dans $D(\mathbb{R})$ vers un processus Gaussien centré et continue $\mathbb{H}$ avec

$$\begin{aligned} \Gamma_{\mathbb{H}}(x_1, x_2, x_1', x_2') &= \mathrm{E}\{\mathbb{H}(x_1, x_2)\mathbb{H}(x_1', x_2')\} \\ &= H(x_1 \wedge x_1', x_2 \wedge x_2') - H(x_1, x_2)H(x_1', x_2'). \quad (4.1) \end{aligned}$$

Sous $\mathcal{H}_0$, on a que $H(x, \infty) = H(\infty, x)$, alors $\mathbb{A}_n(x) = \mathbb{H}_n(x, \infty) - \mathbb{H}_n(\infty, x)$. Donc, il est clair que

$$\mathbb{A}_n(x) \rightsquigarrow \mathbb{A}(x) = \mathbb{H}(x, \infty) - \mathbb{H}(\infty, x),$$

où $\mathbb{A}$ est un processus Gaussien centré et continue tel que

$$\begin{aligned} \mathrm{E}\{\mathbb{A}(x)\mathbb{A}(x')\} &= \mathrm{E}\{(\mathbb{H}(x, \infty) - \mathbb{H}(\infty, x)) \times (\mathbb{H}(x', \infty) - \mathbb{H}(\infty, x'))\} \\ &= \Gamma_{\mathbb{H}}(x, \infty, x', \infty) - \Gamma_{\mathbb{H}}(x, \infty, \infty, x') \\ &\quad - \Gamma_{\mathbb{H}}(\infty, x, x', \infty) + \Gamma_{\mathbb{H}}(\infty, x, \infty, x'). \end{aligned}$$

Finalement, on utilise l'équation (4.1) et le fait que $H(x_1, x_2) = C\{F(x_1), F(x_2)\}$ sous $\mathcal{H}_0$ pour retrouver le résultat annoncé.

4.5.2 Proposition 4.2

Par le Théorème de la limite centrale du multiplicateur (Kosorok, 2008; van der Vaart & Wellner, 1996), les processus empiriques

$$\widehat{\mathbb{H}}^{(h)}(x_1, x_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\xi_i^{(h)}}{\bar{\xi}^{(h)}} - 1 \right) \mathbb{I}\left(X_{i1} \leq x_1, X_{i2} \leq x_2\right), \quad h \in \{1, \ldots, M\},$$

sont tels que

$$\left(\mathbb{H}_n, \widehat{\mathbb{H}}^{(1)}, \ldots, \widehat{\mathbb{H}}^{(M)}\right) \rightsquigarrow \left(\mathbb{H}, \mathbb{H}^{(1)}, \ldots, \mathbb{H}^{(M)}\right)$$

dans l'espace $D(\mathbb{R})^{\otimes(M+1)}$, où $\mathbb{H}^{(1)}, \ldots, \mathbb{H}^{(M)}$ sont des copies indépendantes de $\mathbb{H}$. Le résultat s'obtient en notant que $\widehat{\mathbb{A}}_n^{(h)}(x) = \widehat{\mathbb{H}}^{(h)}(x, \infty) - \widehat{\mathbb{H}}^{(h)}(\infty, x)$.

4.5.3 Lemme 4.2

En posant $a \vee b = \max(a, b)$, on a

$$\begin{aligned}
S_n &= n \int_{\mathbb{R}} \left\{ \frac{1}{n} \sum_{p=1}^{n} \left(\mathbb{I}\left(X_{p1} \leq x\right) - \mathbb{I}\left(X_{p2} \leq x\right) \right) \right\}^2 \mathrm{d}x \\
&= \frac{1}{n} \int_{\mathbb{R}} \sum_{p,q=1}^{n} \left\{ \mathbb{I}\left(X_{p1} \leq x\right) - \mathbb{I}\left(X_{p2} \leq x\right) \right\} \left\{ \mathbb{I}\left(X_{q1} \leq x\right) - \mathbb{I}\left(X_{q2} \leq x\right) \right\} \mathrm{d}x \\
&= \frac{1}{n} \sum_{p,q=1}^{n} \int_{\mathbb{R}} \left\{ \mathbb{I}\left(X_{p1} \vee X_{q1} \leq x\right) - \mathbb{I}\left(X_{p1} \vee X_{q2} \leq x\right) \right. \\
&\qquad \left. - \mathbb{I}\left(X_{p2} \vee X_{q2} \leq x\right) + \mathbb{I}\left(X_{p2} \vee X_{q2} \leq x\right) \right\} \mathrm{d}x \\
&= \lim_{K \to \infty} \frac{1}{n} \sum_{p,q=1}^{n} \left(\int_{X_{p1} \vee X_{q1}}^{K} \mathrm{d}x - \int_{X_{p1} \vee X_{q2}}^{K} \mathrm{d}x - \int_{X_{p2} \vee X_{q1}}^{K} \mathrm{d}x + \int_{X_{p2} \vee X_{q2}}^{K} \mathrm{d}x \right) \\
&= \lim_{K \to \infty} \frac{1}{n} \sum_{p,q=1}^{n} \left\{ K - \left(X_{p1} \vee X_{q1}\right) - K + \left(X_{p1} \vee X_{q2}\right) - K + \left(X_{p2} \vee X_{q1}\right) \right. \\
&\qquad \left. + K - \left(X_{p2} \vee X_{q2}\right) \right\} \\
&= \frac{1}{n} \sum_{p,q=1}^{n} \left\{ 2\left(X_{p1} \vee X_{q2}\right) - \left(X_{p1} \vee X_{q1}\right) - \left(X_{p2} \vee X_{q2}\right) \right\}.
\end{aligned}$$

En terme de matrice, cela est équivalent à $S_n = \mathbf{1}\, L\, \mathbf{1}^\top / n$. Les calculs sont semblables pour les versions multiplicateurs de S_n.

4.5.4 Lemme 4.3

Comme $e^{ix} = \cos(x) + i \sin(x)$, alors

$$\begin{aligned}
\Phi_n(t) &= \frac{1}{\sqrt{n}} \sum_{p=1}^{n} \left\{ \cos(tX_{p1}) - \cos(tX_{p2}) \right\} \\
&\quad + i \frac{1}{\sqrt{n}} \sum_{p=1}^{n} \left\{ \sin(tX_{p1}) - \sin(tX_{p2}) \right\}.
\end{aligned}$$

En utilisant l'identité trigonométrique

$$\cos(x)\cos(y) + \sin(x)\sin(y) = \cos(x-y),$$

on montre que

$$|\Phi_n(t)|^2 = \frac{1}{n}\sum_{p,q=1}^{n} Q_{pq}(t),$$

où

$$Q_{pq}(t) = \cos\left\{t\left(X_{p1} - X_{q1}\right)\right\} - 2\cos\left\{t\left(X_{p1} - X_{q2}\right)\right\} + \cos\left\{t\left(X_{p2} - X_{q2}\right)\right\}.$$

En intégrant sur $\mathbb{R}$ par rapport à la fonction de poids $\mathrm{d}\Psi$, alors

$$T_n^{\Psi} = \frac{1}{n}\sum_{p,q=1}^{n} Q_{pq}^{\Psi} = \frac{1}{n}\,\mathbf{1}\,Q^{\Psi}\,\mathbf{1}^{\top}.$$

Pour les versions multiplicateurs, on note que

$$\begin{aligned}\widehat{\Phi}_n^{(h)}(t) &= \frac{1}{\sqrt{n}}\sum_{p=1}^{n}\gamma_p^{(h)}\left\{\cos(tX_{p1}) - \cos(tX_{p2})\right\} \\ &\quad + i\,\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\gamma_p^{(h)}\left\{\sin(tX_{p1}) - \sin(tX_{p2})\right\}.\end{aligned}$$

Ainsi, en utilisant l'identité $\cos(x)\cos(y) + \sin(x)\sin(y) = \cos\{t(x-y)\}$ une seconde fois,

$$\left|\widehat{\Phi}_n^{(h)}(t)\right|^2 = \frac{1}{n}\sum_{p,q=1}^{n}\gamma_p^{(h)}\gamma_q^{(h)}Q_{pq}(t).$$

Enfin,

$$\widehat{\mathbb{T}}_n^{\Psi,(h)} = \frac{1}{n}\sum_{p,q=1}^{n}\gamma_p^{(h)}\gamma_q^{(h)}Q_{pq}^{\Psi} = \gamma^{(h)}\,Q\,\gamma^{(h),\top}.$$

4.5.5 Proposition 4.3

D'abord, $\mathbb{F}_n = \sqrt{n}(\mathbf{F}_n - \mathbf{F})$ converge vers un vecteur Gaussien $\mathbb{F}$ avec une structure de covariance caractérisée par la matrice $G(x, x') = \mathrm{E}\left\{\mathbb{F}(x)\mathbb{F}(x')^\top\right\}$, où l'élément (j, j') est

$$G_{jj'}(x, x') = \mathrm{E}\left\{\mathbb{F}_j(x)\mathbb{F}_{j'}(x')\right\} = \begin{cases} C_{jj'}\left\{F(x), F(x')\right\} - F(x)F(x'), & j \neq j; \\ \\ F(x \wedge x') - F(x)F(x'), & j \neq j'. \end{cases}$$

Comme $O\,\mathbf{F} = \mathbf{0}_k^\top$ sous $\mathcal{H}_0^k$, alors $\mathbb{D}_n(x) = O\,\mathbb{F}_n(x)$, ce qui fait que

$$\mathbb{D}_n(x) \rightsquigarrow \mathbb{D}(x) = O\,\mathbb{F}.$$

De plus, $\mathrm{E}\left\{\mathbb{D}(x)\mathbb{D}(x')^\top\right\} = O\,G(x, x')\,O$.

4.5.6 Lemme 4.4

Pour chaque $(j, j') \in \{1, \ldots, k\}^2$, soit $M^{jj'} \in \mathbb{R}^{n \times n}$ avec

$$M_{pq}^{jj'} = \max\left(X_{pj}, X_{qj'}\right).$$

On pose $e_j = (\mathbf{0}_{j-1}, 1, \mathbf{0}_{k-j})$ et on note que

$$\mathbb{D}_{nj}(x) = e_j\,\mathbb{D}_n = \sqrt{n}\,e_j\,O\,\mathbf{F}_n(x) = \sqrt{n}\,O_j\,\mathbf{F}_n(x),$$

où O_j est la j-ième ligne de O. Ainsi,

$$\left\{\mathbb{D}_{nj}(x)\right\}^2 = n\,O_j\,\mathbf{F}_n(x)\,\mathbf{F}_n(x)^\top\,O_j^\top.$$

Par conséquent,

$$\int_{\mathbb{R}} \{\mathbb{D}_{nj}(x)\}^2 \, \mathrm{d}x = O_j \, B \, O_j,$$

où les éléments de $B \in \mathbb{R}^{k \times k}$ sont

$$\begin{aligned} B_{jj'} &= \int_{\mathbb{R}} n \, F_{nj}(x) F_{nj'}(x) \, \mathrm{d}x \\ &= \lim_{m \to \infty} \frac{1}{n} \sum_{p,q=1}^{n} (m - X_{pj} \vee X_{qj'}) \\ &\equiv -\frac{1}{n} \mathbf{1} \, M^{jj'} \, \mathbf{1}^\top. \end{aligned}$$

Finalement, on remarque que V_n est la somme des éléments diagonaux de $O \, B \, O$, *i.e.* $V_n = \mathbf{1} \, \mathrm{diag}(O \, B \, O)$.

Pour les versions multiplicateurs, on note que $\widehat{\mathbb{D}}_{nj}^{(h)} = e_j \, \widehat{\mathbb{D}}_n^{(h)}$, alors que $\{\widehat{\mathbb{D}}_{nj}^{(h)}(x)\}^2 = O_j \, \widehat{\mathbb{F}}_n^{(h)}(x) \, \widehat{\mathbb{F}}_n^{(h)}(x)^\top \, O_j^\top$. Ainsi,

$$\int_{\mathbb{R}} \left\{ \widehat{\mathbb{D}}_{nj}^{(h)}(x) \right\}^2 \mathrm{d}x = O_j \, \widehat{B}^{(h)} \, O_j,$$

où la matrice $\widehat{B}^{(h)} \in \mathbb{R}^{k \times k}$ est telle que

$$\widehat{B}_{jj'}^{(h)} = \int_{\mathbb{R}} \widehat{\mathbb{F}}_{nj}^{(h)}(x) \widehat{\mathbb{F}}_{nj'}^{(h)}(x) \, \mathrm{d}x = -\gamma^{(h)} \, M^{jj'} \, \gamma^{(h),\top}.$$

Finalement, $\widehat{V}_n^{(h)}$ est la somme des éléments diagonaux de $O \, \widehat{B}^{(h)} \, O$, *i.e.* $\widehat{V}_n^{(h)} = \mathbf{1} \, \mathrm{diag}(O \, \widehat{B}^{(h)} \, O)$.

4.5.7 Lemme 4.5

Soient $\mathrm{Re}(\underline{\phi}_n(t))$ et $\mathrm{Im}(\underline{\phi}_n(t))$, les vecteurs des parties réelles et imaginaires de $\underline{\phi}_n(t)$. Comme $\mathbb{E}_{nj}(t) = \sqrt{n} \, O_j \, \underline{\phi}_n(t)$, on a

$$|\mathbb{E}_{nj}(t)|^2 = O_j \, B(t) \, O_j^\top,$$

où

$$B(t) = n \operatorname{Re}\left(\underline{\phi}_n(t)\right) \operatorname{Re}\left(\underline{\phi}_n(t)\right)^\top + n \operatorname{Im}\left(\underline{\phi}_n(t)\right) \operatorname{Im}\left(\underline{\phi}_n(t)\right)^\top.$$

Maintenant, comme

$$\operatorname{Re}\left(\underline{\phi}_n(t)\right)_j = \frac{1}{n}\sum_{p=1}^{n} \cos\left(tX_{pj}\right) \quad \text{et} \quad \operatorname{Im}\left(\underline{\phi}_n(t)\right)_j = \frac{1}{n}\sum_{p=1}^{n} \sin\left(tX_{pj}\right),$$

on obtient en utilisant l'identité trigonométrique $\cos(x)\cos(y)+\sin(x)\sin(y) = \cos\{t(x-y)\}$ que

$$B_{jj'}(t) = \frac{1}{n}\sum_{p,q=1}^{n} \cos\left\{t\left(X_{pj} - X_{qj'}\right)\right\}.$$

Ainsi,

$$\int_{\mathbb{R}} \left|\mathbb{E}_{nj}(t)\right|^2 \mathrm{d}\Psi(t) = O_j \left\{\int_{\mathbb{R}} B(t)\,\mathrm{d}\Psi(t)\right\} O_j^\top = O_j\, \underline{B}\, O_j^\top,$$

où

$$\underline{B}_{jj'} = \frac{1}{n}\sum_{p,q=1}^{n} \beta^{\Psi}\left(X_{pj} - X_{qj'}\right) = \frac{1}{n}\,\mathbf{1}\,\underline{M}^{jj'}\,\mathbf{1}^\top.$$

On note, enfin, que $O_j\, \underline{B}\, O_j^\top$ est le j-ième élément de $\operatorname{diag}(O\, \underline{B}\, O)$. On utilise des arguments similaires pour les versions multiplicateurs de W_n^{Ψ}.

