

JEAN-CLAUDE BOIES

**UNE MÉTHODE GRAPHIQUE DE DÉTECTION DE LA
DÉPENDANCE**

**Mémoire
présenté
à la Faculté des études supérieures
de l'Université Laval
pour l'obtention
du grade de maître ès sciences (M. Sc.)**

**Département de mathématiques et de statistique
FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL**

Octobre 2003

© Jean-Claude Boies, 2003

RÉSUMÉ

Fisher et Switzer (1985, 2001) ont proposé une méthode graphique appelée “khi-plot,” laquelle permet de détecter la présence d’association entre deux variables continues à partir d’un échantillon aléatoire. Dans ce mémoire, l’auteur décrit une méthode concurrente, le “K-plot” ou “Kendall plot,” qui à l’instar du khi-plot, appuie sa représentation graphique sur les rangs des observations. Parce que la courbure qu’il produit en cas de dépendance est associée de près à la copule sous-jacente à la loi des observations, le K-plot s’avère plus facile à interpréter que le khi-plot. Il a de plus l’avantage de se généraliser aisément au cas de plusieurs variables.

Jean-Claude Boies, étudiant

Christian Genest, directeur

AVANT-PROPOS

Je souhaite tout d'abord exprimer ma gratitude envers mon directeur, M. Christian Genest, pour l'aide qu'il m'a apportée au cours de ce projet. Ma collaboration avec lui fut très enrichissante et fort agréable.

Je désire également remercier ma famille pour le soutien moral qu'elle m'a accordé durant mes études. Son appui m'a été très secourable et je lui en suis reconnaissant. Je remercie aussi M. Jean-François Plante, qui m'a prodigué de précieux conseils.

Une partie de ce travail a été financée par des octrois individuels et collectifs accordés au professeur Genest et à l'équipe du *Fond québécois de la recherche sur la nature et les technologies* qu'il coordonne. Je sais gré aux membres du groupe pour leur aide financière, ainsi qu'au *Conseil de recherches en sciences naturelles et en génie du Canada*, pour la bourse d'études supérieures qu'il m'a octroyée.

TABLE DES MATIÈRES

RÉSUMÉ	ii
AVANT-PROPOS	iii
LISTE DES FIGURES	vi
INTRODUCTION	1
CHAPITRE I. DETECTING DEPENDENCE WITH KENDALL	
PLOTS	3
1.1 Introduction	3
1.2 Chi-plots	7
1.3 K-plots	9
1.4 Properties of K and K-plots	12
1.4.1 Properties of K	12
1.4.2 Properties of K-plots	13
1.5 Examples of K-plots	15
1.5.1 Artificial data	15
1.5.2 Real data	20
1.6 Extensions	21
1.7 Conclusion	25
CHAPITRE II. EXEMPLES SUPPLÉMENTAIRES	27
2.1 Une application au baseball	27

2.2	Une application au hockey	30
2.3	Une application à caractère social	32
2.4	Une application à caractère médical	34
CHAPITRE III. UN RÉSULTAT DE CONVERGENCE		37
CONCLUSION		39
ANNEXE		41
BIBLIOGRAPHIE		65

LISTE DES FIGURES

1.1	Two scatterplots for a random sample of a pair of independent exponential random variables	5
1.2	Chi-plot for a random sample of a pair of independent exponential random variables	6
1.3	K-plot for a random sample of a pair of independent exponential random variables	11
1.4	K-plots for random samples of comonotonic uniform random variables	15
1.5	K-plots for random samples of bivariate normal distributions with different degrees of association	17
1.6	K-plots for random samples of a bivariate distribution with underlying Clayton copula and with different degrees of association	18
1.7	Rank scatterplot, chi-plot, and K-plot for an automobile exhaust data set	21
1.8	Rank scatterplot, chi-plot, and K-plot for a fish population data set	22
1.9	K-plots for random samples from a trivariate uniform distribution with comonotonic components	24
1.10	K-plots for a random sample from a triple of normal variables that are pairwise independent but not mutually independent	25

2.1	Graphiques mettant en relation les variables AVG, BBAB et SOAB deux à deux	28
2.2	Application des K-plots sur les variables AVG, BBAB et SOAB	28
2.3	Graphiques mettant en relation les variables AVG, BBAB et AGE deux à deux	29
2.4	Application des K-plots sur les variables AVG, BBAB et AGE	30
2.5	PIMPG versus PTSPG, et rangs(PIMPG) versus rangs(PTSPG)	31
2.6	K-plot illustrant la structure de dépendance entre PIMPG et PTSPG	31
2.7	Matrice de diagrammes de dispersion et de K-plots pour cinq variables sociales mesurées à Chicago	33
2.8	Khi-plots et K-plots de toutes les paires de variables biochimiques provenant de Reaven et Miller (1979).	36

INTRODUCTION

Une manière simple d'examiner la relation entre deux variables, observées simultanément un certain nombre de fois, consiste à représenter les paires de données par des points sur un plan cartésien. Le nuage qui en résulte peut alors fournir de précieux indices quant à la nature et à l'ampleur de la relation entre la première variable, en abscisse, et la seconde, en ordonnée.

Il peut cependant arriver que la relation de dépendance entre deux variables ne soit pas facilement identifiable à l'œil. Ce mémoire présente un outil graphique, nommé K-plot ou "Kendall plot," qui peut alors s'avérer utile pour détecter la présence d'association ou, le cas échéant, en confirmer l'absence.

Confrontés à ce problème, Fisher et Switzer (1985, 2001) avaient déjà proposé une méthode graphique appelée "khi-plot." Leur technique s'appuyait exclusivement sur les couples de rangs des observations et était donc non paramétrique par nature. Ceci se justifiait du fait que la structure de dépendance entre deux variables continues est caractérisée par la copule sous-jacente au modèle et que les paires de rangs sont des statistiques maximumment invariantes par rapport à toute transformation monotone croissante des marges.

Les K-plots proposés ici sont eux aussi fonctions des rangs des observations, mais alors que les khi-plots sont inspirés de la statistique du khi-deux d'indépendance et sont apparentés à la notion de carte de contrôle, les K-plots sont issus de la transformation intégrale de probabilité multivariée et

découlent, sur le plan graphique, du principe de la droite de Henry, ou “Q–Q plot.” Ainsi l’absence d’indépendance se manifeste-t-elle dans le K-plot par la présence d’une courbure caractéristique de la copule sous-jacente au modèle.

Après une brève introduction, le chapitre I rappelle la notion de khi-plot et présente le concept nouveau de K-plot. Le texte, rédigé en langue anglaise, paraîtra intégralement sous peu dans la revue *The American Statistician*. On y décrit les fondements théoriques des deux méthodes, ainsi que leur motivation; on en illustre aussi les principales propriétés au moyen de simulations et d’applications concrètes. Quelques exemples complémentaires sont fournis au chapitre II. Quant au chapitre III, il renferme une démonstration détaillée d’un résultat énoncé dans l’article (c’est-à-dire au chapitre I). Une courte conclusion est donnée au chapitre IV.

CHAPITRE I

DETECTING DEPENDENCE WITH KENDALL PLOTS

Christian GENEST and Jean-Claude BOIES

Earlier literature proposed a rank-based graphical tool called a chi-plot which, in conjunction with a traditional scatterplot of the raw data, can help detect the presence of association in a random sample from some continuous bivariate distribution. In this note, an alternative display called a Kendall plot, or K-plot for short, is suggested which adapts the concept of probability plot to the detection of dependence. The new procedure, which is rooted in the probability integral transformation, retains the chi-plot's key property of invariance with respect to monotone transformations of the marginal distributions. K-plots are easier to interpret than chi-plots, however, because the curvature that they display in cases of association is related in a definite way to the copula characterizing the underlying dependence structure. In addition, K-plots have the advantage of being readily extendible to the multivariate context.

1.1 Introduction

Given a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of size $n \geq 2$ from a continuous bivariate cumulative distribution function H with margins F and G ,

the simplest possible graphical means of assessing whether the variables X and Y are stochastically independent or not is to draw a scatterplot of the observations.

The left panel of Figure 1.1 provides an illustration based on $n = 100$ pseudo-random pairs from independent exponential distributions $F(x) = 1 - \exp(-2x)$ and $G(y) = 1 - \exp(-10y)$. The right panel displays the pairs $(F(X_i), G(Y_i))$, $1 \leq i \leq n$, which result from a probability integral transformation of the components in order to make them uniform on the interval $[0, 1]$. As can be seen readily, the form of the graph is contingent on the margins, and this can interfere with the analyst's assessment of the degree of association (or absence thereof) in the data set.

Although the marginal distributions are typically unknown, a simple way to control their influence is to plot the pairs $(\hat{F}_n(X_i), \hat{G}_n(Y_i))$, $1 \leq i \leq n$, where \hat{F}_n and \hat{G}_n are the empirical distribution functions of the X_i and the Y_i , respectively. Note that since

$$\hat{F}_n(t) = \frac{1}{n} \# \{i : X_i \leq t\} \quad \text{and} \quad \hat{G}_n(t) = \frac{1}{n} \# \{i : Y_i \leq t\},$$

this amounts to plotting the pairs $(R_i/n, S_i/n)$, where R_i stands for the rank of X_i in the set $\{X_1, \dots, X_n\}$, and S_i is the rank of Y_i in the set $\{Y_1, \dots, Y_n\}$. Because ranks are maximally invariant under monotone transformations of the marginal distributions, the transformed data may be regarded as observations from the unique underlying copula

$$C(u, v) = H \{F^{-1}(u), G^{-1}(v)\}, \quad 0 \leq u, v \leq 1$$

associated with H (see, for example, Sklar 1959 or Nelsen 1999). Since $H = F \times G$ occurs if and only if $C(u, v) = uv$ on its entire domain, no

loss of information ensues from the rank transformation, but the graph is of somewhat limited use in assessing the null hypothesis of independence, because randomness is a difficult characteristic for the human eye to judge.

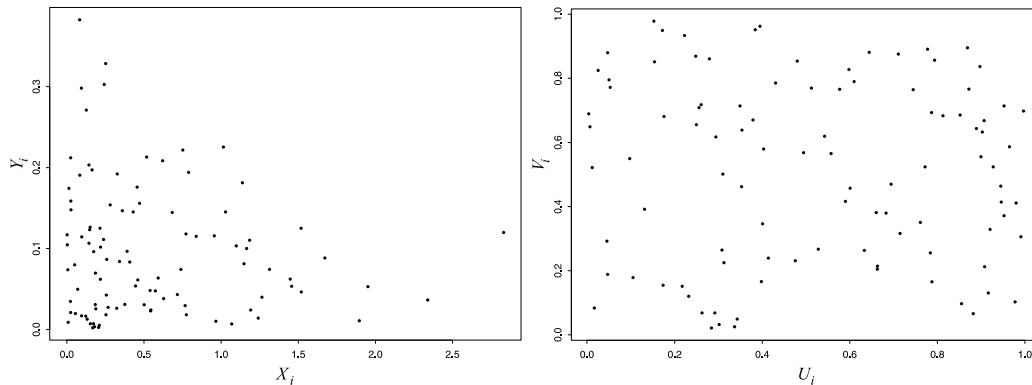


Figure 1.1: Left panel: Scatter plot of $n = 100$ pairs (X_i, Y_i) from independent exponential random variables X_i and Y_i with means $E(X_i) = 1/2$ and $E(Y_i) = 1/10$. Right panel: Scatter plot of the same sample, upon transformation of the data to make the margins uniform on the interval $[0, 1]$.

Motivated by the need for a graphical method in which independence manifests itself in a more characteristic fashion than in scatterplots, Fisher and Switzer (1985, 2001) introduced chi-plots, whose definition and properties are briefly recalled in Section 2. These graphs, which only depend on the data through their ranks, produce diagrams that are approximately horizontal under independence. Such is the case in Figure 1.2, for example, which shows the chi-plot associated with the same data as in Figure 1.1. Nearly all the points fall in between two lines that play a similar role to control limits in an \bar{X} -chart.

