Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Uma Nova Abordagem Para Análise De Dependência Bivariada

Vitor Alex Alves de Marchi
Orientador: Prof. Dr. Francisco Louzada Neto
Co-Orientador: Prof. Dr. Francisco Antonio Rojas Rojas

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de São Carlos PPGEs / UFSCar, como parte dos requisitos necessários para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos
Junho/2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M317na

Marchi, Vitor Alex Alves de.
Uma nova abordagem para análise de dependência bivariada / Vitor Alex Alves de Marchi. -- São Carlos : UFSCar, 2010.
108 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Correlação (Estatística). 2. Cópula. 3. Função de Sibuya. 4. Chi-plot. 5. Dados censurados. I. Título.

CDD: 519.537 (20ª)

Vitor Alex Alves de Marchi

Uma Nova Abordagem para Análise de Dependência Bivariada

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 23 de abril de 2010.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Francisco Louzada Neto (DEs-UFSCar/Orientador)

1° Examinador _____
Prof. Dr. Francisco Antonio Rojas Rojas (DEs-UFSCar/Co-Orientador)

2° Examinador _____
Prof. Dr. Nikolai Valtchev Kolev (IME-USP)

3° Examinador _____
Prof. Dr. Vicente Garibay Cancho (ICMC-USP)

Agradecimentos

Agradeço aos meus pais, Pedro e Rosilda, que me criaram e formaram meu caráter, tanto escolar quanto moral, e que me incentivaram na árdua tarefa de estudar, e a minha namorada Sara pela compreensão e estímulo durante os estudos.

A todos os meus amigos e professores do Departamento de Estatística da UFSCar, que sempre estiveram presentes contribuindo com críticas e sugestões para o aprimoramento deste trabalho.

Ao Prof. Dr. Francisco Antonio Rojas Rojas pela fundamental organização e orientação do trabalho, construção, discussão e tempo compartilhado durante o desenvolvimento deste trabalho.

Ao Prof. Dr. Francisco Louzada Neto pelo acolhimento para orientação, sugestões e toda disposição para discussão do trabalho em desenvolvimento.

Agradeço também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio concedido durante o mestrado.

Resumo

Nesta dissertação descrevemos e implementamos procedimentos para estimação não paramétrica da cópula e da função de Sibuya, e também procedimentos para análise de dependência bivariada, baseados no comportamento das suas curvas de nível. Também, descrevemos e implementamos o procedimento chi-plot e um procedimento para a análise de dependência bivariada com presença de censura na amostra.

Particularmente, propomos formas de usá-los em análise de correlação local. O desempenho dos procedimentos propostos são ilustrados e avaliados em casos de estruturas de correlação simples, mas também em esquemas de correlação mais complexa.

Palavras-chave: dependência bivariada, cópula, função de Sibuya, chi-plot, dados censurados.

Abstract

In this dissertation we describe and implement procedures for nonparametric estimation of copulas and Sibuya function, and also procedures for bivariate analysis of dependence based on the behavior of their contours plot. Besisdes, we describe and implement the chi-plot procedure and as well as a procedure for analising bivariate dependence in presence of censoring in the sample.

Particularly, we propose a way to use it in a local correlation analysis. The performance of the proposed procedures are illustrated and evaluated in cases of very simple correlation, but also in a more complex correlation schemes.

Keywords: bivariate dependence, copula, Sibuya function, chi-plot, censored data.

Sumário

1 Introdução

A análise de dados bivariados tem um papel fundamental em várias áreas da estatística, bem como outras áreas do conhecimento, em que as variáveis de interesse são obtidas de forma pareada e se tem interesse em determinar o grau de dependência entre estas variáveis. Como por exemplo podemos citar, a análise multivariada, em que podemos ter duas variáveis de interesse sendo medidas para o mesmo objeto, a análise de sobrevivência, em que podemos ter dados do tempo de ocorrência de um determinado fenômeno em cada um de dois órgãos de um paciente (olhos, rins etc), a confiabilidade industrial, em que podemos ter dados de tempo de falha de dois componentes diferentes que juntos compõem um determinado equipamento, dentre outras.

Em geral, para a análise da dependência entre duas variáveis aleatórias, são utilizadas medidas estatísticas conhecidas como coeficientes de correlação.

As medidas mais difundidas para quantificar a correlação são os coeficiente de correlação: de Pearson, de Spearman e de Kendall. Entretanto, estes coeficientes resumem as relações entre todos os pares em um único valor, o qual é comparado com uma escala de referência. Estes coeficientes falham em situações de dependência como as da Figura 1.1 a) e b).

O coeficiente de correlação de Pearson está ligado diretamente ao ângulo entre os vetores dos dados e tem sua representação correta quando os dados são oriundos de duas distribuições normais. Por sua vez o coeficiente de correlação de Spearman considera os ranks da amostra e é calculado da mesma forma que o coeficiente de correlação de Pearson.

Ambos os coeficientes de Pearson e Spearman consideram somente uma direção para a correlação em toda a amostra, enquanto que o coeficiente de correlação de Kendall conta as quantidades de pares que crescem e que decrescem, resumindo a prevalência do sinal da correlação e quando esta for única se tem que a magnitude é representativa.

Para exemplificar quão falhos estes procedimentos podem ser, consideremos três casos que são de fácil visualização da existência da correlação entre duas variáveis, enquanto que os coeficientes de correlação estimados são iguais a zero.

A Figura 1.1 a) apresenta um aglomerado de pontos gerados sob uma relação quadrática, $y_i = x_i^2 + e_i$ com valores x segundo uma distribuição uniforme com média 10 e ruídos e_i com distribuição normal.

A Figura 1.1 b) apresenta quatro aglomerados de pontos (x, y), onde cada aglome-

rado é gerado de normais independentes.

Na Figura 1.1 c), os dados são provenientes de duas distribuições exponenciais com correlação linear de -0.1.

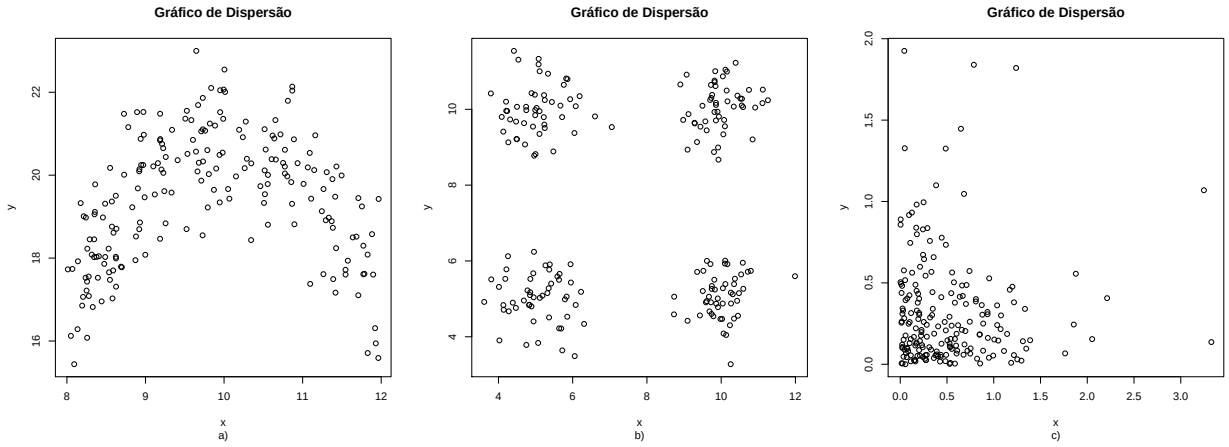

Figura 1.1: Dados com coeficientes de correlação estimados iguais a zero.

Podemos perceber assim que para conjuntos de dados como os da Figura 1.1 há uma necessidade em buscar melhores estimadores com medida de correlação que determine a dependência localmente ou determine regiões para estas dependências.

Na análise de dependência entre duas v.a's pode-se considerar vários aspectos, tal que os coeficientes de correlação de Pearson, de Spearman e de Kendall nem sempre refletem quando a estrutura da dependência é mais complexa. Essas medidas não evidenciam se o grau de correlação muda sua magnitude ou, até mesmo, onde ocorre a transição de uma correlação positiva para uma correlação negativa.

Uma alternativa para análise de dependência que tem se tornado comum em algumas áreas como economia e finanças, para estudar as relações de dependência entre variáveis aleatórias contínuas é a análise via Cópulas. A dependência entre variáveis aleatórias está ligada à função de distribuição conjunta dessas variáveis, mas nem sempre a função de distribuição conjunta informa de que tipo é essa dependência, e a cópula tem como incumbência carregar a estrutura de dependência entre as variáveis (Sklar 1959).

A análise de dependência desenvolvida por Gonçalves (2008), permite identificação de dependência local entre as variáveis como negativa, positiva ou nula, em uma determinada região através da Função Dependência de Sibuya, chamada daqui em diante por Função de Sibuya. A função de Sibuya somente revela com clareza o sinal da dependência local mas não tem capacidade de apresentar uma indicação clara da magnitude da correlação.

Outro procedimento desenvolvido para a análise de dependência é encontrado em Fisher & Switzer (1985), baseado no gráfico denominado Chi-plot, definido sobre as funções de distribuição empíricas dos dados que permite, além da verificação do sinal da correlação, quantificar a força da dependência entre as variáveis aleatórias.

1.1 Estrutura Organizacional do Trabalho

Na Seção 1.2 a seguir apresentaremos os coeficientes de correlação de Pearson, de Spearman e de Kendall, com suas estatísticas para os testes de significância e alguns exemplos de aplicação.

No decorrer do trabalho, com objetivo estabelecido, nos focamos na análise gráfica dos resultados dos procedimentos de estimação não paramétricos que são introduzidos ao longo do trabalho e na distribuição assintótica para χ do Chi-plot no Capítulo 6.

No Capítulo 2 é introduzida a análise por meio de cópulas onde o trabalho se concentra em estimar a Cópula por métodos não paramétricos para obter uma superfície suavizada e analisar suas curvas de nível.

No Capítulo 3 é introduzida a Função de Sibuya e sua estimação é obtida por métodos não paramétricos de forma a obter uma superfície suavizada e analisar suas curvas de nível.

No Capítulo 4 é introduzido o procedimento de análise de dependência pelo Chi-plot e suas características intrínsecas segundo Fisher & Switzer (1985) e Fisher & Switzer (2001).

O Capítulo 5 é reservado para comparação dos resultados obtidos em quatro tipos de dependência denominadas “linear”, “quadrática”, “normais”, “exponenciais”, e de independência entre dados normais através da Cópula estimada, da Função de Sibuya estimada e do Chi-plot.

Para o Chi-plot é apresentado no Capítulo 6, um estudo da consistência das suas propriedades inerentes apresentadas no Capítulo 4, afim de constatar quais destas propriedades são eficientes para se concluir sobre a independência. Ainda, apresentamos um estudo para sobre a distribuição assintótica do Chi-plot e estudamos sua consistência. Os estudos são realizados sobre os cinco conjuntos de dados apresentados no Capítulo 5.

No Capítulo 7 apresentamos um estudo que verifica a dependência local para dados censurados a partir do coeficiente de correlação geral introduzido por Gumbel (1960). Também são apresentados três exemplos para ilustração e a aplicação em um conjunto de

dados reais.

O Capítulo 8 é reservado para comentários e conclusões gerais sobre todos os aspectos desenvolvidos neste trabalho.

As principais rotinas dos procedimentos utilizadas neste trabalho se encontram no Apêndice B.

1.2 Coeficientes de Dependência Usuais

De acordo com Conover (1980), a medida de correlação indica a força e a direção da dependência entre duas variáveis aleatórias, sendo utilizada para quantificar o grau da correlação linear ou simplesmente o grau de dependência geral da amostra.

Tradicionalmente, uma medida de correlação entre duas variáveis aleatórias X e Y contém como características:

(i) A medida da correlação assume valores no intervalo $(-1, 1)$;

(ii) Se os maiores valores de X tendem a ser pareados com os maiores valores de Y, e os menores valores de X aos menores valores de Y, então a medida de correlação é positiva e se aproxima de 1;

(iii) Se os maiores valores de X tendem a ser pareados com os menores valores de Y e vice-versa, então a medida de correlação deve ser negativa e é próxima de -1;

(iv) Se valores de X são aleatoriamente pareados com os valores de Y, então, a medida de correlação deve estar próxima de zero.

Três medidas de correlação são discutidos nesta dissertação. Trata-se dos coeficientes de correlação desenvolvidos por Pearson, Spearman ou por Kendall.

1.2.1 Correlação de Pearson

A medida de correlação mais conhecida e utilizada, é o coeficiente de correlação linear momento-produto de Pearson, denotada por r e definida por (Pearson 1896)

$$r = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{1.2.1}$$

onde $E(.)$ é finita e $\sigma_.$ é positivo.

Para uma amostra $(x_i, y_i), \ldots, (x_n, y_n)$ aleatória de (X, Y), o coeficiente de correlação linear de Pearson é estimado por

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2]^{\frac{1}{2}}}. \tag{1.2.2}$$

Exemplo 1.2.1 Considere a seguinte amostra bivariada de (X, Y) (ver Tabela 1.1) com tamanho $n = 20$, cujo gráfico de dispersão é apresentado na Figura 1.2.

Tabela 1.1: Amostra aleatória bivariada de (X, Y) do exemplo 1.2.1.

n	1	2	3	4	5	6	7	8	9	10
x	2.85	2.48	3.01	3.41	3.09	3.31	0.48	4.07	4.40	3.92
y	2.09	2.01	0.61	1.65	1.89	1.45	1.13	2.61	2.66	1.81
n	11	12	13	14	15	16	17	18	19	20
x	2.24	3.69	4.15	2.77	2.14	3.62	3.17	3.37	2.15	3.04
y	1.69	0.22	3.30	2.39	0.90	1.12	1.36	1.24	2.66	2.42

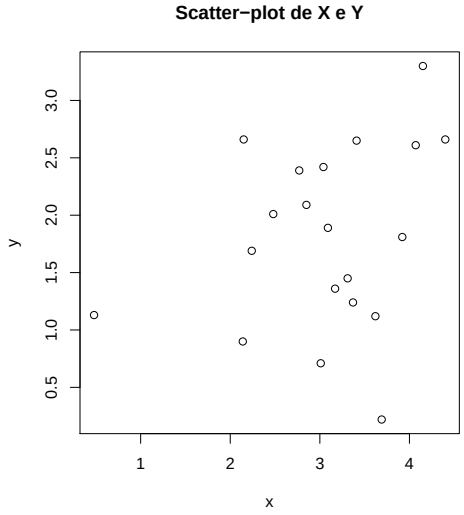

Figura 1.2: Gráfico de dispersão do Exemplo 1.2.1.

Na Tabela 1.1 são apresentados o valores dos dados e o coeficiente de correlação de Pearson estimado de acordo com a Equação (1.2.2) é $\hat{r} = 0.2793$.

O coeficiente de correlação proposto por Pearson para uma amostra, pode ser utilizado em qualquer conjunto de dados numéricos sem qualquer conhecimento da distribuição

dos dados, entretanto r é uma v.a e possui uma distribuição que depende da função distribuição conjunta de (X, Y), como pode ser visto da Equação (1.2.1). Este fato, em princípio, pode ser visto como um problema, pois não se conhece a distribuição conjunta de (X, Y), contornado pela aproximação à transformada Z de Fisher, para se obter um intervalo de confiança para r.

O software R estima um IC para o coeficiente de correlação de Pearson considerando uma aproximação t de Student com $n-2$ graus de liberdade. Por esta aproximação, a partir do intervalo de 95% de confiança $[-0.20, 0.63]$, chegamos a conclusão que $\hat{r} = 0.2599$ não é diferente de zero. Portanto X e Y são independentes.

Exemplo 1.2.2 Considere uma amostra bivariada de (X, Y) de tamanho $n = 13$ (ver Tabela 1.2).

Tabela 1.2: Amostra aleatória bivariada de (X, Y) do exemplo 1.2.2.

n	1	2	3	4	5	6	7	8	9	10	11	12	13
x	3.72	4.51	3.99	4.23	3.24	4.32	3.26	2.50	5.55	4.20	3.63	4.00	5.15
y	5.66	6.83	6.09	6.21	5.42	6.35	5.66	5.24	7.55	6.45	6.09	6.19	7.01

O valor do coeficiente de correlação de Pearson estimado é $\hat{r} = 0.9650$ e é considerado significante, uma vez que seu IC de 95% é dado por $(0.8827, 0.9896)$.

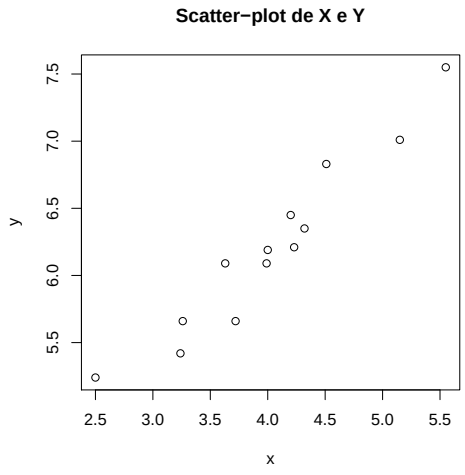

Figura 1.3: Gráfico de dispersão do exemplo 1.2.2.

Para contornar o problema do coeficiente r depender da f.d.p de (X, Y), foram desenvolvidas duas medidas não paramétricas denominadas coeficiente de correlação de Spearman

e coeficiente de correlação de Kendall.

1.2.2 Correlação de Spearman

A medida de correlação de Spearman, denotado ρ, não faz nenhuma suposição sobre a f.d.p de (X, Y) e utiliza somente os ranks e o tamanho da amostra.

A medida do coeficiente de correlação de Spearman é dado por (Conover 1980)

$$\rho = \frac{\sum_{i=1}^{n} \left[R(X_i) - \frac{n+1}{2}\right] \left[R(Y_i) - \frac{n+1}{2}\right]}{\frac{n(n^2-1)}{12}}, \tag{1.2.3}$$

sendo $R(X_i)$ e $R(Y_i)$ os ranks de x_i e y_i, e n o tamanho da amostra.

Em princípio, o coeficiente de correlação de Spearman é um caso particular do coeficiente de correlação de Pearson, pois ao se tomar os ranks da amostra, sua esperança e variância são as mesmas de variáveis uniformes discretas.

Uma expressão equivalente para ρ, porém mais fácil de calcular, é dada por (Conover, 1980)

$$\rho = 1 - 6\frac{\sum_{i=1}^{n} D_i^2}{n(n^2-1)}, \tag{1.2.4}$$

onde $D_i = R(X_i) - R(Y_i)$.

De acordo com Splent & Smeeton (2001), a aproximação de ρ à distribuição *t de Student* é calculada como

$$t_{n-2} = \frac{\rho}{\sqrt{(1-\rho^2)/(n-2)}}. \tag{1.2.5}$$

O coeficiente de correlação de Spearman para os dados do Exemplo 1.2.1 é $\hat{\rho} = 0.06$. A aproximação de t_{n-2} segundo a Equação (1.2.5) é 0.262 com $P[t \geq 0.262] = 0.40$. Portanto X e Y devem ser consideradas independentes. O coeficiente de correlação de Spearman para os dados do Exemplo (1.2.2) é $\hat{\rho} = 0.51$, a aproximação de t_{n-2} é 1.96 com $P[t \geq 1.96] = 0.04$ e portanto X e Y devem ser consideradas dependentes.

1.2.3 Correlação de Kendall

O coeficiente de correlação desenvolvido por Kendall, denotado pela letra τ, requer mais cálculos que o coeficiente ρ de Spearman, entretanto o τ de Kendall converge mais rapidamente em distribuição para a normal padrão quando a hipótese de independência entre X e Y é satisfeita.

Seja $(x_1, y_1), \ldots, (x_n, y_n)$ *uma amostra aleatória. Duas observações,* (x_k, y_k) *e* (x_i, y_i) *com* $1 \leq i, k \leq n$*, são chamados de concordantes se* $(x_k - y_k)(x_i - y_i) > 0$ *e discordantes se* $(x_k - y_k)(x_i - y_i) < 0$*. Caso* $x_k = x_i$ *ou* $y_k = y_i$ *para quaisquer* $1 \leq i, k \leq n$*, estes pares não são nem concordantes nem discordantes.*

O coeficiente de correlação proposto por Kendall é dado por (Conover 1980)

$$\tau = \frac{N_c - N_d}{n(n-1)/2}, \tag{1.2.6}$$

onde N_c é o numero de pares concordantes e N_d é o número de pares discordantes.

Uma interpretação intuitiva do τ de Kendall, em termos de probabilidades, é que este coeficiente determina a diferença entre o número de pares concordantes e discordantes e divide por todas as combinações possíveis dos pares dois a dois, tornando-se assim em uma proporção da diferença entre os pares concordantes e discordantes.

De acordo com Splent & Smeeton (2001), para se determinar se o coeficiente τ de Kendall é significante para valores de $n < 60$ é aconselhável utilizar a estatística $T = N_c - N_d$ tabelada pelos quantis w_p para confiança p. Entretanto, para amostras de tamanho $n \geq 60$ pode-se utilizar a aproximação

$$w_p = z_{1-p/2}\sqrt{\frac{n(n-1)(2n+5)}{18}}, \tag{1.2.7}$$

onde $z_{1-p/2}$ é o quantil da normal padrão com confiança $1 - p$.

Abaixo são reproduzidos de Conover (1980) os valores para w_p com valores de n entre 10 e 30, e com valores de p iguais a 0.900, 0.950, 0.975 e 0.990.

O coeficiente de correlação de Kendall para os dados do Exemplo 1.2.1 é $\hat{\tau} = 0.11$ com $T = 21$. O valor de $w_p = 50$ com n=20 e $p = 0.95$. Portanto X e Y são consideradas independentes. Enquanto, o coeficiente de correlação de Kendall para os dados do Exemplo 1.2.2 é $\hat{\tau} = 0.89$ com $T = 70$. O valor de $w_p = 20$ com n=13 e $p = 0.95$. Portanto X e Y não são consideradas independentes.

Tabela 1.3: Quantis da Statistica Test w_p de Kendall.

n	p=0.900	0.950	0.975	0.990
10	15	19	21	25
11	17	21	25	29
12	18	24	28	34
13	22	26	32	38
14	23	31	35	41
15	27	33	39	47
16	28	36	44	50
17	32	40	48	56
18	35	43	51	61
19	37	47	55	65
20	40	50	60	70
21	42	54	64	76
22	45	59	69	81
23	49	63	73	87
24	52	66	78	92
25	56	70	84	98
26	59	75	89	105
27	61	79	93	111
28	66	84	98	116
29	68	88	104	124
30	73	93	109	129

2 Análise via Cópulas

2.1 Introdução

A cópula, como função de dependência entre variáveis aleatórias, surgiu quando Sklar estudava uma função distribuição tridimensional conjunta e introduziu funções auxiliares definidas no suporte unitário que ligavam a função distribuição às suas marginais. Foi quando notou que tais funções poderiam ser construídas para casos n-dimensionais (Nelsen, 2006). Por ser uma função que liga a função distribuição conjunta às suas marginais e ser invariante através de transformação linear, a cópula captura uma estrutura envolvida na relação entre as variáveis contida na distribuição conjunta, e tem sido utilizada para análise de dependência em muitos estudos.

Uma cópula equivale a uma função distribuição multivariada com marginais uniformes em $[0, 1]$, que contém a estrutura de dependência entre as variáveis aleatórias envolvidas. A teoria de cópulas tem sido amplamente utilizada para casos bivariados, tendo em vista que uma estrutura de dependência entre duas v.a's é muito comum em várias áreas do conhecimento. Neste caso, a cópula $C(u, v)$ é uma função distribuição conjunta bivariada de duas v.a's u e v, ambas tendo distribuição uniforme em $[0, 1]$.

Uma cópula bidimensional, para quaisquer $0 \leq u_1 \leq u_2 \leq 1$ e $0 \leq v_1 \leq v_2 \leq 1$, possui as seguintes características:

1. $C(u, 0) = \int_0^u \int_0^0 c(u, v)dvdu = 0 = \int_0^v \int_0^0 c(u, v)dudv = C(0, v)$;

2. $C(u, 1) = u = \int_0^u \int_0^1 c(u, v)dvdu$ e $C(1, v) = v = \int_0^v \int_0^1 c(u, v)dudv$;

3. $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$.

As características (1) e (2) mostram que a cópula tem distribuições marginais uniformes em $[0, 1]$. A característica (3) mostra que a cópula é uma função crescente. Isto pode ser verificado por meio da Figura 2.1 que apresenta uma representação gráfica da característica (3). Temos que as regiões mais escuras representam o valor da cópula a serem retirados do valor da cópula $C(u_2, v_2)$ e, portanto no quadrado formado pela intersecção das retas, a cópula tem volume sempre maior ou igual a zero.

O teorema a seguir é um resultado obtido por Fréchet durante suas correspondências com Sklar e é utilizado para a interpretação do comportamento da dependência da cópula.

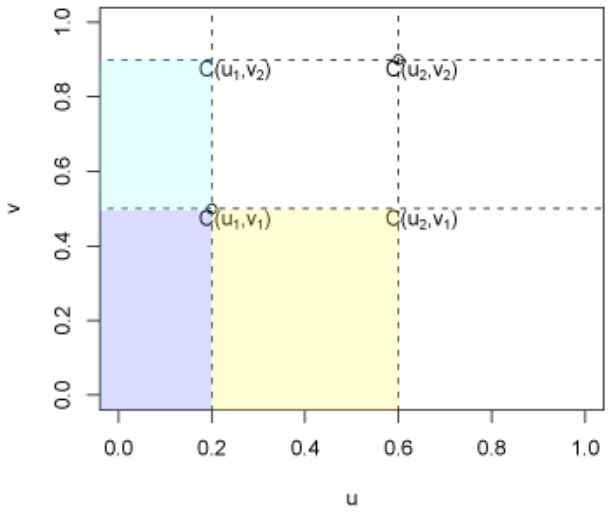

Figura 2.1: Representação gráfica da característica (3).

Teorema 2.1 (Limites de Fréchet) *Seja $C(u,v)$ uma cópula, então para todo $(u,v) \in \mathbf{I}^2$*

$$\max\{u+v-1,0\} \leq C(u,v) \leq \min\{u,v\}. \tag{2.1.1}$$

Denotamos $W(u,v) = \max\{u+v-1,0\}$ e $M(u,v) = \min\{u,v\}$.

Os limites de Fréchet são uteis na comparação das curvas de nível da cópula em estudo, com os limitantes W e M e o caso de independência (Corolário 2.1.2), pois estes limitantes são a máxima correlação positiva e negativa.

Teorema 2.2 *Seja (X,Y) com cópula $W(u,v)$ ou $M(u,v)$. Então existem funções monótonas $\alpha(z)$ e $\beta(z)$ de $\mathbb{R}$ em $\mathbb{R}$, e uma variável aleatória Z tal que*

$$(X,Y) =_d (\alpha(Z), \beta(Z)), \tag{2.1.2}$$

onde o sinal, $=_d$, indica igualdade em distribuição. Note que $\alpha(z)$ e $\beta(z)$ são ambas crescentes se, e somente se, (X,Y) tem cópula $M(u,v)$, ou temos $\alpha(z)$ crescente e $\beta(z)$ decrescente se, e somente se, (X,Y) tem cópula $W(u,v)$.

A teoria sobre cópulas está fundamentada no teorema de Sklar, o qual mostra a relação entre a distribuição dos dados, suas marginais e a cópula. Este teorema surgiu em 1959 e o nome cópula foi escolhido para destacar a ligação das funções de distribuição marginais à função distribuição conjunta das v.a's envolvidas Nelsen (2006).

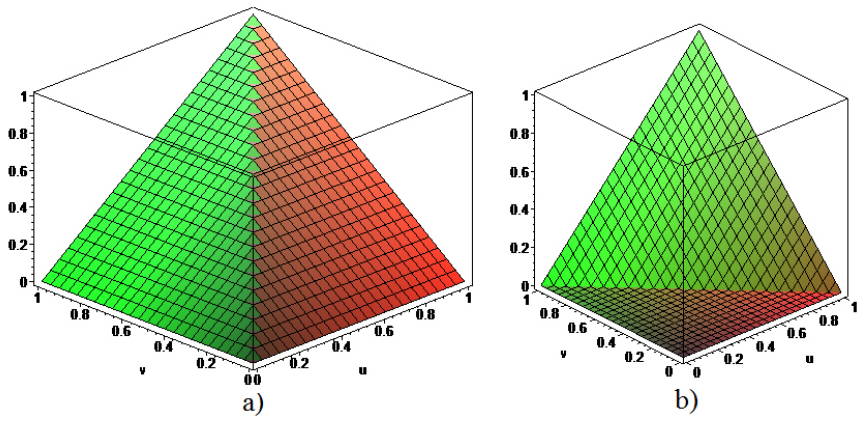

Figura 2.2: Cópula M(u,v) em a) e W(u,v) em b).