Table 4.2: Estimation, basée sur 10 000 itérations, de la probabilité de rejeter $\mathcal{H}_0$ sous différentes structures de dépendance et des marges différentes quand $n = 100$ et $M = 1\ 000$

			$\tau = -\frac{2}{3}$		$\tau = -\frac{1}{3}$		$\tau = \frac{1}{3}$		$\tau = \frac{2}{3}$	
C	F_1	F_2	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}
C^{CL}	Exp(1)	Exp $(\frac{3}{2})$	57.1	52.1	66.2	59.3	90.7	84.3	99.7	98.7
	Exp(1)	G $(\frac{3}{2}, 1)$	75.8	73.8	83.2	80.6	99.1	98.1	100	100
	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	74.2	71.6	81.9	77.8	98.4	97.0	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	91.5	70.3	96.6	80.5	100	98.9	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100
C^{N}	Exp(1)	Exp $(\frac{3}{2})$	55.7	49.1	62.9	57.2	94.1	88.4	100	100
	Exp(1)	G $(\frac{3}{2}, 1)$	72.8	71.2	81.4	79.2	99.6	99.0	100	100
	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	73.5	67.8	81.1	76.7	99.4	98.1	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	88.9	68.2	95.8	80.6	100	99.2	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100
K^{CL}	Exp(1)	Exp $(\frac{3}{2})$	67.0	60.3	71.4	64.8	84.2	77.7	91.2	86.0
	Exp(1)	G $(\frac{3}{2}, 1)$	83.3	81.7	87.3	85.5	96.6	95.0	98.8	98.0
	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	82.9	78.5	87.2	82.9	95.4	92.3	98.4	96.8
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	97.2	82.2	98.4	86.7	99.9	96.6	100	99.4
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100
K^{N}	Exp(1)	Exp $(\frac{3}{2})$	64.3	58.6	68.2	62.5	87.7	79.5	95.0	88.1
	Exp(1)	G $(\frac{3}{2}, 1)$	80.2	79.3	86.3	83.8	97.5	96.1	99.3	98.6
	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	81.1	77.2	85.1	81.2	96.9	94.4	99.3	97.5
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	96.4	79.3	97.7	86.0	99.9	97.5	100	99.8
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100

Table 4.3: Estimation, basée sur 10 000 itérations, de la probabilité de rejeter $\mathcal{H}_0$ sous différentes structures de dépendance et des marges différentes quand $n = 250$ et $M = 1\ 000$

			$\tau = -\frac{2}{3}$		$\tau = -\frac{1}{3}$		$\tau = \frac{1}{3}$		$\tau = \frac{2}{3}$	
C	F_1	F_2	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}	S_n	T_n^{Ψ}
	Exp(1)	Exp $(\frac{3}{2})$	93.2	89.6	96.5	94.1	99.9	99.7	100	100
	Exp(1)	G $(\frac{3}{2}, 1)$	98.8	98.1	99.7	99.3	100	100	100	100
C^{CL}	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	98.6	97.8	99.6	99.3	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	100	99.9	100	100	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100
	Exp(1)	Exp $(\frac{3}{2})$	92.2	87.7	95.2	92.8	100	100	100	100
	Exp(1)	G $(\frac{3}{2}, 1)$	98.2	97.8	99.6	99.3	100	100	100	100
C^{N}	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	98.2	97.1	99.3	98.7	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	100	99.9	100	100	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100
	Exp(1)	Exp $(\frac{3}{2})$	97.1	94.5	98.1	96.1	100	99.1	100	100
	Exp(1)	G $(\frac{3}{2}, 1)$	99.7	99.3	99.9	99.7	100	100	100	100
K^{CL}	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	99.6	99.4	99.9	99.7	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	100	100	100	100	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100
	Exp(1)	Exp $(\frac{3}{2})$	95.9	93.5	97.4	95.0	99.9	99.5	100	99.9
	Exp(1)	G $(\frac{3}{2}, 1)$	99.6	99.4	99.9	99.7	100	100	100	100
K^{N}	Exp(1)	G $(\frac{5}{4}, \frac{5}{4})$	99.4	99.0	99.7	99.6	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	100	100	100	100	100	100	100	100
	$\mathcal{N}(0,1)$	$\mathcal{N}(0, \frac{3}{2})$	100	100	100	100	100	100	100	100

Table 4.4: Estimation, basée sur 10 000 itérations, de la probabilité de rejeter $\mathcal{H}_0$ sous les copules Normale symétrique et asymétrique à $k = 3$ et $k = 4$ dimensions dont les $k-1$ premières marges sont F et la k-ème est $\tilde{F}$ ($n = 100$)

			$\rho = \frac{1}{3}$		$\rho = \frac{2}{3}$	
C	F	$\tilde{F}$	V_n	W_n^{Ψ}	V_n	W_n^{Ψ}
	Exp(1)	Exp(1)	4.4	4.7	3.7	4.0
	Exp(1)	Exp $(\frac{3}{2})$	89.7	83.2	99.2	96.4
C^{N}	Exp(1)	G $(\frac{3}{2}, 1)$	98.9	97.9	100.0	99.9
	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	4.9	5.2	4.3	5.2
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	98.8	98.2	100.0	100.0
	Exp(1)	Exp(1)	4.7	4.7	4.5	4.5
	Exp(1)	Exp $(\frac{3}{2})$	87.5	80.3	96.5	90.9
K^{N}	Exp(1)	G $(\frac{3}{2}, 1)$	98.2	97.0	99.9	99.6
	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	4.6	5.0	5.2	4.8
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	98.1	97.5	99.9	99.7
	Exp(1)	Exp(1)	4.0	4.4	2.9	3.4
	Exp(1)	Exp $(\frac{3}{2})$	90.0	82.4	99.1	96.3
C^{N}	Exp(1)	G $(\frac{3}{2}, 1)$	98.9	97.9	100.0	100.0
	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	4.4	4.8	3.7	4.5
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	99.0	98.4	100.0	100.0
	Exp(1)	Exp(1)	4.3	4.3	3.9	4.3
	Exp(1)	Exp $(\frac{3}{2})$	87.9	81.2	97.2	92.3
K^{N}	Exp(1)	G $(\frac{3}{2}, 1)$	98.5	97.4	100.0	99.7
	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	4.3	4.8	4.5	4.9
	$\mathcal{N}(0,1)$	$\mathcal{N}(\frac{1}{2}, 1)$	98.4	97.7	100.0	99.9

Table 4.5: Tests d'égalité de lois basés sur S_n et T_n^{Ψ} ($\Psi(t) = e^{-t^2}$) pour toutes les paires de {Ca, Fe, Pr, vA, vB}

(X_1, X_2)	S_n	p-valeur	T_n^{Ψ}	p-valeur
(Ca, Fe)	146540	0.00	140,80	0.00
(Ca, Pr)	121690	0.00	27,14	0.00
(Ca, vA)	3780	0.00	1,96	0.11
(Ca, vC)	110120	0.00	17,13	0.00
(Fe, Pr)	13060	0.00	157,34	0.00
(Fe, vA)	142370	0.00	139,34	0.00
(Fe, vC)	10430	0.00	115,97	0.00
(Pr, vA)	118370	0.00	25,58	0.00
(Pr, vC)	1470	0.00	12,72	0.00
(vA, vC)	107400	0.00	15,57	0.00

Table 4.6: Tests d'égalité de lois basés sur S_n et T_n^{Ψ} ($\Psi(t) = e^{-t^2}$) pour toutes les paires du jeu de données de Cook & Johnson (1986)

(X_1, X_2)	S_n	p-valeur	T_n^{Ψ}	p-valeur
(U, Li)	117.0	0	104.6	0
(U, Co)	29.4	0	9.9	0
(U, K)	1015.2	0	1039.9	0
(U, Ce)	281.9	0	310.1	0
(U, Sc)	23.9	0	9.1	0
(U, Ti)	825.4	0	938.6	0
(Li, Co)	91.5	0	60	0
(Li, K)	825.6	0	948.9	0
(Li, Ce)	94.8	0	75.6	0
(Li, Sc)	86.9	0	60.7	0
(Li, Ti)	635.8	0	769.1	0
(Co, K)	998.6	0	1055.8	0
(Co, Ce)	264.7	0	250.2	0
(Co, Sc)	0.6	0	0.0235	0.0440
(Co, Ti)	808.7	0	937.7	0
(K, Ce)	647.5	0	777.2	0
(K, Sc)	993	0	1052.1	0
(K, Ti)	125.7	0	81.6	0
(Ce, Sc)	259.1	0	250.5	0
(Ce, Ti)	457.6	0	534.9	0
(Sc, Ti)	803.1	0	934.3	0

CHAPITRE 5

TESTS OF EQUALITY OF DISTRIBUTIONS UP TO LOCATION AND SCALE FACTORS

5.1 Introduction

Testing for the equality in distribution of two real-valued random variables X and Y has been extensively investigated in the statistical literature. The most popular procedures are original or variants of the sign and Wilcoxon rank-sum tests, Kolmogorov–Smirnov and Cramér–von Mises functional distances between empirical distribution functions, see (Anderson (1962)) or characteristic function tests (Alba Fernández et al. (2008); Epps & Singleton (1986)). These tests are particularly powerful when the two distributions differ only by location and / or a scale factors.

In some situations, one may be interested in testing if two random variables have the same distribution *up to* an affine transformation. For example, it may be expected that the distribution of the salaries of male and female

in a given country are *similar*, up to a translation and / or a change of scale. In other words, the null and alternative hypotheses of interest could be formulated as

$$\mathcal{H}_0 : \frac{X - \mu_X}{\sigma_X} \stackrel{d}{=} \frac{Y - \mu_Y}{\sigma_Y} \quad \text{vs} \quad \mathcal{H}_1 : \frac{X - \mu_X}{\sigma_X} \stackrel{d}{\neq} \frac{Y - \mu_Y}{\sigma_Y},$$

where $\mu_X = \mathrm{E}(X)$, $\sigma_X^2 = \mathrm{var}(X)$, $\mu_Y = \mathrm{E}(Y)$ and $\sigma_Y^2 = \mathrm{var}(Y)$. Of course, the problem reduces to the testing of the classical hypotheses $\mathcal{H}_0 : \tilde{F} = \tilde{G}$ and $\mathcal{H}_1 : \tilde{F} \neq \tilde{G}$ in the case when the parameters are known, where

$$\tilde{F}(x) = \mathrm{P}\left(\frac{X - \mu_X}{\sigma_X} \leq x\right) \quad \text{and} \quad \tilde{G}(y) = \mathrm{P}\left(\frac{Y - \mu_Y}{\sigma_Y} \leq y\right).$$

Here, an empirical process approach based on the empirically standardized observations is adopted. Even if the hypotheses are very simple and the approach is quite natural, the fact that the means and variances are unknown renders this problem challenging. Indeed, one must be very careful in the computation of p-values, since a naive approach based on a classical bootstrap of the standardized observations would fail.

In this paper, two strategies for the computation of p-values are proposed and validated. The latter will prove to be useful for the current problem, but could also but exploited in situations where empirically modified observations are considered, *e.g.* residuals of linear models. Hence, they are of an independent interest. Moreover, the case of paired samples, as well as a non-trivial extension to the situation of $\mathbb{R}^d$-valued random vectors, are fully investigated.

Note that the problem treated herein can be seen as a special case of Pardo-

Fernández (2007), where it is supposed that

$$X = \mu_X(R_X) + \sigma_X(R_X)\epsilon_X \quad \text{and} \quad Y = \mu_Y(R_Y) + \sigma_Y(R_Y)\epsilon_Y,$$

where R_X, R_Y are regressors and $\mu_X(\cdot)$, $\sigma_X(\cdot)$, $\mu_Y(\cdot)$, $\sigma_Y(\cdot)$ are functions that have to be estimated. They considered testing for the equality in distribution of ϵ_X and ϵ_Y. Here, the fact that a simplified version of the model is considered enables to obtain tractable expressions for the limiting empirical process and easier-to-implement re-sampling procedures.

The article is structured as follows. In section 5.2, the necessary asymptotic results are obtained in order to establish the weak convergence of the proposed Cramér–von Mises test statistics; both the cases of independent and paired samples are studied. In section 5.3, two methods for the computation of asymptotically valid p-values are described. In section 5.4, the methodologies are extended to the case of multivariate observations. Section 5.5 is devoted to the numerical investigation of the power of the tests. All the proofs and explicit computations of the test statistics and their bootstrapped versions are relegated to Appendix B.

5.2 Test statistics and asymptotic distributions

As stated in the Introduction, suppose it is desired to test for the equality in distribution of the standardized random variables

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X} \quad \text{and} \quad \tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y}. \tag{5.1}$$

The null and alternative hypotheses associated to this problem can then be stated as $\mathcal{H}_0 : \tilde{F} = \tilde{G}$ and $\mathcal{H}_1 : \tilde{F} \neq \tilde{G}$, where $\tilde{F}(x) = \mathrm{P}(\tilde{X} \leq x)$ and $\tilde{G}(y) = \mathrm{P}(\tilde{Y} \leq y)$. The test statistics that will be introduced in this section are based on natural empirical versions of $\tilde{F}$ and $\tilde{G}$. Since the parameters appearing in Equation (5.1) are generally unknown, however, these empirical functions will be based on *pseudo-observations*; as a consequence, the large-sample results are not obtainable from classical methods. Such situations appear, for example, in residual-based goodness-of-fit statistics (Bai (1994); Loynes (1980); Pierce & Kopecky (1979)) and tests of symmetry (Bai & Ng (2001)); see also Ghoudi & Rémillard (1998); Ghoudi & Rémillard (2004) and van der Vaart & Wellner (2007) for a general treatment of empirical processes based on pseudo-observations.

In what follows, the cases of independent and paired samples are treated separately, even if they share many similarities.

5.2.1 Independent samples

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent samples of $\mathbb{R}$-valued observations from two populations with respective continuous cumulative distribution functions F and G. Denote by $\bar{X}$ and S_X (resp. $\bar{Y}$ and S_Y) the usual empirical mean and standard deviation of $X_1, \ldots, X_n$ (resp. $Y_1, \ldots, Y_m$). Consider the samples of empirically standardized observations $X_{1,n}, \ldots, X_{n,n}$ and $Y_{1,m}, \ldots, X_{m,m}$, where

$$X_{i,n} = \frac{X_i - \bar{X}}{S_X} \quad \text{and} \quad Y_{i,m} = \frac{Y_i - \bar{Y}}{S_Y}\,.$$

The statistical methodologies for $\mathcal{H}_0$ and $\mathcal{H}_1$ developed herein will be based on the empirical distribution functions of these transformed observations, namely

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(X_{i,n} \leq x\right) \quad \text{and} \quad \tilde{G}_m(y) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(Y_{i,m} \leq y\right).$$

A possible measure of discrepancy between $\tilde{F}$ and $\tilde{G}$ is the Cramér–von Mises functional, which leads to the test statistic

$$T_{n,m} = \frac{nm}{n+m}\int_{\mathbb{R}} \left\{\tilde{F}_n(x) - \tilde{G}_m(x)\right\}^2 \mathrm{d}x.$$

Another possible choice is the Kolmogorov–Smirnov distance, but the latter generally yields less powerful tests; moreover, it proves to be harder to compute, especially its bootstrapped versions. Indeed, letting $a \vee b = \max(a, b)$, one can show that $T_{n,m}$ admits the tractable expression

$$\begin{aligned} T_{n,m} \;=\; & \frac{2}{n+m}\sum_{i=1}^{n}\sum_{j=1}^{m} \max\left(X_{i,n}, Y_{j,m}\right) \\ & - \frac{m}{n(n+m)}\sum_{i=1}^{n}\sum_{j=1}^{n} \max\left(X_{i,n}, X_{j,n}\right) \\ & - \frac{n}{m(n+m)}\sum_{i=1}^{m}\sum_{j=1}^{m} \max\left(Y_{i,m}, Y_{j,m}\right). \end{aligned} \tag{5.2}$$

In order to obtain the weak convergence of $T_{n,m}$, first note that

$$T_{n,m} = \int_{\mathbb{R}} \left\{\mathbb{D}_{n,m}(x)\right\}^2 \mathrm{d}x,$$

where

$$\mathbb{D}_{n,m}(x) = \sqrt{\frac{nm}{n+m}}\left\{\tilde{F}_n(x) - \tilde{G}_m(x)\right\}.$$

Clearly, the asymptotic distribution of $T_{n,m}$ under the null hypothesis will be a consequence of the large-sample behavior of $\mathbb{D}_{n,m}$. This is the subject of the next proposition, whose proof is to be found in Appendix B. Here and in the sequel, F_0 refers to the common distribution function of $\tilde{X}$ and $\tilde{Y}$ under $\mathcal{H}_0$, and $f_0 = \mathrm{d}F_0$ to its associated density, for which it is assumed that

$$\sup_{|(a-1,b)|<\epsilon} \sup_{x\in\mathbb{R}} |f_0(ax+b) - f_0(x)| \text{ and } \sup_{|(a-1,b)|<\epsilon} \sup_{x\in\mathbb{R}} |x f_0(ax+b) - x f_0(x)| \quad (5.3)$$

converge to zero as $\epsilon \to 0$. Also, let $\tilde{\mu}_0^j$ be the j-th moment of f_0 and

$$\tilde{\Sigma}_0 = \begin{pmatrix} \dfrac{\tilde{\mu}_0^4 - (\tilde{\mu}_0^2)^2}{4} & \dfrac{\tilde{\mu}_0^3}{2} \\ \dfrac{\tilde{\mu}_0^3}{2} & 1 \end{pmatrix}.$$

The latter corresponds to the asymptotic covariance matrix of $(Z_{n,X,1}, Z_{n,X,2})$ and $(Z_{n,Y,1}, Z_{n,Y,2})$ under $\mathcal{H}_0$, where

$$Z_{n,X,1} = \sqrt{n}\left(\frac{S_X}{\sigma_X} - 1\right), \quad Z_{n,X,2} = \sqrt{n}\left(\frac{\bar{X} - \mu_X}{\sigma_X}\right)$$
$$Z_{m,Y,1} = \sqrt{m}\left(\frac{S_Y}{\sigma_Y} - 1\right), \quad Z_{m,Y,2} = \sqrt{m}\left(\frac{\bar{Y} - \mu_Y}{\sigma_Y}\right).$$

Proposition 5.1. *Suppose $n/(n+m) \to \lambda \in (0,1)$ as $n, m \to \infty$. Then, if (5.3) holds and the fourth moment of f_0 exists, the empirical process $\mathbb{D}_{n,m}$ converges weakly, under $\mathcal{H}_0$, to*

$$\mathbb{D}(x) = \sqrt{1-\lambda}\,\{\mathbb{B}_X(x) + f_0(x)\Theta_X(x)\} - \sqrt{\lambda}\,\{\mathbb{B}_Y(x) + f_0(x)\Theta_Y(x)\}.$$

Here, $\mathbb{B}_X$ and $\mathbb{B}_Y$ are independent F_0-Brownian bridges and $\Theta_X(x) = Z_{X,1}\,x + Z_{X,2}$, $\Theta_Y(x) = Z_{Y,1}\,x + Z_{Y,2}$, where $(Z_{X,1}, Z_{X,2}), (Z_{Y,1}, Z_{Y,2}) \sim \mathcal{N}_2(\mathbf{0}, \tilde{\Sigma}_0)$.