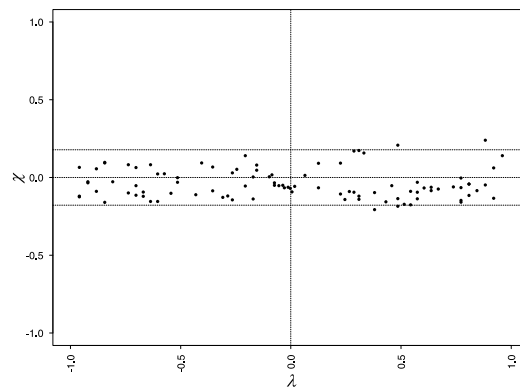


Figure 1.2: Chi-plot of the same pairs $(X_1, Y_1), \dots, (X_{100}, Y_{100})$ of independent exponential random variables as in Figure 1.1.

The purpose of this article is to describe an alternative rank-based procedure which adapts the familiar concept of probability plot (Wilk and Gnanadesikan 1968) to the detection of dependence. Just as lack of linearity is a sign of non-normality in a standard Q–Q plot, the amount of curvature in the proposed graph is characteristic of the degree of association in the data. The method, which is successively detailed, studied, and illustrated in Sections 3–5, is rooted in the probability integral transformation, and closely related to Kendall’s tau statistic. For this reason, this specific type of graphical display might be referred to as a Kendall plot, or as a K-plot for short. As with a chi-plot, the new technique can be adapted easily to time-series contexts and to situations where one of the variables is non-stochastic, but as explained in Section 6, it has the additional advantage of extending readily to the multivariate case.

1.2 Chi-plots

For a given pair (X_i, Y_i) with $1 \leq i \leq n$, let

$$H_i = \frac{1}{n-1} \# \{j \neq i : X_j \leq X_i, Y_j \leq Y_i\}, \quad (1)$$

and

$$F_i = \frac{1}{n-1} \# \{j \neq i : X_j \leq X_i\}, \quad G_i = \frac{1}{n-1} \# \{j \neq i : Y_j \leq Y_i\}.$$

Under independence, one would expect to have $H_i = F_i \times G_i$, up to sampling variation. Accordingly, Fisher and Switzer (1985, 2001) proposed to plot the pairs (λ_i, χ_i) , where

$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}} \quad (2)$$

and

$$\lambda_i = 4 \operatorname{sign} \left(\tilde{F}_i \tilde{G}_i \right) \max \left(\tilde{F}_i^2, \tilde{G}_i^2 \right),$$

where $\tilde{F}_i = F_i - 1/2$ and $\tilde{G}_i = G_i - 1/2$ for $1 \leq i \leq n$. The resulting graph is what they call a chi-plot.

Here, $\lambda_i \in [-1, 1]$ is a measure of the distance of the pair (X_i, Y_i) from the center of the data set, and to avoid spurious observations, the authors recommend that only pairs for which $|\lambda_i| < 4\{1/(n-1) - 1/2\}^2$ be plotted. As for the right-hand side of (2), it may be recognized as the correlation coefficient that would be associated with the $n-1$ pairs (X_{ij}, Y_{ij}) of dichotomous random variables derived from the original sample by fixing both X_i and Y_i , and then setting

$$X_{ij} = \begin{cases} 1 & \text{if } X_j \leq X_i, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad Y_{ij} = \begin{cases} 1 & \text{if } Y_j \leq Y_i, \\ 0 & \text{otherwise,} \end{cases}$$

for all $j \neq i$. Thus $-1 \leq \chi_i \leq 1$ for $1 \leq i \leq n$. Furthermore, $\sqrt{n}\chi_i$ is the signed square root of the chi-square statistic that would typically be used for testing independence in the 2×2 contingency table generated by the cut-point (X_i, Y_i) .

Figure 1.2 shows the chi-plot associated with the 100 pairs of independent exponential random variables plotted in Figure 1.1. To help assess random variation in the observed values of χ , horizontal guidelines are superimposed on the plot, as recommended by Fisher and Switzer (1985, 2001). These “control limits” are placed at $\chi = \pm c_p/\sqrt{n}$, where c_p is determined by Monte Carlo simulations in such a way that approximately $p \times 100\%$ of the pairs (λ_i, χ_i) lie between the lines. In Figure 1.2, $c_p \approx 1.78$ was used, which corresponds to $p = 0.95$, and this is indeed roughly the proportion of the observations falling within the horizontal strip, as might be expected under independence. The fact that positive and negative values of the λ_i are in roughly equal number is also an indication that the variables X and Y do not have a propensity to be simultaneously either large or small relative to their distribution’s median, as when they are positively associated, or to lie simultaneously on opposite sides of their respective median, as when they are negatively associated.

Chi-plots possess several desirable properties, including an adaptability to time-series contexts and to situations where one of the variables is non-stochastic. As pointed out by Fisher and Switzer (1985), the unscaled numerators of the χ_i are also connected to standard nonparametric tests of independence based on Spearman’s empirical rank-order correlation coefficient, ρ_n , and Kendall’s sample measure of concordance, τ_n . More specifically, one

has

$$\sum_{i=1}^n (H_i - F_i G_i) = \frac{n}{12} \left(3\tau_n - \frac{n+1}{n-1} \rho_n \right).$$

Furthermore, the host of examples given by Fisher and Switzer (1985, 2001) suggests that patterns of dependence observed on chi-plots may be useful in identifying the underlying copula. For correlated bivariate normal data, for instance, positive values of Pearson's correlation, r , yield a frown-like scatter of pairs (λ_i, χ_i) with $\chi_i \approx r_n$ when $\lambda_i \approx 0$. Other copulas yield other sorts of patterns, however, and since the exact connections between the two is far from obvious, the richness of the graphs often proves an impediment to their interpretation.

In the following section, a simpler type of dependence graph is presented, based on an adaptation of the notion of Q-Q plot. As will be seen, the proposed technique retains several of the desirable characteristics of the chi-plot, including its reliance on ranks and its filiation with nonparametric tests of independence. In addition, however, the pattern that it displays in case of association is more directly related to the underlying copula function, and multivariate extensions are straightforward.

1.3 K-plots

When faced with a univariate random sample X_1, \dots, X_n , a common way of assessing its Gaussian character graphically is to draw a Q-Q plot, which consists of pairs $(Z_{i:n}, X_{(i)})$, where $X_{(1)} \leq \dots \leq X_{(n)}$ denote the ordered sample and $Z_{i:n}$ is the i th *normal rankit* associated with a sample of size n ,

that is,

$$Z_{i:n} = \mathbf{E}(Z_{(i)}), \quad 1 \leq i \leq n$$

where $Z_{(1)} \leq \dots \leq Z_{(n)}$ are the order statistics of a random sample Z_1, \dots, Z_n from a standard normal distribution.

Similarly, a visual tool for assessing dependence in a bivariate random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ might thus be constructed as follows:

- a) For each $1 \leq i \leq n$, compute H_i as defined in (1).
- b) Order the H_i to get $H_{(1)} \leq \dots \leq H_{(n)}$; equalities are possible, especially in small samples.
- c) Plot the pairs $(W_{i:n}, H_{(i)})$, $1 \leq i \leq n$, where $W_{i:n}$ represents the expectation of the i th order statistic in a random sample of size n from the distribution K_0 of the H_i under the null hypothesis of independence. For convenience, K_0 is taken to be the *asymptotic* null distribution in the sequel.

The resulting graph, called a Kendall plot, or a K-plot for short, is illustrated in Figure 1.3 for the same random sample of size 100 from independent exponential variables used for Figures 1.1 and 1.2; as can be seen, the lack of association here translates into a nearly straight line.

To complete the description of the procedure, one need only determine the form of K_0 under the null hypothesis of independence. For, by definition of the density of an order statistic, one then has

$$W_{i:n} = n \binom{n-1}{i-1} \int_0^1 w \{K_0(w)\}^{i-1} \{1 - K_0(w)\}^{n-i} dK_0(w) \quad (3)$$

for all $1 \leq i \leq n$.

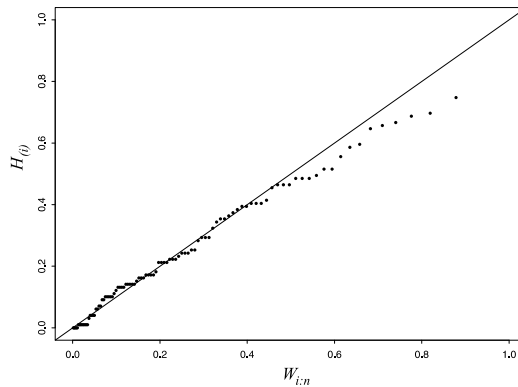


Figure 1.3: K-plot of the same pairs $(X_1, Y_1), \dots, (X_{100}, Y_{100})$ of independent exponential random variables as in Figures 1.1 and 1.2.

Now it follows from Proposition 2.1 in Genest and Rivest (1993) that under mild regularity conditions, the empirical distribution function K_n of the pseudo-observations H_1, \dots, H_n is an asymptotically Gaussian, \sqrt{n} -consistent estimator of

$$K(w) = \mathbb{P}\{H(X, Y) \leq w\}, \quad 0 \leq w \leq 1 \quad (4)$$

a conclusion that may be intuited from the fact that $H_i = \hat{H}_n(X_i, Y_i)$ where \hat{H}_n , the empirical distribution function based on the (X_j, Y_j) , $j \neq i$, converges to H as $n \rightarrow \infty$. Although the result is somewhat delicate to prove, because the H_i are not stochastically independent of each other, this difficulty can be overcome, and a simple calculation shows that when $H = F \times G$,

$$K(w) = K_0(w) = \mathbb{P}(UV \leq w) = w - w \log(w), \quad 0 \leq w \leq 1$$

where U and V are independent uniform random variables on the interval $[0, 1]$. All that remains to do, therefore, is to plug this choice of K into (3) in order to compute the $W_{i:n}$ required for the K-plot.

1.4 Properties of K and \mathbf{K} -plots

1.4.1 Properties of K

Before looking at examples of \mathbf{K} -plots, it is worth listing a few facts about the distribution K that have bearing on the interpretation of the graphs. The following information is excerpted from Genest and Rivest (2001), who survey the literature in the area and give several illustrations.

4.1.1: K is the cumulative distribution function of the random variable $W = H(X, Y)$ obtained through the bivariate probability integral transformation of the random pair (X, Y) with cumulative distribution function H . Since $K(w)$ represents the probability of the event $\{H(X, Y) \leq w\}$, see (4), the distribution puts no mass outside the interval $[0, 1]$.

4.1.2: K depends only on the copula associated with H , and hence not on the margins F and G of H ; indeed, if $H(x, y) = C\{F(x), G(y)\}$ and $\mathbf{1}(A)$ denotes the indicator of the set A , then

$$\begin{aligned} K(w) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{1}\{H(x, y) \leq w\} \, dH(x, y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{1}[C\{F(x), G(y)\} \leq w] \, dC\{F(x), G(y)\} \\ &= \int_0^1 \int_0^1 \mathbf{1}\{C(u, v) \leq w\} \, dC(u, v) \\ &= \mathbf{P}\{C(U, V) \leq w\}, \end{aligned}$$

where (U, V) has cumulative distribution function (and copula) C .

4.1.3: K is a univariate summary of the dependence embodied in C ; this led Capéraà, Fougères and Genest (1997a) to suggest that a stochastic ordering could be defined by declaring a random pair (X, Y) with distribution H less positive dependent than another pair (X^*, Y^*) distributed as H^* , denoted $(X, Y) \prec_K (X^*, Y^*)$, if and only if $K(w) \geq K^*(w)$ for all $0 \leq w \leq 1$, where K and K^* are the cumulative distribution functions of the random variables $H(X, Y)$ and $H^*(X^*, Y^*)$, respectively.