Teorema 2.3 (Teorema de Sklar) *Seja H(x,y) a função distribuição conjunta de X e Y, com função distribuição marginais F(x) e G(y), respectivamente. Então, existe uma cópula C(u,v) tal que, para todo* $(x, y) \in \mathbb{R}^2$,

$$H(x, y) = C(F(x), G(y)). \tag{2.1.3}$$

Ainda, se F(x) e G(y) são contínuas, a cópula C(u,v) é única.

O Teorema 2.3 mostra que para cada função distribuição conjunta existe uma cópula associada e, como consequência, é possível obter a cópula associada a esta função distribuição conjunta.

Corolário 2.1.1 (Ulisses et al 2004) *Seja H(x,y) a f.d.a conjunta de X e Y com f.d.a F(x) e G(y), respectivamente. Então, para quaisquer* $(u, v) \in \mathbb{R}^2$

$$C(u, v) = H(F^{-1}(u), G^{-1}(v)). \tag{2.1.4}$$

Exemplo 2.1.1 Seja a função distribuição bivariada dada por

$$\begin{aligned} H(x, y) &= [1 + \exp(-x) + \exp(-y)]^{-1}, x, y \in \Re, \text{ com} \\ F(x) &= [1 + \exp(-x)]^{-1}, x \in \Re \text{ e } F^{-1}(u) = -log(u^{-1} - 1); \\ G(y) &= [1 + \exp(-y)]^{-1}, y \in \Re \text{ e } G^{-1}(v) = -log(v^{-1} - 1). \end{aligned} \tag{2.1.5}$$

Então, a cópula associada a $H(x, y)$ é determinada por

$$H(F^{-1}(u), G^{-1}(v)) = \left[u^{-1} + v^{-1} - 1\right]^{-1}, u, v \in [0, 1]. \tag{2.1.6}$$

Corolário 2.1.2 (Ulisses et. al 2004) *Sejam X e Y variáveis aleatórias contínuas, então X e Y serão independentes se, e somente se,*

$$C(u, v) = uv. \tag{2.1.7}$$

O Teorema 2.4 a seguir mostra que a cópula obtida de uma transformação estritamente crescente em cada variável marginal, é invariante. Este resultado é utilizado para automatizar o processo de suavização da cópula empírica, apresentado no final do capítulo.

Teorema 2.4 *Seja (X, Y) v.a.'s contínuas com cópula $C(u, v)$, Se α_1, α_2 são funções estritamente crescentes, então $(\alpha_1(X), \alpha_2(Y))$ também tem cópula $C(u, v)$.*

Se α_1, α_2 são funções monótonas, suponha α_1 estritamente decrescente, então

$$C_{\alpha_1(X),\alpha_2(Y)}(u, v) = C_{\alpha_2(Y)}(v) - C_{X,\alpha_2(Y)}(1 - u, v).$$

Para determinar a cópula a ser utilizada para analisar um conjunto de dados é comum considerar uma cópula com características conhecidas, ou que tenha as funções marginais conhecidas. No entanto, como um dos objetivos desta dissertação, apresentamos um procedimento não paramétrico de estimação da cópula associada à um conjunto de dados bivariados, que utiliza apenas a informação dos próprios dados, para o qual o procedimento de suavização se torna imprescindível visando estabelecer se há dependência, ou não, entre os valores $x's$ e $y's$ das amostras, e, se houver dependência, se a mesma é positiva ou negativa.

A análise da cópula empírica suavizada com curvas de nível, associada à amostra, tem seu gráfico apresentado em formas de curvas de nível correspondentes às cópulas de dependência positiva perfeita em a), de independência em b) e de dependência negativa perfeita em c). A Figura 2.3 apresenta as curvas de nível. O processo de suavização da cópula empírica aqui adotado e implementado segue o proposto por Fermanian et al (2004).

2.2 Cópula empírica

Seja (x_i, y_i); $i = 1, \ldots, n$ uma amostra aleatória de (X, Y) com função distribuição $H(x, y)$. As funções:

$$F_i = \frac{\#\{x \leq x_i\}}{n};\ G_i = \frac{\#\{y \leq y_i\}}{n} \text{ e } H_{ij} = \frac{\#\{x \leq x_i, y \leq y_j\}}{n}, \tag{2.2.1}$$

representam as funções de distribuição acumuladas empíricas, de X no ponto $x = x_i$, de Y no ponto $y = y_i$, e de (X, Y) nos pontos $x = x_i$ e $y = y_j$. Como em Nelsen (2006), a cópula

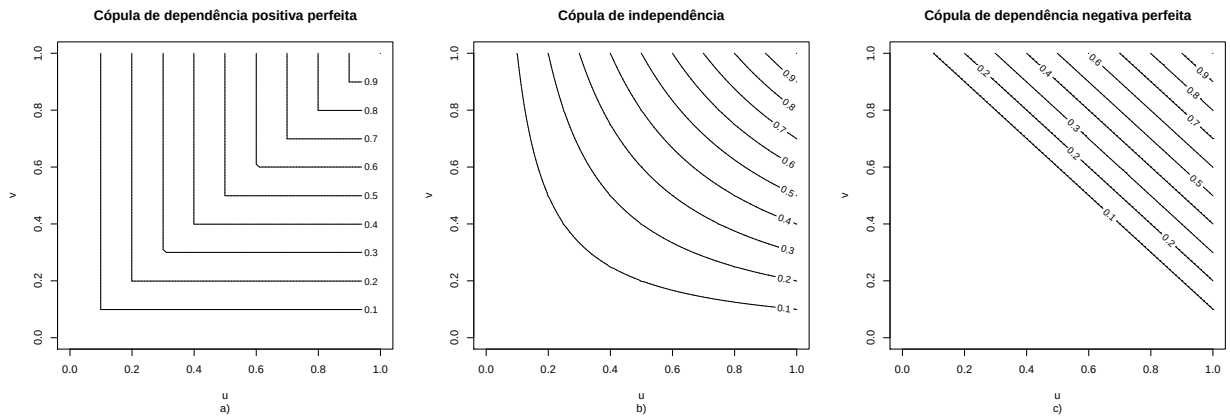

Figura 2.3: Curvas de nível da cópula de a) dependência perfeita positiva, b) independência e c) dependência perfeita negativa.

empírica associada à $H(x, y)$, é definida por

$$C(F_i, G_j) = H_{ij}, \tag{2.2.2}$$

possibilitando visualizar uma estimativa gráfica do comportamento da cópula empírica para os dados da amostra.

Exemplo 2.2.1 Considere a seguinte amostra de (X, Y) apresentada na Tabela 2.1.

Tabela 2.1: Dados do exemplo 2.2.1.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x	0.12	-0.76	-0.58	0.86	1.06	-1.54	-0.33	-0.03	0.68	0.36	-2.33	-0.97	-0.25	0.03	0.97
y	1.04	1.10	0.40	-0.81	0.18	0.20	-0.82	1.34	-1.25	1.02	-0.18	1.39	-0.72	-1.26	1.45

A Figura 2.4 apresenta o gráfico de dispersão, a cópula empírica e suas respectivas curvas de nível referentes ao conjunto de dados da Tabela 2.1.

Pela Figura 2.4 b), observamos que a Cópula empírica não parece apresentar uma boa aproximação para dados contínuos. Esta aproximação no entanto parece ser muito ruim se observarmos suas curvas de nível da Figura 2.4 c).

A análise de dependência pelas curvas de nível não suavizadas da cópula empírica pode ser pouco conclusiva, pela dificuldade de comparação com as curvas teóricas. Por isso, a suavização da cópula empírica é indispensável.

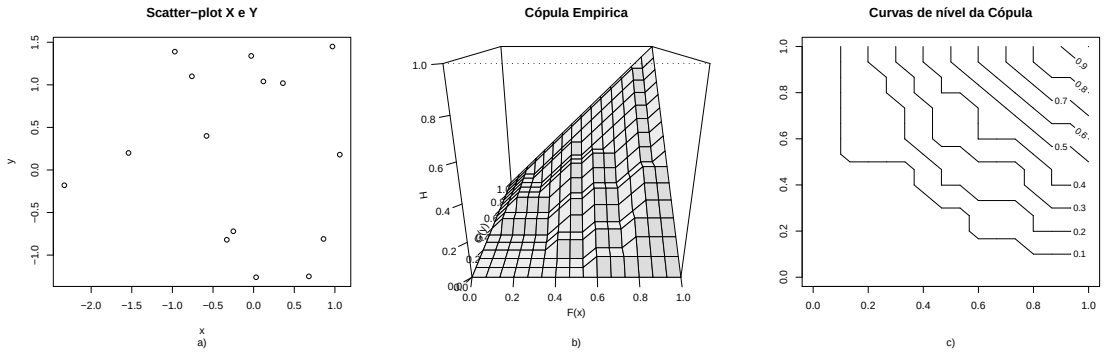

Figura 2.4: a) Scatter-plot de X e Y b)Cópula empírica c) Curvas de nível da cópula empírica.

2.3 Suavização Para Distribuições Empíricas

O estimador não paramétrico para cópulas contínuas proposto em Fermanian et al (2004), propõe-se a suavizar as distribuições empíricas conjunta e marginais para obter, por meio de uma função kernel, a cópula empírica suavizada.

Seja

$$\widehat{H}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i, y - Y_i), \tag{2.3.1}$$

a suavização empírica da função distribuição $H(x,y)$, onde $K_n(x,y) = K(h_n^{-1}x, h_n^{-1}y)$ tal que

$$K(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} k(u,v)dudv, \tag{2.3.2}$$

para uma função kernel k(função densidade de probabilidade), $\int\int k(u,v)dudv = 1$ e $h_n \to 0$ quando $n \to \infty$ uma sequencia de n.

A sequência h_n está ligada diretamente a diferença entre a esperança do estimador proposto e a função de distribuição real dos dados. Portanto, para um valor de h_n muito pequeno, o valor estimado e o valor da função de distribuição real são muito próximas.

Podemos propor um estimador para $\widehat{F}(x)$ usando uma função kernel univariada, tal que

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i), \tag{2.3.3}$$

para uma função kernel k, com $\int k(u)du = 1$, $K_n(x) = k(h_n^{-1}x)$ e $h_n \to 0$ quando $n \to \infty$.

Similarmente, podemos aplicar o mesmo procedimento para estimar a função de densidade acumulada (f.d.a) $\widehat{G}(y)$. Então, podemos estimar a cópula $C(u,v)$ utilizando a

suavização por kernel como sendo

$$\widehat{C}_n(u,v) = \widehat{H}_n(\widehat{F}_n^{-1}(u), \widehat{G}_n^{-1}(v)), \quad 0 \leq u, v \leq 1. \tag{2.3.4}$$

Os coeficientes de correlação de Spearman e de Kendall, obtidos através de cópulas analíticas, não são utilizados neste trabalho porque envolvem os procedimentos para a estimação da cópula de densidade, como em Charpentier et al. (2006), que não são de fácil implementação e nem sempre apresentam uma boa estimação. Portanto os cálculos podem apresentar ineficiência e/ou uma estimativa não confiável. A avaliação da correlação através da cópula se concentra basicamente na avaliação subjetiva do analista sobre as curvas de nível.

Para o exemplo a seguir é feita uma padronização dos dados para contornar problemas na suavização que podem surgir por escolha de uma função kernel que não suavize a estimação devido a magnitude dos dados. Esta padronização não afeta o procedimento de estimação de acordo com o Teorema 2.4.

Para a função kernel da Equação (2.3.2), foi escolhida a função de distribuição acumulada normal bivariada com vetor de médias iguais a zero e matriz de covariância Σ dada por

$$\Sigma = \begin{bmatrix} 0.3^2 & 0 \\ 0 & 0.3^2 \end{bmatrix}. \tag{2.3.5}$$

A função kernel é simétrica com relação ao ponto $(0,0)$ e a escolha da função normal é de uso comum. Os pontos para estimação da cópula são escolhidos de acordo com a distribuição acumulada empírica e a função h_n foi escolhida como $1/\sqrt{n}$.

Exemplo 2.3.1 Considerando os dados do Exemplo 2.2.1, a Figura 2.5 a) apresenta o scatter-plot de (X, Y), a cópula suavizada e as curvas de nível na Figura 2.5 b) e c), respectivamente. Pela Figura 2.5 b), podemos perceber que a cópula apresenta uma superfície bem suave se comparada com a cópula da Figura 2.4 b). As curvas de nível da cópula na Figura 2.5 c) também são bem mais suaves que as da Figura 2.4 c), o que facilita a interpretação da dependência através da comparação das curvas de nível.

A convergência do procedimento proposto por Fermanian el al (2004), está baseado na convergência assintótica, sendo assim, quanto maior for o tamanho da amostra, mais

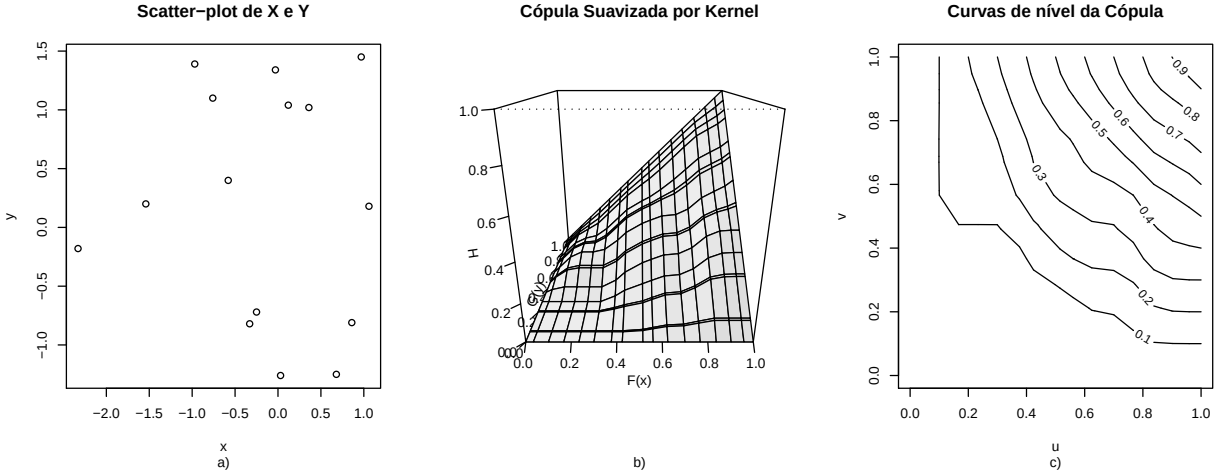

Figura 2.5: a) Scatter-plot de X e Y b)Cópula suavizada c) Curvas de nível da cópula suavizada.

próximas das funções reais de distribuição envolvidas, as cópulas e as curvas de nível suavizadas ficarão das correspondentes estimadas.

Para ilustrar o processo de suavização de Fermanian para amostras maiores, mostramos a seguir dois exemplos.

Exemplo 2.3.2 Seja $X \sim Exp(3)$ e $Y \sim Exp(6)$, onde $Exp(\alpha)$ se refere a distribuição exponencial com taxa α. Então, a f.d.a conjunta como em Gumbel (1960) e sua cópula são dadas, respectivamente, por:

$$\begin{aligned} H(x,y) &= [(1-\exp(-3x))(1-\exp(-6y))]\,[1-0.5\exp(-3x-6y)], (x,y) \in (0,\infty)\times(0,\infty) \\ C(u,v) &= [uv][1-0.5(1-u)(1-v)],\ (u,v) \in [0,1]\times[0,1]. \end{aligned}$$

com correlação linear de $\frac{-0.5}{4} = -0.125$.

Para (x_i, y_i); com $i = 1, \ldots, 100$; gerados através do método da transformada inversa (Luc Devroye, 1986), estimamos sua cópula através do processo de suavização proposto por Fermanian. Na Figura 2.6 são apresentadas a cópula verdadeira e a cópula estimada.

A estimativa do coeficiente de correlação de *Pearson* de acordo com a Equação 1.2.2 é $\hat{r} = -0.1042$, a do coeficiente de correlação *rho* de *Spearman* de acordo com a Equação 1.2.3 é $\hat{\rho} = -0.087$, e a do coeficiente de correlação de *Kendall* de acordo com a Equação 1.2.6 é $\hat{\tau} = -0.056$, estando o coeficiente de correlação de *Pearson* mais próximo da magnitude esperada, mas todos são não significantes. Observando as curvas de nível da

cópula suavizada, juntamente com as da cópula de independência na Figura 2.7, nota-se que são bem próximas. Pelas curvas de nível das cópulas verdadeira e suavizada pode-se perceber que existe uma correlação negativa de acordo com as curvas de nível da cópula de independência.

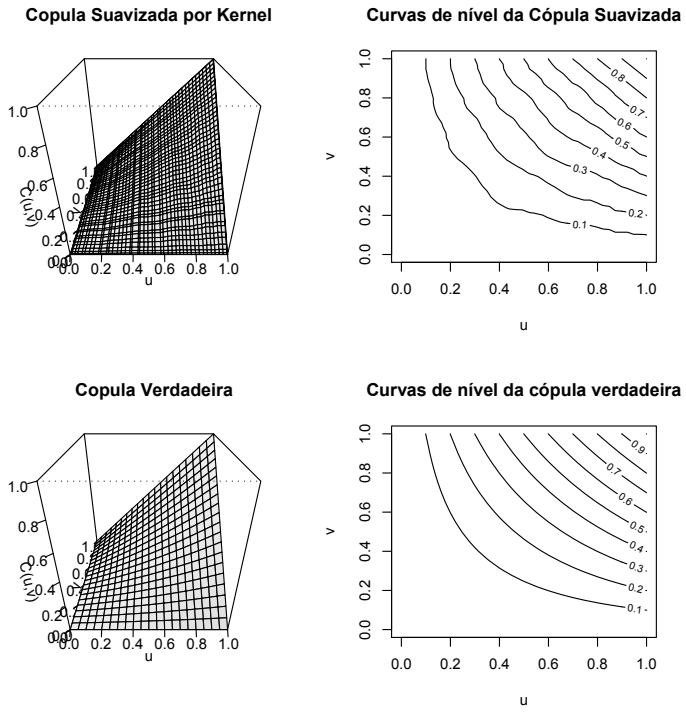

Figura 2.6: Cópulas suavizada e verdadeira e respectivas curvas de nível associadas a H(x,y) do Exemplo 2.3.2.

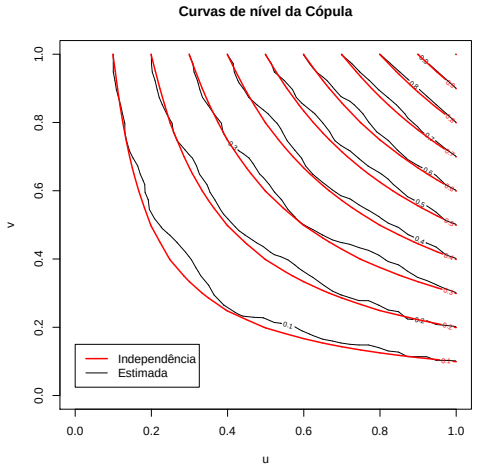

Figura 2.7: Curvas de nível das cópulas estimada para H(x,y) e de independência para o Exemplo 2.3.2.

Exemplo 2.3.3 Sejam x_i e y_i, o peso e a altura em centímetros para 507 indivíduos, com idade entre 18 e 67 e distribuídos entre homens e mulheres (Grete et al 2003).

Na Figura 2.8 são apresentadas as curvas de nível da cópula estimada e verificamos que apresentam o comportamento característico das curvas de nível da estrutura de dependência positiva muito forte. Os valores estimados dos coeficientes de correlação de *Spearman*, *Kendall* e *Pearson* são 0.7318, 0.5438 e 0.7173, repectivamente.

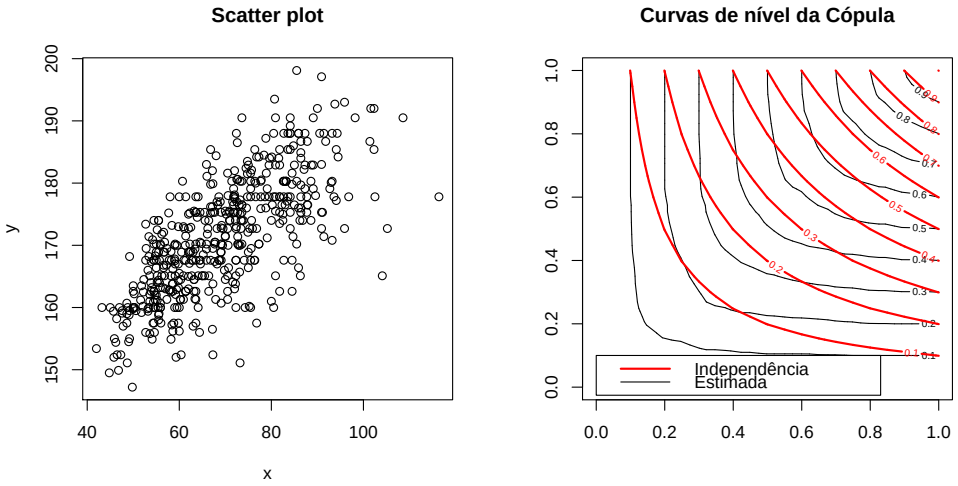

Figura 2.8: Amostras de X e Y como peso e altura respectivamente.

Apesar da intensa utilização de cópulas para representação do relacionamento conjunto de duas variáveis aleatórias, Nelsen (2006) mostrou que uma cópula pode corresponder a pelo menos duas funções distribuições diferentes, o que o próprio Nelsen caracteriza com um problema. Uma alternativa para contornar este problema consiste da utilização da Função Dependência de Sibuya (Gonçalves 2008) o qual descrevemos no próximo capítulo.

3 Análise via Função de Sibuya

3.1 Introdução

Em 1960, Sibuya e muitos outros autores estavam desenvolvendo estudos relacionados aos comportamentos das funções limites, chamadas de valores extremos, definidas sobre as estatísticas de máximo ou mínimo valor de uma amostra. Seu interesse principal era de estudar a relação entre as funções de distribuição marginais bivariadas, criando assim uma função de dependência incorporada à função de distribuição bivariada.

A partir do estudo desenvolvido por Sibuya, Gonçalves (2008) estende seu uso para todo o espaço paramétrico, sendo assim, não é mais de interesse somente o estudo baseado em valores extremos. Esta extensão surge da necessidade de que a cópula seja única. Ou seja, de evitar que possamos ter duas funções de distribuição: $H_1(X,Y)$ e $H_2(X,Y)$, tais que as marginais F_1 e G_1 de $H_1(X,Y)$ não coincidam com as marginais F_2 e G_2 de $H_2(X,Y)$ e, mesmo assim, tenham a mesma cópula associada. Nelsen (2006) refere um caso em que $H_1(X,Y)$ e $H_2(X,Y)$ tem a mesma cópula associada, que apresentamos no seguinte exemplo.

Exemplo 3.1.1 Considere $H_1(X,Y)$ e $H_2(X,Y)$ tais que:

$$H_1(x,y) = \begin{cases} \frac{(x+1)(\exp(y)-1)}{x+2\exp(y)-1}, & (x,y) \in [-1,1] \times [0,\infty]; \\ 1-\exp(-y), & (x,y) \in (1,\infty) \times [0,\infty]; \\ 0 & \text{caso contrário.} \end{cases} \tag{3.1.1}$$

$$H_2(x,y) = (1+\exp(-x)+\exp(-y))^{-1}, (x,y) \in \Re^2. \tag{3.1.2}$$

Então,

$$H_1(F_1^{-1}(u), G_1^{-1}(v)) = \frac{uv}{u+v-uv} = H_2(F_2^{-1}(u), G_2^{-1}(v)) \text{ logo}$$

$$C^* = \frac{uv}{u+v-uv}, \tag{3.1.3}$$

onde $F_1^{-1}, G_1^{-1}, F_2^{-1}$ e G_2^{-1} são as inversas de F_1, G_1, F_2 e G_2 respectivamente dadas por:

$$\begin{aligned} F_1 &= \frac{(x+1)}{2}; \quad x \in [-1,1); \\ G_1 &= 1-\exp(-y); \quad y \in [0,\infty); \\ F_2 &= [1+\exp(-x)]^{-1}; \quad x \in \Re; \\ G_2 &= [1+\exp(-y)]^{-1}; \quad y \in \Re. \end{aligned} \tag{3.1.4}$$

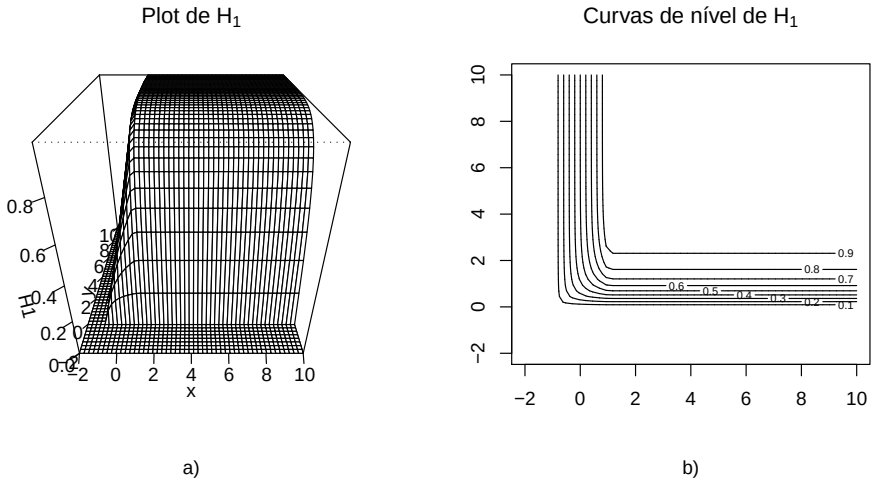

Figura 3.1: Em a) o gráfico de $H_1(x, y)$ e em b) as curvas de nível de $H_1(x, y)$.

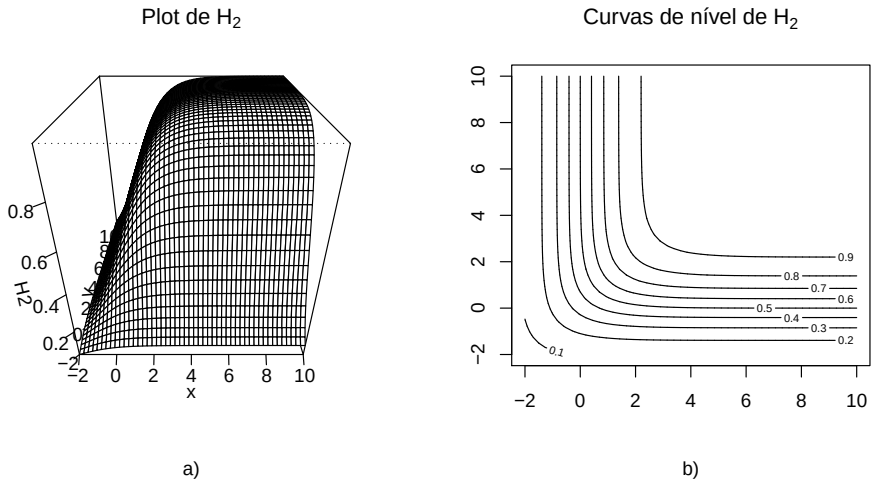

Figura 3.2: Em a), o gráfico de $H_2(x, y)$ e em b) as curvas de nível de $H_2(x, y)$.

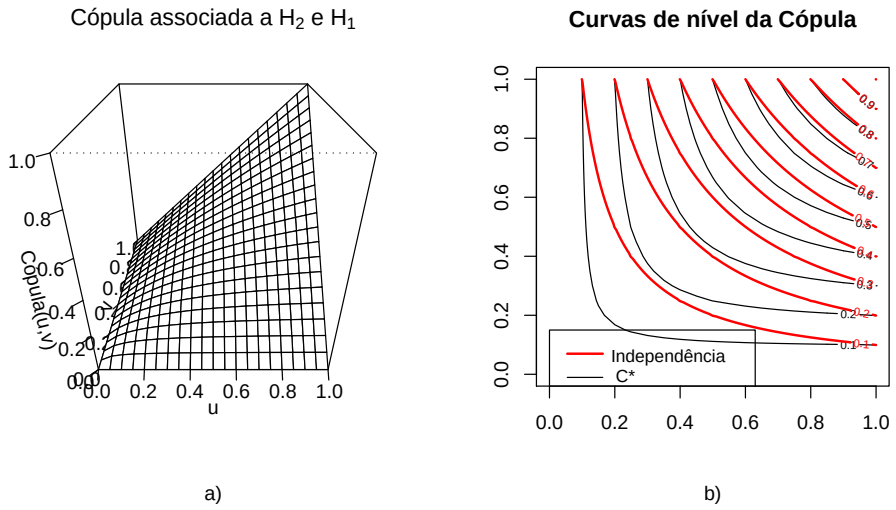

Figura 3.3: Em a) o gráfico da Cópula C^* e em b) as suas curvas de nível.