If the parameters in (5.1) were known, the limit in Proposition 5.1 would have been of the form $\mathbb{D} = \sqrt{1-\lambda}\,\mathbb{B}_X - \sqrt{\lambda}\,\mathbb{B}_Y$. Hence, $f_0(x)\{\sqrt{1-\lambda}\,\Theta_X(x) - \sqrt{\lambda}\,\Theta_Y(x)\}$ is the *price to pay* for the lack of information about the first two moments. Observe also that since $\tilde{X}, \tilde{Y} \sim f_0$ under $\mathcal{H}_0$, the limiting process $\mathbb{D}$ depends only on F_0.

A simple application of the continuous mapping theorem combined with Proposition 5.1 entails that as $n, m \to \infty$,

$$T_{n,m} \rightsquigarrow T = \int_{\mathbb{R}} \{\mathbb{D}(x)\}^2 \, \mathrm{d}x.$$

The distribution of T under $\mathcal{H}_0$, however, depends on F_0, which is unknown. Hence, suitable re-sampling methods must be employed in order to conceive appropriate testing procedures. It will be the subject of Section 5.3, where two strategies based on the *multiplier central limit theorem* are proposed.

5.2.2 Paired samples

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a given joint cumulative distribution function H with continuous marginal distributions F and G. It is now well established that there exists a unique copula $C : [0,1]^d \to [0,1]$ such that $H(x,y) = C\{F(x), G(y)\}$. This representation has been discovered by Sklar (1959) in the theoretical framework of probabilistic metric spaces. Statistical applications of this result has been subject to intense researches over the last fifteen years or so, especially in estimation of copula parameters (Genest et al. (1995); Shih & Louis (1995)), goodness-of-fit testing (Fermanian (2005); Genest et al. (2006, 2009)) and composite hypotheses Rémillard

& Scaillet (2009); Kojadinovic & Yan (2010); Quessy (2011)). For more details on the foundations of copulas, the reader is referred to Nelsen (2006).

If $\tilde{H}$ denotes the joint distribution of $(\tilde{X}, \tilde{Y})$, then one obtains easily that $\tilde{H}(x, y) = C\{\tilde{F}(x), \tilde{G}(y)\}$; this is a consequence of the fact that the copula of a random couple is invariant under monotone increasing transformations of its components. Consider its sample version based on the joint empirical distribution function of the sample of standardized observations $(X_{1,n}, Y_{1,n})$, $\ldots$, $(X_{n,n}, Y_{n,n})$, *i.e.*

$$\tilde{H}_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(X_{i,n} \leq x, Y_{i,n} \leq y\right).$$

The weak convergence of the empirical process

$$\tilde{\mathbb{H}}_n(x, y) = \sqrt{n} \left\{ \tilde{H}_n(x, y) - \tilde{H}(x, y) \right\}$$

toward a centered Gaussian limit $\mathbb{H}$ is first obtained in the general case. The proof is deferred to Appendix B. As a by-product of this result, one deduces that $\tilde{H}_n$ is a uniformly consistent estimator of $\tilde{H}$. Before stating it, let

$$\tilde{\Sigma}_{X,Y} = \begin{pmatrix} \mathrm{E}(\tilde{X}\tilde{Y}) & \frac{1}{2}\mathrm{E}(\tilde{X}\tilde{Y}^2) \\ \\ \frac{1}{2}\mathrm{E}(\tilde{X}^2\tilde{Y}) & \frac{1}{4}\mathrm{E}(\tilde{X}^2\tilde{Y}^2) \end{pmatrix}.$$

Letting $\tilde{H}_{10}(x, y) = \partial \tilde{H}(x, y)/\partial x$ and $\tilde{H}_{01}(x, y) = \partial \tilde{H}(x, y)/\partial y$, it is also

assumed that

$$\begin{aligned}
&\sup_{S_\epsilon}\sup_{(x,y)\in\mathbb{R}^2}\left|\tilde{H}_{10}\left(a_X x+b_X, a_Y y+b_Y\right)-\tilde{H}_{10}(x,y)\right|\to 0,\\
&\sup_{S_\epsilon}\sup_{(x,y)\in\mathbb{R}^2}\left|x\tilde{H}_{10}\left(a_X x+b_X, a_Y y+b_Y\right)-x\tilde{H}_{10}(x,y)\right|\to 0,\\
&\sup_{S_\epsilon}\sup_{(x,y)\in\mathbb{R}^2}\left|\tilde{H}_{01}\left(a_X x+b_X, a_Y y+b_Y\right)-\tilde{H}_{01}(x,y)\right|\to 0,\\
&\sup_{S_\epsilon}\sup_{(x,y)\in\mathbb{R}^2}\left|y\tilde{H}_{01}\left(a_X x+b_X, a_Y y+b_Y\right)-y\tilde{H}_{01}(x,y)\right|\to 0 \qquad (5.4)
\end{aligned}$$

as $\epsilon \to 0$, where $S_\epsilon = \{(a_X, b_X, a_Y, b_Y) : |(a_x - 1, b_x, a_Y - 1, b_Y)| < \epsilon\}$. Exploiting Sklar's representation of $\tilde{H}$, these conditions could also be stated as assumptions on its underlying copula C and on its marginal distributions.

Proposition 5.2. *Let H be a bivariate distribution function with finite fourth moments, absolutely continuous marginal distributions and such that the associated normalized version $\tilde{H}$ satisfies* (5.4)*. Then*

$$\tilde{\mathbb{H}}_n(x,y) \rightsquigarrow \tilde{\mathbb{H}}(x,y) \;=\; \mathbb{B}(x,y) + \tilde{H}_{10}(x,y)\Theta_X(x) + \tilde{H}_{01}(x,y)\Theta_Y(y),$$

where $\mathbb{B}$ is a $\tilde{H}$-Brownian sheet, $\Theta_X(x) = Z_{X,1}\,x + Z_{X,2}$ and $\Theta_Y(y) = Z_{Y,1}\,y + Z_{Y,2}$, with

$$(Z_{X,1}, Z_{X,2}, Z_{Y,1}, Z_{Y,2}) \sim \mathcal{N}\left(\mathbf{0}, \tilde{\Sigma}\right), \quad \tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_X & \tilde{\Sigma}_{X,Y} \\ \tilde{\Sigma}_{X,Y} & \tilde{\Sigma}_Y \end{pmatrix}.$$

Since the null hypothesis can be stated as $\mathcal{H}_0 : \tilde{H}(x,\infty) = \tilde{H}(\infty,x)$, natural test statistics can be build from the empirical process

$$\mathbb{E}_n(x) = \sqrt{n}\left\{\tilde{H}_n(x,\infty) - \tilde{H}_n(\infty,x)\right\}.$$

Its asymptotic behavior is given in the next proposition.

Proposition 5.3. *Let H be a bivariate distribution function with finite fourth moments and such that the associated normalized version $\tilde{H}$ satisfies* (5.4). *Then, if $\tilde{H}(x,y) = C\{F_0(x), F_0(y)\}$,*

$$\mathbb{E}_n(x) \rightsquigarrow \mathbb{E}(x) = \mathbb{B}(x,\infty) - \mathbb{B}(\infty,x) + f_0(x)\left\{\Theta_X(x) - \Theta_Y(x)\right\},$$

where $\mathbb{B}$ is a $\tilde{H}$-Brownian sheet, $f_0 = \mathrm{d}F_0$, $\Theta_X(x) = Z_{X,1}\,x + Z_{X,2}$ and $\Theta_Y(y) = Z_{Y,1}\,y + Z_{Y,2}$, with

$$(Z_{X,1}, Z_{X,2}, Z_{Y,1}, Z_{Y,2}) \sim \mathcal{N}\left(\mathbf{0}, \tilde{\Sigma}\right), \quad \tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_0 & \tilde{\Sigma}_{X,Y} \\ \tilde{\Sigma}_{X,Y} & \tilde{\Sigma}_0 \end{pmatrix}.$$

As in the case of independent samples, the process $\mathbb{E}$ depends on the marginal distributions through the common law F_0 of the standardized random variables; it also depends on the unknown underlying copula C of $\tilde{H}$.

The proposed test statistic will be the Cramér–von Mises functional computed from $\mathbb{E}_n$. As a consequence of Proposition 5.3 and of the continuous mapping theorem, one has

$$T_n = \int_{\mathbb{R}} \left\{\mathbb{E}_n(x)\right\}^2 \mathrm{d}x \rightsquigarrow T' = \int_{\mathbb{R}} \left\{\mathbb{E}(x)\right\}^2 \mathrm{d}x.$$

For practical purposes, note that the statistic T_n is equivalent to $2\,T_{n,m}$ in equation (5.2) with $m = n$.

5.3 Computation of p-values

5.3.1 Preliminaries

Let F_n and G_m be the sample distribution functions based on the original observations $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively. Donsker's theorem (see Billingsley (1999), for instance) ensures that $\mathbb{F}_n = \sqrt{n}(F_n - F) \rightsquigarrow \mathbb{F}$ and $\mathbb{G}_m = \sqrt{m}(G_m - G) \rightsquigarrow \mathbb{G}$, where $\mathbb{F}$ and $\mathbb{G}$ are F- and G-Brownian bridges, respectively. The multiplier central limit theorem for empirical processes enables to generate asymptotically independent copies of these limiting processes. To describe the idea, let $(\xi_1^{(h)}, \ldots, \xi_n^{(h)})$ and $(\kappa_1^{(h)}, \ldots, \kappa_m^{(h)})$, $h \in \{1, \ldots, M\}$, be independent vectors of i.i.d. random variables with unit mean and variance. Following van der Vaart & Wellner (1996) and Kosorok (2008), multiplier versions of $(\mathbb{F}_n, \mathbb{G}_m)$ are $(\mathbb{F}_n^{(1)}, \mathbb{G}_m^{(1)}), \ldots, (\mathbb{F}_n^{(M)}, \mathbb{G}_m^{(M)})$, where

$$\begin{aligned} \mathbb{F}_n^{(h)}(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\tilde{\xi}_i^{(h)} - 1 \right) \mathbb{I}(X_i \leq x), \\ \mathbb{G}_m^{(h)}(x) &= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left(\tilde{\kappa}_i^{(h)} - 1 \right) \mathbb{I}(Y_i \leq x). \end{aligned}$$

Here, $\tilde{\xi}_i^{(h)} = \xi_i^{(h)} / \bar{\xi}^{(h)}$, $i \in \{1, \ldots, n\}$ and $\tilde{\kappa}_i^{(h)} = \kappa_i^{(h)} / \bar{\kappa}^{(h)}$, $i \in \{1, \ldots, m\}$, where $\bar{\xi}^{(h)} = \sum_{i=1}^{n} \xi_i^{(h)} / n$ and $\bar{\kappa}^{(h)} = \sum_{i=1}^{m} \kappa_i^{(h)} / m$. This way of re-scaling the multiplier variables is called the Bayesian bootstrap in Kosorok (2008). Formally, one can invoke the multiplier central limit theorem to establish that the vector

$$\left((\mathbb{F}_n, \mathbb{G}_m), (\mathbb{F}_n^{(1)}, \mathbb{G}_m^{(1)}), \ldots, (\mathbb{F}_n^{(M)}, \mathbb{G}_m^{(M)}) \right)$$

converges weakly to

$$\left((\mathbb{F},\mathbb{G}),(\mathbb{F}^{(1)},\mathbb{G}^{(1)}),\ldots,(\mathbb{F}^{(M)},\mathbb{G}^{(M)})\right),$$

where $(\mathbb{F}^{(1)},\mathbb{G}^{(1)}),\ldots,(\mathbb{F}^{(M)},\mathbb{G}^{(M)})$ are independent copies of $(\mathbb{F},\mathbb{G})$. As a consequence, it is possible to replicate, at least asymptotically, the behavior of statistics based on functionals of $(\mathbb{F}_n,\mathbb{G}_m)$.

As a particular case of the multiplier method for empirical processes, one deduces a formula for the sample mean. Indeed, from the fact that $Z_{n,X,2} = \sqrt{n}(\bar{X}-\mu_X)/\sigma_X = -(1/\sigma_X)\int_{\mathbb{R}}\mathbb{F}_n(x)\mathrm{d}x$, one defines, for $h\in\{1,\ldots,M\}$,

$$\begin{aligned} Z_{n,X,2}^{(h)} = -\frac{1}{S_X}\int_{\mathbb{R}}\mathbb{F}_n^{(h)}(x)\mathrm{d}x &= \frac{1}{S_X}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{\xi}_i^{(h)}-1\right)X_i \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{\xi}_i^{(h)}-1\right)X_{i,n}. \end{aligned}$$

The continuous mapping theorem and Slutsky's lemma ensure that $Z_{n,X,2}^{(1)}$, ..., $Z_{n,X,2}^{(M)}$ are independent copies, asymptotically, of $Z_{n,X,2}$. One proceeds similarly in order to obtain multiplier versions $Z_{m,Y,2}^{(1)},\ldots,Z_{m,Y,2}^{(M)}$ of $Z_{m,Y,2}$.

For the sample variances, an application of the Delta-method in $\mathbb{R}$ (see van der Vaart (1998), for instance) yields

$$\begin{aligned} \sqrt{n}\left(S_X^2-\sigma_X^2\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{(X_i-\mu_X)^2-\sigma_X^2\right\}+o_{\mathrm{P}}(1) \\ &= \int_{\mathbb{R}}(x-\mu_X)^2\,\mathrm{d}\mathbb{F}_n(x)+o_{\mathrm{P}}(1) \\ &= -2\int_{\mathbb{R}}(x-\mu_X)\,\mathbb{F}_n(x)\mathrm{d}x+o_{\mathrm{P}}(1), \end{aligned}$$

where the last equality arises from a simple integration by parts. This rep-

resentation suggests the multiplier versions

$$\begin{aligned}\sqrt{n}\left(S_X^2-\sigma_X^2\right)^{(h)} &= -2\int_{\mathbb{R}}\left(x-\bar{X}\right)\mathbb{F}_n^{(h)}(x)\mathrm{d}x \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{\xi}_i^{(h)}-1\right)\left(X_i-\bar{X}\right)^2.\end{aligned}$$

Another application of the Delta-method enables to write

$$Z_{n,X,1}=\frac{1}{\sigma_X}\sqrt{n}\left(\sqrt{S_X^2}-\sqrt{\sigma_X^2}\right)=\frac{1}{2\sigma_X^2}\sqrt{n}\left(S_X^2-\sigma_X^2\right)+o_{\mathrm{P}}(1),$$

so that multiplier versions of $Z_{n,X,1}$ are given by

$$Z_{n,X,1}^{(h)}=\frac{1}{2S_X^2}\sqrt{n}\left(S_X^2-\sigma_X^2\right)^{(h)}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{\xi}_i^{(h)}-1\right)\frac{X_{i,n}^2}{2},$$

and similarly for the multiplier versions $Z_{m,Y,1}^{(1)},\ldots,Z_{m,Y,1}^{(M)}$ of $Z_{m,Y,1}$.

In the next two subsections, multiplier versions of the empirical processes $\mathbb{D}_{n,m}$ and $\mathbb{E}_n$ will be described and formally justified. The problem has not a straightforward solution since the empirical measures are not based on the original observations, but on empirically standardized versions. Two strategies will be developed : the first exploits the limiting representation of the processes encountered in Propositions 5.1 and 5.3; the second is rooted in an application of the functional Delta-method.

5.3.2 Strategy I: exploiting the form of the limits of $\mathbb{D}_{n,m}$ and $\mathbb{E}_n$

The idea will first be described for the case of independent samples, *i.e.* bootstrapped versions of $\mathbb{D}_{n,m}$ will be considered. Only slight modifications are required for $\mathbb{E}_n$ in the case of paired samples.

It appears from the proof of Proposition 5.1 that $\mathbb{B}_X$ and $\mathbb{B}_Y$ are the limits, respectively, of $\mathbb{F}_n^\star = \sqrt{n}(F_n^\star - \tilde{F})$ and $\mathbb{G}_m^\star = \sqrt{m}(G_m^\star - \tilde{G})$, where

$$F_n^\star(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\frac{X_i - \mu_X}{\sigma_X} \le x\right) \quad \text{and} \quad G_m^\star(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(\frac{Y_i - \mu_Y}{\sigma_Y} \le x\right).$$

Since $\mathbb{F}_n^\star(x) = \mathbb{F}_n(\sigma_X\, x + \mu_X)$ and $G_m^\star(x) = \mathbb{G}_m(\sigma_Y\, x + \mu_Y)$, it seems natural to define the multiplier versions of $\mathbb{B}_X$ and $\mathbb{B}_Y$, for $h \in \{1, \ldots, M\}$, by

$$\begin{aligned}
\mathbb{F}_n^{(h)}\left(S_X\, x + \bar{X}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{\xi}_i^{(h)} - 1\right)\mathbb{I}\left(X_{i,n} \le x\right), \\
\mathbb{G}_m^{(h)}\left(S_Y\, x + \bar{Y}\right) &= \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left(\tilde{\kappa}_i^{(h)} - 1\right)\mathbb{I}\left(Y_{i,n} \le x\right).
\end{aligned}$$

Next, note that $\Theta_X(x)$ and $\Theta_Y(x)$ are the limits, respectively, of $\Theta_{n,X}(x) = Z_{n,X,1}\, x + Z_{n,X,2}$ and $\Theta_{m,Y}(x) = Z_{m,Y,1}\, x + Z_{m,Y,2}$. Their multiplier versions are then defined by

$$\begin{aligned}
\Theta_{n,X}^{(h)}(x) &= Z_{n,X,1}^{(h)}\, x + Z_{n,X,2}^{(h)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{\xi}_i^{(h)} - 1\right)\left(\frac{X_{i,n}^2}{2}\, x + X_{i,n}\right), \\
\Theta_{m,Y}^{(h)}(x) &= Z_{m,Y,1}^{(h)}\, x + Z_{m,Y,2}^{(h)} = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left(\tilde{\kappa}_i^{(h)} - 1\right)\left(\frac{Y_{i,m}^2}{2}\, x + Y_{i,m}\right).
\end{aligned}$$

Finally, for each $h \in \{1, \ldots, M\}$, let

$$\begin{aligned}
\mathbb{D}_{n,m}^{(h)}(x) &= \sqrt{1 - \lambda_{n,m}}\left\{\mathbb{F}_n^{(h)}\left(S_X\, x + \bar{X}\right) + \hat{f}_0(x)\Theta_{n,X}^{(h)}(x)\right\} \\
&\quad - \sqrt{\lambda_{n,m}}\left\{\mathbb{G}_m^{(h)}\left(S_Y\, x + \bar{Y}\right) + \hat{f}_0(x)\Theta_{m,Y}^{(h)}(x)\right\},
\end{aligned}$$

where $\lambda_{n,m} = n/(n+m)$ and $\hat{f}_0$ is some uniformly consistent nonparametric estimator of f_0. Such an estimator is described in 5.7.1 in the case when the support of F and G is the whole real line. The next proposition establishes that $\mathbb{D}_{n,m}^{(1)}, \ldots, \mathbb{D}_{n,m}^{(M)}$ are valid multiplier versions of $\mathbb{D}_{n,m}$. The proof is deferred to Appendix B.