4.1.4: The relation \prec_K orders population values of Kendall's τ , because for a given pair (X, Y) distributed as H ,

$$\tau(X, Y) = 4E\{H(X, Y)\} - 1 = 3 - 4 \int_0^1 K(w) \, dw.$$

Similarly, the estimate τ_n of Kendall's coefficient of concordance is just $4\bar{H} - 1$, with $\bar{H} = (H_1 + \dots + H_n)/n$.

4.1.5: The variables (X, Y) are said to be comonotonic whenever $\tau(X, Y) = \pm 1$, which is equivalent to saying that Y is (almost surely) a monotone increasing or decreasing function of X . Thus when $\tau = 1$, one finds $Y = G^{-1}\{F(X)\}$ with probability one and hence $K(w) = w$ for all $0 \leq w \leq 1$, while when $\tau = -1$, $Y = G^{-1}\{1 - F(X)\}$ almost everywhere and $K \equiv 1$ on its domain (in other words, it is the distribution function of a point mass at the origin).

1.4.2 Properties of K-plots

It is a simple matter to deduce from Proposition 2.1 of Genest and Rivest (1993) that as $n \rightarrow \infty$, $K_n(w) \rightarrow K(w)$ in probability for all $0 \leq w \leq 1$, and

hence that $K_n^{-1}(p) \rightarrow K^{-1}(p)$ in probability for all $0 \leq p \leq 1$ as well. Since K_n and its inverse are bounded, this convergence naturally extends to their expectations, with the following consequences.

4.2.1: For arbitrary integer $n \geq 1$ and $0 \leq p \leq 1$, let $\lceil np \rceil$ denote the smallest integer greater than or equal to np . Then

$$H_{(\lceil np \rceil)} = K_n^{-1}(p) \rightarrow K^{-1}(p)$$

and hence also

$$\lim_{n \rightarrow \infty} \mathbb{E} (H_{(\lceil np \rceil)}) = \lim_{n \rightarrow \infty} W_{\lceil np \rceil; n} = K_0^{-1}(p)$$

under the null hypothesis of independence.

4.2.2: For large enough sample size n , the pairs $(W_{i;n}, H_{(i)})$ will tend to concentrate along the curve $p \mapsto (K_0^{-1}(p), K^{-1}(p))$; in other words, the points on the K-plot will look like a plot of $w \mapsto K^{-1}\{K_0(w)\}$.

4.2.3: The graph will tend to be linear when $K = K_0$, as under the null hypothesis of independence;

4.2.4: All points on the graph will fall on the horizontal axis ($p \equiv 0$) when the variables X and Y are comonotonic with $\tau(X, Y) = -1$, because then $K^{-1}(p) = 0$ for all possible values of $0 \leq p \leq 1$;

4.2.5: All points will fall on the curve $K_0(p)$ when X and Y are comonotonic with $\tau(X, Y) = 1$, since then $K^{-1}(p) \equiv p$ on $[0, 1]$.

1.5 Examples of K-plots

Prototypical examples of K-plots based on artificial data are presented in Section 5.1. Real-life applications can be found in Section 5.2.

1.5.1 Artificial data

As a first example, consider a pair (X, Y) of random variables that are comonotonic in the sense given to that term in 4.1.5. If $Y = G^{-1}\{F(X)\}$, their underlying copula is then the upper Fréchet bound, $M(u, v) = \min(u, v)$. If $Y = G^{-1}\{1 - F(X)\}$, their copula is the Fréchet lower bound, $W(u, v) = \max(0, u + v - 1)$. Figure 1.4 shows K-plots associated with a sample of 100 pseudo-random observations from such pairs when F and G are uniform on $[0, 1]$, so that the joint distribution of (X, Y) is then the copula itself.

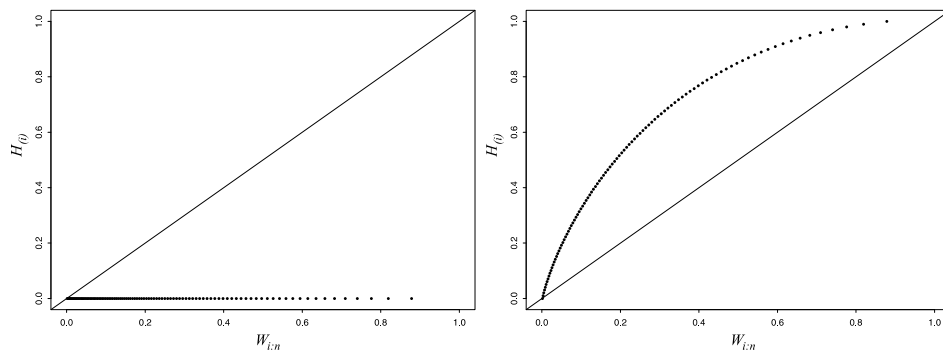


Figure 1.4: K-plots based on pseudo-random samples of size 100 from a bivariate vector (X, Y) in which X is uniform on $[0, 1]$ and $Y = 1 - X$ (left panel) or $Y = X$ (right panel).

To fix ideas, the scatterplot corresponding to the left panel of Figure 1.4 would show 100 points randomly positioned on the line $y = 1 - x$. On the scatterplot associated with the right panel, the points would fall on the line $y = x$ instead. In accordance with Properties 4.2.4 and 4.2.5, the plot is a flat line at height 0 in the case of perfect negative functional dependence (lower Fréchet bound, W). It can also be seen to match nearly the curve $K_0(p) = p - p \log(p)$ in the case of perfect positive functional dependence (upper Fréchet bound, M).

Next, Figure 1.5 depicts K-plots associated with pseudo-random samples of size 100 from a bivariate normal distribution N_r with zero means, unit variances, and Pearson correlation r chosen in such a way that Kendall's coefficient of concordance, namely

$$\tau(X, Y) = \frac{2}{\pi} \arcsin(r),$$

is successively equal to $1/4$, $1/2$ and $3/4$. The corresponding approximate values of r are 0.383, 0.707, and 0.924, respectively. The fact that these distributions are ordered by \prec_K , namely

$$r \leq r^* \quad \Rightarrow \quad N_r \prec_K N_{r^*},$$

translates into plots that are further and further away from a straight line as $r \rightarrow 1$, as might be expected in the light of Property 4.2.2.

Property 4.2.2 also implies that sufficiently large sample sizes from distributions H and H^* having different probability integral transformations K and K^* could be distinguished graphically. To illustrate this fact, consider Clayton's family of copulas, defined by

$$C_\alpha(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad 0 \leq u, v \leq 1$$

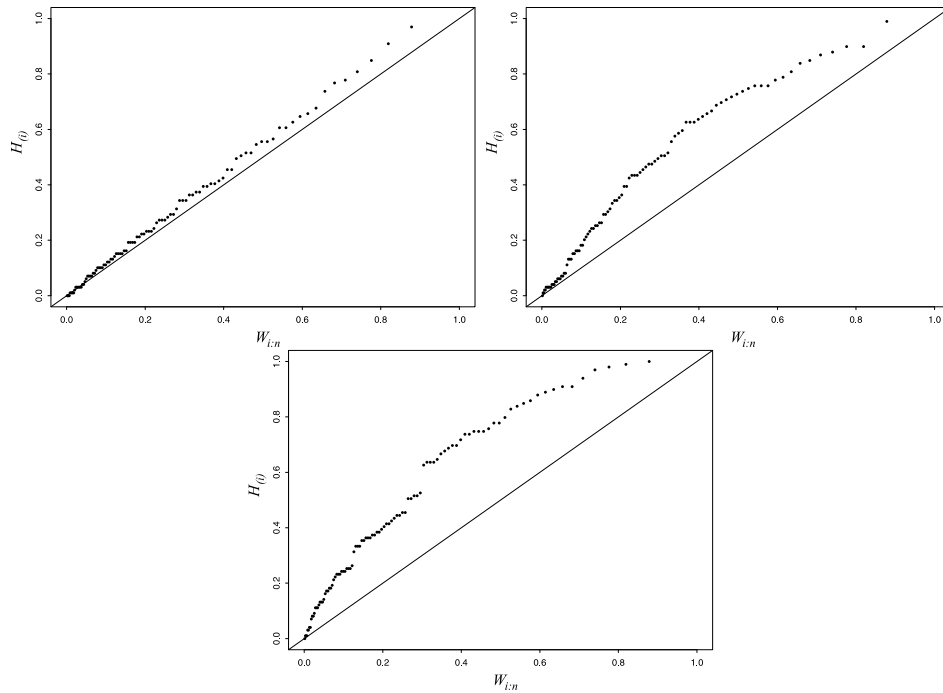


Figure 1.5: K-plots based on pseudo-random samples of size 100 from a bivariate normal vector (X, Y) with Kendall's tau equal to $1/4$ (top left panel), $1/2$ (top right panel), and $3/4$ (bottom panel).

in terms of a parameter α related to Kendall's tau through $\tau(X, Y) = \alpha/(\alpha + 2)$. This class of distributions is a prime example of frailty model (Clayton 1978, Oakes 1986) used for the study of association in bivariate survival data. As in the normal case, one has

$$\alpha \leq \alpha^* \quad \Rightarrow \quad C_\alpha \prec_K C_{\alpha^*},$$

as can be readily checked using the closed-form expression

$$K_\alpha(w) = w + \frac{w}{\alpha} (1 - w^\alpha), \quad 0 \leq w \leq 1$$

which derives from the Archimedean character of C_α (Genest and MacKay 1986; Genest and Rivest 2001).

Figure 1.6 shows K-plots based on pseudo-random samples of size 100 from C_α with $\tau = 1/4, 1/2, 3/4$. The curves are rather different from those displayed in Figure 1.5. One should keep in mind, however, that bivariate distributions with distinct copulas C and C^* do not necessarily have different distributions K and K^* associated with their probability integral transformations.

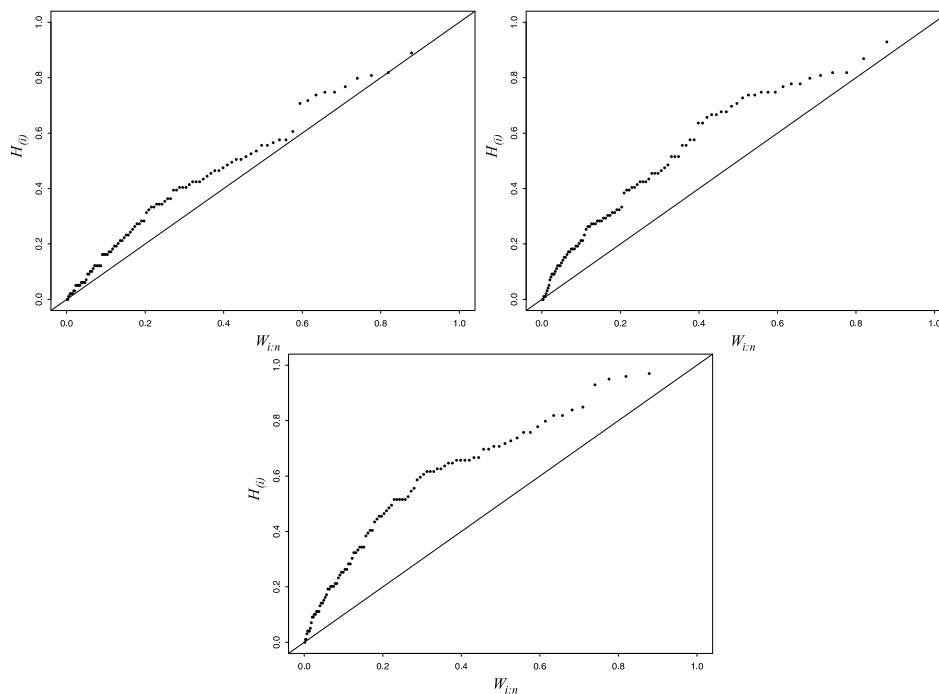


Figure 1.6: K-plots based on pseudo-random samples of size 100 from a bivariate vector (X, Y) with Clayton copula and Kendall's tau equal to $1/4$ (top left panel), $1/2$ (top right panel), and $3/4$ (bottom panel).