Observando as Figuras 3.1 b) e 3.2 b), podemos perceber que as f.d.a's conjuntas têm curvas de nível distintas mas com a mesma estrutura, o que pode ser visto pela cópula e suas curvas de nível na Figura 3.3.

3.2 Propriedades da Função de Sibuya

Definição 1 (Gonçalves 2008) *A Função de Sibuya, Ω, é definida com base na relação*

$$H(x,y) = \Omega(F(x), G(y))F(x)G(y), \tag{3.2.1}$$

tal que

(1) $\Omega(F(x), G(y)) = 1$ se, e somente se, X e Y são v.a's independentes;

(2) Se $\alpha(x)$ e $\beta(y)$ são funções monótonas não decrescentes, as funções $\Omega(F(x), G(y))$ e $\Omega(F(\alpha(x)), G(\beta(y)))$ são idênticas;

(3) $\max\left\{0, \frac{F(x)+G(y)-1}{F(x)G(y)}\right\} \leq \Omega(F(x), G(y)) \leq \min\left\{\frac{1}{F(x)}, \frac{1}{G(y)}\right\}$.

A propriedade (2) é utilizada para uma transformação linear sobre os dados com o intuito de automatizar o processo de suavização, visto que dependendo da magnitude dos dados a suavização por kernel pode falhar, pela escolha da função h_n da Equação (2.3.1). A propriedade (3) é utilizada para corrigir erros nos valores da suavização obtida.

Da Equação (3.2.1) pode-se reescrever $\Omega(F(x), G(y))$, tal que

$$\Omega(F(x), G(y)) = \begin{cases} \frac{H(x,y)}{F(x)G(y)}, & F(x), G(y) > 0 \\ 0, \quad \text{se} & F(x), G(y) = 0 \end{cases}$$

$$= \begin{cases} \frac{C(F(x),G(y))}{F(x)G(y)}, & F(x), G(y) > 0 \\ 0 \quad \text{se} & F(x), G(y) = 0 \end{cases},$$

o que vale somente para o caso contínuo.

A função de Sibuya pode ser utilizada para classificar a dependência de uma região como positiva ou negativa, mas para esta classificação é necessário enunciar um critério chamado de "Quadrant Dependence".

Definição 2 *A função distribuição conjunta $H(x, y)$ das variáveis aleatórias X e Y, com f.d.a marginais $F(x)$ e $G(y)$ é "positive quadrant dependent" se $H(x, y) > F(x)G(y)$ para algum $(x, y) \in \mathbb{R}^2$, ou é "negative quadrant dependent" se $H(x, y) < F(x)G(y)$ para algum $(x, y) \in \mathbb{R}^2$.*

Este tipo de dependência pode ser avaliada considerando o fato de podermos escrever

$$H(x, y) = F(x|Y = y)G(y) = G(y|X = x)F(x), \tag{3.2.2}$$

ou seja, se considerarmos $H(x, y) = F(x|Y = y)G(y)$ com $H(x, y)$ sendo "positive quadrant dependent", temos que $F(x|Y = y) > F(x)$.

Exemplo 3.2.1 Considere o caso onde o suporte de $H(x, y)$ é determinado pela forma retangular como na figura 3.4 e os valores indicando a área em porcentagem de cada região, limitadas pelas retas de $Y = y$ e x. Ao considerarmos a restrição que somente os valores menores que $Y = y$ são possíveis de ocorrer, temos $F(x|Y = y) = 0.55$, entretanto $F(x) = 0.50$. Logo $F(x|Y = y) > F(x)$, $H(x, y)$ é "positive quadrant dependent".

Segue então da Definição (2) que

1. Se $H(x, y)$ é 'Negative Quadrant Dependent'(NQD), então $\Omega(F(x), G(y)) < 1$.

2. Se $H(x, y)$ é 'Positive Quadrant Dependent'(PQD), então $\Omega(F(x), G(y)) > 1$.

3. Se $F(x)$ e $G(y)$ são independentes, então $\Omega(F(x), G(y)) = 1$.

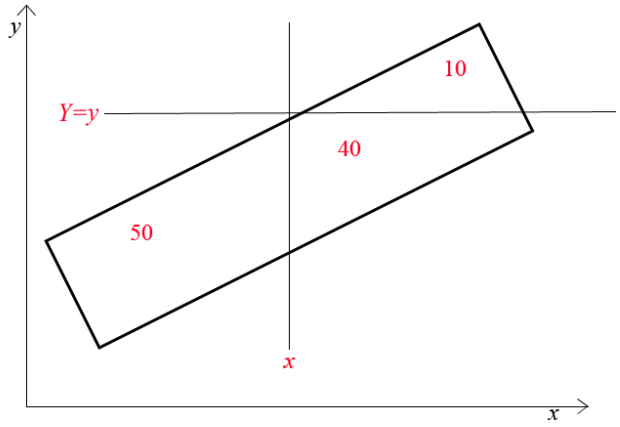

Figura 3.4: Exemplo de uma relação "positive quadrant dependent".

3.3 Função Dependência de Sibuya

Definição 3 (Gonçalves 2008) *Para quaisquer pares de variáveis aleatórias contínuas X e Y, com distribuição conjunta $H(x,y)$ e marginais $F(x)$ e $G(y)$, temos que*

$$\Omega(x,y) = \frac{H(x,y)}{F(x)G(y)} \tag{3.3.1}$$

é chamada de Função Dependência de Sibuya(Gonçalves 2008). Nesta dissertação chamamos simplesmente Função de Sibuya.

A diferença entre as Equações (3.2.1) e (3.3.1) é a extensão do quadrado unitário determinado por $(F(x), G(y))$ para todo o suporte da distribuição $H(x,y)$.

A Função de Sibuya descreve a dependência entre X e Y levando em consideração a f.d.a conjunta $H(x,y)$ e as contribuições das marginais $F(x)$ e $G(y)$. Como visto anteriormente se $\Omega(x,y) \neq 1$, então a Função de Sibuya representa o distanciamento de $H(x,y)$ com relação à estrutura de independência (Gonçalves 2008). No entanto, $\Omega(x,y) \in [0,\infty)$, logo não se tem uma clara interpretação da magnitude da dependência entre X e Y.

Exemplo 3.3.1 Considere $H_1(x,y)$ e $H_2(x,y)$, como descritas no exemplo (3.1.1). Então, $\Omega_1(x,y)$ e $\Omega_2(x,y)$ são:

$$\Omega_1(x,y) = \begin{cases} \frac{2}{2-(1-x)\exp(-y)} > 1, & (x,y) \in [-1,1] \times [0,\infty]; \\ 1, & (x,y) \in (1,\infty) \times [0,\infty]; \\ 0 & \text{caso contrário.} \end{cases}$$

$$\Omega_2(x,y) = 1 + \frac{\exp\left[-(x+y)\right]}{1+\exp\left(-y\right)+\exp\left(-x\right)}, (x,y) \in \Re^2. \tag{3.3.2}$$

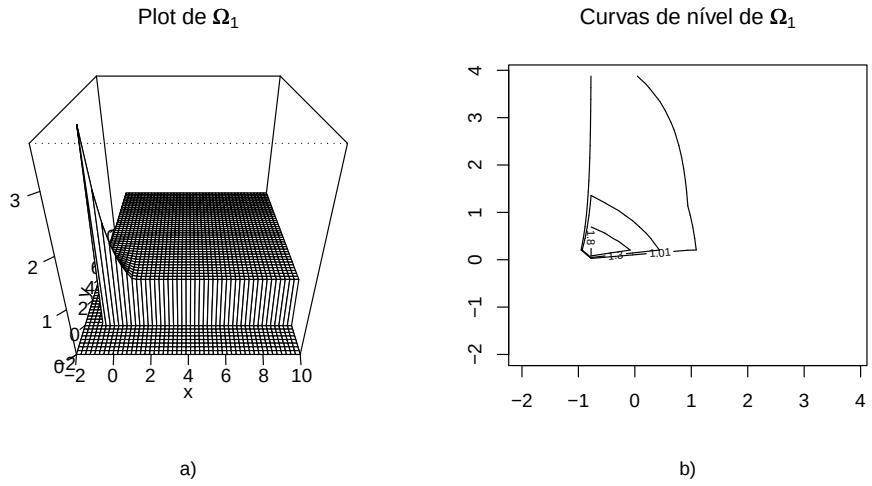

Figura 3.5: Plot de $\Omega_1(x,y)$em a), Curvas de nível de $\Omega_1(x,y)$ em b).

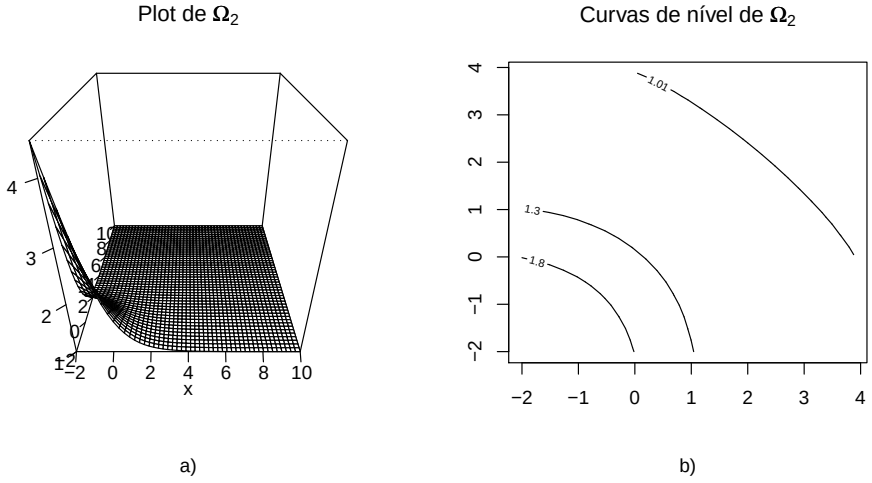

Figura 3.6: Plot de $\Omega_2(x,y)$em a), Curvas de nível de $\Omega_2(x,y)$ em b).

Observando as Figuras 3.5 e 3.6 e as respectivas curvas de nível, podemos perceber que as funções de dependência de Sibuya Ω_1 e Ω_2, são diferentes, enquanto que há uma mesma cópula associada a H_1 e H_2.

Da análise das curvas de nível da Figura 3.5 b), conjuntamente com a Equação (3.3.1), conclue-se que, no intervalo $(x,y) \in [-1,1] \times [0,\infty]$, a dependência é positiva. No restante do suporte de $H_1(x,y)$, as variáveis X e Y são independentes. Ainda, conforme se aproxima do ponto $(-1,0)$ aumenta a magnitude da correlação. Da observação das curvas de nível da Figura 3.6 b), conjuntamente com a Equação (3.3.2), conclue-se que as variáveis

X e Y são positivamente dependentes em todo o suporte de $H_2(x, y)$ e com magnitude da dependência maior para x e y decrescendo.

3.4 Função de Sibuya Empírica

Seja (x_i, y_i); com $i = 1, \ldots, n$; uma amostra aleatória de (X, Y). Considere as funções de distribuição empírica de X, de Y e de (X, Y) como em (2.2.1), respectivamente. Então,

$$\Omega_i(x, y) = \frac{H_i(x, y)}{F_i(x)G_i(y)}, \tag{3.4.1}$$

é a função empírica de Sibuya $\Omega(X, Y)$. Observe que quando $n \to \infty$, $F_i(x) \to F(x)$, $G_i(y) \to G(y)$, $H_i(x, y) \to H(x, y)$ e $\Omega_i(x, y) \to \Omega(x, y)$ para todo (x, y). Observe também que se X e Y são variáveis aleatórias independentes, para $n \to \infty$, $\Omega_i(x, y) \to 1$ quase certamente e , além disso, $E[\Omega_i(x, y)] \to 1$ (Gonçalves 2008).

Com esses resultados (Gonçalves 2008), desenvolveu um procedimento para testar independência entre as variáveis aleatórias X e Y com base na amostra (x_i, y_i); com $i = 1, \ldots, n$, calculando as estimativas de Ω, de $E[\Omega]$, de $\sigma^2[\Omega]$ e de $z = \dfrac{\Omega(x, y) - E[\Omega]}{\sigma[\Omega]}$, respectivamente,

$$\Omega_i = \frac{H_i(x, y)}{F_i(x)G_i(y)}, \quad \Omega_n = \frac{\sum_{i=1}^{n} \Omega_i}{n}, \tag{3.4.2}$$

$$s^2_{\Omega_n} = \frac{\sum_{k=1}^{n} (\Omega_i - \Omega_n)^2}{n - 1} \text{ e } \hat{z}_n = \sqrt{n}\frac{\Omega_n - 1}{s_{\Omega_n}}. \tag{3.4.3}$$

e considerando que, sob a hipótese de independência, $\lim_{n\to\infty} \hat{z_n}$ tem distribuição aproximadamente $N(0, 1)$, a verificação da hipótese de independência entre X e Y é imediata. Note que $E[\Omega(X, Y)] = 1$. Ainda um intervalo com $100(1 - \alpha)\%$ de confiança para $E[\Omega(X, Y)]$ pode ser definido por,

$$\left[\Omega_n - z_{1-\frac{\alpha}{2}}\frac{s_{\Omega_n}}{\sqrt{n}}; \Omega_n + z_{1-\frac{\alpha}{2}}\frac{s_{\Omega_n}}{\sqrt{n}}\right], \tag{3.4.4}$$

sendo $z_{1-\alpha/2}$ o quartil $(1 - \alpha/2)$ da distribuição $N(0, 1)$.

3.5 Suavização da função de Sibuya empírica

Como na Seção 2.3, podemos usar os estimadores para a função distribuição $H(x, y)$ como o da Equação (2.3.1), e para $F(x)$ e $G(y)$ os da Equação (2.3.3). Lembrando que os "Kernels" utilizados não são necessariamente idênticos.

Para ilustrar o uso da Função de Sibuya, apresentaremos alguns gráficos de dispersão, as funções de Sibuya e as curvas de nível correspondentes.

Exemplo 3.5.1 Sejam 200 pares gerados de (X, Y), onde $X \sim U(8, 12)$, $Y_i = (X_i - 10)^2 + 20 + e_i$ e $e_i \sim N(0.8, 1)$, cujo gráfico de dispersão aparece na Figura 3.7 a). Claramente há a dependência entre os valores de x_i e y_i, pois os pontos em y tais que para $x_i \leq 10$ têm correlação positiva, e para $x_i \geq 10$ têm correlação negativa. Olhando para a Função de Sibuya, pode-se reparar um comportamento atípico próximo as menores valores de x_i e de y_i, nesta região a f.d.a conjunta tem o valor estimado próximo de zero, e portanto não devemos considerar este comportamento, pois caso contrário seríamos levados a crer que nesta região a correlação positiva é muito mais forte que nas demais regiões.

Pelas curvas de nível da Figura 3.7 d), e) e f), pode-se perceber uma separação nítida em duas regiões, uma com correlação positiva do lado esquerdo e correlação negativa do lado direito, da curva de nível $\Omega = 1$ a qual passa próximo de $x = 10$.

O coeficiente de correlação de Pearson $\hat{r}$ é estimado como zero, sugere independência, enquanto que o coeficiente de correlação de Spearman $\hat{\rho}$ indica uma correlação bem fraca de 0.195, como também o coeficiente de correlação de Kendall $\hat{\tau}$ indica uma correlação bem fraca de 0.13. Por outro lado, as estimativas Ω_n, $\hat{z}_n$ e o IC, também sugerem que haja dependência entre X e Y:

$$\Omega_n = 2.03;\ \hat{z}_n = 3.71;\ IC = [1.49, 2.589].$$

Se separarmos a amostra nas duas regiões sugeridas pelas curvas de nível, ou seja, uma amostra A, que contenha os valores com $x_i \leq 10$ e o restante em uma amostra B, então calculando os coeficientes de correlação para cada região obtemos para a amostra A, $\hat{r} = 0.754$, $\hat{\rho} = 0.763$ e $\hat{\tau} = 0.553$. Para a amostra B, $\hat{r} = -0.697$, $\hat{\rho} = -0.676$ e $\hat{\tau} = -0.485$.

Nota-se então que a Função de Sibuya pode trazer informação importante quanto a estrutura de dependência geral e local.

Exemplo 3.5.2 Seja (x_i, y_i); com $i = 1, \ldots, 200$; gerados de $X \sim N(0, 1)$ e $Y \sim N(0, 1)$, sendo X e Y independentes. O gráfico de dispersão dos (x_i, y_i), na Figura 3.8 a), sugere que não existe dependência entre os valores de x e y. Do gráfico da função de Sibuya (ver Figura 3.8 b)), pode-se reparar um comportamento isolado atípico na região onde não há pontos no scatter-plot, pois a f.d.a conjunta e as f.d.a's marginais tem valores muito próximos de zero. Como nestes casos mencionados, devemos desconsiderar tais comportamento.

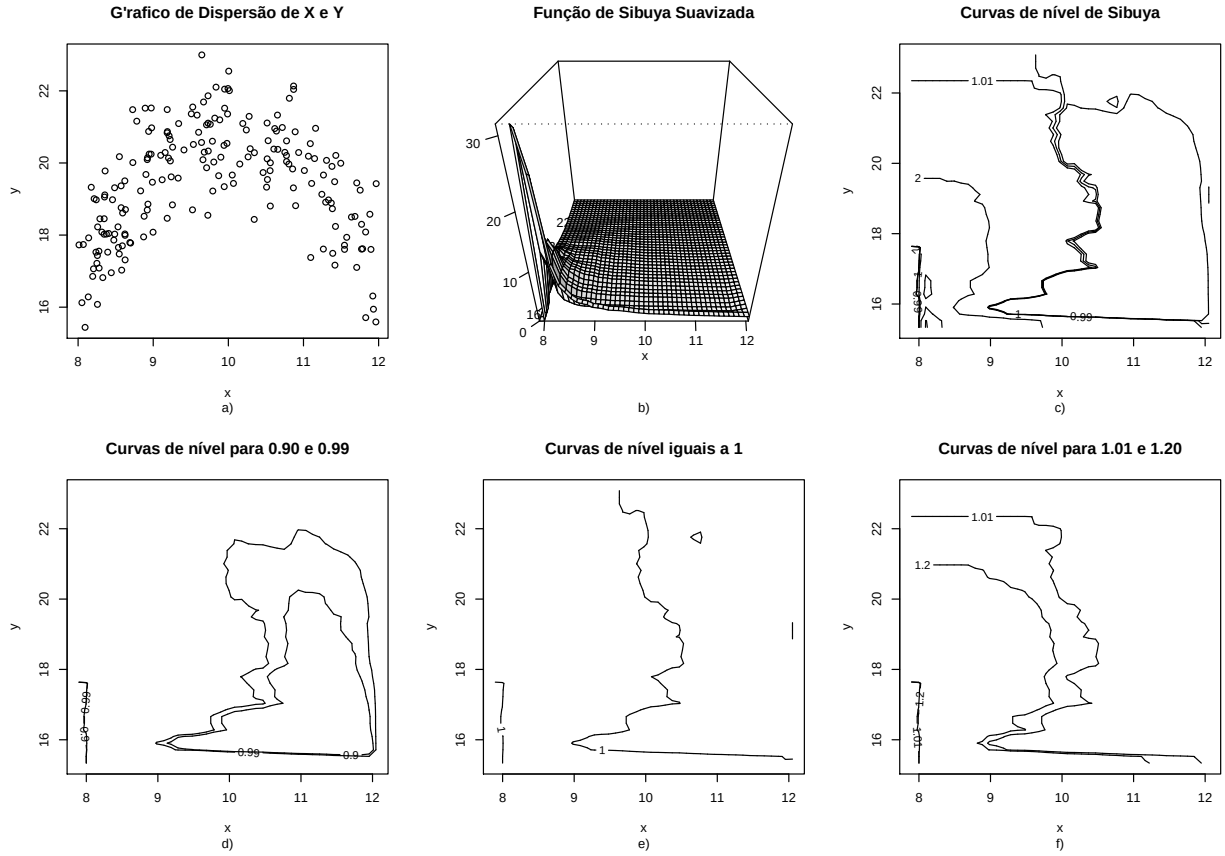

Figura 3.7: a)Scatter-plot b)Função de Sibuya c)Curvas de nível da Função de Sibuya do exemplo 3.5.1.

Na região abaixo e à esquerda, da Figura 3.8 c), a função de Sibuya sempre permanece abaixo do valor 1, oscilando entre 0.97 e 0.99, o que pode ser observado pelas curvas de nível da Figura 3.8 d). Isso nos leva a acreditar que nessa região o caso de independência é pouco evidente. Observando a Figura 3.8 a) e c), pode-se observar que há uma concentração de pontos à direita que são classificados com dependência positiva, como pode ser constatado pelas curvas de nível da Figura 3.8 f). Ainda, pode ser observado pela curva de nível da Figura 3.8 e) uma região que separa as regiões de dependência positiva e negativa.

Os coeficientes de correlação de Pearson, Spearman e Kendall iguais a 0, também sugerem independência entre x e y. Além disso, $\Omega n = 1.1315$, $IC = [0.9842, 1.2787]$, com $\hat{z}_n = 1.7500$. Esses resultados também sugerem independência entre X e Y.

Como mencionado anteriormente, não é possível uma interpretação da magnitude da função de Sibuya, portanto em cada caso as curvas de nível têm valores escolhidos para uma melhor visualização. As estimativas de $\hat{\rho}$, $\hat{\tau}$ e $\hat{r}$ foram obtidas através do software R e a medida $\hat{z}_n$ do teste de independência como na Equação (3.4.3).

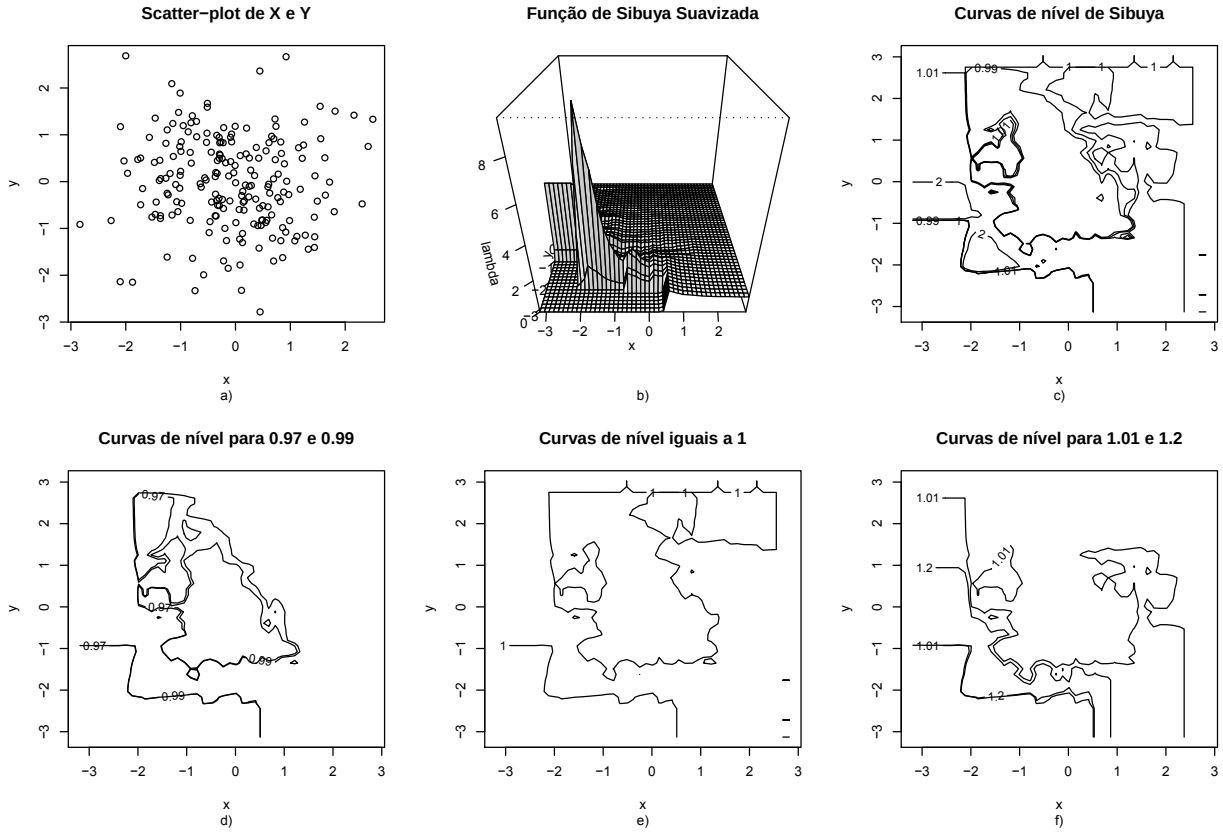

Figura 3.8: a): Scatter-plot; b): Função de Sibuya; c) d) e) e f): Curvas de nível da Função de Sibuya, do Exemplo 3.5.2.

Nos gráficos da Função de Sibuya suavizada é importante notar que $\Omega(x, y) = 0$, para $H(x, y) = 0$ e, ainda quando $H(x, y)$ é pequena, $\Omega(x, y)$ poderá assumir valores extremamente grandes em comparação com os demais valores de $\Omega(x, y)$.

Alguns valores atípicos de $\Omega(x, y)$ podem ser estimados também quando os valores das f.d.a's marginais forem próximos de zero. Também vale lembrar que, quando as funções acumuladas marginais são próximas do valor 1, a Função $\Omega(x, y)$ se aproxima do valor 1, mas não deve ser interpretado como uma região em que ocorra a independência. Salvo esta última observação, os outros casos atípicos ocorrem por causa do processo de suavização.

Em ambos os dois últimos exemplos, a análise de dependência na Função de Sibuya no gráfico que contém curvas de nível de dependências positiva e negativa e de independência, não se mostra como melhor opção, pois pela quantidade de curvas que apresenta dificulta a análise. A maneira mais eficaz de analisar a dependência por meio das curvas de nível da Função de Sibuya é analisar as curvas de nível para os casos isolados de dependência positiva, negativa e de independência.

4 Análise via Chi-Plot

4.1 Introdução

A estrutura de dependência entre pontos de uma amostra aleatória bivariada pode conter diversos aspectos, que podem ser interessantes de serem identificados. Porém algumas medidas de dependência nem sempre os refletem, como por exemplo na Figura 1.1, onde as estruturas de dependência são mais complexas, com regiões de dependência com magnitudes e/ou sinais distintos.

Para situações desse tipo, uma ferramenta de análise gráfica que parece apropriada é conhecida como Chi-plot e foi desenvolvida por Fisher & Switzer (1985), e consiste na representação gráfica dos valores de duas medidas de dependência local associadas a cada ponto da amostra.

Os Chi-plot proporcionam maior informação e facilidade de interpretação, com relação aos coeficientes de correlação usuais ou outros procedimentos de análise de dependência. Ainda, sua utilização pode ser vantajosa em relação à função de Sibuya porque, além de mostrar o sinal da dependência local, o valor da medida pontual de dependência pode ser interpretado.

4.2 Construção do Chi-Plot

Sejam (x_i, y_i) com $i = 1, \ldots n$ uma amostra aleatória de (X, Y). Considere respectivamente a *Função Distribuição Empírica* conjunta $H(x, y)$ e marginais $F(x)$ e $G(y)$ da seguinte forma:

$$H_i = \frac{\sum_{j \neq i} I\{x_j \leq x_i, y_j \leq y_i\}}{(n-1)}, \tag{4.2.1}$$

$$F_i = \frac{\sum_{j \neq i} I\{x_j \leq x_i\}}{(n-1)}, \tag{4.2.2}$$

$$G_i = \frac{\sum_{j \neq i} I\{y_j \leq y_i\}}{(n-1)}, \tag{4.2.3}$$

onde $I(A)$ é a função indicadora sobre o conjunto A.

Considere também

$$S_i = \text{sinal}\{(F_i - 0.5)(G_i - 0.5)\}, \tag{4.2.4}$$

$$\chi_i = (H_i - F_i G_i)/\sqrt{F_i(1-F_i)G_i(1-G_i)}, \tag{4.2.5}$$

$$\lambda_i = 4S_i \max\{(F_i - 0.5)^2, (G_i - 0.5)^2\}. \tag{4.2.6}$$

O Chi-plot é o scatter-plot de (χ_i, λ_i), pata todos os $|\lambda_i| < 4\left(\frac{1}{n-1} - 0.5\right)^2$(Fisher, & Switzer, 1985). Esta restrição vem do fato que os pontos que estão nas extremidades da amostra influenciam a aproximação assintótica do Chi-plot, segundo os autores, ou simplesmente pelo chi-plot não estar definido para alguns pontos das extremidades.