Proposition 5.4. *Under the conditions of Proposition 5.1, one has under* $\mathcal{H}_0$ *that*

$$\left(\mathbb{D}_{n,m}, \mathbb{D}_{n,m}^{(1)}, \ldots, \mathbb{D}_{n,m}^{(M)}\right) \rightsquigarrow \left(\mathbb{D}, \mathbb{D}^{(1)}, \ldots, \mathbb{D}^{(M)}\right),$$

where $\mathbb{D}^{(1)}, \ldots, \mathbb{D}^{(M)}$ *are independent copies of* $\mathbb{D}$.

In order to obtain asymptotically independent copies of T, simply define

$$T_{n,m}^{(h)} = \int_{\mathbb{R}} \left\{\mathbb{D}_{n,m}^{(h)}(x)\right\}^2 \mathrm{d}x, \quad h \in \{1, \ldots, M\}.$$

From Proposition 5.4 and the continuous mapping theorem,

$$(T_{n,m}, T_{n,m}^{(1)}, \ldots, T_{n,m}^{(M)}) \rightsquigarrow (T', T'^{,(1)}, \ldots, T'^{,(M)}),$$

where $T'^{,(1)}, \ldots, T'^{,(M)}$ are independent copies of T'. Simple formulas for $T_{n,m}^{(h)}$ are possible by first writing

$$\mathbb{D}_{n,m}^{(h)}(x) = \sqrt{\frac{1-\lambda_{n,m}}{n}}\, \xi^{(h)} A(x) - \sqrt{\frac{\lambda_{n,m}}{\sqrt{m}}}\, \kappa^{(h)} B(x),$$

where

$$\xi^{(h)} = \left(\tilde{\xi}_1^{(h)} - 1, \ldots, \tilde{\xi}_n^{(h)} - 1\right), \quad \kappa^{(h)} = \left(\tilde{\kappa}_1^{(h)} - 1, \ldots, \tilde{\kappa}_n^{(h)} - 1\right),$$

and for each $x \in \mathbb{R}$, $A(x) \in \mathbb{R}^{n\times 1}$ and $B(x) \in \mathbb{R}^{m\times 1}$ with

$$\begin{aligned} A_i(x) &= \mathbb{I}\left(X_{i,n} \leq x\right) + x\hat{f}_0(x)\frac{X_{i,n}^2}{2} + \hat{f}_0(x) X_{i,n}, \\ B_i(x) &= \mathbb{I}\left(Y_{i,m} \leq x\right) + x\hat{f}_0(x)\frac{Y_{i,m}^2}{2} + \hat{f}_0(x) Y_{i,m}. \end{aligned}$$

One can then show that

$$\begin{aligned} T_{n,m}^{(h)} &= \left(\frac{1-\lambda_{n,m}}{n}\right) \xi^{(h)} \mathbf{A} \left(\xi^{(h)}\right)^\top + \frac{\lambda_{n,m}}{m}\, \kappa^{(h)} \mathbf{B} \left(\kappa^{(h)}\right)^\top \\ &\quad - 2\sqrt{\frac{\lambda_{n,m}(1-\lambda_{n,m})}{nm}}\, \xi^{(h)} \mathbf{C} \left(\kappa^{(h)}\right)^\top, \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{m\times m}$, $\mathbf{C} \in \mathbb{R}^{n\times m}$ with

$$\mathbf{A}_{ij} = \int_{\mathbb{R}} A_i(x)A_j(x)\mathrm{d}x, \quad \mathbf{B}_{ij} = \int_{\mathbb{R}} B_i(x)B_j(x)\mathrm{d}x, \quad \mathbf{C}_{ij} = \int_{\mathbb{R}} A_i(x)B_j(x)\mathrm{d}x.$$

The entries of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are computed in Appendix B in the case when f_0 is estimated by the simple estimator described in Appendix B.

For paired samples, the main difference is the necessity to replicate the joint process $\mathbb{H}_n$ with

$$\mathbb{H}_n^{(h)}(x,y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\tilde{\xi}_i^{(h)} - 1\right) \mathbb{I}\left(X_i \le x, Y_i \le y\right), \quad h \in \{1,\dots,M\}.$$

Practically, it suffices to put $n = m$ and $\kappa_i^{(h)} = \xi_i^{(h)}$ in the formulas for independent samples. The multiplier versions of $\mathbb{E}_n$ are then given by $\mathbb{E}_n^{(h)}(x) = \xi R(x)/\sqrt{n}$, $h \in \{1,\dots,M\}$, where for each $x \in \mathbb{R}$, $R(x) \in \mathbb{R}^{n\times 1}$ with

$$\begin{aligned} R_i(x) &= I\left(X_{i,n} \le x\right) - \mathbb{I}\left(Y_{i,n} \le x\right) \\ &\quad + x\hat{f}_0(x)\left(\frac{X_{i,n}^2}{2} - \frac{Y_{i,n}^2}{2}\right) + \hat{f}_0(x)\left(X_{i,n} - Y_{i,n}\right). \end{aligned}$$

Thus,

$$T_n^{(h)} = \int_{\mathbb{R}} \left\{\mathbb{E}_n^{(h)}(x)\right\}^2 \mathrm{d}x = \frac{1}{n}\,\xi^{(h)}\,\mathbf{R}\left(\xi^{(h)}\right)^{\top},$$

where $\mathbf{R} \in \mathbb{R}^{n\times n}$ with

$$\mathbf{R}_{ij} = \int_{\mathbb{R}} R_i(x)R_j(x)\mathrm{d}x.$$

The asymptotic validity of the method is stated in the next proposition. The proof, which is essentially the same as that of Proposition 5.4, is omitted.

Proposition 5.5. *Under* $\mathcal{H}_0$,

$$\left(\mathbb{E}_n, \mathbb{E}_n^{(1)}, \dots, \mathbb{E}_n^{(M)}\right) \rightsquigarrow \left(\mathbb{E}, \mathbb{E}^{(1)}, \dots, \mathbb{E}^{(M)}\right),$$

where $\mathbb{E}^{(1)}, \dots, \mathbb{E}^{(M)}$ *are independent copies of* $\mathbb{E}$.

5.3.3 Strategy II: application of the functional Delta-method

The conclusion of Proposition 5.1 could have been obtained by the theory of Hadamard derivative. This approach will enable to justify the following bootstrap procedure. To this end, consider the functional

$$\mathcal{L}(F, \theta) = F\{\theta(x)\},$$

where $(F, \theta) \in D(\mathbb{R}) \times \mathcal{S}$, with $\mathcal{S} = \{sx + m; x \in \mathbb{R}, s \in \mathbb{R}^+, m \in \mathbb{R}\}$. The space $D(\mathbb{R})$ is endowed by the sup norm, while the norm on $\mathcal{S}$ is defined by

$$||\theta||_{\mathcal{S}} = \sup_{x \in \mathbb{R}} \left| \frac{\theta(x)}{1 + |x|} \right|.$$

Hence, if $\theta_j(x) = s_j\, x + m_j$, $j = 1, 2$,

$$||\theta_1 - \theta_2||_{\mathcal{S}} \leq |s_1 - s_2| \sup_{x \in \mathbb{R}} \left| \frac{x}{1 + |x|} \right| + |m_1 - m_2| \leq |s_1 - s_2| + |m_1 - m_2|.$$

If F is absolutely continuous, one can show that $\mathcal{L}$ is Hadamard differentiable with derivative at (F, θ) given by

$$\mathcal{L}'_{F,\theta}(\Delta, \delta) = \Delta\{\theta(x)\} + \delta(x) F'\{\theta(x)\}.$$

With this notation, one can write

$$\mathbb{D}_{n,m}(x) = \sqrt{\frac{nm}{n+m}} \left\{ \mathcal{L}\left(F_n, \tilde{\theta}_{n,X}\right) - \mathcal{L}\left(G_m, \tilde{\theta}_{m,Y}\right) \right\}, \tag{5.5}$$

where $\tilde{\theta}_{n,X}(x) = S_X\, x + \bar{X}$ and $\tilde{\theta}_{m,Y}(x) = S_Y\, x + \bar{Y}$. Letting $\tilde{\theta}_X(x) = \sigma_X\, x + \mu_X$ and $\tilde{\theta}_Y(x) = \sigma_Y\, x + \mu_Y$, the Hadamard differentiability of $\mathcal{L}$

entails that

$$\begin{aligned}\mathcal{L}(F_n, \tilde{\theta}_{n,X}) - \mathcal{L}(F, \tilde{\theta}_X) &= \mathcal{L}'_{F,\tilde{\theta}_X}\left(\frac{\mathbb{F}_n}{\sqrt{n}}, \frac{\tilde{\Theta}_{n,X}}{\sqrt{n}}\right) + o_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\right),\\ \mathcal{L}(G_m, \tilde{\theta}_{m,Y}) - \mathcal{L}(G, \tilde{\theta}_Y) &= \mathcal{L}'_{G,\tilde{\theta}_Y}\left(\frac{\mathbb{G}_m}{\sqrt{m}}, \frac{\tilde{\Theta}_{m,Y}}{\sqrt{m}}\right) + o_{\mathrm{P}}\left(\frac{1}{\sqrt{m}}\right),\end{aligned}$$

where

$$\begin{aligned}\tilde{\Theta}_{n,X} &= \sqrt{n}(\tilde{\theta}_{n,X} - \tilde{\theta}_X) = S_X\,\Theta_{n,X} \rightsquigarrow \sigma_X\,\Theta_X,\\ \tilde{\Theta}_{m,Y} &= \sqrt{m}(\tilde{\theta}_{m,Y} - \tilde{\theta}_Y) = S_Y\,\Theta_{m,Y} \rightsquigarrow \sigma_Y\,\Theta_Y.\end{aligned}$$

Since $\mathcal{L}(F, \tilde{\theta}_X) = \mathcal{L}(G, \tilde{\theta}_Y)$ under $\mathcal{H}_0$, one deduces that

$$\begin{aligned}\mathbb{D}_{n,m}(x) &= \sqrt{1-\lambda_{n,m}}\,\mathcal{L}'_{F_n,\tilde{\theta}_{n,X}}(\mathbb{F}_n, \tilde{\Theta}_{n,X})\\ &\quad - \sqrt{\lambda_{n,m}}\,\mathcal{L}'_{G_m,\tilde{\theta}_{m,Y}}(\mathbb{G}_m, \tilde{\Theta}_{m,Y}) + o_{\mathrm{P}}(1)\\ &\rightsquigarrow \sqrt{1-\lambda}\,\mathcal{L}'_{F,\tilde{\theta}_X}(\mathbb{B}_X, \sigma_X\,\Theta_X) - \sqrt{\lambda}\,\mathcal{L}'_{G,\tilde{\theta}_Y}(\mathbb{B}_Y, \sigma_Y\,\Theta_Y),\end{aligned}$$

this last expression being equivalent to $\mathbb{D}(x)$ presented in Proposition 5.1.

The idea of the multiplier method developed herein is to exploit the representation in Equation (5.5). To this end, define the multiplier versions of $\bar{X}$, $\bar{Y}$, S_X^2, S_Y^2, F_n and G_m as

$$\begin{aligned}&\bar{X}^{(h)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\xi}_i^{(h)}X_i, \quad \bar{Y}^{(h)} = \frac{1}{m}\sum_{i=1}^{m}\tilde{\kappa}_i^{(h)}Y_i,\\ &(S_X^2)^{(h)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\xi}_i^{(h)}\left(X_i - \bar{X}\right)^2, \quad (S_Y^2)^{(h)} = \frac{1}{m}\sum_{i=1}^{m}\tilde{\kappa}_i^{(h)}\left(Y_i - \bar{Y}\right)^2,\\ &F_n^{(h)}(x) = \frac{1}{n}\sum_{i=1}^{n}\tilde{\xi}_i^{(h)}\mathbb{I}\left(X_i \le x\right), \quad G_m^{(h)}(x) = \frac{1}{m}\sum_{i=1}^{m}\tilde{\kappa}_i^{(h)}\mathbb{I}\left(Y_i \le x\right).\end{aligned}$$

Hence, a direct way to bootstrap $\mathbb{D}_{n,m}$ consists in defining, for $h \in \{1, \ldots, M\}$,

$$\begin{aligned}\tilde{\mathbb{D}}_{n,m}^{(h)}(x) &= \sqrt{\frac{nm}{n+m}} \left\{ \mathcal{L}\left(F_n^{(h)}, \tilde{\theta}_{n,X}^{(h)}\right) - \mathcal{L}\left(F_n, \tilde{\theta}_{n,X}\right) \right\} \\ &\quad - \sqrt{\frac{nm}{n+m}} \left\{ \mathcal{L}\left(G_m^{(h)}, \tilde{\theta}_{m,Y}^{(h)}\right) - \mathcal{L}\left(G_m, \tilde{\theta}_{m,Y}\right) \right\},\end{aligned}$$

where $\tilde{\theta}_{n,X}^{(h)}(x) = S_X^{(h)} x + \bar{X}^{(h)}$ and $\tilde{\theta}_{m,Y}^{(h)}(x) = S_Y^{(h)} x + \bar{Y}^{(h)}$. Finally,

$$\begin{aligned}\tilde{\mathbb{D}}_{n,m}^{(h)}(x) &= \sqrt{\frac{1-\lambda_{n,m}}{n}} \sum_{i=1}^{n} \left\{ \tilde{\xi}_i^{(h)} \mathbb{I}\left(X_{i,n}^{(h)} \leq x\right) - \mathbb{I}\left(X_{i,n} \leq x\right) \right\} \\ &\quad - \sqrt{\frac{\lambda_{n,m}}{m}} \sum_{i=1}^{m} \left\{ \tilde{\kappa}_i^{(h)} \mathbb{I}\left(Y_{i,m}^{(h)} \leq x\right) - \mathbb{I}\left(Y_{i,m} \leq x\right) \right\},\end{aligned}$$

where $X_{i,n}^{(h)} = (X_i - \bar{X}^{(h)})/S_X^{(h)}$ and $Y_{i,m}^{(h)} = (Y_i - \bar{Y}^{(h)})/S_Y^{(h)}$. In Appendix B, explicit expressions for

$$\tilde{T}_{n,m}^{(h)} = \int_{\mathbb{R}} \left\{ \tilde{\mathbb{D}}_{n,m}^{(h)}(x) \right\}^2 \mathrm{d}x, \quad h \in \{1, \ldots, M\},$$

are derived. The validity of the method is established in the next proposition, whose proof is deferred to Appendix B.

Proposition 5.6. *Under* $\mathcal{H}_0$,

$$\left(\mathbb{D}_{n,m}, \tilde{\mathbb{D}}_{n,m}^{(1)}, \ldots, \tilde{\mathbb{D}}_{n,m}^{(M)}\right) \rightsquigarrow \left(\mathbb{D}, \tilde{\mathbb{D}}^{(1)}, \ldots, \tilde{\mathbb{D}}^{(M)}\right),$$

where $\tilde{\mathbb{D}}^{(1)}, \ldots, \tilde{\mathbb{D}}^{(M)}$ *are independent copies of* $\mathbb{D}$.