That the implication $C \neq C^* \Rightarrow K \neq K^*$ is *false* may be verified easily using the class of bivariate extreme-value distributions, whose underlying copulas (see for example Capéraà, Fougères and Genest 1997b, 2000) are of the form

$$C_A(u, v) = \exp \left[\log(uv) A \left\{ \frac{\log(u)}{\log(uv)} \right\} \right]$$

for some convex function $A : [0, 1] \rightarrow [1/2, 1]$ such that $A(0) = A(1) = 0$ and $A(w) \geq \max(w, 1 - w)$ for all $0 \leq w \leq 1$. As shown by Ghoudi, Khoudraji and Rivest (1998), $W = C_A(U, V)$ is distributed as

$$K_A(w) = w - (1 - \tau_A)w \log(w), \quad 0 \leq w \leq 1$$

where

$$\tau_A = \int_0^1 \frac{w(1-w)}{A(w)} dA'(w)$$

is the population value of Kendall's tau (which, by the way, is always positive for an extreme-value distribution, since A is convex; for a stronger result along these lines, see Garralda-Guillem 2000).

Thus if two extreme-value distributions with generators $A \neq A^*$ verify $\tau_A = \tau_{A^*}$, then clearly $K_A = K_{A^*}$. A fortiori, one could not hope to catalogue with complete precision the types of dependence implied by various K-plots. Of course, chi-plots suffer from the same limitation, though presumably to a lesser extent, given their bivariate nature. At the same time, however, the exact features of the copula function that chi-plots depict seem hard to pin down.

1.5.2 Real data

Figure 1.7 displays a scatterplot, a chi-plot, and a K-plot of the ranks for 88 pairs of measurements used to investigate the relationship between the equivalence ratio (NO_x, the concentration of nitric oxide NO and nitrogen dioxide NO₂ in engine exhaust, normalized by the work done by the engine), and a measure of the richness of the air/ethanol mix. The data, considered by Kallenberg and Ledwina (1999), are used by Fisher and Switzer (2001) to illustrate some of the properties of their chi-plots in a blatant case of non-monotone association in which the χ_i coordinates of the chi-plot tend to be centered at zero, but are abnormally distributed. The K-plot, provided as a complement, suggests the presence of mild negative association in the data. This is not obvious from the scatterplot but in line with the empirical values of Spearman's rho and Kendall's tau, which are both of the order of $-.14$. A third, undetermined factor may well explain the basic data pattern.

A rank scatterplot, a chi-plot and a K-plot may also be found in Figure 1.8 for the second example considered by Kallenberg and Ledwina (1999) and Fisher and Switzer (2001). This data set consists of 28 measurements of size of the annual spawning stock of salmon and corresponding production of new catchable-sized fish in the Skeena River (BC, Canada). While the presence of overall positive association can be detected on all plots ($\rho_n \approx .55$, $\tau_n \approx .41$), the plateau observed in the K-plot highlights the presence of a cluster in the lower left corner of the rank scatterplot. This phenomenon translates into the equality of several H_i corresponding to points roughly located along the line $y = 1 - x$ in the rank scatterplot. Fisher and Switzer (2001) apparently did not notice this feature of the data on their chi-plot.

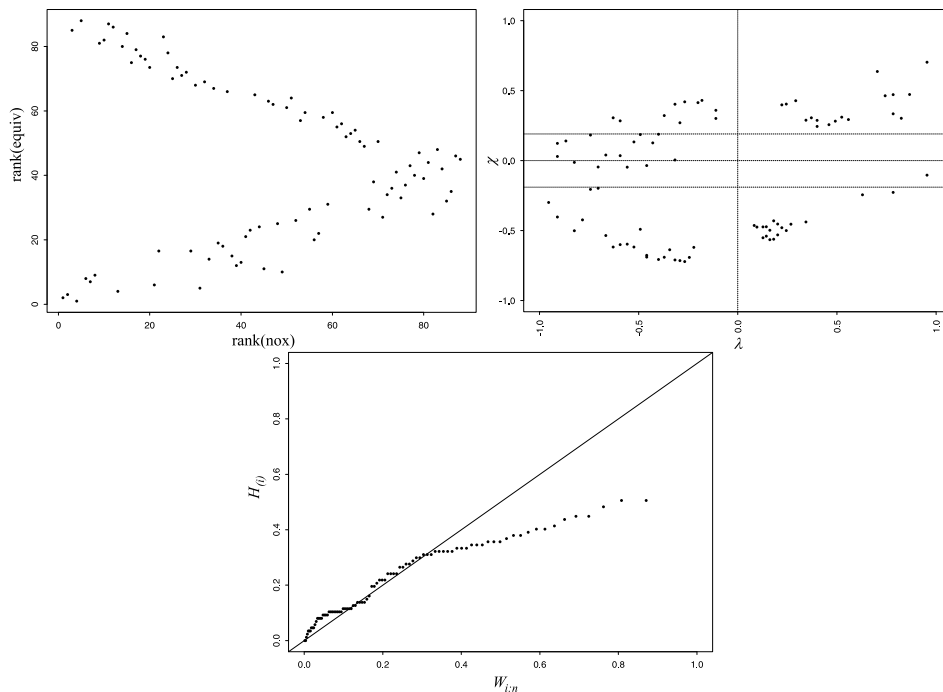


Figure 1.7: Rank scatterplot, chi-plot, and K-plot for an automobile exhaust data set involving 88 measurements of equivalence ratio and corresponding air-ethanol.

1.6 Extensions

In their papers, Fisher and Switzer (1985, 2001) mention that chi-plots can be adapted easily to detect the presence of autocorrelation in a stationary time series Z_1, \dots, Z_{n+m} . Thus to detect dependence at lag $\ell \in \{1, \dots, m\}$, their graphical procedure could simply be applied to the pairs $(X_i, Y_i) = (Z_i, Z_{i+\ell})$, $1 \leq i \leq n + m - \ell$.

Obviously, the same can be done with K-plots, and recent results of Genest, Quessy and Rémillard (2002) imply that properties 4.2.1–4.2.5 of

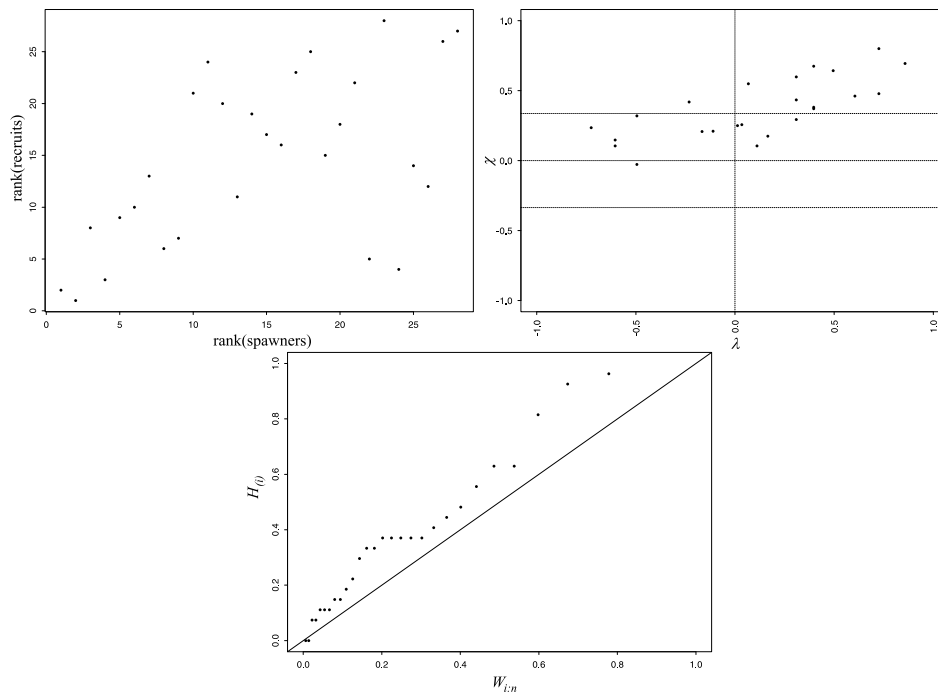


Figure 1.8: Rank scatterplot, chi-plot, and K-plot for 28 measurements of size of the annual spawning stock of salmon and corresponding production of new catchable-sized fish in the Skeena River (BC, Canada).

K-plots remain true with the same $K_0(w) = w - w \log(w)$ under the null hypothesis of randomness.

There is of course no reason why K-plots could not also be used in situations where one of the variables is treated as fixed. But contrary to chi-plots, the concept of K-plot can be extended straightforwardly to the multivariate case. Given a p -variate random sample $(X_{11}, \dots, X_{1p}), \dots, (X_{n1}, \dots, X_{np})$, the procedure is as follows:

a) For each $1 \leq i \leq n$, compute

$$H_i = \frac{1}{n-1} \#\{j \neq i : (X_{j1}, \dots, X_{jp}) \leq (X_{i1}, \dots, X_{ip})\},$$

where an inequality between vectors is interpreted to hold component-wise.

b) Order the H_i to get $H_{(1)} \leq \dots \leq H_{(n)}$; as in the case $p = 2$, equalities are possible.

c) Plot the pairs $(W_{i:n}, H_{(i)})$, $1 \leq i \leq n$, where $W_{i:n}$ is the i th rankit from a random sample of size n from the asymptotic distribution of the H_i under the null hypothesis of mutual independence between the p components.

In the light of Example 1 of Barbe, Genest, Ghoudi and Rémillard (1996), $W_{i:n}$ is thus computed as in (3), but with

$$K_0(w) = w + w \sum_{k=1}^{p-1} \frac{\log^k(1/w)}{k!}, \quad 0 \leq w \leq 1.$$

Figure 1.9 illustrates the above procedure using two pseudo-random samples of size 100 from a vector (X_1, X_2, X_3) whose components are uniformly distributed on the interval $[0, 1]$. In the right panel, $X_1 = X_2 = X_3$ almost everywhere, which represents the most extreme case of positive dependence, whose associated copula is the Fréchet upper bound $M(x_1, x_2, x_3) = \min(x_1, x_2, x_3)$. In the left panel, $X_1 = X_2 = 1 - X_3$ with probability one, which illustrates a situation of strong negative dependence, though not the only one possible because while

$$C(u_1, \dots, u_p) \geq W(u_1, \dots, u_p) \equiv \max(0, u_1 + \dots + u_p + 1 - p)$$

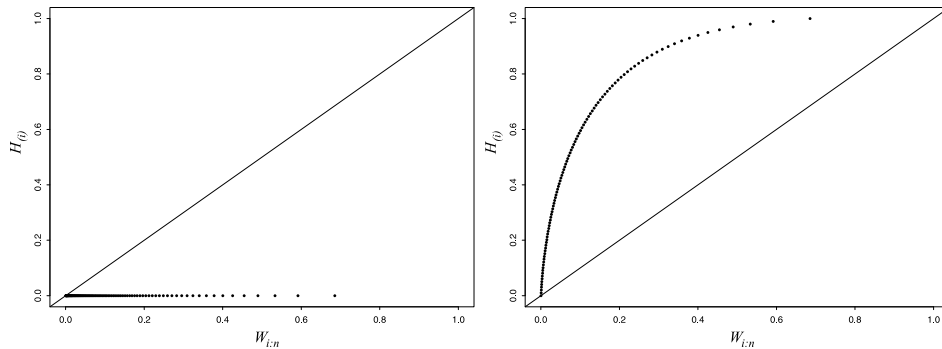


Figure 1.9: K-plot based on a pseudo-random sample of size 100 from a vector (X_1, X_2, X_3) of uniform random variables with $X_1 = X_2 = 1 - X_3$ (left panel) or $X_1 = X_2 = X_3$ (right panel).

everywhere for any p -variate copula C , the Fréchet lower bound W is no longer a distribution function in dimension $p \geq 3$.