Fisher, N. & Switzer, P.(2001) propõe considerar

$$\left(-\frac{c_p}{\sqrt{n}}, \frac{c_p}{\sqrt{n}}\right), \tag{4.2.7}$$

como intervalo de confiança para χ_i, com valores c_p obtidos através de simulação de Monte Carlo para o caso da independência entre X e Y, tais que $c_p = 1.54, c_p = 1.78$ e $c_p = 2.18$, para níveis de confiança de $90\%, 95\%$ e 99%, respectivamente.

4.3 Propriedades de χ_i e λ_i

Existem pontos (x_i, y_i) onde a medida χ_i não está definida. Estes ocorrem quando temos $F_i = 0$, $G_i = 0$, $F_i = 1$ e $G_i = 1$. Note que $\chi \in [-1, 1]$.

Se y_i é uma função estritamente crescente de x_i, temos que $\chi_i = 1$, e se y_i é uma função estritamente decrescente de x_i, temos que $\chi_i = -1$. Por outro lado, $\lambda_i \in [-1, 1]$, pois $\max\{(F_i - 0.5)^2, (G_i - 0.5)^2\}$ tem valor máximo igual a 0.25. Se as v.a's são independentes, λ_i é uniformemente distribuído. Além disso, $\lambda_i/4$ é o valor da maior distância entre (x_i, y_i) e $(\tilde{x}, \tilde{y})$, sendo $\tilde{x}$ e $\tilde{y}$ as medianas dos valores x_i e dos y_i, respectivamente.

Quando há independência entre os valores x_i e y_i, então χ_i é aleatoriamente distribuído em torno de zero. Se y_i é crescente(decrescente) em relação a x_i, temos que $\lambda_i > 0 (< 0)$. Se Y for positivamente(negativamente) associada com X, ou seja, $Cov(Y, X) > 0 (< 0)$, há tendência de a maioria dos valores de λ serem maiores(menores) que zero.

Para considerar a dependência entre os valores x_i e y_i pelo IC, temos que ter pelo menos $p.n$ pontos fora do IC para o limite de confiança de $(1-p/2)100\%$, onde n é o números de pontos (λ_i, χ_i).

Se $p.n$ pontos não estiverem fora do IC para o limite de confiança de $(1-p/2)100\%$ e forem distribuídos aleatoriamente, com média dos χ_i acima(abaixo) de zero, então é considerada a dependência positiva(negatica) entre os valores x_i e y_i.

Exemplo 4.3.1 Considere a Tabela 4.1 de uma amostra aleatória de X e Y independentes, e seus valores de χ_i e λ_i. Na figura 4.1 a) e b) são apresentados seu diagrama de dispersão e o Chi-plot.

Tabela 4.1: Dados do Exemplo 4.3.1.

x_i	0.30	-1.16	-0.94	-1.71	-2.43	0.95	-1.60	0.54	1.27	0.29
y_i	-0.11	-0.12	1.52	-1.89	-1.20	0.12	2.75	-1.23	-1.62	-1.51
χ_i	0.00	0.15	-0.39	-	-	-0.18	-	-0.18	-	-0.05
λ_i	0.11	-0.11	-0.60	1.00	1.00	0.60	-1.00	-0.30	-1.00	-0.30

Todos os pontos onde o Chi-plot está definido permanecem dentro do intervalo de confiança, sendo assim concluímos que a os valores de x_i são independentes dos valores de y_i.

Exemplo 4.3.2 Considere a Tabela 4.2 para uma amostra aleatória de X e Y positivamente dependentes e seus valores de χ_i e λ_i. Na Figura 4.1 c) e d) são apresentados seu diagrama de dispersão e o Chi-plot. Seis pontos entre sete estão acima do limite superior de confiança, sendo assim o chi-plot indica que os valores de x_i e y_i são positivamente dependentes e por estarem próximos do valor 1, concluímos que a dependência é forte.

Tabela 4.2: Dados do Exemplo 4.3.2.

x_i	-1.25	-0.83	1.05	0.63	0.95	1.99	0.49	0.63	0.62	2.01
y_i	-1.31	-1.01	0.81	0.54	0.95	2.09	0.59	0.64	0.77	2.06
χ_i	-	1.00	0.75	0.47	0.75	-	0.75	0.80	0.63	-
λ_i	1.00	0.60	0.31	-0.31	0.31	1.00	0.31	-0.01	-0.11	1.00

Exemplo 4.3.3 Considere a Tabela 4.3 para uma amostra aleatória de X e Y negativamente dependentes e seus valores de χ_i e λ_i. Na Figura 4.1 e) e f) são apresentados seu diagrama de dispersão e o Chi-plot. Temos todos os oito valores abaixo do limite inferior de confiança, sendo assim o chi-plot sugere que a os valores de x_i e y_i são negativamente dependentes e por estarem muito próximos do valor -1 e por temos uma grande concentração de $\lambda_i < 0$ concluímos que a dependência entre os valores de x_i e y_i é muito forte.

Para ilustrar melhor a interpretação do Chi-plot e mostrar algumas de suas características com relação ao diagrama de dispersão, a seguir apresentaremos mais exemplos com o gráfico de dispersão com os pontos retirados onde o Chi-plot não está definido. Os gráficos de dispersão contém uma linha vertical passando por $\tilde{x}$ e uma horizontal passando por $\tilde{y}$, e

Tabela 4.3: Dados do Exemplo 4.3.3.

x_i	-0.36	-0.31	1.36	-0.52	-0.40	0.47	0.93	-1.26	-0.15	-0.42
y_i	0.25	0.36	-1.18	0.45	0.38	-0.53	-1.05	1.17	0.28	0.53
χ_i	-0.63	-0.80	-	-0.66	-1.00	-1.00	-1.00	-	-0.79	-0.66
λ_i	0.11	0.01	-1.00	-0.60	-0.11	-0.31	-0.60	-1.00	-0.11	-0.60

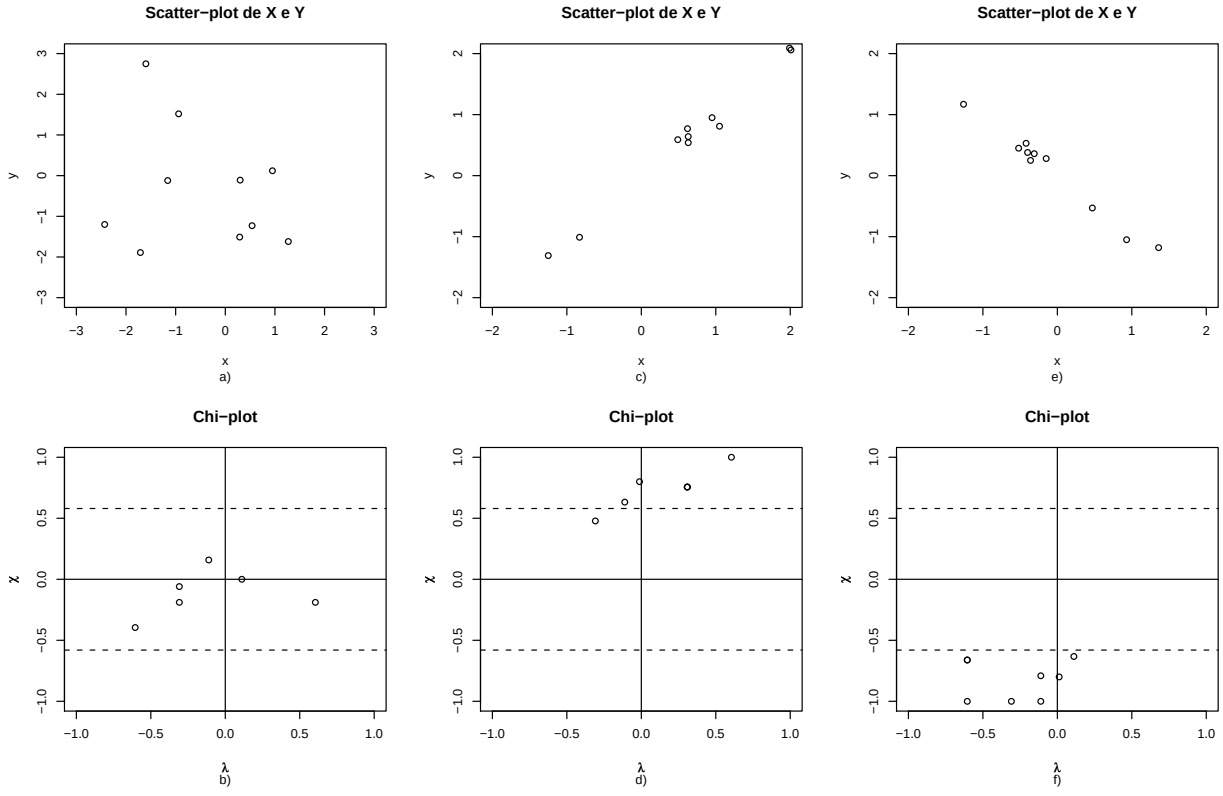

Figura 4.1: Scatter-plots e Chi-plots dos exemplos 4.3.1, 4.3.2 e 4.3.3.

para uma melhor visualização dos comportamentos do Chi-plot, os pontos são representados pelas figuras: círculo, quadrado, losango e triângulo para os pontos cujos $\lambda_{>}0$ e $\chi_i > 0$, $\lambda_{<}0$ e $\chi_i > 0$, $\lambda_{<}0$ e $\chi_i < 0$ e $\lambda_{>}0$ e $\chi_i < 0$, respectivamente.

Exemplo 4.3.4 Para 100 pares gerados de $X \sim N(0,1)$ e $Y = X + e$, com $e \sim N(0, 0.1)$, ou seja, tais que há uma correlação linear positiva muito forte entre esses pontos. Na Figura 4.2 são apresentados o diagrama de dispersão e o Chi-plot onde vemos que é evidente a existência de uma correlação linear muito forte. Temos também os valores $\hat{r} = 0.9948$, $\hat{\tau} = 0.929$ e $\hat{\rho} = 0.992$.

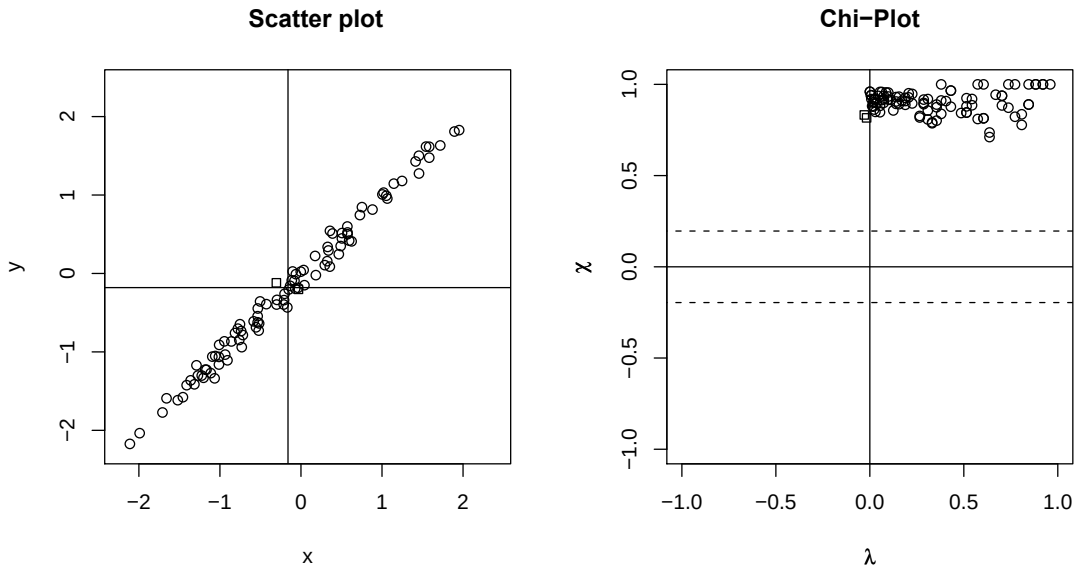

Figura 4.2: Diagrama de Dispersão e o Chi-plot do Exemplo 4.3.4.

Analisando o Chi-plot, podemos perceber que praticamente todos os pares (λ_i, χ_i) se encontram no quadrante onde $\chi_i > 0$, $\lambda_i > 0$ e, ainda, os valores de χ_i são próximos de 1, refletindo a correlação muito forte entre os valores de x_i e y_i.

Exemplo 4.3.5 Considere 42 pares de valores (x_i, y_i), onde x_i é a medida do peso em quilogramas e y_i é a medida da circunferência do pescoço em centímetros, para indivíduos adultos jovens, homens e mulheres. Na Figura 4.3 são apresentados o diagrama de dispersão e o chi-plot correspondentes. Analisando o chi-plot da Figura 4.3, podemos notar que a maioria dos valores de χ_i se encontram acima do limite superior de confiança, indicando a existência de correlação forte. Além disso podemos verificar, também, que a dependência é

positiva, já que para quase todos os pares (x_i, y_i) com valores $\lambda_i > 0$. De outro lado, note que os valores $\hat{r} = 0.837$, $\hat{\tau} = 0.538$ e $\hat{\rho} = 0.732$ diferem relativamente quanto à magnitude da correlação.

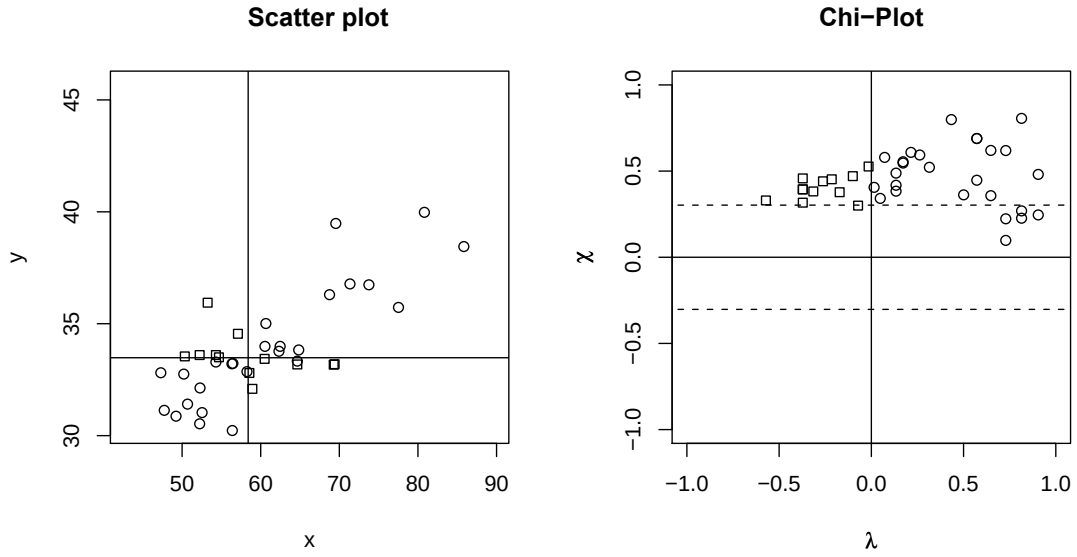

Figura 4.3: Diagrama de Dispersão e o Chi-plot do exemplo 4.3.5.

Exemplo 4.3.6 Considere 100 observações de X e Y geradas com correlação linear de 0.3 e tais que $X \sim N(0,1)$ e $Y \sim N(4 + 0.6 * X; 4 * (1 - 0.3^2))$, onde $4 * (1 - 0.3^2)$ denota que a variância de Y é quatro vezes um menos a correlação linear de 0.3 ao quadrado e $4 + 0.6 * X$ é a média da variável Y mais a correlação linear multiplicando a razão do desvio padrão de Y pelo desvio padrão de X.

O gráfico de dispersão da Figura 4.4 sugere uma correlação positiva, entretanto, os valores $\hat{r} = 0.178$, $\hat{\tau} = 0.116$ e $\hat{\rho} = 0.17$ não são significativos. Por outro lado, analisando o Chi-plot da Figura 4.4 constatamos que a maioria dos valores de χ_i são maiores que zero, indicando uma possível correlação positiva, embora praticamente metade dos pontos (x_i, y_i) tem seus $\lambda_i > 0$. Desta forma pode-se concluir que a correlação predominante é positiva e de baixa magnitude.

Exemplo 4.3.7 Para (x_i, y_i) com $X_i \sim N(0,1)$ e $Y_i \sim N(0,1)$ se $i = 1, \ldots, 50$, e (x_i, y_i) com $X_i \sim N(4,2)$ e $Y_i \sim N(4,2)$ se $i = 51, \ldots, 100$, o gráfico de dispersão e o chi-plot são apresentados na Figura 4.5. O gráfico de dispersão sugere correlação positiva, assim

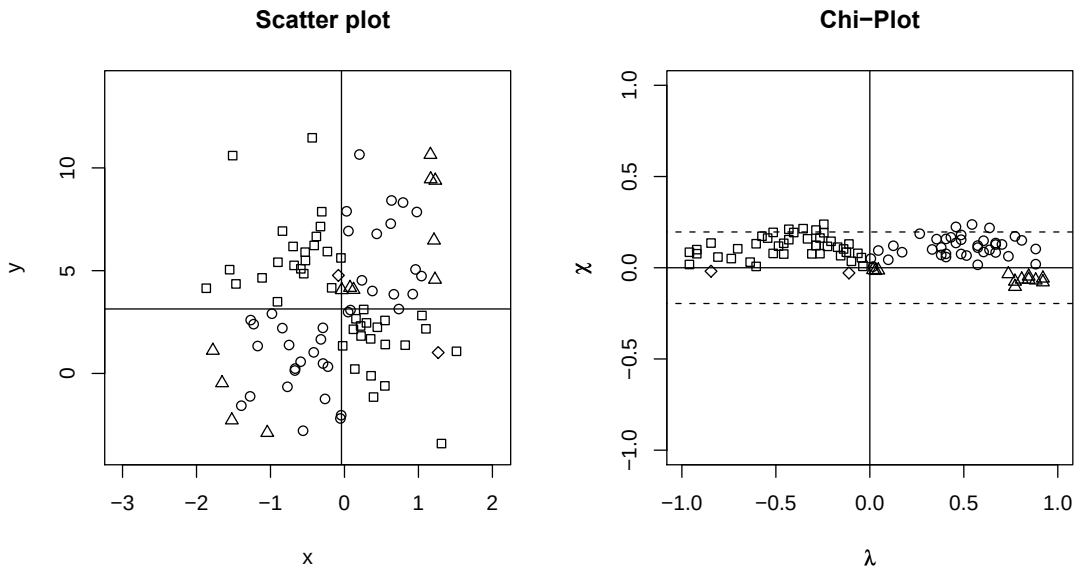

Figura 4.4: Diagrama de Dispersão e o Chi-plot do exemplo 4.3.6.

como os coeficientes de correlação $\hat{r} = 0.613$, $\hat{\tau} = 0.411$ e $\hat{\rho} = 0.633$. O chi-plot mostra que quase todos os pontos χ_i são positivos, sugerindo dependência global positiva. Entretanto, o comportamento característico da situação imposta é notado no chi-plot: o gráfico apresenta um comportamento como o visualizado, pois ao se distanciar das medianas de $\tilde{x}$ e $\tilde{y}$, o ponto no qual será calculado o valor de χ_i, localmente apresenta independência.

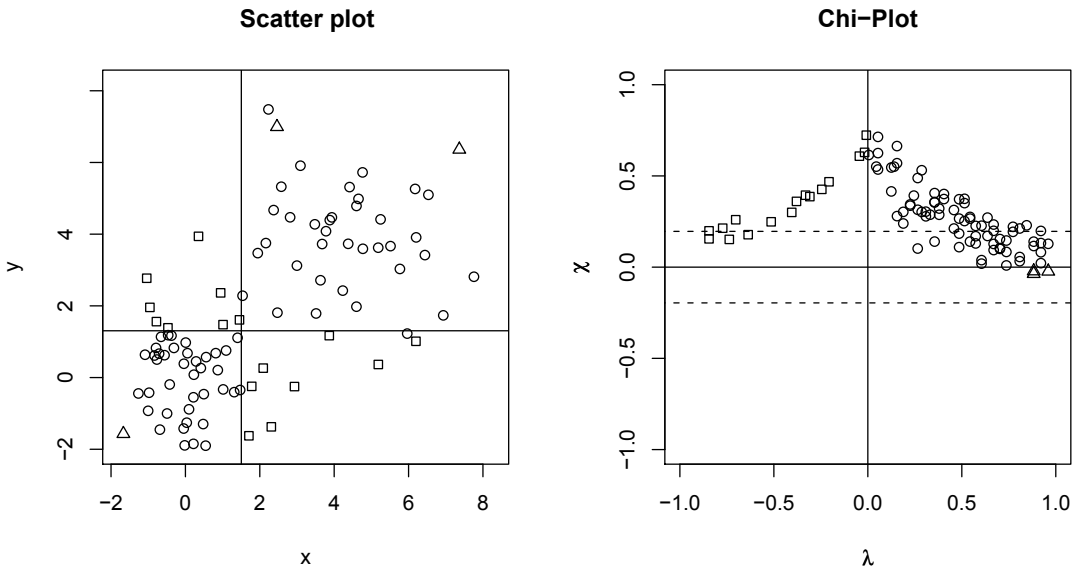

Figura 4.5: Diagrama de Dispersão e o Chi-plot do Exemplo 4.3.7.

5 Ilustração dos Procedimentos

Neste capítulo ilustraremos os procedimentos de análise de dependência abordados nos capítulos anteriores e apresentamos comentários à respeito dos cuidados para a utilização em cada caso. São considerados os resultados obtidos de quatro tipos de dependência nas amostras, denominadas "linear", "quadrática", "normais" e exponenciais", e uma de independência em dados normais através da cópula estimada, da função de Sibuya estimada e do Chi-plot.

Uma tabela contendo as estimativas dos coeficientes de correlação de Kendall, de Spearman, e de Pearson, e o gráfico de dispersão de X contra Y para cada amostra são apresentadas primeiramente. Posteriormente são apresentadas as curvas de nível da cópula estimada de (X, Y), as curvas de nível da função de Sibuya estimada e o resultado do teste de Gonçalves (2008) e o Chi-plot dos pares (x_i, y_i) com IC de 95% de confiança.

Os códigos dos procedimentos de estimação da cópula, da função de Sibuya e do Chi-plot estão expostos no Apêndice B. Nas rotinas implementadas da cópula e da Função de Sibuya os dados são padronizados para prevenir uma escolha inadequada da função kernel que não apresente uma suavização adequada devido a magnitude dos dados. Esta padronização não afeta o procedimento de estimação, pois como visto em Nelsen (2006) e Marcelo (2008), transformações crescestes em ambas as componentes marginais preservam as características da cópula e da função de Sibuya.

Para a função kernel, dada pela Equação (2.3.2), foi escolhida a função de distribuição acumulada normal bivariada com vetor de médias iguais a zero e matriz de covariância Σ dada por

$$\Sigma = \begin{bmatrix} 0.3^2 & 0 \\ 0 & 0.3^2 \end{bmatrix}. \tag{5.0.1}$$

Os pontos para estimação da cópula foram escolhidos de acordo com valores da distribuição acumulada empírica, e para a função de Sibuya foram obtidos dividindo o intervalo de variação dos dados por uma quantidade pré determinada, atribuindo um valor acima e abaixo dos pontos de mínimo e máximo respectivamente, de acordo com a média e o desvio padrão dos dados. A função h_n escolhida foi $1/\sqrt{n}$.

Para a estimação da cópula esses valores podem ser selecionados mais de uma vez pelo critério escolhido, portanto foi introduzida uma rotina para identificar estes casos e

subtrair estes pontos. Para uma melhor análise das curvas de nível da função de Sibuya apresentamos curvas de nível para $\theta < 1$, $\theta = 1$ e $\theta > 1$.

5.1 Caso de Independência

Considere uma amostra com $n = 200$ pares de (x, y) com valores de X e Y gerados independentemente, onde os valores de X são gerados de uma distribuição normal com média 50.0, e desvio padrão 7.0 e os valores de Y são gerados de uma distribuição normal com média 30.0 e desvio padrão 2.0.

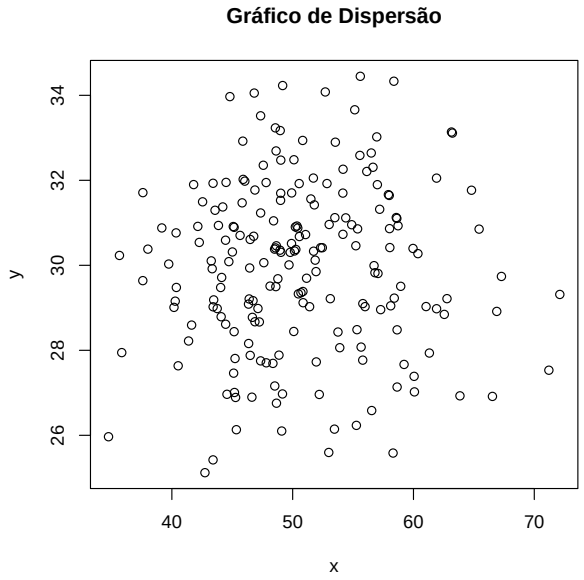

Figura 5.1: Gráfico de dispersão de x contra y.

Através de uma análise do gráfico de dispersão não é possível perceber nenhuma característica que contradiga a suposição inicial de que as amostras são independentes. Na Tabela 5.1 são apresentados os valores das estimativas dos coeficientes de correlação para o conjunto de dados da Figura 5.1 com os valores entre parênteses indicando a correlação assumida considerando um p-value de 5%.

Tabela 5.1: Coeficientes de correlação.

Kendall	Spearman	Pearson
0.0325(0)	0.0499(0)	0.0315(0)

Os resultados da Tabela 5.1 mostram que as hipóteses de que os coeficientes são

nulos não são rejeitadas.

Na Figura 5.2 a cópula estimada se assemelha com a cópula de independência, mas mesmo a cópula não sendo uma estimativa de correlação local, é possível notar que as regiões separadas, destacadas na Figura 5.3 como N e P, sugerem predominância de dependência positiva na região P, e negativa na região N, sendo de magnitudes baixas. Uma relação direta com os dados não pode ser feita de imediato, encontrando assim certa dificuldade para se separar adequadamente os pontos ou regiões de dependência.

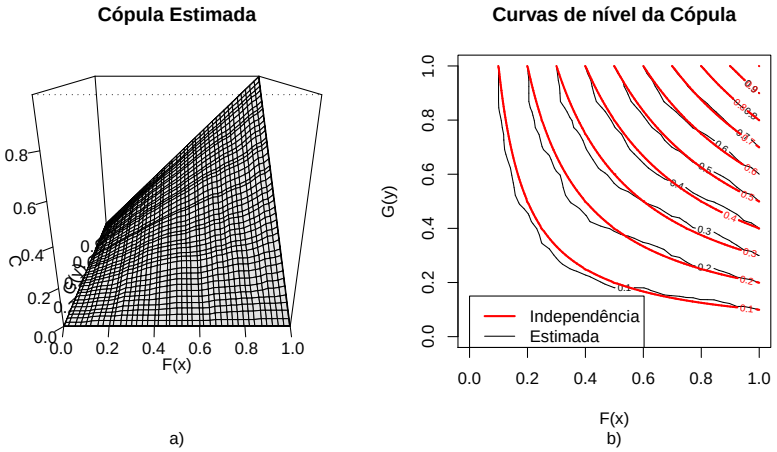

Figura 5.2: Cópula estimada em a) e suas curvas de nível em b).

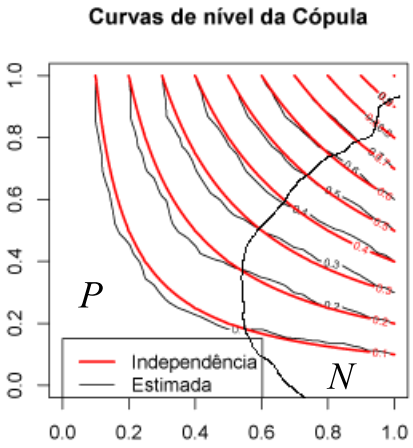

Figura 5.3: Regiões com evidências de correlação.

Para o teste desenvolvido por Gonçalves (2008), obtemos $\Omega_n = 1.3700$, com $\hat{z}_n = 1.8188$, $IC_{90\%} = [1.0363, 1.7036]$ e $IC_{95\%} = [0.9712, 1.7687]$, cujo resultado indica a clara independência entre os valores de $x's$ e $y's$ na amostra para o caso do IC de 95% e de dependência para o IC de 90%.

Os resultados divergentes pelos intervalos de confiança sugerem considerar a análise da função de Sibuya estimada para buscar outros fatos na relação entre os $x's$ e os $y's$. Na Figura 5.4 são apresentadas as curvas de nível da função de Sibuya estimada.