The idea for paired samples is similar; one has to define

$$\mathcal{L}(H, \theta) = F\left\{\theta_X(x)\right\} - G\left\{\theta_Y(x)\right\},$$

where $F(x) = H(x, \infty)$, $G(y) = H(\infty, y)$ and $\theta(x) = (\theta_X(x), \theta_Y(x))$, with $\theta_X(x) = s_X\, x + m_Y$, $\theta_Y(x) = s_Y\, x + m_Y$, $s_X, s_Y \in \mathbb{R}^+$ and $m_X, m_Y \in \mathbb{R}$. It

can be shown that the Hadamard derivative of $\mathcal{L}$ at (H, θ) is

$$\mathcal{L}'_{II,\theta}(\Delta, \delta) = f\{\theta_X(x)\}\delta_X(x) - g\{\theta_Y(x)\}\delta_Y(x) + \Delta_X(\theta_X(x)) - \Delta_Y(\theta_Y(x)),$$

where $f = \mathrm{d}F$, $g = \mathrm{d}G$, $\Delta_X(x) = \Delta(x, \infty)$, $\Delta_Y(x) = \Delta(\infty, x)$ and $\delta(x) = (\delta_X(x), \delta_Y(x))$. Hence, since one can write

$$\mathbb{E}_n(x) = \sqrt{n}\left\{\mathcal{L}\left(H_n, (\tilde{\theta}_{n,X}, \tilde{\theta}_{n,Y})\right) - \mathcal{L}\left(H, (\tilde{\theta}_X, \tilde{\theta}_Y)\right)\right\}$$

under $\mathcal{H}_0$, one has $\mathbb{E}_n(x) \rightsquigarrow \mathcal{L}'_{H,\tilde{\theta}}(\mathbb{H}, (\sigma_X\Theta_X, \sigma_Y\Theta_Y))$. Hence, a direct multiplier method consists in defining

$$\mathbb{E}_n^{(h)}(x) = \sqrt{n}\left\{\mathcal{L}\left(H_n^{(h)}, (\tilde{\theta}_{n,X}^{(h)}, \tilde{\theta}_{n,Y}^{(h)})\right) - \mathcal{L}\left(H_n, (\tilde{\theta}_{n,X}, \tilde{\theta}_{n,Y})\right)\right\},$$

where

$$H_n^{(h)}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\xi}_i^{(h)} \mathbb{I}(X_i \leq x, Y_i \leq y)$$

and

$$\left(\theta_{n,X}^{(h)}(x), \theta_{n,Y}^{(h)}(x)\right) = \left(S_X^{(h)}\, x + \bar{X}^{(h)}, S_Y^{(h)}\, x + \bar{Y}^{(h)}\right).$$

It can be shown that the method is valid asymptotically, using the Hadamard derivative of $\mathcal{L}$.

5.4 Multivariate extension

5.4.1 Hypotheses

Consider the random vectors $\mathbf{X} = (X_1, \ldots, X_d)$ and $\mathbf{Y} = (Y_1, \ldots, Y_d)$ from d-variate distribution functions F and G with finite fourth moments, mean

vectors $\mu_{\mathbf{X}}$, $\mu_{\mathbf{Y}}$ and positive definite variance-covariance matrices Σ_X, Σ_Y. In that case, the standardized versions are given by

$$\tilde{\mathbf{X}} = \Sigma_{\mathbf{X}}^{-1/2}\left(\mathbf{X} - \mu_{\mathbf{X}}\right) \quad \text{and} \quad \tilde{\mathbf{Y}} = \Sigma_{\mathbf{Y}}^{-1/2}\left(\mathbf{Y} - \mu_{\mathbf{Y}}\right).$$

A multivariate extension of $\mathcal{H}_0$ is then $\mathcal{H}_0^\star : \tilde{\mathbf{X}} \stackrel{d}{=} \tilde{\mathbf{Y}}$.

For example, if both $\mathbf{X}$ and $\mathbf{Y}$ are elliptically distributed, then one has the representations

$$\mathbf{X} = \mu_X + R_{\mathbf{X}}\Sigma_{\mathbf{X}}^{1/2}U_{\mathbf{X}} \quad \text{and} \quad \mathbf{Y} = \mu_Y + R_{\mathbf{Y}}\Sigma_{\mathbf{Y}}^{1/2}U_{\mathbf{Y}},$$

where $U_{\mathbf{X}}$, $U_{\mathbf{Y}}$ are uniformly distributed on the unit sphere $\mathbb{R}^d$ and $R_{\mathbf{X}}$, $R_{\mathbf{Y}}$ are positive random variables. In that context, $\mathcal{H}_0^\star$ will be true whenever $R_{\mathbf{X}}$ and $R_{\mathbf{Y}}$ are identically distributed, *i.e.* $\mathbf{X}$ and $\mathbf{Y}$ belong to the same elliptical family.

5.4.2 Empirical process and test statistic

A statistical procedure for $\mathcal{H}_0^\star$ will first be described for independent samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{id})$. The test statistic will be a Cramér–von Mises functional of the empirical distributions of the standardized observations

$$\begin{aligned}
\mathbf{X}_{i,n} &= S_{\mathbf{X}}^{-1/2}\left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right), \quad i \in \{1, \ldots, n\}, \\
\mathbf{Y}_{i,m} &= S_{\mathbf{Y}}^{-1/2}\left(\mathbf{Y}_i - \bar{\mathbf{Y}}_m\right), \quad i \in \{1, \ldots, m\},
\end{aligned}$$

where

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i, \quad S_{\mathbf{X}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^{\top},$$
$$\bar{\mathbf{Y}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{Y}_i, \quad S_{\mathbf{Y}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right)^{\top}.$$

For $\mathbf{x} \in \mathbb{R}^d$, let

$$\tilde{F}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\mathbf{X}_{i,n} \leq \mathbf{x}\right) \quad \text{and} \quad \tilde{G}_m(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(\mathbf{Y}_{i,m} \leq \mathbf{x}\right),$$

and consider the empirical process

$$\mathbb{D}^{\star}_{n,m}(\mathbf{x}) = \sqrt{\frac{nm}{n-m}}\left\{\tilde{F}_n(\mathbf{x}) - \tilde{G}_m(\mathbf{x})\right\}.$$

Its asymptotic behavior is described in the next proposition; the proof can be found in Appendix B. It is a multivariate generalization of Proposition 5.1.

Proposition 5.7. *Under the null hypothesis $\mathcal{H}_0^{\star}$, the empirical process $\mathbb{D}^{\star}_{n,m}$ converges weakly to*

$$\begin{aligned}\mathbb{D}^{\star}(\mathbf{x}) &= \sqrt{1-\lambda}\left\{\mathbb{B}_{\mathbf{X}}(\mathbf{x}) + f_0(\mathbf{x})\left(\mathbf{Z}_{\mathbf{X},1}\mathbf{x} + \mathbf{Z}_{\mathbf{X},2}\right)\right\} \\ &\quad - \sqrt{\lambda}\left\{\mathbb{B}_{\mathbf{Y}}(\mathbf{x}) + f_0(\mathbf{x})\left(\mathbf{Z}_{\mathbf{Y},1}\mathbf{x} + \mathbf{Z}_{\mathbf{Y},2}\right)\right\},\end{aligned}$$

where $\mathbb{B}_{\mathbf{X}}$ and $\mathbb{B}_{\mathbf{Y}}$ are independent F_0-Brownian sheets, f_0 is the density of F_0, and $\mathbf{Z}_{\mathbf{X},1}$, $\mathbf{Z}_{\mathbf{X},2}$, $\mathbf{Z}_{\mathbf{Y},1}$, $\mathbf{Z}_{\mathbf{Y},2}$ are the limits, respectively, of

$$\mathbf{Z}_{n,\mathbf{X},1} = \Sigma_X^{-1/2}\sqrt{n}\left(S_{\mathbf{X}}^{1/2} - \Sigma_X^{1/2}\right), \quad \mathbf{Z}_{n,\mathbf{X},2} = \Sigma_X^{-1/2}\sqrt{n}\left(\bar{\mathbf{X}} - \mu_{\mathbf{X}}\right),$$
$$\mathbf{Z}_{m,\mathbf{Y},1} = \Sigma_Y^{-1/2}\sqrt{m}\left(S_{\mathbf{Y}}^{1/2} - \Sigma_Y^{1/2}\right), \quad \mathbf{Z}_{m,\mathbf{Y},2} = \Sigma_Y^{-1/2}\sqrt{m}\left(\bar{\mathbf{Y}} - \mu_{\mathbf{Y}}\right).$$

The test statistic that will be investigated is

$$T^{\star}_{n,m} = \int_{\mathbb{R}^d}\left\{\mathbb{D}^{\star}_{n,m}(\mathbf{x})\right\}^2 \mathrm{d}\mathbf{x}.$$

If λ_d is Lebesgue's measure in $\mathbb{R}^d$ and $\mathbf{N} = (N, \ldots, N)$, then

$$\begin{aligned} T^{\star}_{n,m} &= \lim_{N\to\infty} \frac{m}{n(n+m)} \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_d\left([\mathbf{X}_{i,n} \vee \mathbf{X}_{j,n}, \mathbf{N}]\right) \\ &\quad - \frac{2}{n+m} \sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_d\left([\mathbf{X}_{i,n} \vee \mathbf{Y}_{j,m}, \mathbf{N}]\right) \\ &\quad + \frac{n}{m(n+m)} \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_d\left([\mathbf{Y}_{i,m} \vee \mathbf{Y}_{j,m}, \mathbf{N}]\right). \end{aligned}$$

Proposition 5.7 entails that $T^{\star}_{n,m}$ converges weakly, under $\mathcal{H}^{\star}_0$, to

$$T^{\star} = \int_{\mathbb{R}^d} \{\mathbb{D}^{\star}(\mathbf{x})\}^2 \, \mathrm{d}\mathbf{x}.$$

5.4.3 Multiplier versions

The re-sampling method developed in the multivariate case will be similar to strategy II proposed in the univariate context. Contrary to strategy I, it has the notable advantage of avoiding the estimation of the density, which is particularly touchy in $\mathbb{R}^d$.

In order to describe the methodology, first note that classical multiplier versions of $\mathbf{W}_{n,\mathbf{X},2} = \sqrt{n}(\bar{\mathbf{X}} - \mu_{\mathbf{X}})$ and $\mathbf{W}_{m,\mathbf{Y},2} = \sqrt{m}(\bar{\mathbf{Y}} - \mu_{\mathbf{Y}})$ are given, for $h \in \{1, \ldots, M\}$, by

$$\mathbf{W}^{(h)}_{n,\mathbf{X},2} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\tilde{\xi}^{(h)}_i - 1)\mathbf{X}_i \quad \text{and} \quad \mathbf{W}^{(h)}_{m,\mathbf{Y},2} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\tilde{\kappa}^{(h)}_i - 1)\mathbf{Y}_i.$$

Hence, multiplier versions of $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are, for $h \in \{1, \ldots, M\}$,

$$\begin{aligned}
\bar{\mathbf{X}}^{(h)} &= \frac{\mathbf{W}^{(h)}_{n,\mathbf{X},2}}{\sqrt{n}} + \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\xi}_i^{(h)} \mathbf{X}_i, \\
\bar{\mathbf{Y}}^{(h)} &= \frac{\mathbf{W}^{(h)}_{m,\mathbf{Y},2}}{\sqrt{m}} + \bar{\mathbf{Y}} = \frac{1}{m}\sum_{i=1}^{m} \tilde{\kappa}_i^{(h)} \mathbf{Y}_i.
\end{aligned}$$

Next, note that the classical multiplier versions of $\mathbf{W}_{n,\mathbf{X},1} = \sqrt{n}(S_{\mathbf{X}} - \Sigma_{\mathbf{X}})$ and $\mathbf{W}_{m,\mathbf{Y},1} = \sqrt{m}(S_{\mathbf{Y}} - \Sigma_{\mathbf{Y}})$ are

$$\begin{aligned}
\mathbf{W}^{(h)}_{n,\mathbf{X},1} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left(\tilde{\xi}_i^{(h)} - 1\right)(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^{\top}, \\
\mathbf{W}^{(h)}_{m,\mathbf{Y},1} &= \frac{1}{\sqrt{m}}\sum_{i=1}^{m} \left(\tilde{\kappa}_i^{(h)} - 1\right)(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^{\top},
\end{aligned}$$

so that

$$\begin{aligned}
S^{(h)}_{\mathbf{X}} &= \frac{\mathbf{W}^{(h)}_{n,\mathbf{X},1}}{\sqrt{n}} + S_{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\xi}_i^{(h)}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^{\top}, \\
S^{(h)}_{\mathbf{Y}} &= \frac{\mathbf{W}^{(h)}_{m,\mathbf{Y},1}}{\sqrt{m}} + S_{\mathbf{Y}} = \frac{1}{m}\sum_{i=1}^{m} \tilde{\kappa}_i^{(h)}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^{\top}.
\end{aligned}$$

Mimicking the procedure described in the univariate case, define the multiplier standardized observations by

$$\mathbf{X}^{(h)}_{i,n} = \left(S^{(h)}_{\mathbf{X}}\right)^{-1/2}\left(\mathbf{X}_i - \bar{\mathbf{X}}^{(h)}\right) \quad \text{and} \quad \mathbf{Y}^{(h)}_{i,m} = \left(S^{(h)}_{\mathbf{Y}}\right)^{-1/2}\left(\mathbf{Y}_i - \bar{\mathbf{Y}}^{(h)}\right),$$

and let

$$\begin{aligned}
\mathbb{D}^{\star(h)}_{n,m}(\mathbf{x}) &= \sqrt{\frac{1-\lambda_{n,m}}{n}}\sum_{i=1}^{n}\left\{\tilde{\xi}_i^{(h)}\mathbb{I}\left(\mathbf{X}^{(h)}_{i,n} \le \mathbf{x}\right) - \mathbb{I}\left(\mathbf{X}_{i,n} \le \mathbf{x}\right)\right\} \\
&\quad - \sqrt{\frac{\lambda_{n,m}}{m}}\sum_{i=1}^{m}\left\{\tilde{\kappa}_i^{(h)}\mathbb{I}\left(\mathbf{Y}^{(h)}_{i,m} \le \mathbf{x}\right) - \mathbb{I}\left(\mathbf{Y}_{i,m} \le \mathbf{x}\right)\right\}.
\end{aligned}$$

The multiplier statistics are given by

$$T_{n,m}^{\star(h)} = \int_{\mathbb{R}} \left\{ \mathbb{D}_{n,m}^{\star(h)}(\mathbf{x}) \right\}^2 \mathrm{d}\mathbf{x}, \quad h \in \{1, \ldots, M\}.$$

An explicit formula is given in Appendix B.

5.4.4 Paired samples

The extension to paired samples can readily be made. In that case, one observes $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ i.i.d. from some joint distribution $H(\mathbf{x}, \mathbf{y}) = \mathrm{P}(\mathbf{X} \leq \mathbf{x}, \mathbf{Y} \leq \mathbf{y})$ with d-variate marginal distributions $F(\mathbf{x}) = \mathrm{P}(\mathbf{X} \leq \mathbf{x})$ and $G(\mathbf{y}) = \mathrm{P}(\mathbf{Y} \leq \mathbf{y})$. The Cramér–von Mises statistic is then based on

$$\mathbb{E}_n^\star(\mathbf{x}) = \sqrt{n} \left\{ \tilde{H}_n(\mathbf{x}, \underline{\infty}) - \tilde{H}_n(\underline{\infty}, \mathbf{x}) \right\},$$

where $\underline{\infty}$ is a d-variate vector with all components being ∞ and $\tilde{H}_n(\mathbf{x}, \mathbf{y})$ is the empirical joint distribution of the standardized observations $(\mathbf{X}_{i,n}, \mathbf{Y}_{i,n})$, $i \in \{1, \ldots, n\}$. Then one can show that $\mathbb{E}_n^\star \rightsquigarrow \mathbb{E}^\star$ under $\mathcal{H}_0^\star$, where

$$\mathbb{E}^\star(\mathbf{x}) = \mathbb{B}(\mathbf{x}, \underline{\infty}) - \mathbb{B}(\underline{\infty}, \mathbf{x}) + f_0(\mathbf{x}) \left\{ \Theta_{\mathbf{X}}(\mathbf{x}) - \Theta_{\mathbf{Y}}(\mathbf{x}) \right\}.$$

Here, $\mathbb{B}$ is a $\tilde{H}$-Brownian sheet, where $\tilde{H}(\mathbf{x}, \mathbf{y})$ is a $2d$-variate distribution with $\tilde{H}(\mathbf{x}, \underline{\infty}) = F_0(\mathbf{x})$ and $\tilde{H}(\underline{\infty}, \mathbf{y}) = \tilde{F}_0(\mathbf{y})$. The computation of p-values is similar to that presented in the case of independent samples; essentially, it suffices to put $\kappa_i^{(h)} = \xi_i^{(h)}$.

5.5 Simulation study

The statistical methodologies developed in this work rely mainly on re-sampling strategies in order to compute p-values. These methods have been shown to be valid asymptotically. It is also important to study their validity in small samples. In this section, the power and size of the tests is investigated.

In the univariate case, three models for the marginal distributions will be considered, namely the Normal (N), double-exponential (DE) and logistic (L) distributions. Since the test statistics are invariant under affine transformations, there is no loss in generality in considering only the standard densities, *i.e.*

$$f_N(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad f_{DE}(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|}, \quad f_L(x) = \frac{\beta e^{-\beta x}}{(1+e^{-\beta x})^2},$$

where $\beta = \pi/\sqrt{3}$.

Firstly, the power of the test based on $T_{n,m}$ for the two re-sampling strategies has been estimated with the help of 1 000 pairs of independent samples from the above distributions. The results are in Table 5.1.