As a final example, consider a vector (X_1, X_2, X_3) with

$$X_1 = |Z| \operatorname{sign}(X_2 X_3),$$

where X_2 , X_3 and Z are mutually independent standard normal random variables. As mentioned by Romano and Siegel (1986, p. 33), the X_i are then pairwise independent, but not jointly independent.

Figure 1.10 shows four K-plots based on a pseudo-random sample of size $n = 100$ from this distribution. The plots are based on the trivariate data (lower right corner) and on bivariate data obtained by ignoring either variable 3 (upper left corner), variable 2 (upper right corner), or variable 1 (lower left corner). Only the K-plot based on the full data could possibly detect the lack of mutual independence, and one might hope to see a hint of that in the graph, but the pattern is not significant. Note that in this particular model,

zero is the value of the p -variate extension of Kendall's tau (see, for example, Jouini and Clemen, 1996), namely

$$\tau = \frac{2^p \mathbb{E}\{H(X_1, \dots, X_p)\} - 1}{2^{p-1} - 1} .$$

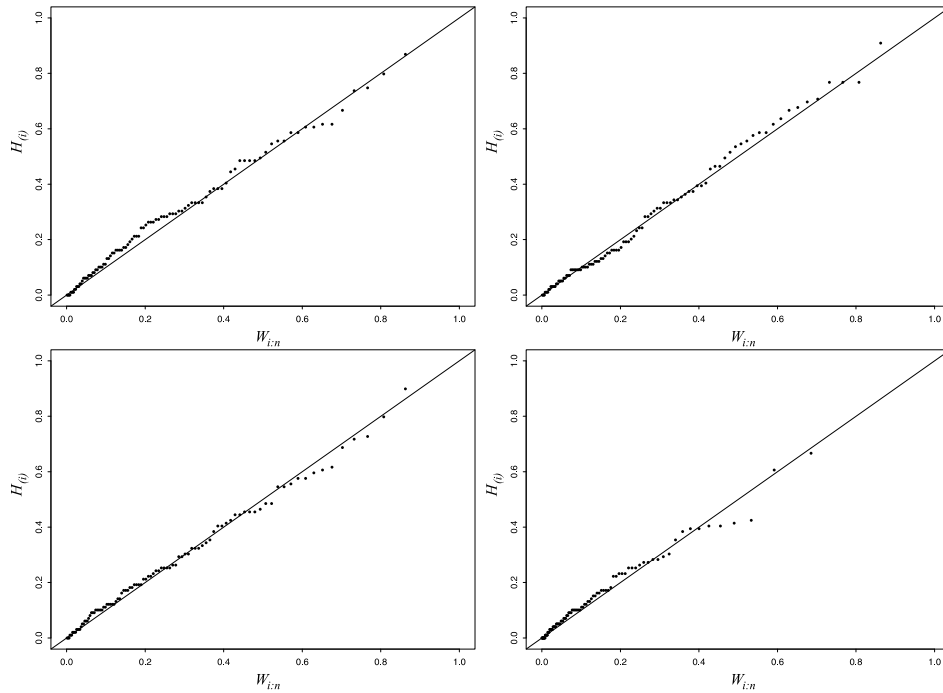


Figure 1.10: K-plot based on a pseudo-random sample of size 100 from a vector (X_1, X_2, X_3) whose components are pairwise independent normal random variables, but which are not trivariate Gaussian.

1.7 Conclusion

For multivariate continuous data, stochastic independence is a characteristic of the underlying copula and hence both graphical and formal tests of association should be based on ranks, since the latter are the maximally invariant

statistics under transformations that leave the copula unchanged. Chi-plots and K-plots both meet this basic requirement, and enjoy other properties that make them complementary tools in visual inspection of data. Because they are germane to Q–Q plots and connected in an explicit way to the underlying copula structure, K-plots are easier to interpret and more readily extendible from the bivariate to the multivariate case. They are, however, an inherently univariate object. Chi-plots are two-dimensional and hence clearly convey more information about possible dependence, but their relation to the copula is somewhat elusive. As there is no single best way of looking at a multivariate object, analysts will derive advantages from a joint utilization of both types of graphics.

CHAPITRE II

EXEMPLES SUPPLÉMENTAIRES

Comme son titre l'indique, ce chapitre propose quelques illustrations supplémentaires de l'emploi du K-plot, à raison d'une application par section.

2.1 Une application au baseball

Le site internet *www.sportslines.com* rapporte les statistiques individuelles des 98 joueurs de la *Ligue nationale de baseball* ayant eu au moins 400 présences au bâton lors de la saison 2002. Les variables observées sont la moyenne au bâton (AVG), le nombre moyen de buts sur balles par présence au bâton (BBAB), le nombre moyen de retraits au bâton par présence au bâton (SOAB), le nombre moyen de circuits par présence au bâton (HRAB) et l'âge du joueur en date du 31 décembre 2002 (AGE).

Les relations entre les variables AVG, BBAB et SOAB sont illustrées deux à deux à la figure 2.1. Les trois graphiques présentent des nuages de points à partir desquels il est difficile d'observer des tendances générales. L'application du K-plot (voir figure 2.2) révèle cependant la présence d'une faible relation positive entre les variables AVG et BBAB, ainsi qu'entre les variables BBAB et SOAB. Comme on pourrait s'y attendre, les variables AVG et SOAB sont toutefois en dépendance négative.

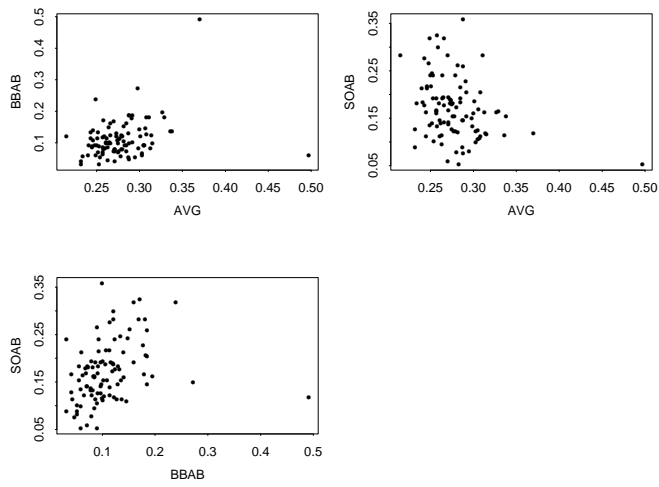


Figure 2.1: Graphiques mettant en relation les variables AVG, BBAB et SOAB deux à deux.

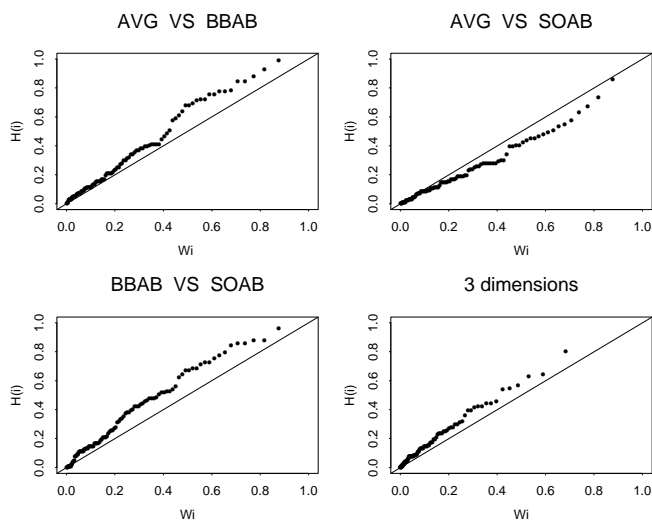


Figure 2.2: Application des K-plots sur les variables AVG, BBAB et SOAB.

La relation positive entre BBAB et SOAB pourrait s'expliquer par le fait qu'un frappeur plus patient, recevant plus de lancers, a plus de chances d'obtenir un but sur balles, mais aura également plus de chances de se faire retirer sur trois prises.

Lorsque l'on répète la même procédure pour les variables AVG, BBAB, et AGE (voir figure 2.3), on constate à nouveau que l'application des K-plots nous aide à mieux discerner des relations entre les variables. Les relations illustrées sur les K-plots de la figure 2.4 sont faibles mais positives entre chacune des variables entre elles, ainsi que pour les trois variables ensemble. En vieillissant, les joueurs ont tendance à voir leur moyenne et leur taux de buts sur balles augmenter très légèrement.

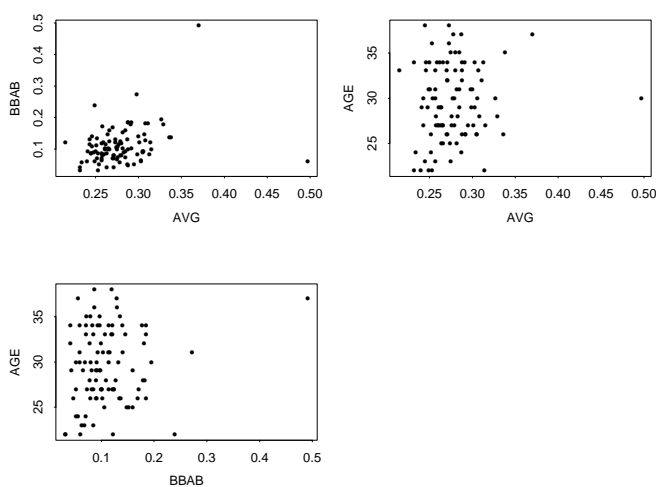


Figure 2.3: Graphiques mettant en relation les variables AVG, BBAB et AGE deux à deux.

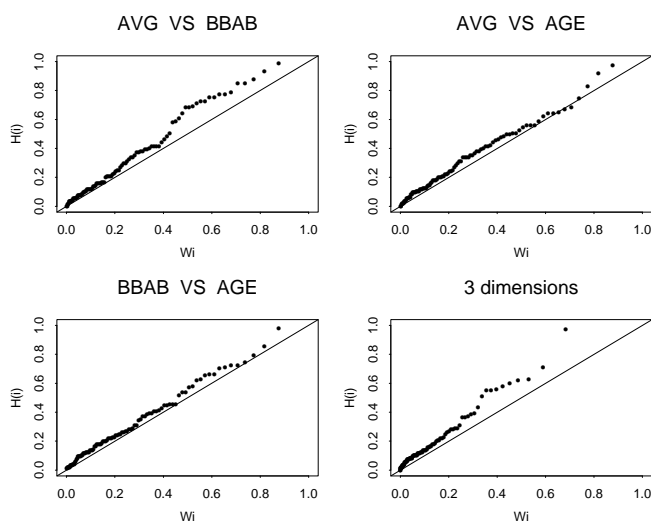


Figure 2.4: Application des K-plots sur les variables AVG, BBAB et AGE.

2.2 Une application au hockey

Il arrive parfois, lorsque le nombre d'observations est grand, que les graphiques mettant deux variables en relation soient trop denses pour qu'on puisse facilement y percevoir une tendance. L'exemple suivant est de ce type. Il porte sur les statistiques, pour la saison 2002–2003, des 888 joueurs de la *Ligue nationale de hockey* ayant participé à des matchs de la saison régulière au cours de cette saison. Les deux variables considérées sont PTSPG (points marqués par partie), PIMPG (minutes de punition par partie). Les données proviennent du site internet *www.sportsline.com*.

Lorsqu'on examine le graphique de PTSPG versus PIMPG et le graphique des rangs de PTSPG versus les rangs de PIMPG (voir figure 2.5), on a de la difficulté à observer une tendance.