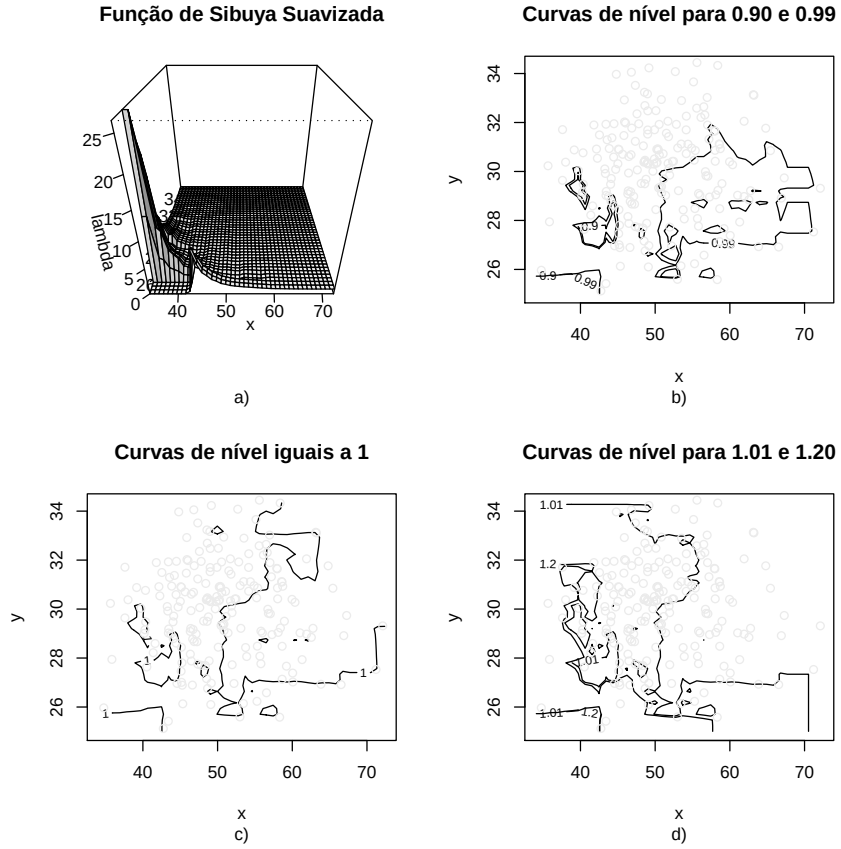

Figura 5.4: Função de Sibuya estimada em a), curvas de nível $\hat{\Omega} = 0.90$ e $\hat{\Omega} = 0.99$ em b), igual a 1 em c) e iguais a $\hat{\Omega} = 1.01$ e $\hat{\Omega} = 1.20$ em d).

Das Figuras 5.4 b) e c) pode-se distinguir que a região inferior direita concentra pontos de dependência negativa, e das Figuras 5.4 c) e d) distingue-se a região de pontos de dependência positiva à esquerda da curva de nível $\theta = 1$.

Da Figura 5.5 também podemos enxergar do Chi-plot e o gráfico de dispersão que os pares de (x, y) que correspondem aos losangos e triângulos têm tendência à dependência local negativa, enquanto que os pares de círculos e quadrados têm tendência de dependência local positiva.

Também do Chi-plot na Figura 5.5 é possível calcular a quantidade de pontos que se encontram em cada quadrante, determinados pelas medianas de x e de y. Cerca de 63% dos pontos têm valores de χ positivos e 52% têm valores de λ positivos. Aproximadamente 35% dos pontos se encontram no quadrante superior direito e 29% no quadrante superior

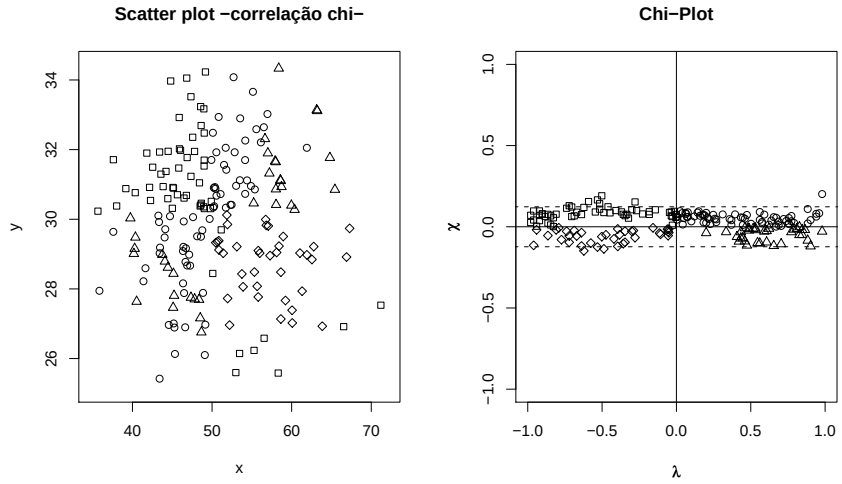

Figura 5.5: Gráfico de dispersão com pontos classificados segundo os valores χ e o Chi-plot.

Tabela 5.2: Porcentagem de acertos do teste de Gonçalves (2008).

$\alpha =$	10%	5%	1%		10%	5%	1%
cem amostras	32%	53%	64%	mil amostras	32.3%	43.7%	62.8%

esquerdo, enquanto que 19% dos pontos se encontram no quadrante inferior esquerdo e 17% no quadrante inferior direito. Por outro lado, 2.5%, 4.1% e 6.1% de pontos fora do intervalo de confiança de 99%, 95% e 90%, respectivamente.

Notamos também que há predominância de valores $\chi > 0$ de magnitude baixa, tal que os pontos se encontram dentro do seu intervalo de confiança, e sugerem uniformidade quanto aos valores de λ. Podemos considerar este caso como um caso de independência.

O teste de Sibuya classificou a amostra como independente considerando um intervalo de confiança de 95% e dependente considerando um intervalo de comfiança de 90%. Com o objetivo de verificar se o teste de Golçalves (2008) apresenta consistência na classificação de amostras independentes, pode se calcular a porcentagem de amostras classificada corretamente. Na Tabela 5.2 são apresentados os resultados da correta classificaçao no procedimento para o qual foram geradas amostras independentes de tamanho $n = 100$, e a classificação é baseada na aproximação normal do teste de Gonçalves (2008).

5.2 Casos de Dependência

Na seção anterior vimos como as metodologias se comportam em um caso particular de independência considerando uma amostra bivariada com as marginais normais

independentes. Nosso interesse não é somente determinar o comportamento nos casos de independência, mas também compreender como cada tipo de metodologia se comporta correspondente aos vários tipos de dependência

Cada característica se comporta de tal maneira que vamos adquirir conhecimento dos comportamentos individuais de cada metodologia e quais dos comportamentos são significativos. Desta maneira, a análise sobre quatro conjuntos de dados denominados "linear", "quadrática", "normais" e exponenciais" são apresentados e estudados a seguir.

5.2.1 Dependência Linear Forte

Consideremos uma amostra bivariada com tamanho $n = 200$ gerada de modo a obter uma correlação linear forte, tal que a variável aleatória X tem distribuição normal com média 50.0, desvio padrão 12.0, e a variável aleatória $Y = X + E$ é tal que $y_i = x_i + e_i$, com $e_i \sim N(4, 7^2)$. Na Figura 5.6 é apresentado o gráfico de dispersão de (x, y).

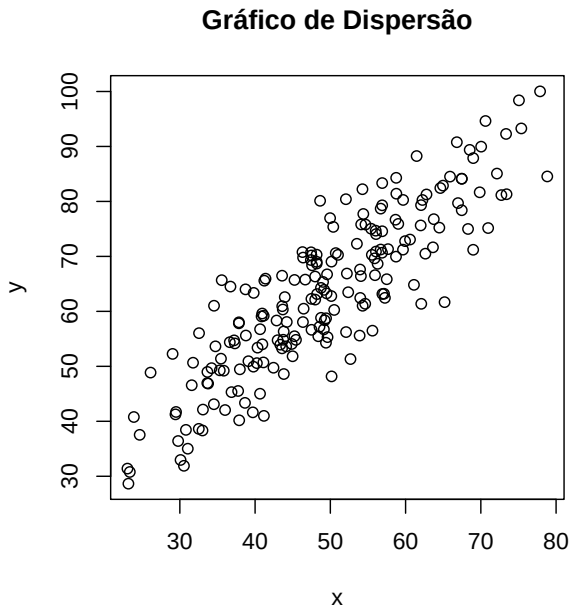

Figura 5.6: Gráfico de Dispersão de x contra y.

A Tabela 5.3 apresenta as estimativas dos coeficientes de correlação usuais, os quais indicam correlação alta.

Na Figura 5.7 é apresentada a cópula estimada da qual podemos notar que, como esperado, as curvas de nível se aproximam muito da curva de nível para a correlação positiva perfeita.

Tabela 5.3: Coeficientes de correlação.

Kendall	Spearman	Pearson
0.6789	0.8620	0.8734

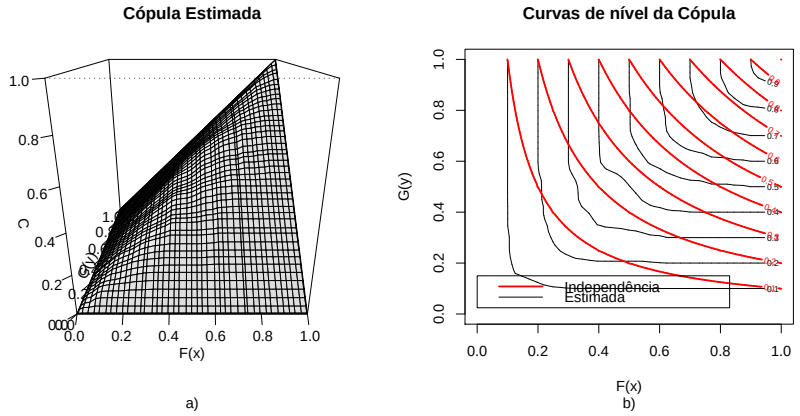

Figura 5.7: Cópula estimada em a) e suas curvas de nível em b).

Da Figura 5.8 a) e d) pode ser vista dos gráficos que a função de Sibuya estimada possui seus valores acima de 1, e das Figuras 5.8 b) e c) as ausências das curvas de nível para $\hat{\Omega} \leq 1$. Estas são claras evidências de haver forte dependência positiva.

Para o teste de Gonçalves (2008), $\Omega_n = 3.7868$, com $\hat{z}_n = 4.0195$, $IC_{95\%} = [2.4279, 5.1458]$ e $P(\hat{z}_n > 4.0915) = 2\ 10^{(-5)}$ indicam um caso de dependência positiva entre os valores de x e y.

Infelizmente, a função de Sibuya estimada apresenta valores atípicos nos extremos de x e y, pela sua própria definição e ainda apresenta incoerência com a magnitude da dependência no espaço amostral.

Da Figura 5.9, pode ser observado do chi-plot correspondente, que todos os χ_i são positivos e têm valores significativos. A maioria dos valores λ_i positivos indicam tendência de correlação positiva.

5.2.2 Dependência de Forma Quadrática

Para ilustrar os três procedimentos de análise de dependência estudados, sobre dados que não apresentam monotonocidade na associação entre as variáveis, foram gerados pares de valores (x_i, y_i) com $X_i \sim U(8, 12)$ e $Y_i = -(X_i - 10)^2 + N(0.8, 1) + 20$; com $i = 1, \ldots, 200$, onde $U(a, b)$ é a distribuição uniforme no intervalo (a, b). Ou seja, da Figura 5.10 pode ser

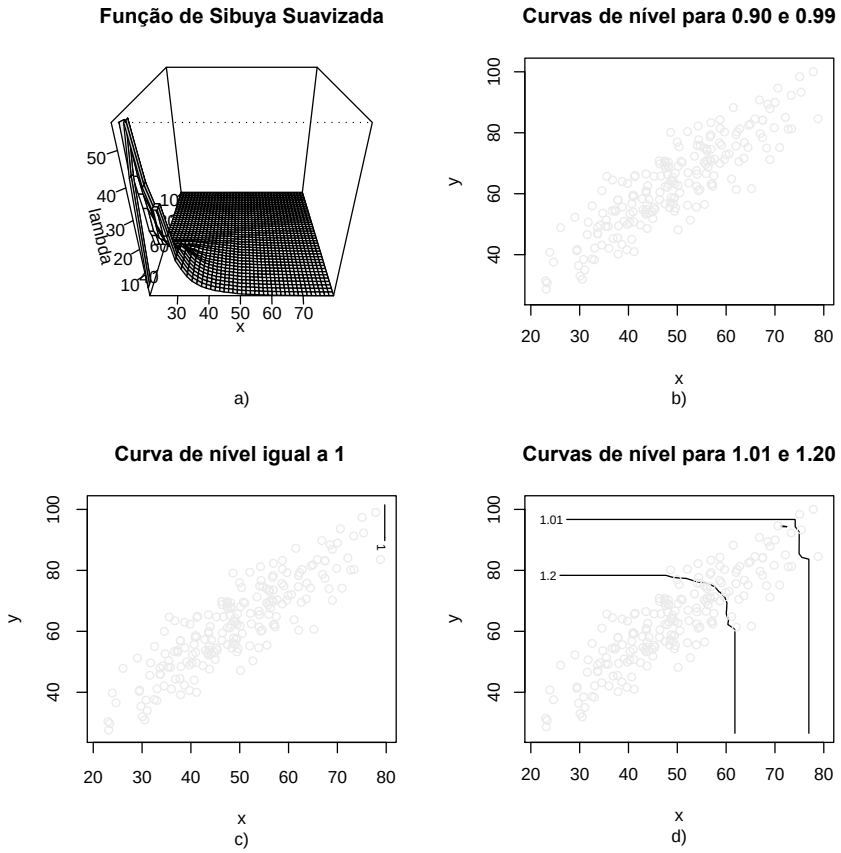

Figura 5.8: Função de Sibuya estimada em a), curvas de nível $\hat{\Omega} = 0.90$ e $\hat{\Omega} = 0.99$ em b), igual a 1 em c) e iguais a $\hat{\Omega} = 1.01$ e $\hat{\Omega} = 1.20$ em d).

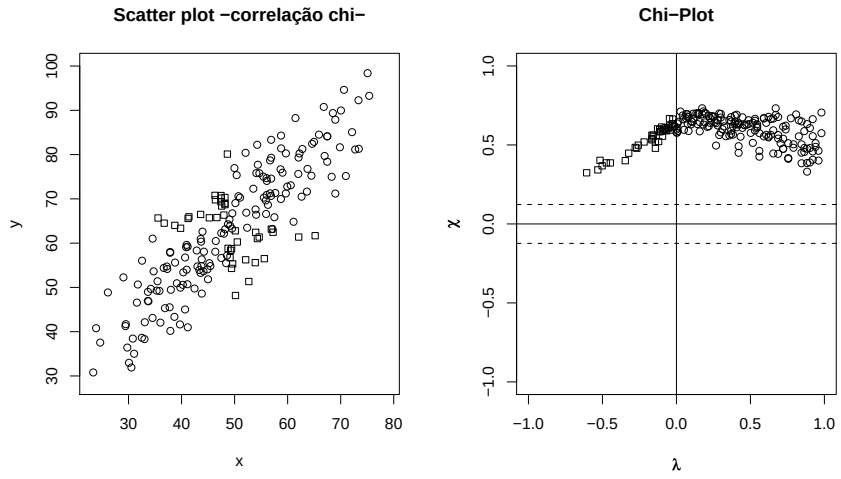

Figura 5.9: Gráfico de dispersão com pontos classificados segundo os valores χ e o Chi-plot.

visto que as variáveis tem correlação positiva e negativa e o ponto especificado de mudança da correlação é $x = 10$, onde podemos observar uma dependência em forma quadrática.

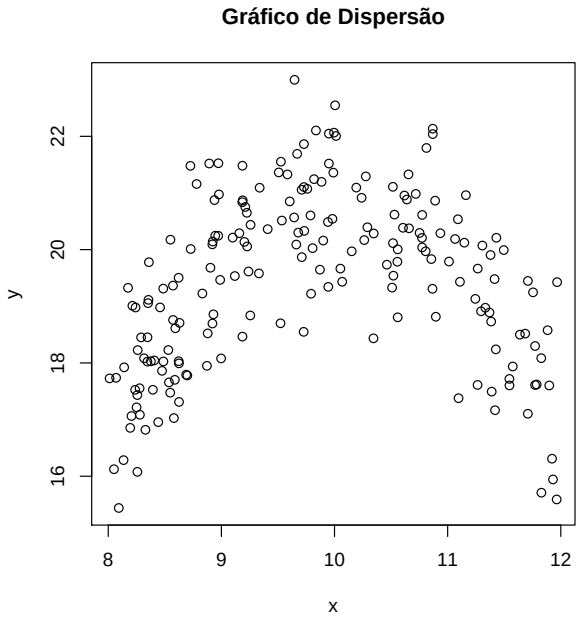

Figura 5.10: Gráfico de Dispersão de x contra y.

A Tabela 5.4 contém as estimativas pontuais dos coeficientes de correlação, dentre as quais resulta não significativo para o coeficiente de correlação de Pearson considerando os p-values de > 0.05.

Tabela 5.4: Coeficientes de correlação.

Kendall	Spearman	Pearson
0,1301	0,1959	0,1269

Na Figura 5.4 percebemos das curvas de nível da cópula, que para valores de $F(x) < 0,4$ aproximadamente, as curvas de nível indicam dependência positiva e que para valores $F(x) > 0,6$ as curvas de nível indicam dependência negativa. Para $0,4 < F(x) < 0,6$ as curvas de nível mudam de comportamento, começando com dependência positiva passando para dependência negativa. Para esta situação, as curvas de nível são bastante informativas sobre a dependência local e global.

Da Figura 5.12 b), c) e d) as curvas de nível da função de Sibuya nos indicam que a dependência é positiva para valores de $x < 10$ e negativa para $x > 10$, e que o grau de dependência vai aumentando na medida que se afasta de $x = 10$. Entretanto, das curvas de

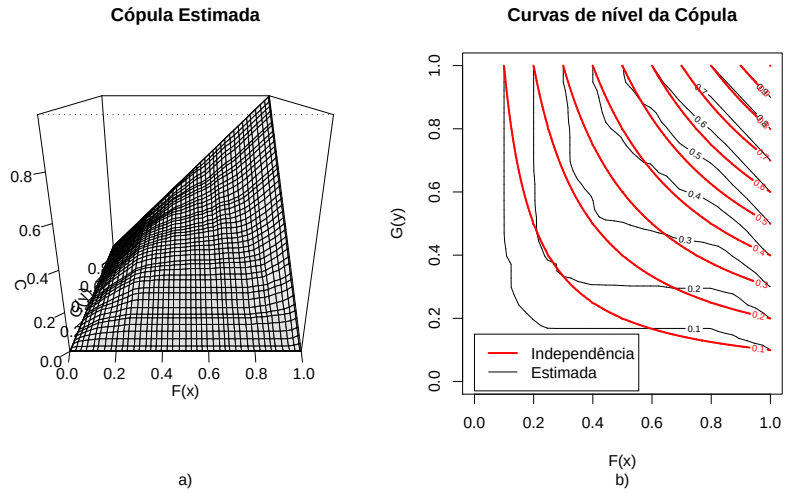

Figura 5.11: Cópula estimada em a) e suas curvas de nível em b).

nível da função de Sibuya para $\hat{\Omega} = 0,9$ e $\hat{\Omega} = 1,2$, por exemplo, não é possível saber qual delas representa um grau de correlação maior. Para o teste de Gonçalves (2008) $\Omega_n = 2.0399$, $\hat{z}_n = 3.7100$, $IC_{95\%} = [1.4905, 2.5893]$, $IC_{99\%} = [1.3195, 2.7603]$ e $P(\hat{z}_n > 3.7100) = 0.0001$ indicam um caso de dependência positiva entre os valores de x e y.

Do Chi-plot na Figura 5.13 observa-se que a grande maioria dos valores de χ_i além de significativos (ao nível de 5%), indicam forte correlação. A classificação para os valores χ_i, segundo sua posição num dos quadrantes, transportada para o gráfico de dispersão dos pares (x, y) na mesma Figura 5.13, permite vizualizar que, exceto os pontos bem próximos do ponto de mudança de sinal da correlação ($x = 10$), os pontos (x, y) com $x < 10$ têm $\chi_i > 0$ e os pontos (x, y) com $x > 10$ têm $\chi_i < 0$. Ou seja, o Chi-plot representa de forma adequada a situação de dependência sugerida pelo gráfico de dispersão.

5.2.3 Dois Fatores Com Dois Níveis

Considere o conjunto de dados gerados de tal forma que x_i e y_i são independentes localmente, e com $X_i \sim N(5, 0.6)$ e $Y_i \sim N(5, 0.6)$; $i = 1, \ldots, 50$, $X_i \sim N(10, 0.6)$ e $Y_i \sim N(5, 0.6)$; $i = 51, \ldots, 100$, $X_i \sim N(5, 0.6)$ e $Y_i \sim N(10, 0.6)$; $i = 101, \ldots, 150$ e $X_i \sim N(10, 0.6)$ e $Y_i \sim N(10, 0.6)$; $i = 151, \ldots, 200$, chamaremos o conjunto de pares (x_i, y_i) com essas características de "Normais".

Na Figura 5.14, observamos quatro estruturas bem definidas. A Tabela 5.5 traz as estimativas dos coeficientes de correlação usuais e entre parênteses a correlação assumida utilizando um p-value de 5%. Os resultados da Tabela 5.5 mostram que as hipóteses de que

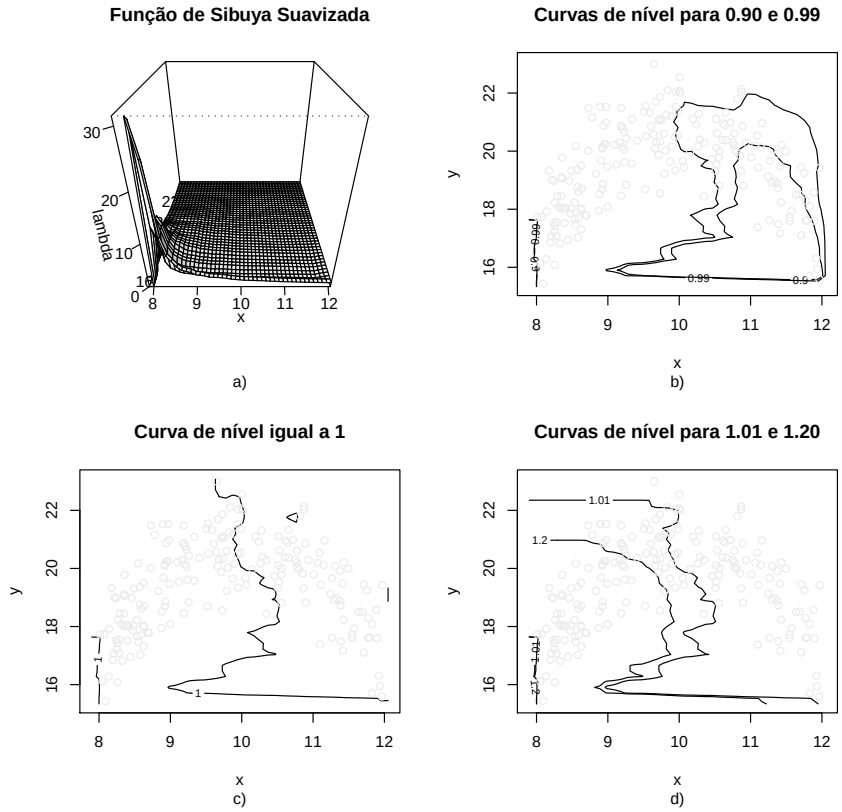

Figura 5.12: Função de Sibuya estimada em a), curvas de nível $\hat{\Omega} = 0.90$ e $\hat{\Omega} = 0.99$ em b), igual a 1 em c) e iguais a $\hat{\Omega} = 1.01$ e $\hat{\Omega} = 1.20$ em d).

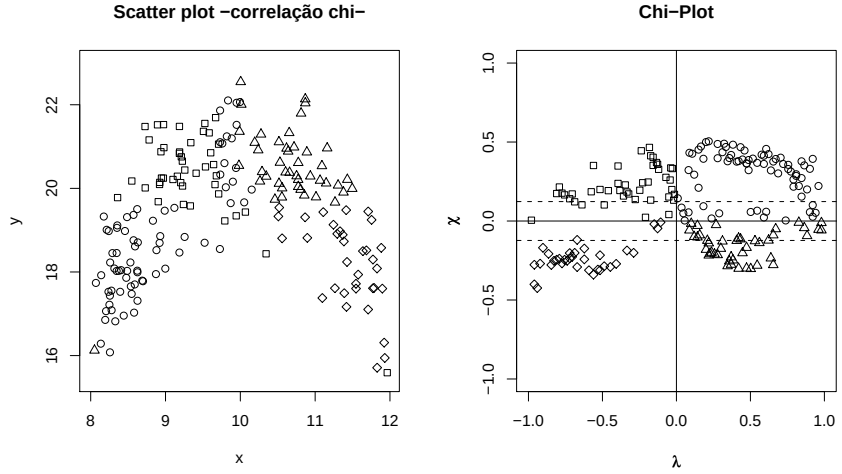

Figura 5.13: Gráfico de dispersão com pontos classificados segundo os valores χ e o Chi-plot.

os coeficientes são nulos não são rejeitadas.

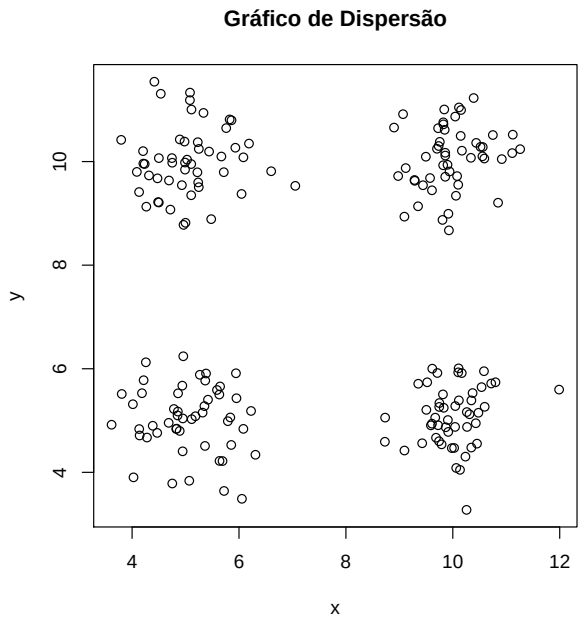

Figura 5.14: Gráfico de Dispersão de x contra y.

Tabela 5.5: Coeficientes de correlação.

Kendall	Spearman	Pearson
0.0309(0)	0.0486(0)	0.0130(0)

Da Figura 5.15 verificamos que a cópula estimada quase coincide com a cópula de independência, se analisadas pelas curvas de nível.

As curvas de nível da função de Sibuya não apresentam uma clara identificação de alguma tendência global ou local. Da estatística do teste de independência de Gonçalves (2008), $\hat{z}_n = 3.2345$, indica dependência global positiva, mas não há uma indicação de que essa dependência corresponda ou seja causada pela presença dos "dois fatores". Para o teste de Gonçalves (2008) $\Omega_n = 1.1140$, $\hat{z}_n = 3.2345$, $IC_{95\%} = [1.0449, 1.1831]$ e $IC_{99\%} = [1.0234, 1.2046]$ indicam um caso de dependência entre os valores de x e y.

O Chi-plot na Figura 5.17 classifica 67.3% dos pontos com $\chi > 0$, e se encontram distribuidos entre zero e o limite superior do intervalo. Temos que 7.1% dos pontos se encontram fora do intervalo de confiança de 95%, sugerindo assim que existe dependência. Os pontos χ_i estão dispostos de maneira que a dependência predominante é positiva e ainda

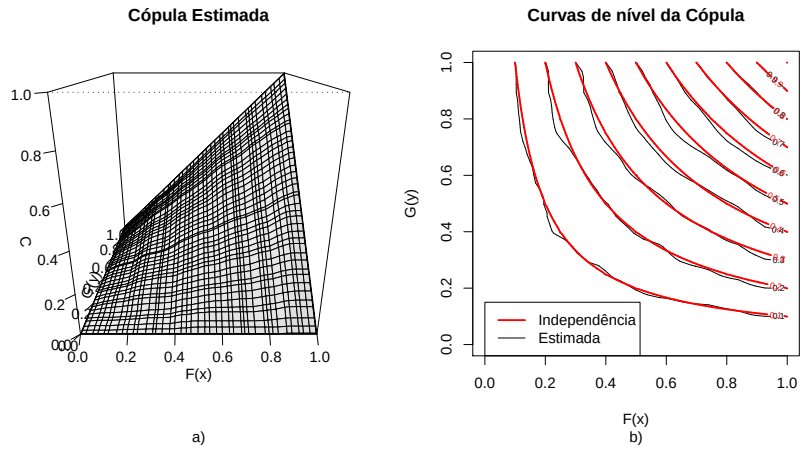

Figura 5.15: Cópula estimada em a) e suas curvas de nível em b).