As it cans be seem in Table 5.1, the tests give excellent results with the bootstrap strategy I under the null hypothesis. However, the bootstrap strategy II give bad results except for the double exponential distribution. Now, under the alternative hypothesis, the bootstraps strategies seem to give similar results. In all cases, the strategy I give good results whereas the strategy II only give a not bad result for the comparaison of the Normal and the double

Table 5.1: Percentage of rejection of $\mathcal{H}_0$, as evaluated from 10 000 replicates, in the case of independent samples, using $M = 500$ copies from the bootstrap strategies I and II

Margins		(n, m) (50, 100)		(n, m) (100, 50)		(n, m) (100, 100)		(n, m) (250, 250)	
F	G	I	II	I	II	I	II	I	II
N	N	0.0427	0.0080	0.0399	0.0087	0.0463	0.0157	0.0574	0.0256
N	DE	0.2720	0.1499	0.3258	0.1058	0.4601	0.2764	0.8870	0.8095
N	L	0.0798	0.0194	0.0941	0.0222	0.1155	0.0396	0.2319	0.1270
DE	DE	0.0681	0.0334	0.0671	0.0294	0.0654	0.0309	0.0665	0.0389
DE	L	0.1434	0.0411	0.1407	0.0714	0.1922	0.0974	0.4422	0.3115
L	L	0.0581	0.0144	0.0626	0.0141	0.0600	0.0198	0.0676	0.0295

exponential distribution. Finally, the size used for the simulation don't affect considerably the power under the null hypothesis. In the opposite, the power increase under the alternative hypothesis.

In Table 5.2, the size and power of the test based on T_n, suitable for the case of dependent samples, is evaluated. To this end, bivariate random samples from distributions with chosen margins F, G being either the Normal, double-exponential or logistic distribution, and a given copula C, are considered. The latter are easily generated as long as one can simulate $(U, V) \sim C$; it then suffices to define $X = F^{-1}(U)$ and $Y = G^{-1}(V)$. The models considered are the Clayton, Normal and Frank copulas. For the seek of comparison among these various dependence structures, the latter are parameterized with

respect to Kendall's measure of association, which can be written as

$$\tau(C) = 4 \int_0^1 \int_0^1 C(u,v) \mathrm{d}C(u,v) - 1.$$

First of all, the used of Clayton, Normal or Frank copula don't change the power under the null and the alternative hypothesis in Table 5.2. However, when the kendall's tau is weakly negative or positive, the results are better. Specifically, both bootstraps strategies are good under H_0. For the normal distribution, the strategy I gives better results when the Kendall's tau is strongly positive or negative. For the double exponential and the logistic distributions, strategy I gives better results than strategy II when the Kendall's tau is negative. At the opposite, strategy II gives better results than strategy I when the Kendall's tau is positive. Now, the power under the alternative seem to be good in the major cases. However, strategy II gives bad results for the comparaison of the double exponential and the logistic when the Kendall's tau is positive. For the comparaison of the normal and the double exponential distributions, strategy II gives better results except when the Kendall's tau is strongly positive. In that case, both strategies seem to give the same conclusion. For the comparaison of the normal and the logistic distribution, strategy I gives better results in all possibility for the Kendall's tau.

5.6 Proofs of the theoretical results

5.6.1 Proposition 5.1

Under $\mathcal{H}_0$, one can write

$$\begin{aligned}\mathbb{D}_{n,m}(x) &= \sqrt{1-\lambda_{n,m}}\,\mathbb{F}_n\left(S_X\,x+\bar{X}\right)-\sqrt{\lambda_{n,m}}\,\mathbb{G}_m\left(S_Y\,x+\bar{Y}\right)\\ &\quad+\sqrt{1-\lambda_{n,m}}\,A_{n,X}(x)-\sqrt{\lambda_{n,m}}\,A_{m,Y}(x),\end{aligned}$$

where $\lambda_{n,m}=n/(n+m)$, $\mathbb{F}_n=\sqrt{n}(F_n-F)$, $\mathbb{G}_m=\sqrt{m}(G_m-G)$,

$$\begin{aligned}A_{n,X}(x) &= \sqrt{n}\left\{F\left(S_X\,x+\bar{X}\right)-F\left(\sigma_X\,x+\mu_X\right)\right\},\\ A_{m,Y}(x) &= \sqrt{m}\left\{G\left(S_Y\,x+\bar{Y}\right)-G\left(\sigma_Y\,x+\mu_Y\right)\right\}.\end{aligned}$$

First, Donsker's theorem entails that

$$\mathbb{F}_n\left(S_X\,x+\bar{X}\right)\rightsquigarrow\mathbb{F}\left(\sigma_X\,x+\mu_X\right)=\mathbb{B}_X(x)$$

and

$$\mathbb{G}_n\left(S_Y\,x+\bar{Y}\right)\rightsquigarrow\mathbb{G}\left(\sigma_Y\,x+\mu_Y\right)=\mathbb{B}_Y(x),$$

where $\mathbb{B}_X$ and $\mathbb{B}_Y$ are independent F_0-Brownian bridges; hence,

$$\sqrt{1-\lambda_{n,m}}\,\mathbb{F}_n\left(S_X\,x+\bar{X}\right)-\sqrt{\lambda_{n,m}}\,\mathbb{G}_m\left(S_Y\,x+\bar{Y}\right)\rightsquigarrow\sqrt{1-\lambda}\,\mathbb{B}_X-\sqrt{\lambda}\,\mathbb{B}_Y.$$

Next, the mean-value theorem entails that there exists $(\sigma^\star,\mu^\star)$ such that

$$\|(\sigma^\star,\mu^\star)-(\sigma_X,\mu_X)\|<\|(S_X,\bar{X})-(\sigma_X,\mu_X)\|=\frac{\sigma_X\|(Z_{n,X,1},Z_{n,X,2})\|}{\sqrt{n}}\quad(5.6)$$

and

$$\begin{aligned} A_{n,X}(x) &= x\, f(\sigma^\star x + \mu^\star)\sqrt{n}\,(S_X - \sigma_X) + f(\sigma^\star x + \mu^\star)\sqrt{n}\,(\bar{X} - \mu_X) \\ &= x\,\sigma_X f(\sigma^\star x + \mu^\star) Z_{n,X,1} + \sigma_X f(\sigma^\star x + \mu^\star) Z_{n,X,2} \\ &= x\, f_0\left(\frac{\sigma^\star}{\sigma_X}\, x + \frac{\mu^\star - \mu_X}{\sigma_X}\right) Z_{n,X,1} + f_0\left(\frac{\sigma^\star}{\sigma_X}\, x + \frac{\mu^\star - \mu_X}{\sigma_X}\right) Z_{n,X,2}, \end{aligned}$$

where the fact that $f(x) = f_0\{(x - \mu_X)/\sigma_X\}/\sigma_X$ was used. Thus,

$$\begin{aligned} A_{n,X}(x) - \{x f_0(x) Z_{n,X,1} + f_0(x) Z_{n,X,2}\} &= \{f_0(a\,x + b) - f_0(x)\} \\ &\quad \times (x\, Z_{n,X,1} + Z_{n,X,2}), \end{aligned}$$

where $a = \sigma^\star/\sigma_X$ and $b = (\mu^\star - \mu_X)/\sigma_X$. Since $(Z_{n,X,1}, Z_{n,X,2})$ is tight, for any $\delta > 0$ there exists $M > 0$ and $n_0 \in \mathbb{N}$ such that $\mathrm{P}(\|(Z_{n,X,1}, Z_{n,X,2})\| > M) < \delta$ for all $n \geq n_0$. As a consequence, since from equation (5.6), $|(a - 1, b)| < |(Z_{n,X,1}, Z_{n,X,2})|/\sqrt{n}$, one has for any $\lambda > 0$ that

$$\begin{aligned} & \mathrm{P}\left\{\sup_{x\in\mathbb{R}} |A_{n,X}(x) - \{x f_0(x) Z_{n,X,1} + f_0(x) Z_{n,X,2}\}| > \lambda\right\} \\ < \; & \mathrm{P}\left\{\sup_{x\in\mathbb{R}} |\{f_0(a\,x + b) - f_0(x)\}\,(x\, Z_{n,X,1} + Z_{n,X,2})| > \lambda, \|(Z_{n,X,1}, Z_{n,X,2})\| \leq M\right\} \\ & + \mathrm{P}\left(|(Z_{n,X,1}, Z_{n,X,2})| > M\right) \\ < \; & \mathrm{P}\left\{Z_{n,X,1} \sup_{|(a-1,b)|\leq M/\sqrt{n}} \sup_{x\in\mathbb{R}} |x f_0(ax + b) - x f_0(x)| > \lambda\right\} \\ & + \mathrm{P}\left\{Z_{n,X,2} \sup_{|(a-1,b)|\leq M/\sqrt{n}} \sup_{x\in\mathbb{R}} \left|f_0(ax + b) - \tilde{f}_0(x)\right| > \lambda\right\} + \delta. \end{aligned}$$

This last expression can be made arbitrarily small since $\delta > 0$ can be chosen arbitrarily small and from the fact that (5.3) holds with $\epsilon = M/\sqrt{n}$. As a consequence, $A_{n,X}(x) \rightsquigarrow x f_0(x) Z_{X,1} + f_0(x) Z_{X,2}$. One proceeds in a similar manner to show that $A_{m,Y}(x) \rightsquigarrow x f_0(x) Z_{Y,1} + f_0(x) Z_{Y,2}$.

5.6.2 Proposition 5.2

One can write $\tilde{\mathbb{H}}_n = \mathbb{A}_{n1} + \mathbb{A}_{n2}$, where

$$\begin{aligned}
\mathbb{A}_{n1}(x,y) &= \mathbb{H}_n\left(S_X\, x + \bar{X}, S_Y\, y + \bar{Y}\right), \\
\mathbb{A}_{n2}(x,y) &= \sqrt{n}\left\{H\left(S_X\, x + \bar{X}, S_Y\, y + \bar{Y}\right) - H\left(\sigma_X\, x + \mu_X, \sigma_Y\, x + \mu_Y\right)\right\}.
\end{aligned}$$

By standard empirical process theory, $\mathbb{H}_n$ converges weakly to $\mathbb{B}\{F(\cdot), G(\cdot)\}$, where $\mathbb{B}$ is a C-Brownian bridge; hence $\mathbb{A}_{n1}(x,y)$ converges weakly to

$$\mathbb{A}_1(x,y) = \mathbb{B}\left\{F\left(\sigma_X\, x + \mu_X\right), G\left(\sigma_Y\, y + \mu_Y\right)\right\} = \mathbb{B}\left\{\tilde{F}(x), \tilde{G}(y)\right\},$$

by the continuity of $\mathbb{B}$, F and G. Next, the mean-value theorem ensures that there exists $(\sigma_X^\star, \mu_X^\star, \sigma_Y^\star, \mu_Y^\star)$ such that

$$\begin{aligned}
&\left|(\sigma_X^\star, \mu_X^\star, \sigma_Y^\star, \mu_Y^\star) - (\sigma_X, \mu_X, \sigma_Y, \mu_Y)\right| \\
&\qquad < \left|(S_X, \bar{X}, S_Y, \bar{Y}) - (\sigma_X, \mu_X, \sigma_Y, \mu_Y)\right| \\
&\qquad = \frac{\left|(\sigma_X Z_{n,X,1}, \sigma_X Z_{n,X,2}, \sigma_Y Z_{n,Y,1}, \sigma_Y Z_{n,Y,2})\right|}{\sqrt{n}}
\end{aligned}$$

and

$$\begin{aligned}
A_{n2}(x,y) &= x\sigma_X H_{10}\left(\sigma_X^\star x+\mu_X^\star, \sigma_Y^\star x+\mu_Y^\star\right) Z_{n,X,1} \\
&\quad + \sigma_X H_{10}\left(\sigma_X^\star x+\mu_X^\star, \sigma_Y^\star x+\mu_Y^\star\right) Z_{n,X,2} \\
&\quad + y\sigma_Y H_{01}\left(\sigma_X^\star x+\mu_X^\star, \sigma_Y^\star x+\mu_Y^\star\right) Z_{n,Y,1} \\
&\quad + \sigma_Y H_{01}\left(\sigma_X^\star x+\mu_X^\star, \sigma_Y^\star x+\mu_Y^\star\right) Z_{n,Y,2} \\
&= x\tilde{H}_{10}\left(\frac{\sigma_X^\star}{\sigma_X}x+\frac{\mu_X^\star-\mu_X}{\sigma_X}, \frac{\sigma_Y^\star}{\sigma_Y}y+\frac{\mu_Y^\star-\mu_Y}{\sigma_Y}\right) Z_{n,X,1} \\
&\quad + \tilde{H}_{10}\left(\frac{\sigma_X^\star}{\sigma_X}x+\frac{\mu_X^\star-\mu_X}{\sigma_X}, \frac{\sigma_Y^\star}{\sigma_Y}y+\frac{\mu_Y^\star-\mu_Y}{\sigma_Y}\right) Z_{n,X,2} \\
&\quad + y\tilde{H}_{01}\left(\frac{\sigma_X^\star}{\sigma_X}x+\frac{\mu_X^\star-\mu_X}{\sigma_X}, \frac{\sigma_Y^\star}{\sigma_Y}y+\frac{\mu_Y^\star-\mu_Y}{\sigma_Y}\right) Z_{n,Y,1} \\
&\quad + \tilde{H}_{01}\left(\frac{\sigma_X^\star}{\sigma_X}x+\frac{\mu_X^\star-\mu_X}{\sigma_X}, \frac{\sigma_Y^\star}{\sigma_Y}y+\frac{\mu_Y^\star-\mu_Y}{\sigma_Y}\right) Z_{n,Y,2}.
\end{aligned}$$

The remaining of the proof is similar to that of Proposition 5.1, where one uses the tightness of $(Z_{n,X,1}, Z_{n,X,2}, Z_{n,Y,1}, Z_{n,Y,2})$ and the fact that (5.4) is assumed to hold.

5.6.3 Proposition 5.3

Using the notation introduced in the proof of Proposition 5.2, one can write $\mathbb{E}_n(x) = \tilde{\mathbb{H}}_n(x,\infty) - \tilde{\mathbb{H}}_n(\infty,x)$ under $\mathcal{H}_0$, where $\tilde{\mathbb{H}}_n(x,\infty) = \mathbb{A}_{n1}(x,\infty) + \mathbb{A}_{n2}(x,\infty)$ and $\tilde{\mathbb{H}}_n(\infty,x) = \mathbb{A}_{n1}(\infty,x) + \mathbb{A}_{n2}(\infty,x)$. One can then deduce that

$$\begin{aligned}
\left(\mathbb{A}_{n1}(x,\infty), \mathbb{A}_{n1}(\infty,x)\right) &\rightsquigarrow \left(\mathbb{A}_1(x,\infty), \mathbb{A}_1(\infty,x)\right) = \left(\mathbb{B}(x,\infty), \mathbb{B}(\infty,x)\right), \\
\left(\mathbb{A}_{n2}(x,\infty), \mathbb{A}_{n2}(\infty,x)\right) &\rightsquigarrow f_0(x)\left(\Theta_X(x), \Theta_Y(x)\right).
\end{aligned}$$

Finally, the continuous mapping theorem entails that

$$\begin{aligned}\mathbb{E}_n(x) \rightsquigarrow \mathbb{E}(x) &= \mathbb{A}_1(x,\infty) + \mathbb{A}_2(x,\infty) - \mathbb{A}_1(\infty,x) - \mathbb{A}_2(\infty,x) \\ &= \mathbb{B}(x,\infty) - \mathbb{B}(\infty,x) + f_0(x)\left\{\Theta_X(x) - \Theta_Y(x)\right\}.\end{aligned}$$

5.6.4 Proposition 5.4

An application of the multiplier central limit theorem entails that the vector

$$\left(\mathbb{F}_n(\sigma_X\, x + \mu_X), \mathbb{F}_n^{(1)}(S_X\, x + \bar{X}), \ldots, \mathbb{F}_n^{(M)}(S_X\, x + \bar{X})\right)$$

converges weakly to

$$\left(\mathbb{B}_X(x), \mathbb{B}_X^{(1)}(x), \ldots, \mathbb{B}_X^{(M)}(x)\right),$$

where $\mathbb{B}_X^{(1)}, \ldots, \mathbb{B}_X^{(M)}$ are independent copies of $\mathbb{B}_X$, *i.e.* independent F_0-Brownian bridges. Next, note that

$$S_X\, Z_{n,X,1}^{(h)} = \frac{1}{2S_X\sqrt{n}} \sum_{i=1}^{n} \left(\tilde{\xi}_i^{(h)} - 1\right)\left(X_i - \bar{X}\right)^2$$

and

$$S_X\, Z_{n,X,2}^{(h)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\tilde{\xi}_i^{(h)} - 1\right) X_i.$$

By the classical multiplier central limit theorem and Slutsky's lemma,

$$\left(\sigma_X(Z_{n,X,1}, Z_{n,X,2}), S_X(Z_{n,X,1}^{(1)}, Z_{n,X,2}^{(1)}), \ldots, S_X(Z_{n,X,1}^{(M)}, Z_{n,X,2}^{(M)})\right)$$

converges weakly to

$$\sigma_X\left((Z_{X,1}, Z_{X,2}), (Z_{X,1}^{(1)}, Z_{X,2}^{(1)}), \ldots, (Z_{X,1}^{(M)}, Z_{X,2}^{(M)})\right),$$

where $(Z_{X,1}^{(1)}, Z_{X,2}^{(1)}), \ldots, (Z_{X,1}^{(M)}, Z_{X,2}^{(M)})$ are independent copies of $(Z_{X,1}, Z_{X,2})$. By assumption, $\sup_{x\in\mathbb{R}} \left|\hat{f}_0(x) - f_0(x)\right|$ converge to 0 in probability. Moreover, since $x f_0(x)$ is bounded, there exists $0 < N < \infty$ such that

$$\sup_{x\in\mathbb{R}} \left|x\hat{f}_0(x) - x f_0(x)\right| \leq N \sup_{x\in\mathbb{R}} \left|\frac{\hat{f}_0(x) - f_0(x)}{f_0(x)}\right|.$$