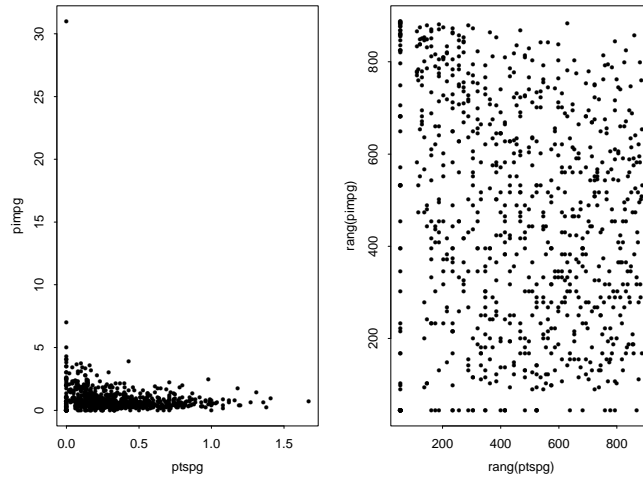


Figure 2.5: PIMPG versus PTSPG, et rangs(PIMPG) versus rangs(PTSPG).

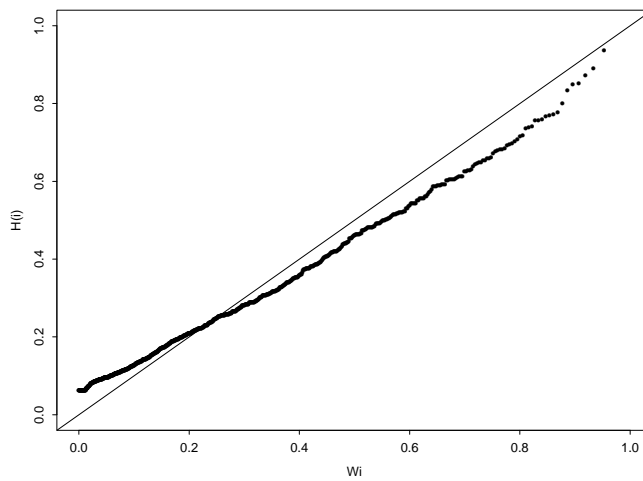


Figure 2.6: K-plot illustrant la structure de dépendance entre PIMPG et PTSPG.

Pour sa part, le K-plot (figure 2.6) suggère qu'il existe peut-être une faible relation positive pour les observations (X_i, Y_i) pour lesquelles $H(X_i, Y_i)$ est petit, alors que cette relation deviendrait négative à mesure que $H(X_i, Y_i)$ grandit. Une telle relation semble difficile à expliquer et ne saurait en tout cas permettre de conclure que la loi H des observations est globalement en dépendance positive ou négative dans le sens de l'ordre \prec_K mentionné à la section 1.4.1.

2.3 Une application à caractère social

L'exemple suivant porte sur 47 zones résidentielles de la ville de Chicago (Illinois). Chacune d'elles correspond à un code postal différent. Les variables mesurées sont RACE (composition raciale, correspondant au pourcentage de minorités, en 1975), FIRE (incendies par 1000 résidences, pour l'année 1975), THEFT (vols par 1000 résidences en 1975), AGE (pourcentage de résidences construites avant 1940) et INCOME (revenu moyen en 1975). Les données sont extraites de Andrews et Herzberg (1985).

On trouve à la figure 2.7 des diagrammes de dispersion et des K-plots illustrant la relation de dépendance observée entre les différentes paires de variables. Les diagrammes de dispersion se situent au-dessus de la diagonale principale, tandis que les K-plots correspondants se situent sous la diagonale.

Un simple coup d'œil permet de voir qu'en général, les variables sont toutes associées positivement, sauf lorsque la variable INCOME est en jeu. On constate que les ménages à revenu élevé sont moins sujettes aux incendies et aux vols, que les quartiers les plus riches sont celles où les minorités sont

le moins présentes, et que les quartiers les moins cossus sont aussi les plus anciens.

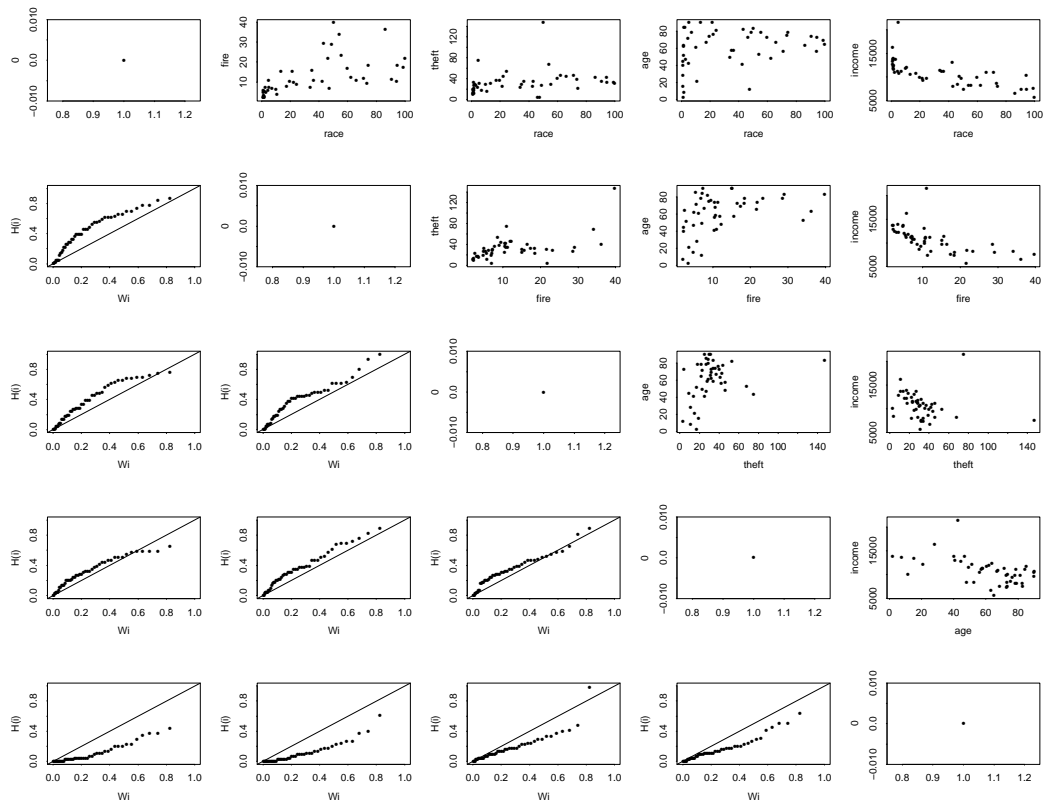


Figure 2.7: Matrice de diagrammes de dispersion et de K-plots pour cinq variables sociales mesurées à Chicago.

2.4 Une application à caractère médical

Ce dernier exemple, qui porte sur des données biomédicales rapportées par Reaven et Miller (1979), met en évidence certaines différences entre le K-plot et le khi-plot, en ce qui concerne leur aptitude à détecter divers types d'association.

Les données visent à étudier la relation entre le poids relatif et quatre variables biochimiques mesurées à partir d'échantillons de sang prélevés sur 145 sujets adultes non-obèses classifiés comme normaux ou comme diabétiques symptomatiques et asymptomatiques.

La figure 2.8 représente un tableau 5×5 dans lequel se retrouvent les khi-plots et les K-plots de toutes les paires de variables. Les K-plots et les khi-plots se situent respectivement au-dessus et au-dessous de la diagonale principale. Les variables considérées sont

- (1) Poids relatif
- (2) Glucose du plasma en régime
- (3) Glucose
- (4) Insuline
- (5) Glucose du plasma en situation normale

Le lecteur peut se rapporter à l'article de Reaven et Miller (1979) ou au livre de Andrews et Herzberg (1985) pour de plus amples détails concernant ces variables, mesurées au Centre de recherche clinique de l'Université Stanford, en Californie.

Les khi-plots et les K-plots font tous les deux ressortir une forte association positive entre les variables 2, 3 et 5. Les deux outils graphiques permettent aussi d'affirmer que la variable 1 est positivement associée à toutes les autres, quoique faiblement. Cependant, les K-plots ne détectent pas de relation apparente entre la variable 4 et les autres. Les khi-plots suggèrent pourtant une forme d'association entre les paires (2, 4), (3, 4) et (4, 5).

Parce qu'ils sont bidimensionnels, les khi-plots fournissent vraisemblablement plus d'information que les K-plots quant à la nature d'une éventuelle dépendance entre deux variables. En revanche, ils sont beaucoup plus difficiles à interpréter que les K-plots, au moins pour un néophyte. Par ailleurs, le lien ténu entre les éléments du khi-plot et la copule sous-jacente au modèle rend hasardeuse l'interprétation des patrons de ce type de graphique en terme de la structure de dépendance entre les variables.

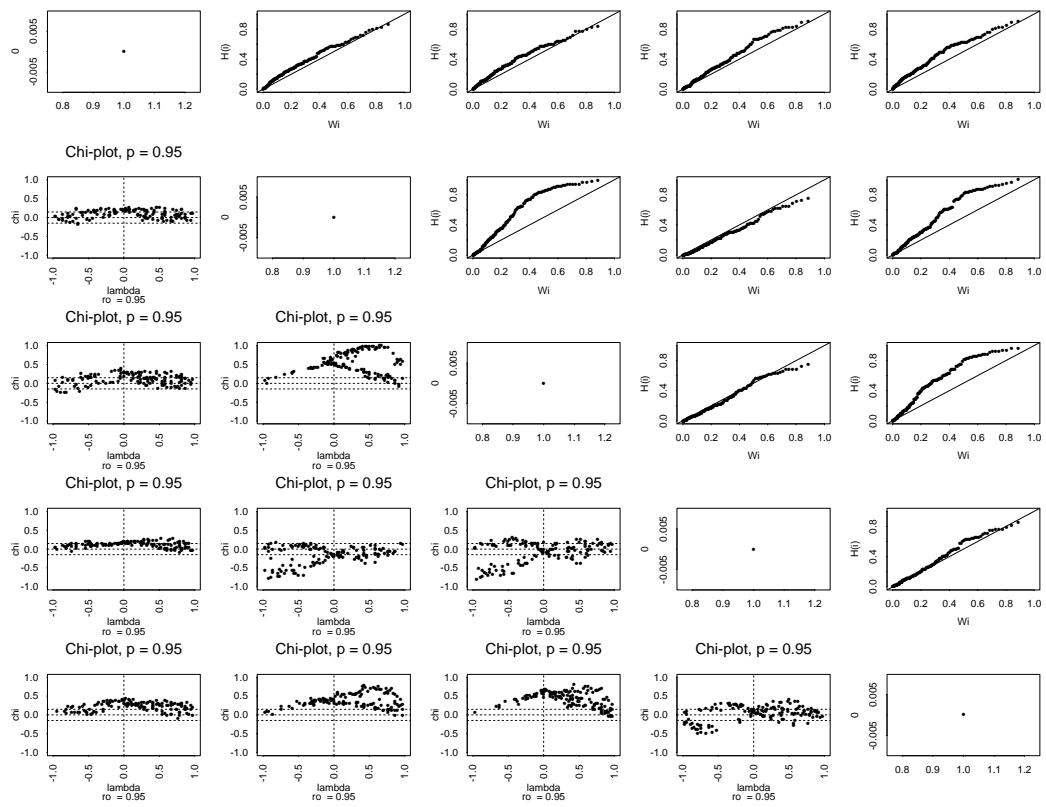


Figure 2.8: Khi-plots et K-plots de toutes les paires de variables biochimiques provenant de Reaven et Miller (1979).

CHAPITRE III

UN RÉSULTAT DE CONVERGENCE

L'objectif du présent chapitre est de fournir une démonstration rigoureuse du résultat suivant, évoqué au début du paragraphe 1.4.2 du chapitre I.