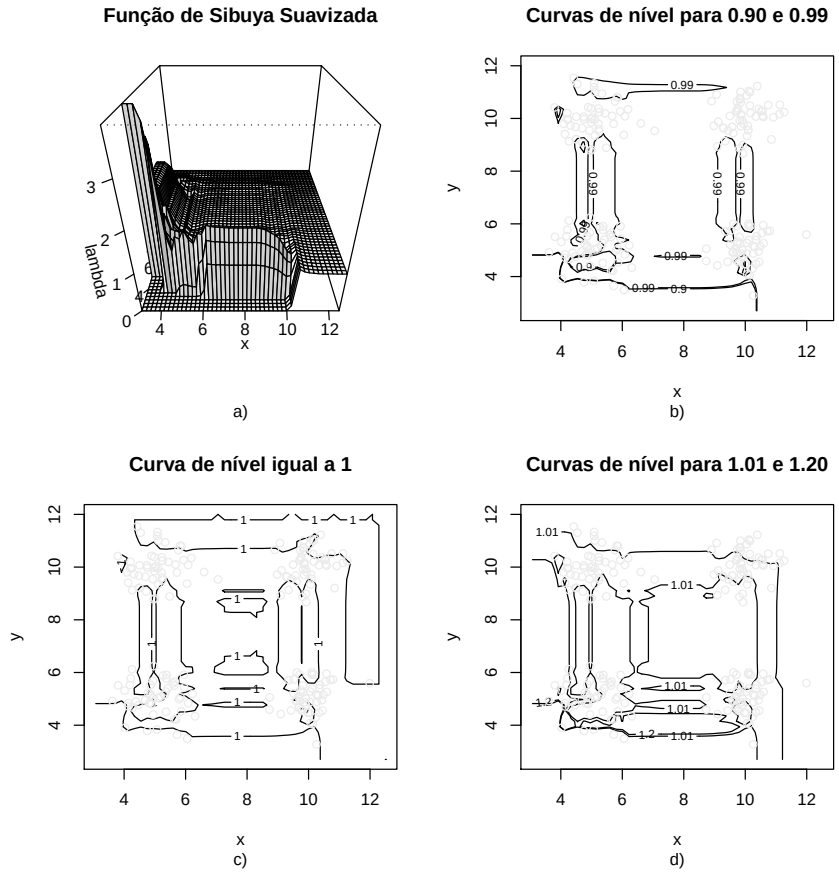

Figura 5.16: Função de Sibuya estimada em a), curvas de nível $\hat{\Omega} = 0.90$ e $\hat{\Omega} = 0.99$ em b), igual a 1 em c) e iguais a $\hat{\Omega} = 1.01$ e $\hat{\Omega} = 1.20$ em d).

dentre os quatorze pontos que se encontram fora do IC de 95% treze destes pontos têm $\chi > 0$. Ainda, a grande maioria dos pares (x_i, y_i) nos 1°, 2°e 3°quadrantes têm $\chi_i > 0$.

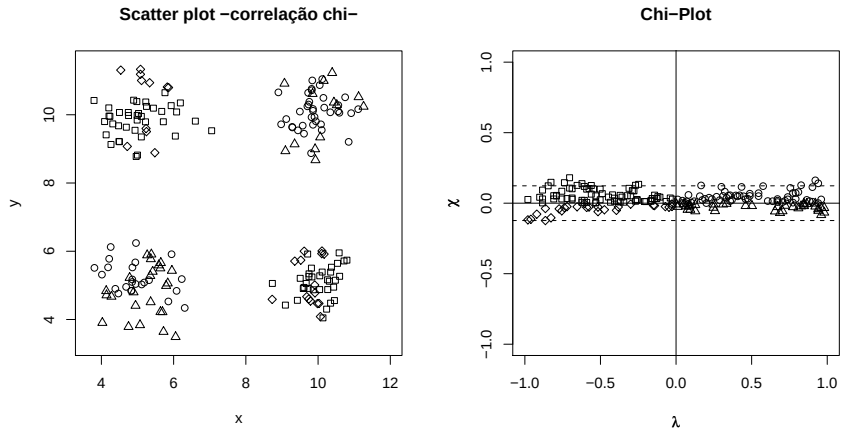

Figura 5.17: Gráfico de dispersão com pontos classificados segundo os valores χ e o Chi-plot.

5.2.4 Dependência Exponencial Bivariada

Considerando o conjunto de 200 pares de pontos gerados a partir de uma distribuição exponencial bivariada de (Gumbel 1960), com correlação linear de -0.125 e valores de taxas iguais a 2 e 3 para as variáveis X e Y, respectivamente. Chamaremos o conjunto de dados com essas características de "Exponenciais".

Da Figura 5.18 pelo gráfico de dispersão não é possível discernir sobre a correlação envolvida, ou mesmo se há correlação. A Tabela 5.6 traz as estimativas dos coeficientes de correlação usuais, os quais resultam significativos.

Tabela 5.6: Coeficientes de correlação.

Kendall	Spearman	Pearson
-0.1283	-0.1890	-0.1619

Na Figura 5.19 pelo gráficos da cópula e suas curvas de nível, observamos que o valor das curvas de nível da cópula se afastam das curvas de nível da cópula de independência na direção das curvas de nível da cópula de correlação perfeita negativa.

Dos gráficos das curvas de nível da função de Sibuya (ver Figura 5.20), particularmente, em b), é claro que praticamente todos os pontos (x, y) estão numa região de dependência negativa e apenas alguns poucos pontos em regiões de dependência positiva.

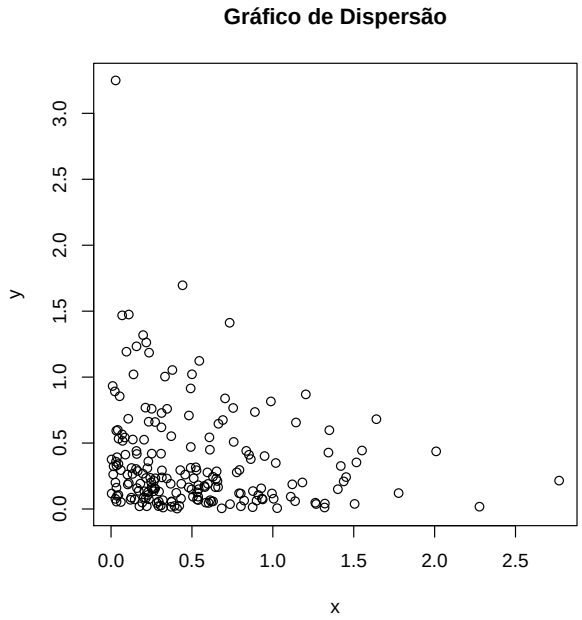

Figura 5.18: Gráfico de Dispersão de x contra y.

O resultado do teste de Gonçalves (2008), são tais que $\Omega_n = 0.8804$, $\hat{z}_n = -9.0849$, $IC_{95\%} = [0.8546, 0.9062]$, $IC_{99\%} = [0.8466, 0.9142]$ e $P(-9.0849) < 410^{-20}$, indicando a existência da correlação negativa.

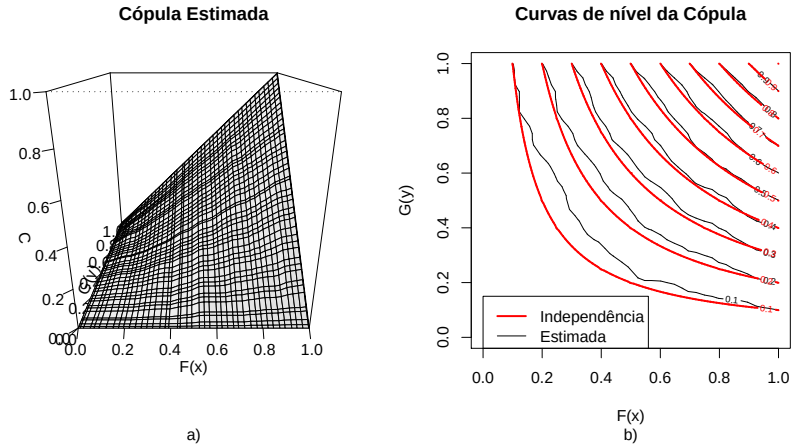

Figura 5.19: Cópula estimada em a) e suas curvas de nível em b).

O Chi-plot na Figura 5.21 mostra claramente uma tendência negativa, ao apresentar praticamente todos os valores de χ menores que zero e bem mais de 5% dos pontos abaixo do limite inferior de confiança de 95%. Em praticamente todo o domínio dos dados a correlação permanece próxima em magnitude.

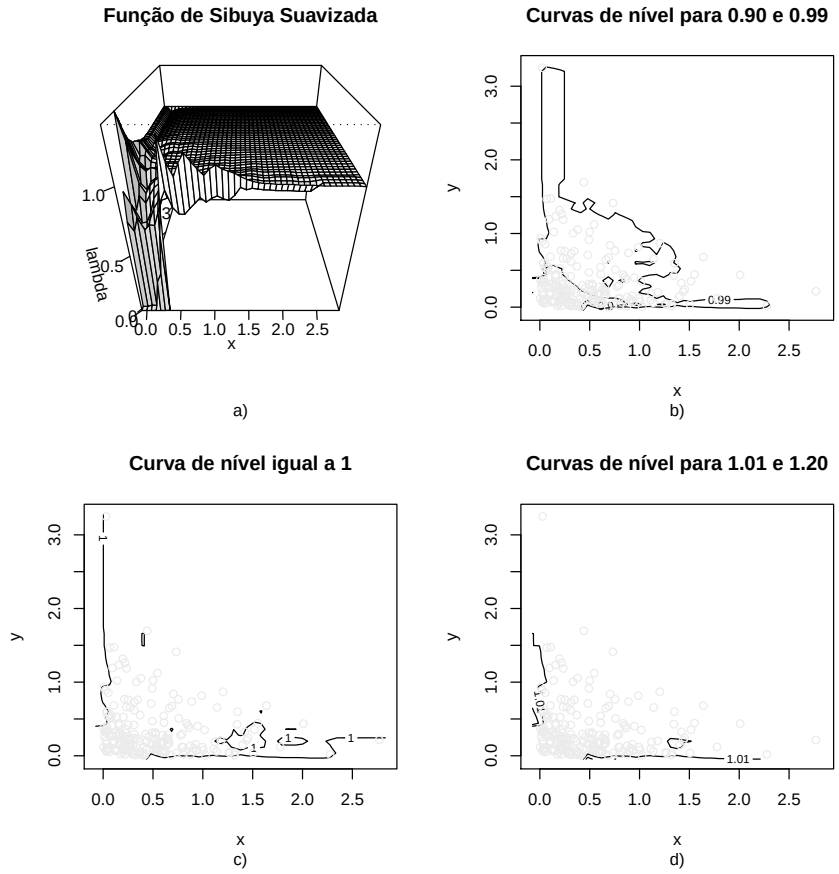

Figura 5.20: Função de Sibuya estimada em a), curvas de nível $\hat{\Omega} = 0.90$ e $\hat{\Omega} = 0.99$ em b), igual a 1 em c) e iguais a $\hat{\Omega} = 1.01$ e $\hat{\Omega} = 1.20$ em d).

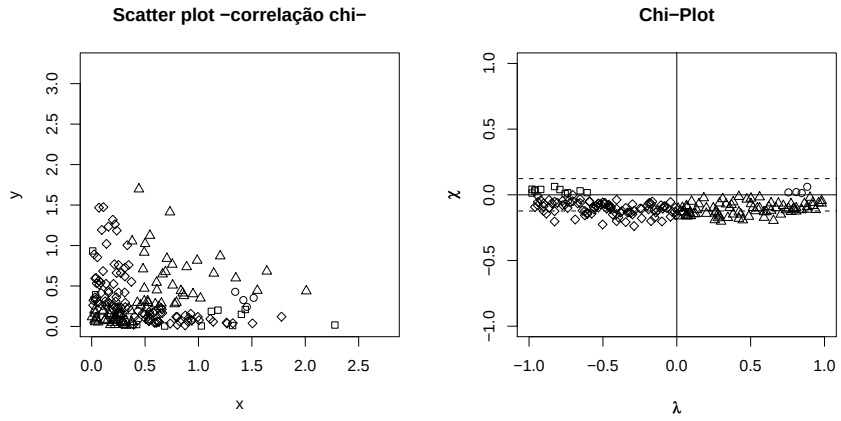

Figura 5.21: Gráfico de dispersão com pontos classificados segundo os valores χ e o Chi-plot.

5.3 Conclusões

Das ilustrações dos procedimentos de análise de dependência apresentados neste capítulo, observamos que:

1 - Os coeficientes de correlação de Pearson, Spearman e Kendall podem não detectar a presença e magnitude da correlação entre x e y em estruturas mais complexas, como nos casos de dependência quadrática e dois fatores apresentados.

2 - As curvas de nível das cópulas apresentam indicações sobre a existência e a magnitude de correlação global e local em situações mediamente complexas, mas podem não dar indicações em situações mais complexas (como no caso de dois fatores).

3 - As curvas de nível da função de Sibuya também apresentam boas indicações sobre a presença de correlação global e local em situações medianas e complexas. Em situações mais complexas, como no caso de dois fatores, a identificação de regiões de dependência local pode se tornar confusa ou inviável.

4 - No Chi-plot as indicações da presença de alguma tendência ou correlação global ou local, são bastante evidentes nos comportamentos dos respectivos χ_i e/ou λ_i. Da mesma forma, as indicações quanto à magnitude da dependência local são bem claras, pelo distanciamento dos valores χ_i com relação aos limites superior e inferior de confiança, as indicações da dependência global são claras pela presença de alguma tendência dos valores χ_i e/ou λ_i.

Em situações de dependência mais complexas, como no caso da dependência quadrática ou dois fatores, as tendências (na estrutura global) podem só se tornarem claras quando analisamos o chi-plot e o gráfico de dispersão com seus pontos identificados segundo a posição de (χ_i, λ_i) em um dos quadrantes do Chi-plot.

6 Estudo do Chi-Plot

No Capítulo 5 ilustramos a análise gráfica de dependência bivariada, por meio dos procedimentos de Cópula, Função de Sibuya e do Chi-plot, em cinco situações práticas distintas. Verificamos maior eficácia do Chi-plot com relação aos outros dois procedimentos para a detecção de dependência local entre os valores das variáveis X e Y.

De acordo com Fisher & Switzer (2001), cinco características são próprias do chi-plot sob a condição de independência das variáveis X e Y. Tais características **C1** a **C5** podem ser descritas da seguinte maneira:

C1. Os limites de 90%, 95% e 99% confiança para os χ_i são $\pm c_p/\sqrt{n}$, com c_p igual a 1.54, 1.78 e 2.18, respectivamente;

C2. Cada um dos quatro quadrantes do chi-plot, determinados por $\chi = 0$ e $\lambda = 0$ contém 25% dos pares da amostra;

C3. 50% dos valores χ_i são maiores do que zero;

C4. 50% dos valores λ_i são maiores que zero;

C5. λ tem distribuição uniforme no intervalo $\pm 4\left(\frac{1}{n-1} - 0.5\right)^2$. As interpretações destas características foram apresentados nas seções 4.2 e 4.3.

Como parte desta dissertação, foram realizados estudos de simulação para avaliar a validade destas características na situação de amostras com variáveis X e Y independentes. Os resultados dos estudos são apresentados neste capítulo.

Nas seções 6.1-6.3 consideramos somente os intervalos de confiança do Chi-plot. Na Seção 6.1 verificamos a baixa coerência da classificação de amostras de variáveis independentes através dos intervalos de confiança de Fisher & Switzer (2001). Na Seção 6.2 determinamos a distribuição assintótica de χ e consequentemente obtemos os seus intervalos de confiança assintóticos. Na Seção 6.3 comparamos os intervalos de confiança de Fisher & Switzer (2001) com os intervalos de confiança assintóticos para verificar p^* (proporção de pontos que permanece fora do intervalo), e ainda, obtemos intervalos para χ que deixam em média o valor p^* para os níveis especificados de 90%, 95% e 99%. Na Seção 6.3.2 verificamos a probabilidade de cobertura para p^* considerando os intervalos de confiança de Fisher & Switzer (2001) e os intervalos de confiança assintóticos.

Na Seção 6.4 apresentamos uma breve análise da coerência das características **C2-C5** para classificar uma amostra bivariada com variáveis independentes como uma amostra independente. Apontamos ainda, uma nova estratégia para classificar a amostra bivariada

como independente com base na distribuição assintótica de χ do Chi-plot e verificamos sua coerência e bem como a combinação desta estratégia com os intervalos de confiança de Fisher & Switzer (2001) e os intervalos assintóticos.

No Apêndice A são apresentadas tabelas com os resultados das aplicações descritas acima para os exemplos com situações de dependência "linear forte", "quadrática", "normais" e "exponenciais", que são utilizadas para fazer as conclusões dos comportamentos das características **C2-C5** e da combinação da distribuição assintótica de χ, com os intervalos de confiança de Fisher & Switzer (2001) e os intervalos de confiança assintóticos, para se classificar uma amostra bivariada como independente.

6.1 Intervalo de Confiança Para χ

Os intervalos de confiança de **C1** foram definidos a partir dos valores c_p's obtidos por simulação de Monte Carlo. Entretanto, na Tabela 6.1 são apresentados resultados de um estudo de simulação com mil amostras de n pares das variáveis X e Y normais independentes, para n igual a 20, 50, 100 e 200, no qual foi avaliada a eficiência do critério **C1** de correta classificação das amostras com pelo menos $100(1-p)\%$ de pontos dentro dos intervalos $\pm\frac{c_p}{\sqrt{n}}$, para $p = 0.05$.

Tabela 6.1: Casos de independência classificados corretamente.

n	20	50	100	200
$c_p = 1.78$	449 (44,9%)	506 (50,6%)	448 (44,8%)	438 (43,8%)

Estes resultados indicam que é alta a porcentagem de amostras que não satisfazem a característica **C1**, independente do tamanho de amostra. Este fato sugere que os intervalos de confiança para χ definidos em **C1** apresentam amplitudes menores do que deveriam ser. Portanto, uma porcentagem de pontos bem maior que o esperado excede os limites e induzem a classificação errônea de dependência entre X e Y para amostras onde X e Y são independentes.

Lembrando que os intervalos em **C1** foram definidos pelos autores utilizando simulação de Monte Carlo, neste trabalho determinamos os intervalos de confiança para χ com base na sua própria distribuição assintótica.

6.2 Distribuição Assintótica de χ

Nesta seção é apresentado o desenvolvimento da distribuição assintótica da estatística χ_i, por semelhança com a distribuição assintótica da estatística qui-quadrado do teste de independência sobre uma tabela de contingência 2×2. A partir dessa distribuição é possível assim obter o intervalo de confiança assintótico.

Teorema 6.1 *Considere o Chi-plot definido por Fisher & Switzer (1985), ou seja, as funções empíricas H_i, F_i e G_i tais que*

$$\begin{aligned} H_i &= H^n(x_i, y_i) = \frac{\sum_{j \neq i} I\{x_j \leq x_i, y_j \leq y_i\}}{(n-1)}, \\ F_i &= F^n(x_i) = \frac{\sum_{j \neq i} I\{x_j \leq x_i\}}{(n-1)}, \\ G_i &= G^n(y_i) = \frac{\sum_{j \neq i} I\{y_j \leq y_i\}}{(n-1)}, \end{aligned}$$

onde $I(A)$ é a função indicadora sobre o conjunto A, e

$$\begin{aligned} S_i &= sinal\{(F_i - 0.5)(G_i - 0.5)\}, \\ \chi_i &= (H_i - F_i G_i)/\sqrt{F_i(1-F_i)G_i(1-G_i)}, \ e \\ \lambda_i &= 4S_i \max\{(F_i - 0.5)^2, (G_i - 0.5)^2\}. \end{aligned}$$

Se as variáveis X e Y forem independentes, então assintoticamente $\chi \sim N(0, 1/(n-1))$, ou seja χ tem distribuição de probabilidade normal com média zero e variância $1/(n-1)$.

Demonstração: Para cada par de pontos (x_i, y_i) de uma amostra $(x_1, y_1), \ldots, (x_n, y_n)$, podemos particionar o plano (X, Y) em quatro quadrantes como ilustrado na Figura 6.1, representados pelos conjuntos A_i, B_i, C_i e D_i, sendo:

$$\begin{aligned} A_i &= \{x \leq x_i \cap y \leq y_i\}, \qquad B_i = \{x \leq x_i \cap y > y_i\}, \\ C_i &= \{x > x_i \cap y > y_i\} \quad \text{e} \quad D_i = \{x > x_i \cap y \leq y_i\}. \end{aligned} \tag{6.2.1}$$

Sejam a_i, b_i, c_i e d_i os números de pontos da amostra nos conjuntos A_i, B_i, C_i e D_i, respectivamente, exceto o ponto (x_i, y_i).

A Tabela 6.2 apresentado como esses números podem ser representados segundo o posicionamento dos ponto (x, y) em relação a (x_i, y_i).

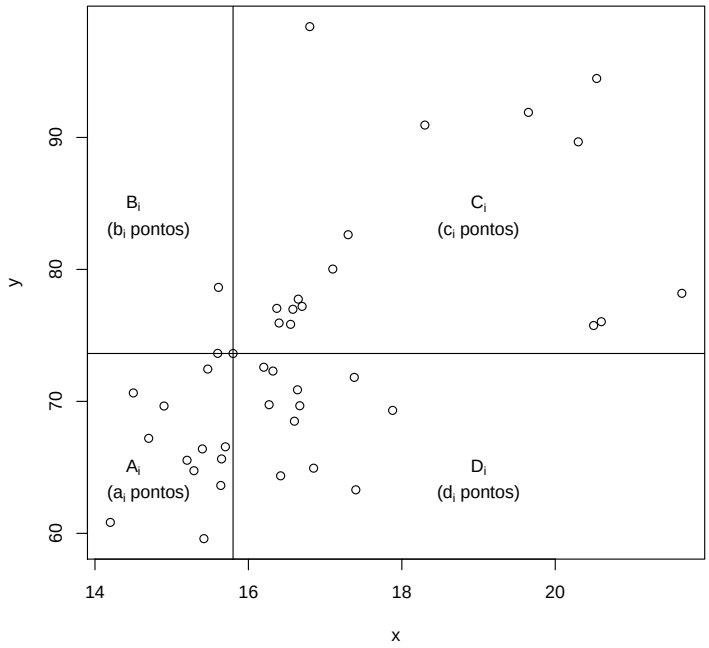

Figura 6.1: Quadrantes definidos pelo ponto (x_i, y_i).

Tabela 6.2: Números de pontos nos conjuntos A_i, B_i, C_i e D_i.

	abaixo ou igual a x_i	acima de x_i
abaixo ou igual a y_i	a_i	d_i
acima de y_i	b_i	c_i

Utilizando as frequências de pontos, podemos reescrever as Equações 4.2.1, 4.2.3 e 4.2.2 como:

$$F_i = \frac{a_i + b_i}{(n-1)}, \;\; G_i = \frac{a_i + d_i}{(n-1)} \;\; \text{e} \;\; H_i = \frac{a_i}{(n-1)}.$$

Ao substituir F_i, G_i e H_i de (4.2.5), obtemos

$$\chi_i = \frac{\dfrac{a_i}{(n-1)} - \dfrac{(a_i+b_i)}{(n-1)}\dfrac{(a_i+d_i)}{(n-1)}}{\sqrt{\dfrac{(a_i+b_i)}{(n-1)}\dfrac{(c_i+d_i)}{(n-1)}\dfrac{(a_i+d_i)}{(n-1)}\dfrac{(b_i+c_i)}{(n-1)}}} \tag{6.2.2}$$

$$= \frac{\dfrac{(n-1)a_i - (a_i+b_i)(a_i+d_i)}{(n-1)^2}}{\dfrac{\sqrt{(a_i+b_i)(c_i+d_i)(a_i+d_i)(b_i+c_i)}}{(n-1)^2}}, \tag{6.2.3}$$

e, dado que $(n-1) = a_i + b_i + c_i + d_i$, então

$$\chi_i = \frac{a_i c_i - b_i d_i}{\sqrt{(a_i+b_i)(c_i+d_i)(a_i+d_i)(b_i+c_i)}}. \tag{6.2.4}$$

Para a Tabela 6.2, a estatística χ_1^2 pode ser escrita como

$$\chi_1^2 = \frac{(n-1)(a_i c_i - b_i d_i)^2}{(a_i+b_i)(c_i+d_i)(a_i+d_i)(b_i+ci)}, \tag{6.2.5}$$

a qual tem distribuição assintótica qui-quadrado com 1 grau de liberdade. Portanto, $(n-1)(\chi_i)^2 = \chi_1^2$, ou seja, $(n-1)(\chi_i)^2$ converge assintoticamente para uma distribuição qui-quadrado com 1 grau de liberdade e, então, $\chi \sim N(0, 1/(n-1))$ assintoticamente.

■

A partir do Teorema 6.1 pode ser definido um *Intervalo de Confiança Assintótico* de $(100 - \frac{\alpha}{2})\%$ de confiança para χ com limites

$$\pm \frac{z_{1-\alpha/2}}{\sqrt{(n-1)}}, \tag{6.2.6}$$

onde $z_{1-\alpha/2}$ é o $(1-\alpha/2)$ quantil da distribuição normal padrão, que para os níveis de confiança de 90%, 95% e 99% tem os valores 1.64, 1.96 e 2.57, respectivamente.

6.3 IC assintótico para χ

Um estudo de simulação análogo ao apresentado na Seção 6.1 foi realizado para avaliar a eficiência do critério de se terem pelo menos 95% dos valores de χ_i dentro do IC assintótico de 95%. Para tal, foram utilizadas as mesmas mil amostras que foram utilizadas no estudo da Seção 6.1, de n pares das variáveis X e Y normais independentes, para n igual a 20, 50, 100 e 200. Na Tabela 6.3 aparecem os números e proporções de amostras que satisfazem o critério acima, junto com os números e proporções de amostras que satisfazem **C1**.

Tabela 6.3: Amostras Classificadas Corretamente Pelos IC das Equações (4.2.7) e (6.2.6).

n	20	50	100	200
$c_p = 1.78$	449 (44,9%)	506 (50,6%)	448 (44,8%)	438 (43,8%)
$z_{1-\alpha/2} = 1.96$	573 (57,3%)	691 (69,1%)	645 (64,5%)	650 (65,0%)

Observa-se números menores de amostras classificadas incorretamente pelo intervalo como na Equação (6.2.6) com relação às classificadas pelo intervalo como na Equação (4.2.7). Entretanto, para ambos os tipos intervalos ocorre uma quantidade grande de amostras onde o número de pontos fora dos intervalos é maior que o esperado.

Resultados dos números de pontos fora dos intervalos das Equações (4.2.7) e (6.2.6) foram obtidos a partir de trezentas amostras bivariadas independentes, geradas com $X \sim N(50, 7^2)$ e $Y \sim N(30, 2^2)$ para os tamanhos de amostra 20, 50, 100 e 200. Para a verificação de independência entre as variáveis X e Y em cada amostra utilizou-se o teste qui-quadrado. Na Tabela 6.4 apresentamos o resumo destes números.

Tabela 6.4: Tabela de Proporção de Pontos Fora dos Intervalos das Equações (4.2.7) e (6.2.6).

		proporção para $c_p = 1.54$			proporção para $c_p = 1.78$			proporção para $c_p = 2.18$		
	n	**moda**	**média**	**mediana**	**moda**	**média**	**mediana**	**moda**	**média**	**mediana**
Fisher & Switzer	20	0,0085	0,1392	0,1250	0,0025	0,0838	0,0625	0,0001	0,0322	0,0001
	50	0,0471	0,1131	0,0870	0,0130	0,0650	0,0435	0,0007	0,0278	0,0001
	100	0,0557	0,1202	0,1064	0,0276	0,0722	0,0532	0,0028	0,0263	0,0106
	200	0,0807	0,1191	0,1071	0,0378	0,0725	0,0561	0,0051	0,0282	0,0179
		proporção para $z_{1-\alpha/2} = 1.64$			proporção para $z_{1-\alpha/2} = 1.96$			proporção para $z_{1-\alpha/2} = 2.57$		
	n	**moda**	**média**	**mediana**	**moda**	**média**	**mediana**	**moda**	**média**	**mediana**
Assintótico	20	0,0136	0,1045	0,0625	0,0001	0,0543	0,0001	0,0001	0,0107	0,0001
	50	0,0340	0,0895	0,0652	0,0060	0,0418	0,0217	0,0003	0,0107	0,0001
	100	0,0436	0,0966	0,0745	0,0149	0,0482	0,0319	0,0001	0,0105	0,0001
	200	0,0528	0,0960	0,0816	0,0201	0,0483	0,0357	0,0006	0,0110	0,0051

Observamos que para todos os casos de α igual a 10%, 5% e 1% os números médios de pontos fora dos intervalos da equação (4.2.7) se mostram maiores que o esperado (10%, 5% e 1%) respectivamente, mais ainda na medida em que o α diminui. Entretanto, os números médios de pontos fora dos intervalos considerando a Equação (6.2.6) são bastante próximos dos valores esperados.