Since

$$\hat{f}_0(x)\Theta_{n,X}^{(h)}(x) = Z_{n,X,1}^{(h)} x\hat{f}_0(x) + Z_{n,X,2}^{(h)}\hat{f}_0(x),$$

one deduces that

$$\left(f_0(x)\Theta_{n,X}(x), \hat{f}_0(x)\Theta_{n,X}^{(1)}(x), \ldots, \hat{f}_0(x)\Theta_{n,X}^{(M)}(x)\right)$$

converges weakly to

$$f_0(x)\left(\Theta_X(x), \Theta_X^{(1)}(x), \ldots, \Theta_X^{(M)}(x)\right),$$

where $\Theta_X^{(h)}(x) = Z_{X,1}^{(h)} x f_0(x) + Z_{X,2}^{(h)} f_0(x)$, $h \in \{1, \ldots, M\}$. The arguments are similar for $\mathbb{G}_m^{(h)}(S_Y\, x + \bar{Y}) + \hat{f}_0(x)\Theta_{m,Y}^{(h)}(x)$; as a consequence,

$$\left(\mathbb{D}_{n,m}, \mathbb{D}_{n,m}^{(1)}, \ldots, \mathbb{D}_{n,m}^{(M)}\right) \rightsquigarrow \left(\mathbb{D}, \mathbb{D}^{(1)}, \ldots, \mathbb{D}^{(M)}\right).$$

5.6.5 Proposition 5.6

By the Hadamard differentiability of $\mathcal{L}$,

$$\begin{aligned}\tilde{\mathbb{D}}_{n,m}^{(h)}(x) &= \sqrt{1-\lambda_{n,m}}\,\mathcal{L}'_{F_n,\theta_{n,X}}\left(\mathbb{F}_n^{(h)}, \Theta_{n,X}^{(h)}\right) \\ &\quad - \sqrt{\lambda_{n,m}}\,\mathcal{L}'_{G_m,\theta_{m,Y}}\left(\mathbb{G}_m^{(h)}, \Theta_{m,Y}^{(h)}\right) + o_{\mathrm{P}}(1),\end{aligned}$$

where $o_{\mathrm{P}}(1)$ arises from the fact that

$$\left\| \left(F_n^{(h)}, \theta_{n,X}^{(h)} \right) - (F_n, \theta_{n,X}) \right\| = \frac{1}{\sqrt{n}} \left\| \left(\mathbb{F}_n^{(h)}, \Theta_{n,X}^{(h)} \right) \right\|$$

and

$$\left\| \left(G_m^{(h)}, \theta_{m,Y}^{(h)} \right) - (G_m, \theta_{m,Y}) \right\| = \frac{1}{\sqrt{m}} \left\| \left(\mathbb{G}_m^{(h)}, \Theta_{m,Y}^{(h)} \right) \right\|$$

converge in probability to 0. The result follows from the fact that

$$\left(\mathbb{F}_n^{(1)}, \Theta_{n,X}^{(1)} \right), \ldots, \left(\mathbb{F}_n^{(M)}, \Theta_{n,X}^{(M)} \right)$$

are asymptotically independent copies of $(\mathbb{B}_X, \Theta_X)$, and

$$\left(\mathbb{G}_m^{(1)}, \Theta_{m,Y}^{(1)} \right), \ldots, \left(\mathbb{G}_m^{(M)}, \Theta_{m,Y}^{(M)} \right)$$

are asymptotically independent copies of $(\mathbb{B}_Y, \Theta_Y)$.

5.6.6 Proposition 5.7

Let's start with a lemma that characterizes the asymptotic behavior of the inverse square root of an empirical covariance matrix.

Lemma 5.1. *Let S be an empirical covariance matrix and Σ its theoretical counterpart. Then*

$$\tilde{\mathbf{Z}}_n = \sqrt{n} \left(S^{-1/2} - \Sigma^{-1/2} \right) \rightsquigarrow \tilde{\mathbf{Z}} = -\Sigma^{-1/2} \mathbf{Z} \Sigma^{-1/2},$$

where $\mathbf{Z}$ is the limit of $\mathbf{Z}_n = \sqrt{n}(S^{1/2} - \Sigma^{1/2})$.

Proof. Simply observe that $\tilde{\mathbf{Z}}_n = -S^{-1/2} \mathbf{Z}_n \Sigma^{-1/2}$, so that $\tilde{\mathbf{Z}}_n \rightsquigarrow \tilde{\mathbf{Z}} = -\Sigma^{-1/2} \mathbf{Z} \Sigma^{-1/2}$, by Slutsky's lemma and the continuous mapping theorem.

Note that it can be shown that the limit $\mathbf{Z}$ is the unique solution of

$$\Sigma^{1/2}\mathbf{Z} + \mathbf{Z}\Sigma^{1/2} = \mathbf{W},$$

where $\mathbf{W}$ is the limit of $\mathbf{W}_n = \sqrt{n}(S - \Sigma)$.

Let

$$\tilde{\mathbb{F}}_n(\mathbf{x}) = \sqrt{n}\left\{\tilde{F}_n(\mathbf{x}) - \tilde{F}(\mathbf{x})\right\} \quad \text{and} \quad \tilde{\mathbb{G}}_m(\mathbf{x}) = \sqrt{m}\left\{\tilde{G}_m(\mathbf{x}) - \tilde{G}(\mathbf{x})\right\},$$

where $\tilde{F}$, $\tilde{G}$ are the distribution functions of $\tilde{\mathbf{X}} = \Sigma_{\mathbf{X}}^{-1/2}(\mathbf{X} - \mu_{\mathbf{X}})$ and $\tilde{\mathbf{Y}} = \Sigma_{\mathbf{Y}}^{-1/2}(\mathbf{Y} - \mu_{\mathbf{Y}})$, respectively. Note that

$$\tilde{\mathbb{F}}_n(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbb{I}\left(\gamma_n(\mathbf{X}_i) \leq \mathbf{x}\right) - \mathrm{E}\,\mathbb{I}\left(\gamma(\mathbf{X}) \leq \mathbf{x}\right)\right\},$$

where

$$\gamma_n(\mathbf{x}) = S_{\mathbf{X}}^{-1/2}\left(\mathbf{x} - \bar{\mathbf{X}}\right) \quad \text{and} \quad \gamma(\mathbf{x}) = \Sigma_{\mathbf{X}}^{-1/2}\left(\mathbf{x} - \mu_{\mathbf{X}}\right),$$

and

$$\mathcal{G}_n(\mathbf{x}) = \sqrt{n}\left\{\gamma_n(\mathbf{x}) - \gamma(\mathbf{x})\right\} = \tilde{\mathbf{Z}}_{n,\mathbf{X},1}(\mathbf{x} - \mu_{\mathbf{X}}) - S_{\mathbf{X}}^{-1/2}\Sigma_X^{-1/2}\mathbf{Z}_{n,\mathbf{X},2},$$

where $\tilde{\mathbf{Z}}_{n,\mathbf{X},1} = \sqrt{n}(S_{\mathbf{X}}^{-1/2} - \Sigma_{\mathbf{X}}^{-1/2})$. From Lemma 5.1,

$$\mathcal{G}_n(\mathbf{x}) \rightsquigarrow \mathcal{G}(\mathbf{x}) = \tilde{\mathbf{Z}}_{\mathbf{X},1}(\mathbf{x} - \mu_{\mathbf{X}}) - \mathbf{Z}_{\mathbf{X},2} = -\mathbf{Z}_{\mathbf{X},1}\Sigma_{\mathbf{X}}^{-1/2}(\mathbf{x} - \mu_{\mathbf{X}}) - \mathbf{Z}_{\mathbf{X},2}.$$

From Theorem 2.4 of Ghoudi & Rémillard (2004), $\tilde{\mathbb{F}}_n(\mathbf{x}) \rightsquigarrow \mathbb{B}_X(\mathbf{x}) - \mu_{\mathbf{x}}(\mathcal{G})$, where $\mathbb{B}_X$ is a $\tilde{F}$-Brownian sheet and

$$\mu_{\mathbf{x}}(\mathcal{G}) = \tilde{F}'(\mathbf{x})\mathrm{E}\left\{\mathcal{G}(\mathbf{X})\,|\gamma(\mathbf{X}) = \mathbf{x}\right\} - f_0(\mathbf{x})\left(\mathbf{Z}_{\mathbf{X},1}\mathbf{x} + \mathbf{Z}_{\mathbf{X},2}\right).$$

As a consequence,

$$\tilde{\mathbb{F}}_n(\mathbf{x}) \rightsquigarrow \mathbb{B}_X(\mathbf{x}) + \tilde{F}'(\mathbf{x}) \left(\mathbf{Z}_{\mathbf{X},1}\mathbf{x} + \mathbf{Z}_{\mathbf{X},2}\right).$$

Similarly,

$$\tilde{\mathbb{G}}_m(\mathbf{x}) \rightsquigarrow \mathbb{B}_Y(\mathbf{x}) + \tilde{G}'(\mathbf{x}) \left(\mathbf{Z}_{\mathbf{Y},1}\mathbf{x} + \mathbf{Z}_{\mathbf{Y},2}\right).$$

Finally, under $\mathcal{H}_0^\star$, one has $\tilde{F} = \tilde{G} = F_0$, so that

$$\mathbb{D}_{n,m}^\star(\mathbf{x}) = \sqrt{\frac{m}{n+m}}\,\tilde{\mathbb{F}}_n(\mathbf{x}) - \sqrt{\frac{n}{n+m}}\,\tilde{\mathbb{G}}_m(\mathbf{x})$$

converges weakly to

$$\begin{aligned}\mathbb{D}^\star(\mathbf{x}) \;=\; & \sqrt{1-\lambda}\left\{\mathbb{B}_{\mathbf{X}}(\mathbf{x}) + f_0(\mathbf{x})\left(\mathbf{Z}_{\mathbf{X},1}\mathbf{x} + \mathbf{Z}_{\mathbf{X},2}\right)\right\} \\ & - \sqrt{\lambda}\left\{\mathbb{B}_{\mathbf{Y}}(\mathbf{x}) + f_0(\mathbf{x})\left(\mathbf{Z}_{\mathbf{Y},1}\mathbf{x} + \mathbf{Z}_{\mathbf{Y},2}\right)\right\}.\end{aligned}$$

5.7 Computation formulas

5.7.1 Estimation of the density

In order to estimate the density function f_0, consider the pooled empirical distribution function based on the standardized observations, *i.e.*

$$\widehat{F}_0(x) = \frac{1}{m+n}\left\{\sum_{i=1}^{n} \mathbb{I}\left(X_{i,n} \leq x\right) + \sum_{i=1}^{m} \mathbb{I}\left(Y_{i,m} \leq x\right)\right\}.$$

One can estimate f_0 with the histogram-type estimator

$$\begin{aligned}\hat{f}_0(x) &= \frac{1}{2\epsilon_{m,n}}\left\{\hat{F}_0\left(x+\epsilon_{m,n}\right)-\hat{F}_0\left(x-\epsilon_{m,n}\right)\right\} \\ &= \frac{\epsilon_{m,n}}{2}\left\{\sum_{k=1}^{n}\mathbb{I}\left(X_{k,n}-\epsilon_{m,n}\leq x\leq X_{k,n}+\epsilon_{m,n}\right)\right. \\ &\qquad \left.+\sum_{k=1}^{m}\mathbb{I}\left(Y_{k,m}-\epsilon_{m,n}\leq x\leq Y_{k,m}+\epsilon_{m,n}\right)\right\}.\end{aligned}$$

where $\epsilon_{m,n}=O(1/\sqrt{m+n})$. In that case, one can show that $\sup_{x\in\mathbb{R}}|\hat{f}_0(x)-f_0(x)|\to 0$ in probability.

5.7.2 Computations involving $\hat{f}_0$

Let $a\wedge b=\min(a,b)$ and $a\vee b=\max(a,b)$. If $\hat{f}_0$ is the simple density estimator described in 5.7.1, one can show that

$$\begin{aligned}R_1(r) &= \int_{\mathbb{R}}\mathbb{I}\left(r\leq x\right)\hat{f}_0(x)\,\mathrm{d}x \\ &= \frac{\epsilon_{m,n}}{2}\left[\sum_{k=1}^{n}\left\{X_{k,n}+\epsilon_{m,n}-\left(X_{k,n}-\epsilon_{m,n}\vee r\right)\right\}\mathbb{I}\left(r<X_{k,n}+\epsilon_{m,n}\right)\right. \\ &\qquad \left.+\sum_{k=1}^{m}\left\{Y_{k,m}+\epsilon_{m,n}-\left(Y_{k,n}-\epsilon_{m,n}\vee r\right)\right\}\mathbb{I}\left(r<Y_{k,m}+\epsilon_{m,n}\right)\right]\end{aligned}$$

and

$$\begin{aligned}R_2(r) &= \int_{\mathbb{R}}\mathbb{I}\left(r\leq x\right)x\hat{f}_0(x)\,\mathrm{d}x \\ &= \frac{\epsilon_{m,n}}{4}\left[\sum_{k=1}^{n}\left\{\left(X_{k,n}+\epsilon_{m,n}\right)^2-\left(X_{k,n}-\epsilon_{m,n}\vee r\right)^2\right\}\mathbb{I}\left(r<X_{k,n}+\epsilon_{m,n}\right)\right. \\ &\qquad \left.+\sum_{k=1}^{m}\left\{\left(Y_{k,m}+\epsilon_{m,n}\right)^2-\left(Y_{k,m}-\epsilon_{m,n}\vee r\right)^2\right\}\mathbb{I}\left(r<Y_{k,m}+\epsilon_{m,n}\right)\right].\end{aligned}$$

Moreover, let

$$R_3 = \int_{\mathbb{R}} \left\{\hat{f}_0(x)\right\}^2 \mathrm{d}x, \quad R_4 = \int_{\mathbb{R}} x \left\{\hat{f}_0(x)\right\}^2 \mathrm{d}x \quad \text{and} \quad R_5 = \int_{\mathbb{R}} x^2 \left\{\hat{f}_0(x)\right\}^2 \mathrm{d}x,$$

and introduce the notation

$$\begin{aligned}
M^{\vee}_{XX,k\ell} &= \max(X_{k,n}, X_{\ell,n}), \quad M^{\wedge}_{XX,k\ell} = \min(X_{k,n}, X_{\ell,n}), \\
M^{\vee}_{XY,k\ell} &= \max(X_{k,n}, Y_{\ell,m}), \quad M^{\wedge}_{XY,k\ell} = \min(X_{k,n}, Y_{\ell,m}), \\
M^{\vee}_{YY,k\ell} &= \max(Y_{k,m}, Y_{\ell,m}), \quad M^{\wedge}_{YY,k\ell} = \min(Y_{k,m}, Y_{\ell,m}).
\end{aligned}$$

One can then write

$$\begin{aligned}
4(m+n)\left\{\hat{f}_0(x)\right\}^2 &= \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathbb{I}\left(M^{\vee}_{XX,k\ell} - \epsilon_{m,n} \leq x \leq M^{\wedge}_{XX,k\ell} + \epsilon_{m,n}\right) \\
&\quad + 2\sum_{k=1}^{n}\sum_{\ell=1}^{m} \mathbb{I}\left(M^{\vee}_{XY,k\ell} - \epsilon_{m,n} \leq x \leq M^{\wedge}_{XY,k\ell} + \epsilon_{m,n}\right) \\
&\quad + \sum_{k=1}^{m}\sum_{\ell=1}^{m} \mathbb{I}\left(M^{\vee}_{YY,k\ell} - \epsilon_{m,n} \leq x \leq M^{\wedge}_{YY,k\ell} + \epsilon_{m,n}\right).
\end{aligned}$$

It follows by simple computations that

$$\begin{aligned}
4(m+n)R_3 &= \sum_{k=1}^{n}\sum_{\ell=1}^{n} \max\left(M^{\wedge}_{XX,k\ell} - M^{\vee}_{XX,k\ell} + 2\epsilon_{m,n}, 0\right) \\
&\quad + 2\sum_{k=1}^{n}\sum_{\ell=1}^{m} \max\left(M^{\wedge}_{XY,k\ell} - M^{\vee}_{XY,k\ell} + 2\epsilon_{m,n}, 0\right) \\
&\quad + \sum_{k=1}^{m}\sum_{\ell=1}^{m} \max\left(M^{\wedge}_{YY,k\ell} - M^{\vee}_{YY,k\ell} + 2\epsilon_{m,n}, 0\right).
\end{aligned}$$