Résultat. Soit X_1, \dots, X_n un échantillon aléatoire de loi K continue et strictement croissante telle que $K(0) = 0$ et $K(1) = 1$. Notons $X_{(i)}$ la i ème statistique d'ordre associée à l'échantillon et, pour tout $p \in (0, 1)$, désignons par $[np]$ le plus petit entier supérieur ou égal à np . Alors

$$\lim_{n \rightarrow \infty} E(X_{([np])}) = K^{-1}(p).$$

Dans un premier temps, il est bien connu (voir par exemple le Théorème III.2.1 en p. 24 de Capéraà et van Cutsem 1988) que la densité de $X_{(i)}$ est égale à

$$n \binom{n-1}{i-1} \int_0^1 k(x) \{K(x)\}^{i-1} \{1-K(x)\}^{n-i} dx,$$

de sorte qu'en appliquant le changement de variable $y = K(x)$, dont le jacobien est $dy = k(x)dx$, on voit que

$$\begin{aligned} E(X_{(i)}) &= n \binom{n-1}{i-1} \int_0^1 x k(x) \{K(x)\}^{i-1} \{1-K(x)\}^{n-i} dx \\ &= n \binom{n-1}{i-1} \int_0^1 K^{-1}(y) y^{i-1} (1-y)^{n-i} dy, \\ &= E\{K^{-1}(Y_n)\}, \end{aligned}$$

où Y_n est une variable aléatoire bêta de paramètres $\alpha = i$ et $\beta = n + 1 - i$, dont la moyenne et l'espérance sont respectivement égales à

$$E(Y_n) = \frac{i}{n+1} \quad \text{et} \quad \text{var}(Y_n) = \frac{i(n+1-i)}{(n+1)^2(n+2)}.$$

On remarque que si $i = \lceil np \rceil$, alors

$$\lim_{n \rightarrow \infty} E(Y_n) = p \quad \text{et} \quad \lim_{n \rightarrow \infty} \text{var}(Y_n) = 0.$$

Il s'ensuit que la suite (Y_n) converge en moyenne quadratique et donc en probabilité vers p . Par conséquent,

$$K^{-1}(Y_n) \Rightarrow K^{-1}(p)$$

par continuité de K^{-1} ; voir par exemple le Théorème 6'(a), p. 42 de Ferguson (1996). Bien entendu, la suite $K^{-1}(Y_n)$ converge aussi en loi. Puisqu'en outre les éléments de la suite prennent leurs valeurs dans l'intervalle $(0, 1)$, on en conclut (voir par exemple le Théorème 3 (c) en p. 13 de Ferguson 1996) que

$$\lim_{n \rightarrow \infty} E(X_{(\lceil np \rceil)}) = \lim_{n \rightarrow \infty} E\{K^{-1}(Y_n)\} = K^{-1}(p),$$

ce qu'il fallait démontrer.

CONCLUSION

Ce mémoire a proposé et décrit une manière simple et efficace de visualiser la dépendance. Étant donné un échantillon $(X_{11}, \dots, X_{1p}), \dots, (X_{n1}, \dots, X_{np})$ de loi inconnue H et H_n la fonction de répartition expérimentale correspondante, le K-plot s'appuie essentiellement sur la comparaison des statistiques d'ordre associées aux pseudo-observations $H_i = H_n(X_{i1}, \dots, X_{ip})$ et sur leurs espérances sous l'hypothèse nulle que les p variables sont mutuellement indépendantes.

Le principe des K-plots est donc le même que celui de la droite de Henry, encore appelée Q-Q plot. Comme les pseudo-observations H_1, \dots, H_n ne dépendent que des rangs des observations originelles, la technique est non paramétrique, au sens où elle ne fait pas intervenir les marges de H . À l'instar de la méthode du khi-plot développée par Fisher et Switzer (1985, 2001), elle ne fait intervenir que la structure de dépendance sous-jacente au modèle, laquelle est représentée par la copule.

Parce qu'elle prend appui sur la transformation intégrale de probabilité et conduit à un graphique unidimensionnel dont le lien avec la copule sous-jacente est explicite, l'interprétation des K-plots est plus immédiate que celle des khi-plots. Ces derniers, cependant, contiennent vraisemblablement plus d'information concernant la nature de la dépendance entre deux variables. Pour que ces renseignements soient vraiment utiles, il importerait toutefois

d'élucider la relation exacte entre le graphique fourni par le khi-plot et la copule. Ce problème ne semble pas facile à résoudre.

L'autre particularité du K-plot concerne sa généralisation immédiate au cas de plus de deux variables. Bien que ceci lui confère un avantage indéniable sur le khi-plot, l'utilisateur doit être conscient que la représentation unidimensionnelle d'une relation de dépendance multivariée complexe perd forcément une partie de son pouvoir discriminant à mesure que le nombre de variables augmente.

Dans l'état actuel des choses, la meilleure pratique consiste sans doute à utiliser les khi-plots et les K-plots en synergie afin d'accroître ses chances de détecter et de bien modéliser toute forme de dépendance entre les variables dans la phase exploratoire d'une analyse de données multivariées.

ANNEXE

Code S-plus

Voici le code réalisé pour créer la fonction nommée `kplot`, qui applique la technique du K-plot pour deux variables.

```
k<-function(alp)
{
  alp-alp*log(alp)
}
k.inv<-function(beta,minx=0,maxx=1,etapes=20)
{
  x<-0.5
  for(k in 2:etapes)
  {
    x<-x-ifelse(k(x*(maxx-minx)+minx)-
beta>0,1,-1)*2^(-k)
  }
  x*(maxx-minx)+minx
}
u_function(n)
{
```

```

    ui_c()

    for(i in 1:n)
        {
            ui[i]_k.inv(i/(n+1))
        }

    ui
}

rankitpremier_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*x*log(x)*(x-x*log(x))^(i-1)
    *(1-x+x*log(x))^(n-i))
    sum(y)/m
}

rankitspetits_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*prod((n-i+1):(n-1))/factorial(i-1)*x*log(x)
    *(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
    sum(y)/m
}

rankitsgrands_function(n,i,m)
{
    x_((1:m)-.5)/m

```

```

    y_(-n*prod(i:(n-1))/factorial(n-i)*x*log(x)
      *(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
    sum(y)/m
  }
rankitdernier_function(n,i,m)
{
  x_((1:m)-.5)/m
  y_(-n*x*log(x)*(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
  sum(y)/m
}

kplot_function(enx,eny)
{
  k<-function(alp)
  {
    alp-alp*log(alp)
  }
  k.inv<-function(beta,minx=0,maxx=1,etapes=20)
  {
    x<-0.5
    for(k in 2:etapes)
    {
      x<-x-ifelse(k(x*(maxx-minx)+minx)-
        beta>0,1,-1)*2^(-k)
    }
  }
}

```

```

        }
        x*(maxx-minx)+minx
    }
u_function(n)
{
    ui_c()

    for(i in 1:n)
    {
        ui[i]_k.inv(i/(n+1))
    }

    ui
}
rankitpremier_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*x*log(x)*(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
    sum(y)/m
}
rankitspetits_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*prod((n-i+1):(n-1))/factorial(i-1)*
    x*log(x)*(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
    sum(y)/m
}

```

```

    }
rankitsgrands_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*prod(i:(n-1))/factorial(n-i)*
    x*log(x)*(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
    sum(y)/m
}
rankitdernier_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*x*log(x)*(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
    sum(y)/m
}
rg_function(n)
{
    if ( n>=4 & n<=268 )
    {
        if( n%%2 < 1 )
        {
            upremier_rankitpremier(n,1,10000)
            upetits_sapply(n=n,2:(n/2),rankitspetits,m=10000)
            ugrands_sapply(n=n,((n/2)+1):
            (n-1),rankitsgrands,m=10000)
            udernier_rankitdernier(n,n,10000)
        }
    }
}

```

```
    }
else if( n%%2 > 0 )
{
    upremier_rankitpremier(n,1,10000)
    upetits_sapply(n=n,2:((n+1)/2),
    rankitspetits,m=10000)
    ugrands_sapply(n=n,((n+3)/2):(n-1),
    rankitsgrands,m=10000)
    udernier_rankitdernier(n,n,10000)
}
}
else if (n==1)
{
    upremier_rankitpremier(1,1,10000)
    upetits_c()
    ugrands_c()
    udernier_c()
}
else if (n==2)
{
    upremier_rankitpremier(2,1,10000)
    upetits_c()
    ugrands_c()
    udernier_rankitdernier(2,2,10000)
}
```

```
else if (n==3)
{
  upremier_rankitpremier(3,1,10000)
  upetits_rankitspetits(3,2,10000)
  ugrands_c()
  udernier_rankitdernier(3,3,10000)
}
else if (n>268)
{
  upremier_u(n)
  upetits_c()
  ugrands_c()
  udernier_c()
}
c(upremier, upetits, ugrands, udernier)
}
if (length(enx) == length(eny))
{
  n_length(enx)
  z_c()
  for(i in 1:length(enx))
  {
    petitx_compare(enx[i],enx[-i])
    petity_compare(eny[i],eny[-i])
    petit_(petitx >= 0 & petity >= 0)
```

```

        z[i]_sum(petit)/(length(enx)-1)
    }
    qqplot(rg(n),z,xlim=c(0,1),ylim=c(0,1))
    abline(0,1)
    paste("OK")
}
else
    paste("length of x and y must be equal")
}

```

Voici maintenant le code réalisé pour créer la fonction nommée `kplot3d`, qui applique la technique du K-plot pour trois variables.

```

rankitpremier_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*x*log(x)*(x-x*log(x))^(i-1)
    *(1-x+x*log(x))^(n-i))
    sum(y)/m
}
rankitspetits_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*prod((n-i+1):(n-1))/factorial(i-1)*
    x*log(x)*(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
    sum(y)/m
}

```

```

}
rankitsgrands_function(n,i,m)
{
  x_((1:m)-.5)/m
  y_(-n*prod(i:(n-1))/factorial(n-i)*
  x*log(x)*(x-x*log(x))^(i-1)*(1-x+x*log(x))^(n-i))
  sum(y)/m
}
rankitdernier_function(n,i,m)
{
  x_((1:m)-.5)/m
  y_(-n*x*log(x)*(x-x*log(x))^(i-1)*
  (1-x+x*log(x))^(n-i))
  sum(y)/m
}
k<-function(alp)
{
  alp-alp*log(alp)
}
k.inv<-function(beta,minx=0,maxx=1,etapes=20)
{
  x<-0.5
  for(k in 2:etapes)
  {
    x<-x-ifelse(k(x*(maxx-minx)+minx)-

```

```
        beta>0,1,-1)*2^(-k)
    }
    x*(maxx-minx)+minx
  }
u_function(n)
{
  ui_c()

  for(i in 1:n)
    {
      ui[i]_k.inv(i/(n+1))
    }

  ui
}
k3d<-function(alp)
{
  alp-alp*log(alp)+0.5*alp*log(alp)*log(alp)
}
k3d.inv<-function(beta,minx=0,maxx=1,etapes=20)
{
  x<-0.5
  for(k in 2:etapes)
    {
      x<-x-ifelse(k3d(x*(maxx-minx)+minx)-
        beta>0,1,-1)*2^(-k)
```

```

        }
        x*(maxx-minx)+minx
    }
u3d_function(n)
{
    ui3d_c()
    for(i in 1:n)
        {
            ui3d[i]_k3d.inv(i/(n+1))
        }
    ui3d
}
rankitpremier3d_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(n*x*.5*(log(x))^2*(x-x*log(x)+
    x*(log(x))^2*.5^(i-1)*(1-x+x*log(x)-x
    *(log(x))^2*.5^(n-i))
    sum(y)/m
}
rankitspetits3d_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(n*prod((n-i+1):(n-1))/factorial(i-1)*
    x*.5*(log(x))^2*(x-x*log(x)

```

```

+x*(log(x))^2*.5^(i-1)*(1-x+x*log(x)
-x*(log(x))^2*.5^(n-i))
sum(y)/m
}
rankitsgrands3d_function(n,i,m)
{
x_((1:m)-.5)/m
y_(n*prod(i:(n-1))/factorial(n-i)*
x*.5*(log(x))^2*(x-x*log(x)
+x*(log(x))^2*.5^(i-1)*
(1-x+x*log(x)-x*(log(x))^2*.5^(n-i))
sum(y)/m
}
rankitdernier3d_function(n,i,m)
{
x_((1:m)-.5)/m
y_(n*x*.5*(log(x))^2*(x-x*log(x)
+x*(log(x))^2*.5^(i-1)
*(1-x+x*log(x)-x*(log(x))^2*.5^(n-i))
sum(y)/m
}