6.3.1 Obtenção do IC através da Proporção de Pontos

Os resultados das Tabelas 6.3 e 6.4 sugerem que os c_p's não sejam coerentes com os valores nominais α iguais à 0.10, 0.05 e 0.01. Por meio de um estudo de simulação com trezentas amostras e utilizando valores de c_p entre 1.50 e 2.70 com acréscimo de 0.01, foi estimada a proporção de pontos no intervalo $[-c_p/\sqrt{n}, c_p/\sqrt{n}]$ e determinado o valor de c_p

correspondente ao nível nominal α iguais a 10%, 5% e 1%. Na Tabela 6.5 são apresentados os resumos dos valores estimados de c_p para os níveis nominais de 10%, 5% e 1%.

Tabela 6.5: Estimativas do Valor de c_p para Níveis de Confiança Estabelecidos.

	estimativas de c_p para 90%			estimativas de c_p para 95%			estimativas de c_p para 99%		
n	**moda**	**média**	**mediana**	**moda**	**média**	**mediana**	**moda**	**média**	**mediana**
20	1,5132	1,7552	1,5900	1,5464	1,9665	1,8750	1,5464	1,9665	1,8750
50	1,5055	1,6759	1,5200	1,5381	1,8269	1,7450	2,6726	2,1751	2,1700
100	1,5098	1,6754	1,5600	1,5497	1,8808	1,8400	2,6841	2,3543	2,4100
200	1,5121	1,6748	1,5650	1,5742	1,9018	1,8400	2,6769	2,3686	2,3900

Dos resultados da Tabela 6.5 é possível notar que os valores estipulados pelos autores Fisher & Switzer (2001), em geral, são menores que os valores médios estimados de tal maneira que os intervalos obtidos considerando a Equação (4.2.7) acabam classificando amostras independentes como amostras dependentes. Este fato pode ser constatado por meio dos resultados expostos na Tabela 6.6 obtidos a partir das mesmas trezentas amostras, que mostram que a porcentagem de classificação correta das amostras de variáveis independentes pelos intervalos obtidos via Equação (4.2.7) é menor do que os intervalos obtidos via Equação (6.2.6).

Tabela 6.6: Contagem e Porcentagem de Classificação Correta.

	Fisher & Switzer			Assintótico		
n	90%	95%	99%	90%	95%	99%
20	140(46,67%)	127(42,33%)	208(69,33%)	173(57,67%)	171(57,00%)	266(88,67%)
50	155(51,67%)	168(56,00%)	152(50,67%)	190(63,33%)	219(73,00%)	226(75,33%)
100	145(48,33%)	137(45,67%)	104(34,67%)	189(63,00%)	194(64,67%)	180(60,00%)
200	141(47,00%)	129(43,00%)	107(35,67%)	181(60,33%)	182(60,67%)	183(61,00%)

A Tabela 6.7 apresenta os valores dos limites superiores do IC Assintótico, do IC segundo Fisher & Switzer (2001) e a média dos intervalos de confiança estimados.

Da Tabela 6.7 é possível constatar que os valores da média do intervalo de confiança encontrados se assemelham com os intervalos de confiança assintóticos, exceto ao nível de confiança de 99% onde se compara ao intervalo segundo Fisher & Switzer (2001) o qual se encontra praticamente abaixo dos valores da média do intervalo de confiança encontrado.

Tabela 6.7: Limites Superiores do IC de Fisher & Switzer (2001) e do IC Assintótico.

	90%			95%			99%		
n	*Assint.*	média	*F&S*	*Assint.*	média	*F&S*	*Assint.*	média	*F&S*
20	0,3762	0,3925	0,3533	0,4397	0,4497	0,4084	0,5896	0,4397	0,5001
50	0,2343	0,2370	0,2200	0,2800	0,2584	0,2543	0,3671	0,3076	0,3114
100	0,1648	0,1675	0,1548	0,1970	0,1881	0,1789	0,2583	0,2354	0,2191
200	0,1163	0,1184	0,1092	0,1389	0,1345	0,1262	0,1822	0,1675	0,1545

6.3.2 Probabilidade de Cobertura do Parâmetro Estimado

Para a análise da probabilidade de cobertura do parâmetro estimado p^* (proporção de pontos que não pertencem ao intervalo de confiança do Chi-plot), consideramos um procedimento bootstrap para cada amostra, onde é contado a proporção de intervalos bootstrap construídos que contém o parâmetro estimado em suas respectivas amostras.

A partir dos gráficos a) e c) da Figura 6.2, fica claro que a probabilidade da proporção de pontos fora do IC obtida via Equação (6.2.6) ser pequena ($\leq \alpha = 5\%$) é maior do que para o IC obtido via Equação (4.2.7). Por outro lado, que os gráficos b) e d) da Figura 6.2 sugerem que os IC obtidos via Equações (4.2.7) e (6.2.6) superestimam a proporção de pontos fora de seus limites, com probabilidade bem maior do que os limites do IC Assintótico. A Figura 6.2 b) e d) mostram a estimação da distribuição bootstap da proporção de pontos fora do intervalo para esta determinada amostra para o IC Assintótico e do proposto por Fisher, N. & Switzer, P. (2001) de 95% respectivamente onde o ponto mostrado é a proporção de pontos que ficam fora dos IC respectivos para uma determinada amostra e as barras representam o intervalo de confiança de 95% obtidos das amostras bootstraps respectivas.

Na Tabela 6.8 é apresentado a probabilidade de cobertura de $(100-\alpha)\%$ do parâmetro estimado p^* que é proporção de pontos que se encontram fora do intervalo de confiança de 99%, 95% e 90% do Chi-plot. Para a determinação da cobertura de p^* considerou-se as estratégias bilateral e unilateral, pois como observado a distribuição concentra os pontos próximo de zero. Se a estratégia for a bicaudal, a proporção esperada de pontos que se encontram fora dos IC for próxima de zero, os intervalos bootstrap para o IC assintótico em grande maioria não conteriam o valor estimado, como para o caso de $z_{1-\alpha/2} = 2,57$, enquanto que para o IC segundo Fisher & Switzer (2001) isso não ocorre pois a distribuição de p^* para os IC obtidos via Equação (4.2.7) tem concentração de massa para valores maiores que o esperado para p^*.

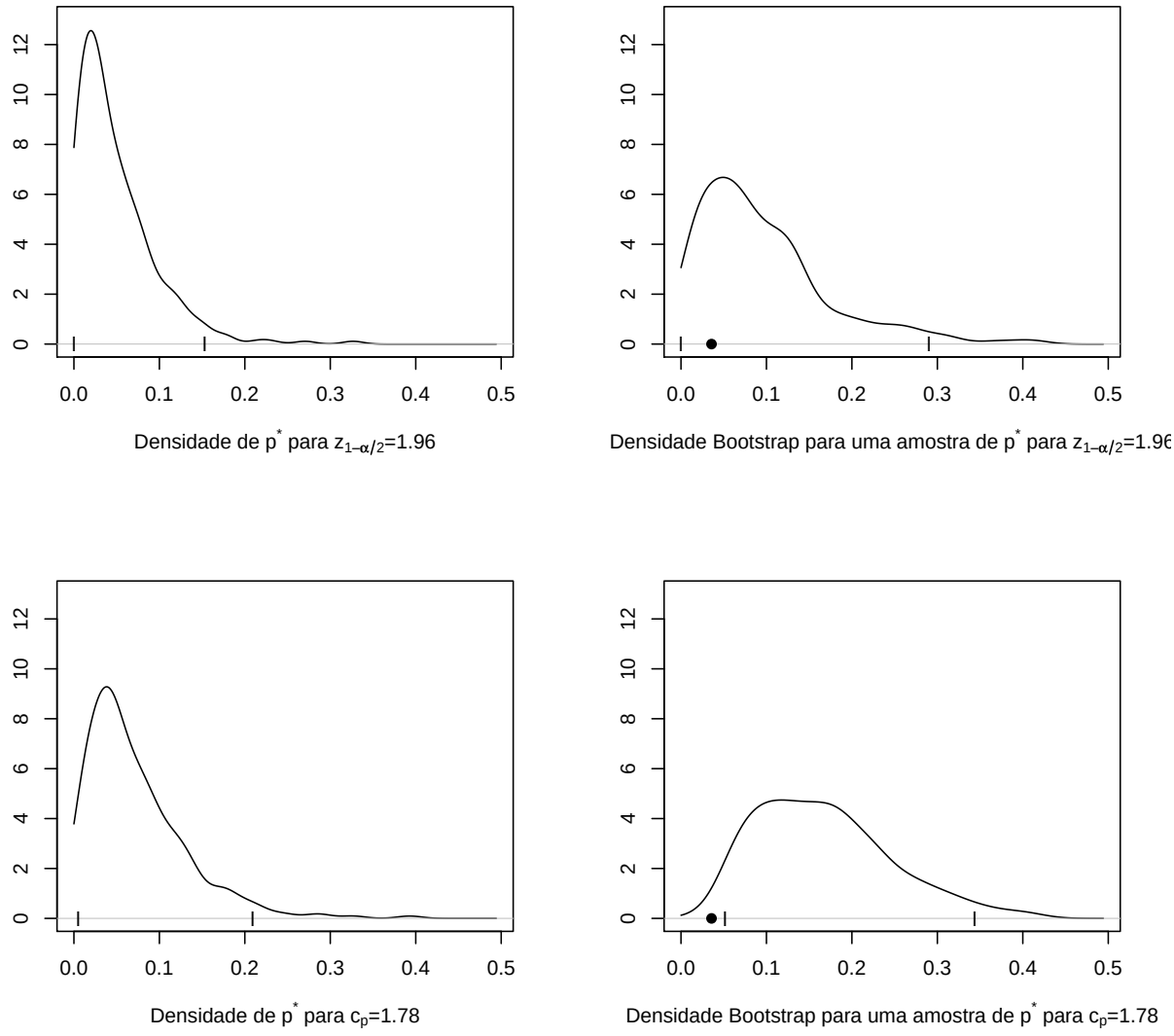

Figura 6.2: Densidades Estimadas de p^* com trezentas amostras para a Equação (6.2.6) em a) e para a Equação (4.2.7) em c). Uma densidade bootstrap para a Equação (6.2.6) em b). Uma densidade bootstrap para a equação (4.2.7) em d).

Tabela 6.8: Coberturas Obtidas Através do Bootstrap para Distribuição Bicaudal e Unicaudal.

		para $z_{1-\alpha/2} = 1,64$			para $z_{1-\alpha/2} = 1,96$			para $z_{1-\alpha/2} = 2,57$		
Bicaudal	n	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**
	20	0,6333	0,6233	0,6033	0,4133	0,4033	0,3900	0,0033	0,0001	0,0001
	50	0,8467	0,8200	0,8000	0,6233	0,6133	0,5933	0,0033	0,0001	0,0001
	100	0,9267	0,9133	0,8800	0,8067	0,7900	0,7833	0,0001	0,0001	0,0001
	200	0,9600	0,9133	0,8467	0,9033	0,8733	0,8133	0,0001	0,0001	0,0001
		para $c_p = 1,54$			para $c_p = 1,78$			para $c_p = 2,18$		
	n	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**
	20	0,7700	0,7667	0,7667	0,7700	0,7667	0,7667	0,3066	0,3066	0,3033
	50	0,9133	0,7800	0,6100	0,9100	0,7833	0,6133	0,4933	0,4933	0,4833
	100	0,8433	0,6133	0,4333	0,8500	0,6133	0,4367	0,6533	0,6200	0,4833
	200	0,7300	0,5000	0,3367	0,7200	0,4900	0,3367	0,7500	0,5233	0,3500
		para $z_{1-\alpha/2} = 1,64$			para $z_{1-\alpha/2} = 1,96$			para $z_{1-\alpha/2} = 2,57$		
Unicaudal	n	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**
	20	0,9767	0,9567	0,9300	0,9733	0,9600	0,9233	0,9999	0,9999	0,9999
	50	0,9667	0,9333	0,9033	0,9800	0,9533	0,9233	0,9999	0,9999	0,9999
	100	0,9500	0,9333	0,9033	0,9633	0,9533	0,9300	0,9999	0,9999	0,9999
	200	0,9500	0,9067	0,8633	0,9733	0,9367	0,8933	0,9999	0,9999	0,9999
		para $c_p = 1,54$			para $c_p = 1,78$			para $c_p = 2,18$		
	n	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**	**99%**	**95%**	**90%**
	20	0,9999	0,9999	0,9967	0,9999	0,9999	0,9967	0,9999	0,9999	0,9999
	50	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
	100	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
	200	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

Da Tabela 6.8 é possível notar que como o esperado a cobertura dos intervalos assintóticos melhoram para os intervalos de 90% e 95% no caso bicaudal conforme o tamanho da amostra aumenta, mas não ocorre o mesmo para o intervalo de 99% com $z_{1-\alpha/2} = 2.57$ e ocorre uma piora na cobertura para todos os três intervalos propostos por Fisher & Switzer (2001).

Da Tabela 6.8, para o caso unicaudal, a probabilidade de cobertura permanece próxima do valor nominal esperado para os intervalos assintóticos, com exceção do intervalo de 99% com $z_{1-\alpha/2} = 2.57$. A probabilidade de cobertura para os intervalos propostos por Fisher & Switzer (2001) é de praticamente 100% para todos os casos assim como para o intervalo Assintótico de 99% com $z_{1-\alpha/2} = 2.57$.

Da Tabela 6.8 concluímos que os intervalos mais próximos dos valores nominais estabelecidos são os intervalos assintóticos com $z_{1-\alpha/2} = 1.64$ e $z_{1-\alpha/2} = 1.96$, porque estes apresentam probabilidades de cobertura próximas aos valores esperados e não se mantém praticamente constantes e/ou com baixa probabilidade de cobertura, como os intervalos propostos por Fisher & Switzer (2001), para o caso bilateral e para o caso bilateral, respectivamente.

6.4 Outras Características do Chi-plot

De acordo com Fisher & Switzer (1985), quatro características são inerentes ao Chi-plot sobre o caso de independência:

1. Cada quadrante do chi-plot deve conter 25% dos pontos (χ_i, λ_i);

2. 50% dos valores λi são positivos;

3. 50% dos valores χ_i são positivos e distribuídos aleatóriamente;

4. λ tem distribuição uniforme no intervalo $\pm 4\left(\frac{1}{n-1} - 0.5\right)^2$.

Com base num estudo de simulação com as trezentas amostras referidas na Seção 6.3, obteve-se resultados da eficiência destes quatro critérios para a verificação da hipótese de independência das variáveis X e Y.

A avaliação da Característica 1 foi realizada com base no teste qui-quadrado, as das Características 2 e 3 com base no teste normal, e a da Característica 4 com base no teste de Kolmogorov-Smirnov, cada um com nível de significância de 5%. Na Tabela 6.9

são apresentados os resultados em termos de proporções de amostras independentes que satisfazem cada um dos critérios, para amostras de tamanho 20, 50 100 e 200.

Tabela 6.9: Testes Estatísticos Baseados no Comportamento do Chi-Plot.

Testes com α=5%				
n	25% cada quadrante	50% de pontos $\lambda > 0$	50% de pontos $\chi > 0$	Uniformidade de λ
20	0,7300	0,8967	0,6233	0,9600
50	0,5033	0,9333	0,4333	0,9767
100	0,3533	0,9467	0,3900	0,9800
200	0,2333	0,9433	0,2700	0,9833

Enquanto os Critérios 1 e 3 tem baixa eficiência, mais ainda para tamanho de amostras maiores, os Critérios 2 e 4 são muito eficientes para os diferentes tamanhos de amostras.

Nos capítulos anteriores o Chi-plot é uma boa ferramenta para identificar e ou determinar a dependência, no entanto uma análise estatística para o comportamento do Chi-plot (Tabela 6.9) ou o intervalo proposto por Fisher & Switzer (2001) (ver Tabela 6.6), não aparenta ser a melhor estratégia para se determinar a existência da dependência. Uma alternativa pode ser construída através da distribuição assintótica do Chi-plot, onde uma análise pode ser desenvolvida para se testar o comportamento dos chi_i para inferir sobre a dependência e posteriormente pode ser construído estratégias para se analisar a existência da dependência.

De acordo com o Teorema 6.1, os valores χ_i tem distribuição assintótica $N(0, 1/(n-1))$. Assim Consideramos de interesse estudar a viabilidade da utilização do Teorema 6.1 para a avaliação de independência entre as duas variáveis de uma amostra bivariada. Para este estudo utilizamos as mesmas trezentas amostras referidas e os testes de aderência de Shapiro, duas variações do teste Cramér Von Mises, o teste de Sherman, o teste Qui-Quadrado e o teste de Komogorov-Smirnov, todos com níveis de significância de 10%, 5% e 1%.

O critério adotado para se concluir sobre a independência entre as duas variáveis de uma amostra bivariado é baseado na afirmação de que se $\chi \sim N(0, 1/(n-1))$ implica que X e Y são independentes. Esta afirmação pode ser informalmente explicada que ao se dizer que se X e Y são dependentes, teríamos uma concentração de pontos no primeiro e terceiro quadrante e/ou no segundo e quarto quadrante do Chi-plot, e pela medida chi_i do Chi-plot ser medida local, esta concentração indica a existência de "dependência linear" local

significativa ou não (com o ponto chi_i permanecendo próximo ao limite de confiança), pois como é conhecido na literatura o valor χ_i é o coeficiente ϕ de Pearson que mede a correlação linear, e desta forma, a variância e/ou a média da distribuição de χ não coincidiria com a variância e ou a média da $N(0, 1/(n-1))$, o que seria uma contradição.

Tabela 6.10: Aceitação de $N(0, 1/(n-1))$ de χ em 300 amostras.

	Shapiro			Cramer vm v.1			Cramer vm v.2		
n	10%	5%	1%	10%	5%	1%	10%	5%	1%
20	109(36,3%)	155(51,7%)	211(70,3%)	140(46,7%)	169(56,3%)	213(71,0%)	172(57,3%)	172(57,3%)	210(70,0%)
50	57(19,0%)	78(26,0%)	111(37,0%)	85(28,3%)	114(38,0%)	141(47,0%)	104(34,7%)	104(34,7%)	137(45,7%)
100	21(7,0%)	37(12,3%)	66(22,0%)	59(19,7%)	74(24,7%)	108(36,0%)	66(22,0%)	66(22,0%)	96(32,0%)
200	8(2,7%)	11(3,7%)	25(8,3%)	28(9,3%)	43(14,3%)	68(22,7%)	29(9,7%)	29(9,7%)	53(17,7%)
	Sherman			Qui-Quadrado			Komogorov-Smirnov		
n	10%	5%	1%	10%	5%	1%	10%	5%	1%
20	214(71,3%)	236(78,7%)	257(85,7%)	198(66,0%)	214(71,3%)	251(83,7%)	150(50,0%)	174(58,0%)	221(73,7%)
50	184(61,3%)	208(69,3%)	243(81,0%)	126(42,0%)	160(53,3%)	195(65,0%)	86(28,7%)	113(37,7%)	150(50,0%)
100	157(52,3%)	185(61,7%)	214(71,3%)	84(28,0%)	98(32,7%)	142(47,3%)	57(19,0%)	80(26,7%)	111(37,0%)
200	113(37,7%)	133(44,3%)	166(55,3%)	38(12,7%)	49(16,3%)	80(26,7%)	31(10,3%)	38(12,7%)	67(22,3%)

A Tabela 6.10 apresenta os números e as porcentagens de aceitação da distribuição $N(0, 1/(n-1))$ para as 300 amostras, pelos seis diferentes testes. Com nenhum dos seis testes individualmente, o critério apresenta eficiência aceitável, pior para os tamanhos de amostra maiores. Entretanto, da Tabela 6.11, com o resultado da avaliação simultânea dos seis testes e considerando o critério satisfeito quando aceito por pelo menos um dos seis testes, observa-se uma maior coerência dos resultados.

A Tabela 6.11 é construída com base na afirmação de que se for aceito em algum dos testes que $\chi \sim N(0, 1/(n-1))$ consiste em aceitar a hipótese de que as variáveis são independentes.

Tabela 6.11: Classificados Como Independentes por Algum Teste da Tabela 6.10.

n	$alpha$ =10%	5%	1%
20	238(79.33%)	250(83.33%)	273(91.00%)
50	201(67.00%)	218(72.67%)	250(83.33%)
100	173(57.67%)	194(64.67%)	221(73.67%)
200	118(39.33%)	138(46.00%)	170(56.67%)

A partir dos resultados da Tabela 6.11, decidiu-se realizar o estudo da eficiência de classificação das amostras independentes pela combinação de qualquer um dos seis testes da normalidade $N(0, 1/(n-1))$ com os intervalos obtidos via Equação (4.2.7) ou obtido via Equação (6.2.6). A Tabela 6.12 é obtida com as mesmas trezentas amostras independentes utilizadas nas simulações anteriores e dos resultados apresentados pode-se observar um apreciável aumento de eficiência de correta avaliação de independência das variáveis X e Y, com relação aos resultados da Tabela 6.11, particularmente para os intervalos assintóticos da Equação (6.2.6).

Tabela 6.12: Classificados Como Independentes em Qualquer Teste ou no IC.

	Fisher & Switzer (2001)			Assintótico		
n	10%	5%	1%	10%	5%	1%
20	266(88,66%)	260(86,66%)	289(96,33%)	274(91,33%)	278(92,66%)	295(98,33%)
50	242(80,66%)	239(79,66%)	277(92,33%)	259(86,33%)	277(92,33%)	286(95,33%)
100	232(77,33%)	214(71,33%)	250(83,33%)	257(85,66%)	265(88,33%)	270(90,00%)
200	192(64,00%)	160(53,33%)	216(72,00%)	226(75,33%)	232(77,33%)	249(83,00%)

As conclusões anteriores relativas à Tabela 6.12 sugerem a verificação da normalidade $N(0, 1/(n-1))$ dos valores χ_i, como forma eficiente de avaliação de independência das variáveis X e Y nas amostras bivariadas.

Nos anteriores estudos de simulação sobre a eficiência das características inerentes ao Chi-plot para a avaliação de independência entre as variáveis X e Y das amostras bivariadas não foram mencionados resultados relativos à eficiência dos coeficientes usuais τ de Kendall, ρ de Spearman e o coeficiente de correlação r de Pearson. Na Tabela 6.13 aparecem o número de amostras com correta avaliação de independência entre as variáveis X e Y, dentre as mesmas trezentas amostras utilizadas nos anteriores estudos de simulação. Estes números indicam alta eficiência para todos os níveis de significância.

Tabela 6.13: Número de Amostras Classificadas como Independentes pelos Testes Usuais.

	p value > 10%			> 5%			> 1%		
n	$\tau = 0$	$\rho = 0$	$r = 0$	$\tau = 0$	$\rho = 0$	$r = 0$	$\tau = 0$	$\rho = 0$	$r = 0$
20	264(88,0%)	265(88,3%)	260(86,7%)	279(93,0%)	283(94,3%)	275(91,6%)	295(98,3%)	295(98,3%)	297(99,0%)
50	274(91,3%)	274(91,3%)	276(92,0%)	288(96,0%)	289(96,3%)	291(97,0%)	299(99,7%)	298(99,3%)	299(99,7%)
100	271(90,3%)	270(90,0%)	268(89,3%)	285(95,0%)	285(95,0%)	287(95,7%)	297(99,0%)	298(99,3%)	299(99,7%)
200	271(90,3%)	270(90,0%)	269(89,7%)	290(96,7%)	290(96,7%)	288(96,0%)	299(99,7%)	299(99,7%)	298(99,3%)

A avaliação das características próprias do Chi-plot para os exemplos de ilustração das situações de dependência "linear forte", "quadrática", "normais" e "exponenciais" são apresentadas no Apêndice A.

6.5 Conclusões da Simulação

A Tabela 9.1 no Apêndice A traz a quantidade dentre trezentas amostras bivariadas, de variáveis com dependência "linear forte", "quadrática", "normais" e "exponenciais" que satisfazem as características inerentes ao Chi-plot sob a hipótese de independência(Seção 6.4). Os resultados encontrados se mostram consistentes somente para o caso de dependência "linear forte", para amostras de tamanho 50 ou mais. Para as outras situações de dependência, os resultados apresentam comportamentos distintos mas em geral inconsistentes com a situação de dependência. Portanto é inviável e não aconselhável a verificação somente destas características para concluir sobre a existência de dependência e consequentemente devemos buscar algum outro método que apresente consistência.

Os resultados da Tabela 9.2 são óbvios, considerando que são relativos à combinações dos resultados da Tabela 9.1.

Os resultados da Tabela 9.3 se referem ao número de amostras classificadas incorretamente como independentes, através dos IC das equações (4.2.7) e (6.2.6), se mostram consistentes somente para o caso de dependência "linear forte", ou "quadrática", para amostras de tamanho 50 ou mais. Para as outras situações de dependência, os resultados apresentam comportamentos distintos mas em geral inconsistentes com a situação de dependência, tornando inviável a verificação destas características como meio para concluir sobre a existência de dependência.

Os resultados das Tabelas 9.4, 9.5 e 9.6 se referem ao número de amostras classificadas incorretamente como independentes, através dos IC das equações (4.2.7) e (6.2.6) combinado com o critério C4 do início do presente capítulo, através de algum teste da $N(0, 1/(n-1))$ e através dos IC das equações (4.2.7) e (6.2.6) ou algum teste da $N(0, 1/(n-1))$ respectivamente, se comportam como os da Tabela 9.3.

Os resultados da Tabela 9.7 se referem ao número de amostras classificadas incorretamente como independentes, através dos testes dos coeficientes usuais τ de Kendall, ρ de Spearman e o coeficiente de correlação r de Pearson se mostram consistentes somente para o caso "linear forte"

Na Tabela 9.6 verifica-se que a eficiência de correta classificação dentre as trezentas amostras com dependência "linear forte", "quadrática", e "exponenciais" é bem inferior aos da Tabela 9.7. Para o caso de dependência das amostras "normais" a baixa eficiência de correta classificação de dependência é consistente com a forma de geração, onde todos os "aglomerados" são gerados com variáveis independentes, portanto com valores dos chi_i próximos de zero e entre os IC das Equações (4.2.7) e (6.2.6) de maneira que passem pelas estratégias adotadas mostrando que a análise gráfica do Chi-plot é imprescindível.

Depois de se analisar os resultados das tabelas do Apêndice A, observa-se que a correta classificação das amostras tem menor eficiência para o nível de significância 1%, do que para 5% e 10% e a melhora nos demais casos conforme o tamanho n da amostra aumenta.

A utilização do critério de uniformidade de λ, junto com critério do intervalo de confiança de χ, não resulta em ganho de eficiência deste último. A uniformidade do λ deve ser verificada pela análise gráfica sobre o Chi-plot, pois somente no caso de fácil visualização da dependência linear através do Gráfico de dispersão o teste de uniformidade do λ se mostrou eficiente.

Observa-se que a classificação das amostras pelo critério de normalidade $N(0, 1/(n-1))$ de chi, por meio dos testes de Shapiro, das duas variações do teste Cramér Von Mises, do teste de Sherman, do teste Qui-Quadrado e do teste de Komogorov-Smirnov é ineficiente e, portanto, desaconselhável para classificar corretamente amostras dependentes para amostras de tamanho 20. Para amostras de tamanho maiores que 50 sua eficiência pode ser alta e sua utilização recomendada dependendo do tipo da dependência, como no caso da dependência "linear" ou "quadrática".

Dos resultados da Tabela 9.6, a utilização conjunta dos IC das equações (4.2.7) e (6.2.6) com o teste de normalidade $N(0, 1/(n-1))$ de χ, como critério para classificação correta das amostras dependentes, observa-se melhora na eficiência com o aumento do tamanho da amostra.