Moreover,

$8(m+n)R_4$

$$
\begin{aligned}
= & \sum_{k=1}^{n}\sum_{\ell=1}^{n}\left\{\left(M^{\wedge}_{XX,k\ell}+\epsilon_{m,n}\right)^2-\left(M^{\vee}_{XX,k\ell}-\epsilon_{m,n}\right)^2\right\} \\
& \qquad \times \mathbb{I}\left(M^{\vee}_{XX,k\ell}-M^{\wedge}_{XX,k\ell}<2\epsilon_{m,n}\right) \\
& +2\sum_{k=1}^{n}\sum_{\ell=1}^{m}\left\{\left(M^{\wedge}_{XY,k\ell}+\epsilon_{m,n}\right)^2-\left(M^{\vee}_{XY,k\ell}-\epsilon_{m,n}\right)^2\right\} \\
& \qquad \times \mathbb{I}\left(M^{\vee}_{XY,k\ell}-M^{\wedge}_{XY,k\ell}<2\epsilon_{m,n}\right) \\
& +\sum_{k=1}^{m}\sum_{\ell=1}^{m}\left\{\left(M^{\wedge}_{YY,k\ell}+\epsilon_{m,n}\right)^2-\left(M^{\vee}_{YY,k\ell}-\epsilon_{m,n}\right)^2\right\} \\
& \qquad \times \mathbb{I}\left(M^{\vee}_{YY,k\ell}-M^{\wedge}_{YY,k\ell}<2\epsilon_{m,n}\right)
\end{aligned}
$$

and

$12(m+n)R_5$

$$
\begin{aligned}
= & \sum_{k=1}^{n}\sum_{\ell=1}^{n}\left\{\left(M^{\wedge}_{XX,k\ell}+\epsilon_{m,n}\right)^3-\left(M^{\vee}_{XX,k\ell}-\epsilon_{m,n}\right)^3\right\} \\
& \qquad \times \mathbb{I}\left(M^{\vee}_{XX,k\ell}-M^{\wedge}_{XX,k\ell}<2\epsilon_{m,n}\right) \\
& +2\sum_{k=1}^{n}\sum_{\ell=1}^{m}\left\{\left(M^{\wedge}_{XY,k\ell}+\epsilon_{m,n}\right)^3-\left(M^{\vee}_{XY,k\ell}-\epsilon_{m,n}\right)^3\right\} \\
& \qquad \times \mathbb{I}\left(M^{\vee}_{XY,k\ell}-M^{\wedge}_{XY,k\ell}<2\epsilon_{m,n}\right) \\
& +\sum_{k=1}^{m}\sum_{\ell=1}^{m}\left\{\left(M^{\wedge}_{YY,k\ell}+\epsilon_{m,n}\right)^3-\left(M^{\vee}_{YY,k\ell}-\epsilon_{m,n}\right)^3\right\} \\
& \qquad \times \mathbb{I}\left(M^{\vee}_{YY,k\ell}-M^{\wedge}_{YY,k\ell}<2\epsilon_{m,n}\right).
\end{aligned}
$$

5.7.3 Computation of A, B and C

One has

$$\begin{aligned}
\mathbf{A}_{ij}(x) &= A_i(x)A_j(x) \\
&= \mathbb{I}\left(X_{i,n} \vee X_{j,n} \leq x\right) + \mathbb{I}\left(X_{i,n} \leq x\right) \hat{f}_0(x) \left(\frac{X_{j,n}^2\, x}{2} + X_{j,n}\right) \\
&\quad + \mathbb{I}\left(X_{j,n} \leq x\right) \hat{f}_0(x) \left(\frac{X_{i,n}^2\, x}{2} + X_{i,n}\right) \\
&\quad + \left\{\hat{f}_0(x)\right\}^2 \left(\frac{X_{i,n}^2\, x}{2} + X_{i,n}\right) \left(\frac{X_{j,n}^2\, x}{2} + X_{j,n}\right),
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{A}_{ij} &= \lim_{M\to\infty} \int_{-\infty}^{M} \mathbf{A}_{ij}(x)\, \mathrm{d}x \\
&= \lim_{M\to\infty} \left\{M - (X_{i,n} \vee X_{j,n})\right\} + \frac{X_{j,n}^2}{2} R_2(X_{i,n}) + X_{j,n} R_1(X_{i,n}) \\
&\quad + \frac{X_{i,n}^2}{2} R_2(X_{j,n}) + X_{i,n} R_1(X_{j,n}) + \frac{X_{i,n}^2 X_{j,n}^2}{4} R_5 + \frac{X_{i,n} X_{j,n}^2}{2} R_4 \\
&\quad + \frac{X_{i,n}^2 X_{j,n}}{2} R_4 + X_{i,n} X_{j,n} R_3.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{B}_{ij} &= \lim_{M\to\infty} \int_{-\infty}^{M} \mathbf{B}_{ij}(x)\, \mathrm{d}x \\
&= \lim_{M\to\infty} \left\{M - (Y_{i,m} \vee Y_{j,m})\right\} + \frac{Y_{j,m}^2}{2} R_2(Y_{i,m}) + Y_{j,m} R_1(Y_{i,m}) \\
&\quad + \frac{Y_{i,m}^2}{2} R_2(Y_{j,m}) + Y_{i,m} R_1(Y_{j,m}) + \frac{Y_{i,m}^2 Y_{j,m}^2}{4} R_5 + \frac{Y_{i,m} Y_{j,m}^2}{2} R_4 \\
&\quad + \frac{Y_{i,m}^2 Y_{j,m}}{2} R_4 + Y_{i,m} Y_{j,m} R_3
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{C}_{ij} &= \lim_{M\to\infty}\int_{-\infty}^{M} \mathbf{C}_{ij}(x)\,\mathrm{d}x \\
&= \lim_{M\to\infty}\{M-(X_{i,n}\vee Y_{j,m})\} + \frac{Y_{j,m}^2}{2}\,R_2(X_{i,n}) + Y_{j,m}\,R_1(X_{i,n}) \\
&\quad + \frac{X_{i,n}^2}{2}\,R_2(Y_{j,m}) + X_{i,n}\,R_1(Y_{j,m}) + \frac{X_{i,n}^2 Y_{i,m}^2}{4}\,R_5 + \frac{X_{i,n}^2 Y_{i,m}}{2}\,R_4 \\
&\quad + \frac{Y_{j,m}^2 X_{i,n}}{2}\,R_4 + X_{i,n}Y_{j,m}\,R_3.
\end{aligned}$$

Table 5.2: Percentage of rejection of $\mathcal{H}_0$, as evaluated from 10 000 replicates, in the case of dependent samples of size $n = 100$, using $M = 500$ copies from the bootstrap strategies I and II

Kendall's	Margins		Clayton		Normal		Frank	
tau	F	G	I	II	I	II	I	II
	N	N	0.0344	0.1107	0.0303	0.0928	0.0290	0.0775
	N	DE	0.2721	0.7219	0.2745	0.6995	0.3244	0.6510
$\tau = -2/3$	N	L	0.0630	0.1977	0.0598	0.1865	0.0646	0.1875
	DE	DE	0.0633	0.1508	0.0553	0.1430	0.0547	0.1230
	DE	L	0.1046	0.1419	0.1057	0.1442	0.1226	0.1242
	L	L	0.0443	0.1231	0.0477	0.1066	0.0439	0.0956
	N	N	0.0326	0.0612	0.0314	0.0584	0.0334	0.0566
	N	DE	0.3844	0.6124	0.3782	0.6230	0.3985	0.6090
$\tau = -1/3$	N	L	0.0832	0.1670	0.0778	0.1689	0.0836	0.1600
	DE	DE	0.0577	0.1099	0.0520	0.1094	0.0554	0.0958
	DE	L	0.1511	0.1176	0.1413	0.1198	0.1579	0.1059
	L	L	0.0519	0.0819	0.0473	0.0816	0.0525	0.0732
	N	N	0.0617	0.0403	0.0610	0.0498	0.0556	0.0490
	N	DE	0.5695	0.6334	0.5355	0.6280	0.5008	0.6221
$\tau = 1/3$	N	L	0.1446	0.1390	0.1374	0.1586	0.1392	0.1651
	DE	DE	0.0826	0.0625	0.0765	0.0724	0.0766	0.0824
	DE	L	0.2557	0.0860	0.2516	0.0959	0.2253	0.0990
	L	L	0.0731	0.0523	0.0797	0.0663	0.0748	0.0723
	N	N	0.0767	0.0201	0.0840	0.0155	0.0790	0.0246
	N	DE	0.7744	0.7138	0.8121	0.7324	0.6665	0.6704
$\tau = 2/3$	N	L	0.2218	0.1196	0.2343	0.1119	0.2053	0.1425
	DE	DE	0.1013	0.0400	0.1055	0.0254	0.0973	0.0662
	DE	L	0.3813	0.0639	0.4499	0.0553	0.3131	0.0823
	L	L	0.1030	0.0306	0.1027	0.0164	0.1041	0.0510

CHAPITRE 6

CONCLUSION

La plupart des tests d'égalité de lois s'appliquent dans un contexte où l'information provient d'échantillons indépendants. Le cas de données pairées a été peu exploré dans la littérature scientifique. Dans ce mémoire, plusieurs nouveaux tests applicables pour des données dépendantes ont été développés. On a d'abord obtenu la convergence en loi d'un processus empirique qui sous-tend chacune des statistiques de test proposées. Ensuite, on a adapté judicieusement la méthode du multiplicateur afin de calculer les p-valeurs des tests. Les études de simulation qui ont été effectuées indiquent clairement que la méthodologie utilisée est excellente, puisque les tests conservent leur seuil nominal sous l'hypothèse nulle. De plus, les tests montrent des puissances élevées sous la plupart des hypothèses alternatives qui ont été considérées. Par surcroît, on a découvert des formules élégantes pour les statistiques de test et leurs versions obtenues par la méthode du multiplicateur, ce qui facilite grandement leur implémentation et améliore la vitesse de calcul.

Mes recherches à la maîtrise se sont également attaqué à un problème plus général que l'égalité stricte en loi, à savoir l'égalité de distributions à une

transformation affine près. Comme la moyenne et la variance sont inconnues, ils ont dû être estimés; cela complique le travail pour l'obtention du comportement asymptotique du processus. La méthode du multiplicateur dans ce cas doit être modifiée afin de tenir compte de l'information manquante sur les moyennes et les variances. À ce titre, deux méthodes de ré-échantillonnage ont été développées. Les résultats de simulation montrent que les tests conservent assez bien leur seuil nominal sous $\mathcal{H}_0$.

L'idée de tester l'égalité en loi de variables aléatoires standardisées pourrait être généralisée. D'abord, notons que l'égalité en loi de X_1 et X_2, à paramètres de moyenne et variance près, revient à supposer que

$$X_1 = \mu_1 + \sigma_1 \epsilon_1 \quad \text{et} \quad X_2 = \mu_2 + \sigma_2 \epsilon_2,$$

où $\epsilon_1 \stackrel{d}{=} \epsilon_2$. De cette façon,

$$\frac{X_1 - \mu_1}{\sigma_1} \stackrel{d}{=} \frac{X_2 - \mu_2}{\sigma_2}.$$

On peut étendre cette idée en supposant que les moyennes et les variances dépendent d'un vecteur de variables explicatives $\mathbf{Z} = (Z_1, \ldots, Z_d)$ tel que

$$X_1 = \mu_1(\mathbf{Z}) + \sigma_1(\mathbf{Z})\epsilon_1 \quad \text{et} \quad X_2 = \mu_2(\mathbf{Z}) + \sigma_2(\mathbf{Z})\epsilon_2,$$

où $\epsilon_1 \stackrel{d}{=} \epsilon_2$. Le problème est cependant complexe, car il faut estimer les fonctions μ_j et σ_j, $j = 1, 2$. Pour cela, deux avenues sont possibles : (i) supposer une forme paramétrique et estimer les paramètres inconnus ou (ii) utiliser des estimateurs non-paramétriques à noyaux. Dans les deux cas, le calcul de p-valeurs serait basé sur des adaptations de la méthode du multiplicateur.

Références

Alba Fernández, V., Jiménez Gamero, M. D. & Muñoz García, J. (2008). A test for the two-sample problem based on empirical characteristic functions. *Comput. Statist. Data Anal.* **52**, 3730–3748.

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *Ann. Math. Statist.* **33**, 1148–1159.

Bai, J. (1994). Weak convergence of the sequential empirical processes of residuals in ARMA models. *Ann. Statist.* **22**, 2051–2061.

Bai, J. & Ng, S. (2001). A consistent test for conditional symmetry in time series models. *J. Econometrics* **103**, 225–258. Studies in estimation and testing.

Bickel, P. J. (1968). A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Statist.* **40**, 1–23.

Billingsley, P. (1999). *Convergence of probability measures.* Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc., 2nd ed. A Wiley-Interscience Publication.

BÜNING, H. (2002). Robustness and power of modified Lepage, Kolmogorov-Smirnov and Cramér-von Mises two-sample tests. *J. Appl. Stat.* **29**, 907–924.

BURKE, M. D. (2000). Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap. *Statist. Probab. Lett.* **46**, 13–20.

CASELLA, G. & BERGER, R. L. (1990). *Statistical inference.* The Wadsworth & Brooks/Cole Statistics/Probability Series. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

COOK, R. D. & JOHNSON, M. E. (1986). Generalized Burr-Pareto-logistic distributions with applications to a uranium exploration data set. *Technometrics* **28**, 123–131.

EPPS, T. W. & SINGLETON, K. J. (1986). An omnibus test for the two-sample problem using the empirical characteristic function. *J. Statist. Comput. Simulation* **26**, 177–203.

FERMANIAN, J.-D. (2005). Goodness-of-fit tests for copulas. *J. Multivariate Anal.* **95**, 119–152.

GENEST, C., GHOUDI, K. & RIVEST, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**, 543–552.

GENEST, C., QUESSY, J.-F. & RÉMILLARD, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Statist.* **33**, 337–366.

GENEST, C., RÉMILLARD, B. & BEAUDOIN, D. (2009). Goodness-of-fit tests for copulas: a review and a power study. *Insurance Math. Econom.* **44**, 199–213.

GHOUDI, K. & RÉMILLARD, B. (1998). Empirical processes based on pseudo-observations. In *Asymptotic methods in probability and statistics (Ottawa, ON, 1997)*. Amsterdam: North-Holland, pp. 171–197.

GHOUDI, K. & RÉMILLARD, B. (2004). Empirical processes based on pseudo-observations. II. The multivariate case. In *Asymptotic methods in stochastics*, vol. 44 of *Fields Inst. Commun.* Providence, RI: Amer. Math. Soc., pp. 381–406.

GOMBAY, E. & HORVÁTH, L. (2002). Rates of convergence for U-statistic processes and their bootstrapped versions. *J. Statist. Plann. Inference* **102**, 247–272. Silver jubilee issue.

GREENWELL, R. N. & FINCH, S. J. (2004). Randomized rejection procedure for the two-sample Kolmogorov-Smirnov statistic. *Comput. Statist. Data Anal.* **46**, 257–267.

KHOUDRAJI, A. (1995). *Contributions à l'étude des copules et à la modélisation de valeurs extrêmes bivariées.* Ph.D. thesis, Université Laval, Québec, Canada.

KOJADINOVIC, I. & YAN, J. (2010). Nonparametric rank-based tests of bivariate extreme-value dependence. *Journal of Multivariate Analysis* .

KOSOROK, M. (2008). *Introduction to empirical processes and semiparametric inference.* New York: Springer.

LEE, A. J. (1990). *U-statistics*, vol. 110 of *Statistics: Textbooks and Monographs*. New York: Marcel Dekker Inc. Theory and practice.

LOYNES, R. M. (1980). The empirical distribution function of residuals from generalised regression. *Ann. Statist.* **8**, 285–298.

NELSEN, R. B. (2006). *An introduction to copulas*. Springer Series in Statistics. New York: Springer, 2nd ed.

ODIASE, J. I. & OGBONMWAN, S. M. (2007). Exact permutation algorithm for paired observations: the challenge of R. A. Fisher. *J. Math. Stat.* **3**, 116–121.

PARDO-FERNÁNDEZ, J. C. (2007). Comparison of error distributions in nonparametric regression. *Statist. Probab. Lett.* **77**, 350–356.

PIERCE, D. A. & KOPECKY, K. J. (1979). Testing goodness of fit for the distribution of errors in regression models. *Biometrika* **66**, 1–5.

PRÆSTGAARD, J. T. (1995). Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions. *Scand. J. Statist.* **22**, 305–322.

QUESSY, J.-F. (2011). Testing for bivariate extreme dependence using Kendall's process. *Scandinavian Journal of Statistics* .

RASCH, D., TEUSCHER, F. & GUIARD, V. (2007). How robust are tests for two independent samples? *J. Statist. Plann. Inference* **137**, 2706–2720.

REICZIGEL, J., ZAKARIÁS, I. & RÓZSA, L. (2005). A bootstrap test of stochastic equality of two populations. *Amer. Statist.* **59**, 156–161.

RÉMILLARD, B. & SCAILLET, O. (2009). Testing for equality between two copulas. *J. Multivariate Anal.* **100**, 377–386.

SCAILLET, O. (2005). A Kolmogorov-Smirnov type test for positive quadrant dependence. *Canad. J. Statist.* **33**, 415–427.

SCHMID, F. & TREDE, M. (1995). A distribution free test for the two sample problem for general alternatives. *Computational Statistics and Data Analysis* **20**, 409 – 419.

SCHRÖER, G. & TRENKLER, D. (1995). Exact and randomization distributions of Kolmogorov-Smirnov tests: two or three samples. *Comput. Statist. Data Anal.* **20**, 185–202.

SHIH, J. H. & LOUIS, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384–1399.

SKLAR, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**, 229–231.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15**, 88.

VAN DER VAART, A. W. (1998). *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. New York: Springer-Verlag. With applications to statistics.

VAN DER VAART, A. W. & WELLNER, J. A. (2007). Empirical processes indexed by estimated functions. In *Asymptotics: particles, processes and inverse problems*, vol. 55 of *IMS Lecture Notes Monogr. Ser.* Beachwood, OH: Inst. Math. Statist., pp. 234–252.

WELCH, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350–362.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 80–93.

YUEN, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika* **61**, 165–170.