```

```
kplot3d_function(enx,eny,enw)
{
  k<-function(alp)
  {
    alp-alp*log(alp)
  }
  k.inv<-function(beta,minx=0,maxx=1,etapes=20)
  {
    x<-0.5
    for(k in 2:etapes)
    {
      x<-x-ifelse(k(x*(maxx-minx)+minx)-
        beta>0,1,-1)*2^(-k)
    }
    x*(maxx-minx)+minx
  }
  u_function(n)
  {
    ui_c()

    for(i in 1:n)
    {
      ui[i]_k.inv(i/(n+1))
    }

    ui
```

```
    }
k3d<-function(alp)
  {
    alp-alp*log(alp)+0.5*alp*log(alp)*log(alp)
  }
k3d.inv<-function(beta,minx=0,maxx=1,etapes=20)
  {
    x<-0.5
    for(k in 2:etapes)
      {
        x<-x-ifelse(k3d(x*(maxx-minx)+minx)-
          beta>0,1,-1)*2^(-k)
      }
    x*(maxx-minx)+minx
  }
u3d_function(n)
  {
    ui3d_c()
    for(i in 1:n)
      {
        ui3d[i]_k3d.inv(i/(n+1))
      }
    ui3d
  }
rankitpremier3d_function(n,i,m)
```

```

{
  x_((1:m)-.5)/m
  y_(n*x*.5*(log(x))^2*(x-x*log(x)+
  x*(log(x))^2*.5^(i-1)*(1-x+x*log(x)-x
  *(log(x))^2*.5^(n-i))
  sum(y)/m
}
rankitspetits3d_function(n,i,m)
{
  x_((1:m)-.5)/m
  y_(n*prod((n-i+1):(n-1))/factorial(i-1)*
  x*.5*(log(x))^2*(x-x*log(x)
  +x*(log(x))^2*.5^(i-1)*(1-x+x*log(x)
  -x*(log(x))^2*.5^(n-i))
  sum(y)/m
}
rankitsgrands3d_function(n,i,m)
{
  x_((1:m)-.5)/m
  y_(n*prod(i:(n-1))/factorial(n-i)*
  x*.5*(log(x))^2*(x-x*log(x)
  +x*(log(x))^2*.5^(i-1)*(1-x+x*log(x)
  -x*(log(x))^2*.5^(n-i))
  sum(y)/m
}

```

```

rankitdernier3d_function(n,i,m)
{
  x_((1:m)-.5)/m
  y_(n*x*.5*(log(x))^2*(x-x*log(x)
+x*(log(x))^2*.5)^(i-1)
*(1-x+x*log(x)-x*(log(x))^2*.5)^(n-i))
  sum(y)/m
}
rg3d_function(n)
{
  if ( n>=4 & n<=268 )
  {
    if( n%%2 < 1 )
    {
      upremier3d_rankitpremier3d(n,1,10000)
      upetits3d_sapply(n=n,2:(n/2),
rankitspetits3d,m=10000)
      ugrands3d_sapply(n=n,((n/2)+1):
(n-1),rankitsgrands3d,m=10000)
      udernier3d_rankitdernier3d(n,n,10000)
    }
    else if( n%%2 > 0 )
    {
      upremier3d_rankitpremier3d(n,1,10000)
      upetits3d_sapply(n=n,2:((n+1)/2),

```

```
        rankitspetits3d,m=10000)
        ugrands3d_sapply(n=n,((n+3)/2):
        (n-1),rankitsgrands3d,m=10000)
        udernier3d_rankitdernier3d(n,n,10000)
    }
}
else if (n==1)
{
    upremier3d_rankitpremier3d(1,1,10000)
    upetits3d_c()
    ugrands3d_c()
    udernier3d_c()
}
else if (n==2)
{
    upremier3d_rankitpremier3d(2,1,10000)
    upetits3d_c()
    ugrands3d_c()
    udernier3d_rankitdernier3d(2,2,10000)
}
else if (n==3)
{
    upremier3d_rankitpremier3d(3,1,10000)
    upetits3d_rankitspetits3d(3,2,10000)
    ugrands3d_c()
}
```

```

        udernier3d_rankitdernier3d(3,3,10000)
    }
else if (n>268)
{
    upremier3d_u3d(n)
    upetits3d_c()
    ugrands3d_c()
    udernier3d_c()
}
c(upremier3d, upetits3d, ugrands3d, udernier3d)
}
rankitpremier_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*x*log(x)*(x-x*log(x))^(i-1)
    *(1-x+x*log(x))^(n-i))
    sum(y)/m
}
rankitspetits_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*prod((n-i+1):(n-1))/factorial(i-1)
    *x*log(x)*(x-x*log(x))^(i-1)
    *(1-x+x*log(x))^(n-i))
    sum(y)/m
}

```

```

    }
rankitsgrands_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*prod(i:(n-1))/factorial(n-i)
    *x*log(x)*(x-x*log(x))^(i-1)
    *(1-x+x*log(x))^(n-i))
    sum(y)/m
}
rankitdernier_function(n,i,m)
{
    x_((1:m)-.5)/m
    y_(-n*x*log(x)*(x-x*log(x))^(i-1)
    *(1-x+x*log(x))^(n-i))
    sum(y)/m
}
rg_function(n)
{
    if ( n>=4 & n<=268 )
    {
        if( n%%2 < 1 )
        {
            upremier_rankitpremier(n,1,10000)
            upetits_sapply(n=n,2:(n/2),
            rankitspetits,m=10000)
        }
    }
}

```

```
    ugrands_sapply(n=n, ((n/2)+1):
      (n-1), rankitsgrands, m=10000)
    udernier_rankitdernier(n,n,10000)
  }
else if( n%%2 > 0 )
{
  upremier_rankitpremier(n,1,10000)
  upetits_sapply(n=n, 2:((n+1)/2),
    rankitspetits, m=10000)
  ugrands_sapply(n=n, ((n+3)/2):
    (n-1), rankitsgrands, m=10000)
  udernier_rankitdernier(n,n,10000)
}

}
else if (n==1)
{
  upremier_rankitpremier(1,1,10000)
  upetits_c()
  ugrands_c()
  udernier_c()
}
else if (n==2)
{
  upremier_rankitpremier(2,1,10000)
```

```
    upetits_c()
    ugrands_c()
    udernier_rankitdernier(2,2,10000)
}
else if (n==3)
{
    upremier_rankitpremier(3,1,10000)
    upetits_rankitspetits(3,2,10000)
    ugrands_c()
    udernier_rankitdernier(3,3,10000)
}
else if (n>268)
{
    upremier_u(n)
    upetits_c()
    ugrands_c()
    udernier_c()
}
c(upremier, upetits, ugrands, udernier)
}
if ( length(enx) == length(eny) &
length(enx) == length(enw))
{
    n_length(enx)
    z12_c()
```

```
for(i in 1:length(enx))
{
  petitx_compare(enx[i],enx[-i])
  petity_compare(eny[i],eny[-i])
  petit_(petitx >= 0 & petity >= 0)
  z12[i]_sum(petit)/(length(enx)-1)
}
z13_c()

for(i in 1:length(enx))
{
  petitx_compare(enx[i],enx[-i])
  petitw_compare(enw[i],enw[-i])
  petit_(petitx >= 0 & petitw >= 0)
  z13[i]_sum(petit)/(length(enx)-1)
}
z23_c()

for(i in 1:length(enx))
{
  petity_compare(eny[i],eny[-i])
  petitw_compare(enw[i],enw[-i])
  petit_(petity >= 0 & petitw >= 0)
  z23[i]_sum(petit)/(length(enx)-1)
}
```

```
z123_c()
  for(i in 1:length(enx))
    {
      petitx_compare(enx[i],enx[-i])
      petity_compare(eny[i],eny[-i])
      petitw_compare(enw[i],enw[-i])
      petit_(petitx >= 0 &
             petity >= 0 & petitw >= 0)
      z123[i]_sum(petit)/(length(enx)-1)
    }
par(mfrow=c(2,2))
qqplot(rg(n),z12,xlim=c(0,1),ylim=c(0,1))
title(main="Vector 1 vs Vector 2")
abline(0,1)
qqplot(rg(n),z13,xlim=c(0,1),ylim=c(0,1))
title(main="Vector 1 vs Vector 3")
abline(0,1)
qqplot(rg(n),z23,xlim=c(0,1),ylim=c(0,1))
title(main="Vector 2 vs Vector 3")
abline(0,1)
qqplot(rg3d(n),z123,xlim=c(0,1),ylim=c(0,1))
title(main="Overall")
abline(0,1)
par(mfrow=c(1,1))
paste("OK")
```

```
    }  
else  
    paste("length of x, y and w must be equal")  
}
```

BIBLIOGRAPHIE

- D. F. Andrews & A. M. Herzberg (1985). *Data*. Springer, New-York.
- P. Barbe, C. Genest, K. Ghoudi & B. Rémillard (1996). On Kendall's process. *Journal of Multivariate Analysis*, 58, 197–229.
- P. Capéraà, A.-L. Fougères & C. Genest (1997a). A stochastic ordering based on a decomposition of Kendall's tau. Dans *Distributions with Given Marginals and Moment Problems*, publié sous la direction de V. Beneš et de J. Štěpán, Kluwer, Dordrecht, pp. 81–86.
- P. Capéraà, A.-L. Fougères & C. Genest (1997b). A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84, 567–577.
- P. Capéraà, A.-L. Fougères & C. Genest (2000). Bivariate distributions with given extreme value attractor. *Journal of Multivariate Analysis*, 72, 30–49.
- P. Capéraà & B. van Cutsem (1988). *Méthodes et modèles en statistique non paramétrique*. Dunod, Paris.
- D. G. Clayton (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141–151.
- T. S. Ferguson (1996). *A Course in Large Sample Theory*. Chapman & Hall, Londres.

- N. I. Fisher & P. Switzer (1985). Chi-plots for assessing dependence. *Biometrika*, 72, 253–265.
- N. I. Fisher & P. Switzer (2001). Graphical assessment of dependence: is a picture worth 100 tests? *The American Statistician*, 55, 233–239.
- A. I. Garralda-Guillem (2000). Structure de dépendance des lois de valeurs extrêmes bivariées. *Comptes rendus de l'Académie des sciences de Paris*, série I, 330, 593–596.
- C. Genest & R. J. MacKay (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *La revue canadienne de statistique*, 14, 145–159.
- C. Genest, J.-F. Quessy & B. Rémillard (2002). Tests of serial independence based on Kendall's process. *La revue canadienne de statistique*, 30, 441–461.
- C. Genest & L.-P. Rivest (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88, 1034–1043.
- C. Genest & L.-P. Rivest (2001). On the multivariate probability integral transformation. *Statistics and Probability Letters*, 53, 391–399.
- K. Ghoudi, A. Khoudraji & L.-P. Rivest (1998). Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. *La revue canadienne de statistique*, 26, 187–197.

- M. N. Jouini & R. T. Clemen (1996). Copula models for aggregating expert opinions. *Management Science*, 44, 444–457.
- W. C. M. Kallenberg & T. Ledwina (1999). Data-driven rank tests for independence. *Journal of the American Statistical Association*, 94, 285–301.
- R. B. Nelsen (1999). *An Introduction to Copulas*. Lecture Notes in Statistics No 139. Springer, New-York.
- D. Oakes (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika*, 73, 353–361.
- G. M. Reaven & R. G. Miller (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16, 17–24.
- J. P. Romano & A. F. Siegel (1986). *Counterexamples in Probability and Statistics*. Wadsworth, Londres.
- A. Sklar (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8, 229–231.
- M. B. Wilk & R. Gnanadesikan (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55, 1–17.