Os resultados das tabelas do presente capítulo e do Apêndice A, sugerem recomendar para a correta classificação de uma amostra como independente ou dependente os seguintes passos:

1. Verificar que o IC Assintótico obtido via Equação (6.2.6) contém a proporção esperada de valores χ_i, correspondentes aos níveis de confiança de 90% ou 95%;

2. Verificar a normalidade $N(0, 1/(n-1))$ de χ;

3. Analisar os gráficos de dispersão e Chi-plot.

7 Dependência com Censuras

7.1 Introdução

Na análise de dados de duas variáveis aleatórias contínuas relativas à características da saúde de pacientes é comum a existência de dependência entre os valores das variáveis e presença de censuras nos dados, por motivos de abandono do paciente ou outras situações ou problemas que não estão vinculados ao caso em estudo.

Neste capítulo ilustramos a utilização de um procedimento gráfico para a identificação de dependência, geral ou localmente, entre os dados das duas variáveis observadas sobre os indivíduos. Os gráficos correspondem as estimativas de funções com base em estimativas de densidades univariadas e bivariadas, e funções de sobrevivência univariadas e bivariadas. Estas funções, na presença de dados censurados, não podem ser estimadas da mesma forma que são estimadas as funções relativas aos procedimentos de Cópula e Função de Sibuya.

Também neste capítulo apresentamos diversos procedimentos de estimação apropriados à presença de censuras. Para a análise de dependência bivariada sob censura adotamos o modelo de Clayton (1978), o qual descrevemos a seguir.

7.2 Modelagem da Dependência

Seja (x_i^0, y_i^0), $i = 1, ..., n$ pares independentes e identicamente distribuídos, com função de Sobrevivência contínua $S(x, y)$. Sejam (c_i, d_i), $i = 1, \ldots, n$, os indicadores de censura independentes e identicamente distribuídos para (x_i^0, y_i^0), e $G(x, y)$ a função de Sobrevivência. O modelo para dados bivariados com censura aleatória tem observações $(x_i, y_i, \delta_{1i}, \delta_{2i})$, $i = 1, \ldots, n$, onde

$$x_i = \min(x_i^0, c_i), y_i = \min(y_i^0, d_i^0), \delta_{1i} = I(x_i^0 < c_i), \delta_{2i} = I(y_i^0 < d_i). \tag{7.2.1}$$

Se (x, y) representa o momento de falha do primeiro e segundo elemento com $f(x, y)$, a função de densidade conjunta, segundo Clayton (1978), é dada por,

$$f(x, y) \int_x^\infty \int_y^\infty f(u, v)dudv = \theta \int_x^\infty f(u, y)du \int_y^\infty f(x, v)dv. \tag{7.2.2}$$

Clayton (1978) também define quatro funções derivadas de $f(x, y)$:

(i) Função de Distribuição conjunta de Sobrevivência dada por,

$$S(x,y) = \int_x^\infty \int_y^\infty f(u,v)dudv; \tag{7.2.3}$$

(ii) Função de Risco Acumulado do primeiro elemento dado que o segundo sobreviveu até y_0,

$$g(x,y_0) = \frac{f(x|y \geq y_0)}{S(x|y \geq y_0)}; \tag{7.2.4}$$

(iii) Função de Risco Acumulado do segundo elemento dado que o primeiro sobreviveu até x_0,

$$h(x_0,y) = \frac{f(y|x \geq x_0)}{S(y|x \geq x_0)}; \tag{7.2.5}$$

(iv) Razão de Falha Bivariada dada por,

$$l(x,y) = \frac{f(x,y)}{S(x,y)} = \theta g(x,y)h(x,y). \tag{7.2.6}$$

O parâmetro θ mede o grau de dependência entre as variáveis X e Y. Quando há independência $\theta = 1$; se houver dependência positiva $\theta > 1$, e se houver dependência negativa $0 < \theta < 1$.

Da Equação (7.2.6) pode-se obter que,

$$\theta = \frac{f(x,y)}{S(x,y)g(x,y)h(x,y)}, \tag{7.2.7}$$

para $S(x,y)$, $g(x,y)$ e $h(x,y)$ diferentes de zero.

Para cada par (x,y) se obtêm uma estimativa de θ, portanto para um intervalo em x e um intervalo em y, se obtêm uma superfície de θ. Para uma amostra de n pares (x,y) podemos obter uma suavização da superfície do θ estimado por meio de obtenção do Kaplan-Meier Bivariado (Lin & Ying 1993)e sua suavização por curvas de Bezier (Bae et al 2005), da estimação da função densidade com presença de censuras, da estimação do Kaplan-Meier univariado e da função densidade univariada com dados censurados. As curvas de nível da superfície suavizada de θ sugerem regiões de correlação positiva ou negativa.

7.3 Kaplan-Meier Bivariado

Para o presente estudo o estimador utilizado é o proposto por Lin & Ying (1993), denominado estimador de Kaplan-Meier bivariado, o qual é de fácil implementação.

O estimador para suavização do Kaplan-Meier univariado, não pode ser utilizado repetidamente para estimação da função de sobrevivência bivariada com censura, pois com os métodos de suavização por kernel apresentam resultados pobres e incoerentes (Bae et al 2005). Portanto, utilizaremos o estimador suavizado pelas curvas de Bezier encontrado em Bae et al (2005).

O estimador de Kaplan-Meier bivariado proposto por Lin and Ying (1993), se baseia na equação de sobrevivência, dada por,

$$S(x,y) = P(X > x, Y > y)/G(x \vee y), \tag{7.3.1}$$

onde $x \vee y = \min\{x, y\}$, para então definir a estimativa da função de sobrevivência,

$$\hat{S}(x,y) = \frac{\frac{1}{n}\sum_{i=1}^{n} I(x_i \geq x, y_i \geq y)}{\prod_{i:x_i \vee y_i \leq x \vee y}\left(1 - \frac{1-\delta_i^{\vee}}{n_i^{\vee}}\right)},$$

onde

$$\delta_i^{\vee} = \delta_{1i} * \delta_{2i}, \text{ e } n_i^{\vee} = \sum_j I(x_j \vee y_j \geq x_i \vee y_i).$$

Contudo, este estimador não está definido se o maior valor entre x_i e y_i $i = 1, \ldots, n$ for censurado, pois neste caso $n_i^{\vee} = 1$ e $\delta_i^{\vee} = 0$, fazendo com que o denominador seja zero. Introduzimos aqui uma modificação no procedimento, somando uma unidade em $n_i^{\vee}$ para evitar a indefinição.

O estimador de Lin & Ying (1993) não produz formas suaves. Para obter formas suaves das estimativas de $\hat{S}(,x,y)$ implementamos aqui o procedimento de suavização de Bezier apresentado em Bae et al (2005).

7.3.1 Estimador de Bezier para KM Bivariado

Para a estimação do Kaplan-Meier bivariado reproduzimos o procedimento apresentado em Bae et al (2005), no qual se assume que não há valores repetidos de x nem de y na amostra. Posteriormente define-se duas sequencias: I(.) e J(.), tal que $x_{I(1)} \leq x_{I(2)} \leq \cdots \leq x_{I(N)}$ e $y_{J(1)} \leq y_{J(2)} \leq \cdots \leq y_{J(N)}$ para as observações não censuradas, ou seja $\delta_i^{\vee} = 1$ e $\sum_{i=1}^{n} \delta_i^{\vee} = N$, o número de pares não censurados na amostra.

Considerando o mesmo procedimento de suavização por curvas de Bezier utilizado em Bae et al (2005), utilizam-se $(N+2)^2$ pontos $\mathbf{b}_{ij}$ de Bezier, como definidos a seguir.

Os pontos $\mathbf{b}_{ij}$ são definidos por:

$$\mathbf{b}_{ij} = \begin{pmatrix} x_{I(i)} \\ y_{J(j)} \\ \hat{S}(x_{I(i)}, y_{J(j)}) \end{pmatrix}, \quad i, j = 0, \ldots, N+1, \tag{7.3.2}$$

onde $x_{I(0)} = 0$, $y_{J(0)} = 0$, $x_{I(N+1)} = (1 + 1/n).x_{I(N)}$ e $y_{J(N+1)} = (1 + 1/n).y_{J(N)}$. Portanto,

$$x(u) = \sum_{i=0}^{N+1} x_{I(i)} B_{N+1,i}(u) \tag{7.3.3}$$

$$y(v) = \sum_{j=0}^{N+1} y_{J(j)} B_{N+1,j}(v) \tag{7.3.4}$$

$$z(u,v) = \sum_{i=0}^{N+1} \sum_{j=0}^{N+1} \hat{S}(x_{I(i)}, y_{J(j)}) B_{N+1,i}(u) B_{N+1,j}(v), \text{onde} \tag{7.3.5}$$

$$B_{M,k}(p) = \binom{M}{k} p^k (1-p)^{M-k}. \tag{7.3.6}$$

Então, a suavização por curvas de Bezier para o estimador Kaplan-Meier é dado por

$$\hat{S}_B = z(u,v). \tag{7.3.7}$$

Para maiores detalhes do procedimento é recomendado uma leitura de Kim et al (2003) e Bae et al (2005).

Na implementação do método os pontos onde será realizada a suavização não são atribuídos, são calculados pelas Equações (7.3.3) e (7.3.4). Os pontos calculados são utilizados nos procedimentos que se seguem para a estimação das diferentes funções até a obtenção de θ.

7.4 Estimação da Densidade Bivariada

Para a estimação da densidade $f(x,y)$ das Equação (7.2.7) são utilizadas as técnicas descritas em Wells & Yeo (2005), pelas quais podemos estimar a função densidade bivariada $f(x,y)$ com presença de censura, como,

$$\hat{f}(x,y) = \frac{1}{nb_1b_2} \sum_{i=1}^{n} \sum_{j=1}^{n} K\left(\frac{x - x_i}{b_1}, \frac{y - y_j}{b_2}\right) * \delta_{2j} \left[\frac{I(j > i-1}{\hat{\bar{G}}(x_{i-1}, y_j)} - \frac{I(j > i}{\hat{\bar{G}}(x_i, y_j)}\right], \tag{7.4.1}$$

onde

$$\hat{\bar{G}}(x,y) = \prod_{i=1}^{n} \left[\frac{N(x_i, 0)}{N(x_i, 0) + 1}\right]^{1-\delta_{1i}} * \prod_{j=1}^{n} \left[\frac{N(x, y_j)}{N(x, y_j) + 1}\right]^{\gamma_j(x,y)}, \tag{7.4.2}$$

com

$$N(x,y) = \sum_{i=1}^{n} (x_i > x, y_j > y), \tag{7.4.3}$$

$$\gamma_j(x,y) = I(x_i > x, y_j \leq y, \delta_{2j} = 0). \tag{7.4.4}$$

Para os casos em que $\hat{\bar{G}}(x,y) = 0$ o processo não está bem definido, neste trabalho adotamos a recomendação do autor de adicionar uma unidade em $\hat{\bar{G}}(x,y)$.

Para a estimação das funções $h(x,y)$ e $g(x,y)$ da Equação (7.2.7) é necessário estimar as funções $S(v|z \geq z_0)$ e $f(v|z \geq z_0)$, por meio dos procedimentos do Kaplan-Meier univariado e de Kernel. Para a estimação das funções $S(v|z \geq z_0)$ e $f(v|z \geq z_0)$ considera-se $S(v)$ para $z \geq z_0$ e $f(v)$ para $z \geq z_0$, respectivamente.

7.5 Estimador de Kaplan-Meier Univariado

O estimador de Kaplan-Meier univariado apresentado por Kaplan & Meier (1958), consiste em estimar a função $S(x) = 1 - F(x)$ por:

$$\hat{S}(x) = \begin{cases} 1, & ,0 < x \leq x_1 \\ \prod_{j=1}^{k-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{1j}} & ,x_{k-1} < x \leq x_k, k = 2, \ldots, n \\ 0, & ,x > x_n \end{cases}, \tag{7.5.1}$$

onde $x_1 \leq x_2 \leq \cdots \leq x_n$.

Os pesos dos saltos de $\hat{S}$ nos pontos x_j são dados por

$$\hat{s}_j = \begin{cases} \hat{S}(x_j) - \hat{S}(x_{j+1}), & ,j = 1, \ldots, n-1 \\ \hat{S}(x_n), & ,j = n \end{cases}. \tag{7.5.2}$$

Como em Berg & Politis (2009), o estimador do Kaplan-Meier univariado, suavizado por kernel, é dado por:

$$\hat{S}_h(x) = 1 - \sum_{i=1}^{n} s_j K\left(\frac{x - x_i}{h_n}\right), \tag{7.5.3}$$

onde K e h_n têm as mesmas restrições do que na Equação (2.3.3).

7.6 Estimação da Densidade Univariada

A densidade de probabilidade univariada com presença de censura, é estimada aqui como $\hat{f}_h(x)$, utilizando o procedimento de Berg & Politis (2007), tal que:

$$\hat{f}_h(x) = \frac{1}{h_n} \sum_{i=1}^{n} s_j k\left(\frac{x - x_i}{h_n}\right), \tag{7.6.1}$$

onde k e h_n têm as mesmas restrições do que na equação (2.3.3) e s_j calculado pela equação (7.5.2).

As rotinas de cálculo das estimativas das funções de densidade e de sobrevivência, e do parâmetro θ foram implementadas no Software R e estão disponíveis no Apêndice B.

7.7 Exemplos com Dados Censurados

Similarmente aos exemplos apresentados no Capítulo 5, são apresentados aqui os exemplos: "Exponenciais", "Forma Quadrática" e "Normais", agora com 10% e 15% de censura nos dados de cada variável, sobre as amostras de tamanho vinte e cinquenta.

Em cada um dos exemplos são apresentados: o gráfico de dispersão com os valores censurados em preto em a), a curva de nível de $\theta = 1$ em b), as curvas de nível de $\theta = \{0.5, 0.8\}$ em c) e as curvas de nível com $\theta = \{1.3, 4\}$ em d). Lembramos que as curvas de nível para $\theta < 1$ determinam regiões de pontos com correlação negativa e as curvas de nível para $\theta > 1$ determinam regiões de pontos com correlação positiva.

Exemplo 7.7.1 Consideramos amostras com censuras uniformes de 10% e 15% em cada variável, primeiro sobre vinte e depois sobre cinquenta pares gerados de uma distribuição exponencial bivariada de Gumbel (Gumbel 1960), com coeficiente de correlação linear de -0.20.

A Figura 7.1 apresenta o gráfico a): de dispersão, e os gráficos b), c) e d) com as curvas de nível estimadas, para o caso de censura de 10% sobre uma amostra de vinte pares. A Figura 7.2 apresenta o gráfico a): de dispersão, e os gráficos b), c) e d) com as curvas de nível estimadas, para o caso de censura de 15% sobre uma amostra de vinte pares. A Figura 7.3 apresenta o gráfico a): de dispersão, e os gráficos b), c) e d) com as curvas de nível estimadas, para o caso de censura de 10% sobre uma amostra de cinquenta pares. A Figura 7.4 apresenta o gráfico a): de dispersão, e os gráficos b), c) e d) com as curvas de nível estimadas, para o caso de censura de 15% sobre uma amostra de cinquenta pares.

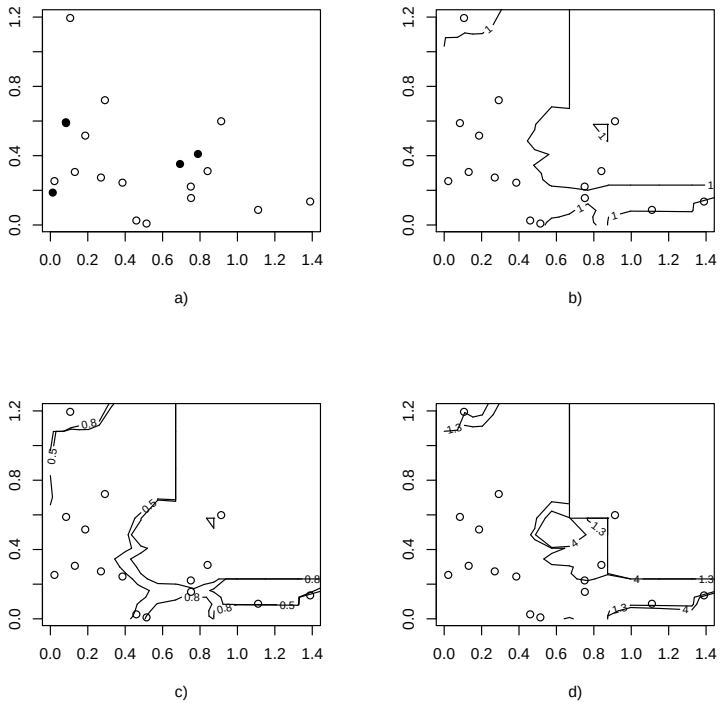

Figura 7.1: Resultados do Exemplo 7.7.1 com 20 pares e 10% de censura em cada variável.

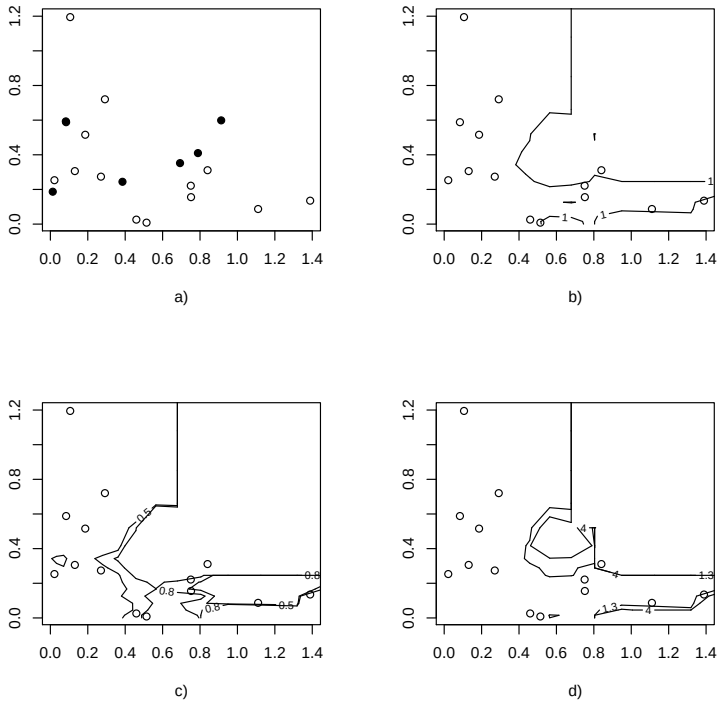

Figura 7.2: Resultados do Exemplo 7.7.1 com 20 pares e 15% de censura em cada variável.

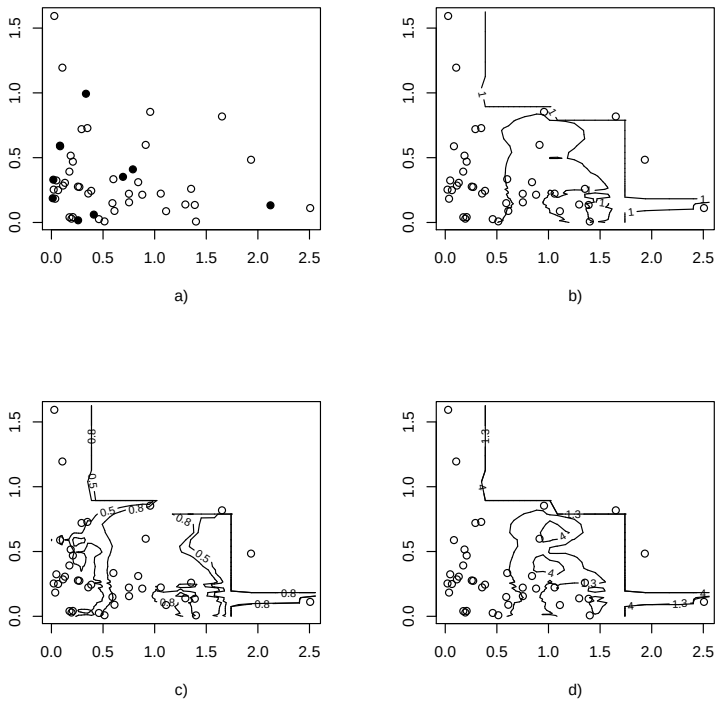

Figura 7.3: Resultados do Exemplo 7.7.1 com 50 pares e 10% de censura em cada variável.

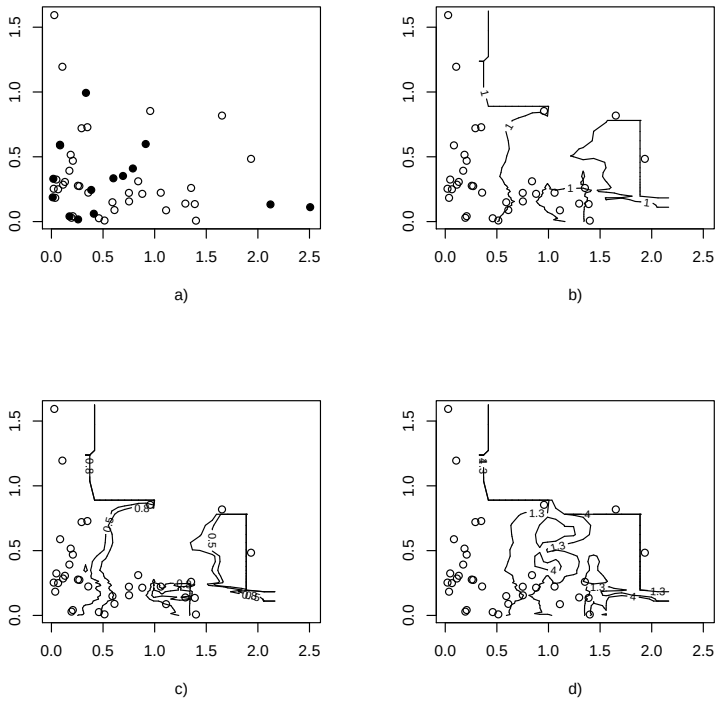

Figura 7.4: Resultados do Exemplo 7.7.1 com 50 pares e 15% de censura em cada variável.

Na Figura 7.1, de acordo com as curvas de nível a grande maioria dos pontos se localizam em regiões de dependência negativa (têm correlação local negativa), a maioria à esquerda da curva de nível de $\theta = 1$. Alguns pontos se encontram posicionados sobre a curva de nível $\theta = 1$ (têm correlação local baixa ou nula). O coeficiente de correlação linear r dos pares que não contém censura é igual a -0.42.

Da Figura 7.2, o comportamento das curvas de nível determinam regiões de correlação local bastante similares às da Figura 7.1. Entretanto, basicamente pela censura do ponto $(0.91, 0.59)$ o coeficiente de correlação linear r passa para -0.52.

Das Figuras 7.3 e 7.4, a grande maioria dos pontos se distribui entre regiões de dependência negativa e regiões de dependência baixa ou nula. Apenas alguns pontos se localizam numa região de correlação local positiva. Em ambos os casos o coeficiente de correlação linear dos pares que não contém censura r é de -0.15.

Exemplo 7.7.2 Consideramos amostras com censuras uniformes de 10% e 15% em cada variável, primeiro sobre vinte e depois sobre cinquenta pares gerados com $X_i \sim U(8, 12)$ e $Y_i = -(X_i - 10)^2 + N(0.8, 1) + 20$, $i = 1, \ldots, n$ para $n = \{20, 50\}$, de tal forma a obter alguns pares de pontos com correlação geral negativa e outros pares de pontos com correlação geral positiva.

Os gráficos das Figuras 7.5 e 7.6 apresentam os resultados relativos aos casos de censuras de 10% e 15%, sobre uma amostra de vinte pares. Os gráficos das Figuras 7.7 e 7.8 apresentam os resultados relativos aos casos de censuras de 10% e 15%, sobre uma amostra de cinquenta pares.

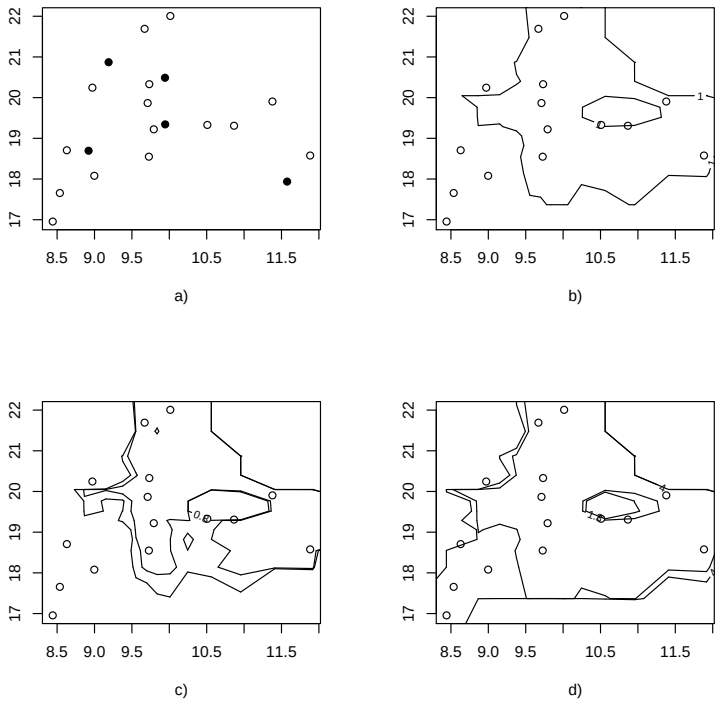

Figura 7.5: Resultados do Exemplo 7.7.2 com tamanho 20 e 10% de censura em cada variável.

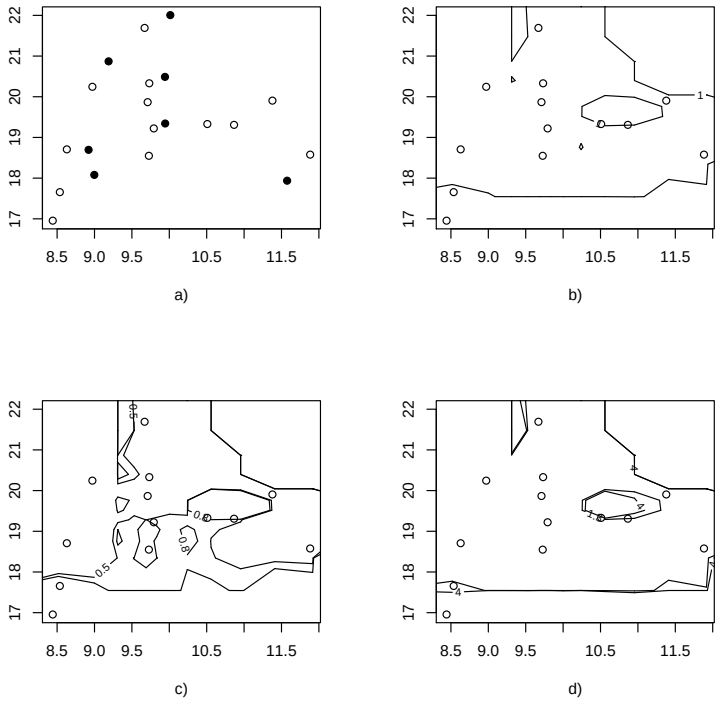

Figura 7.6: Resultados do Exemplo 7.7.2 com tamanho 20 e 15% de censura em cada variável.

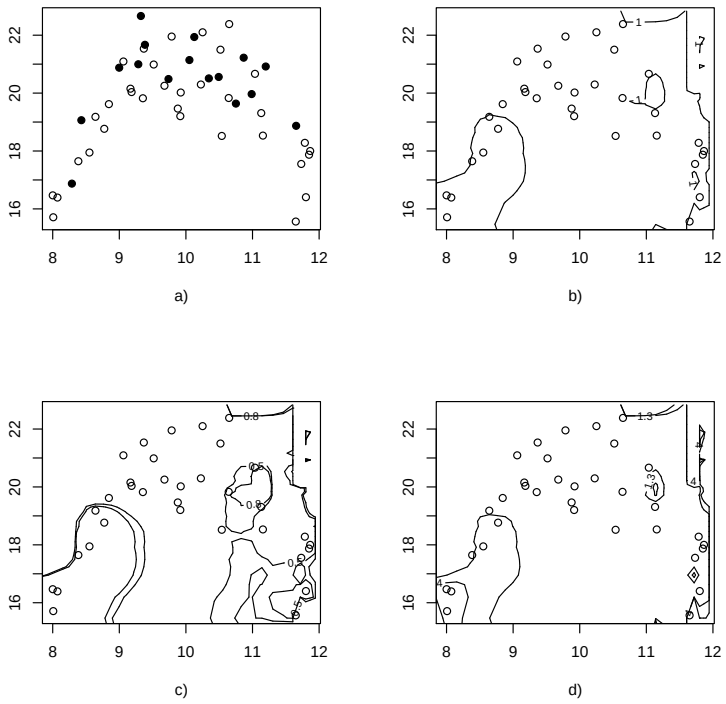

Figura 7.7: Resultados do Exemplo 7.7.2 com tamanho 50 e 10% de censura em cada variável.

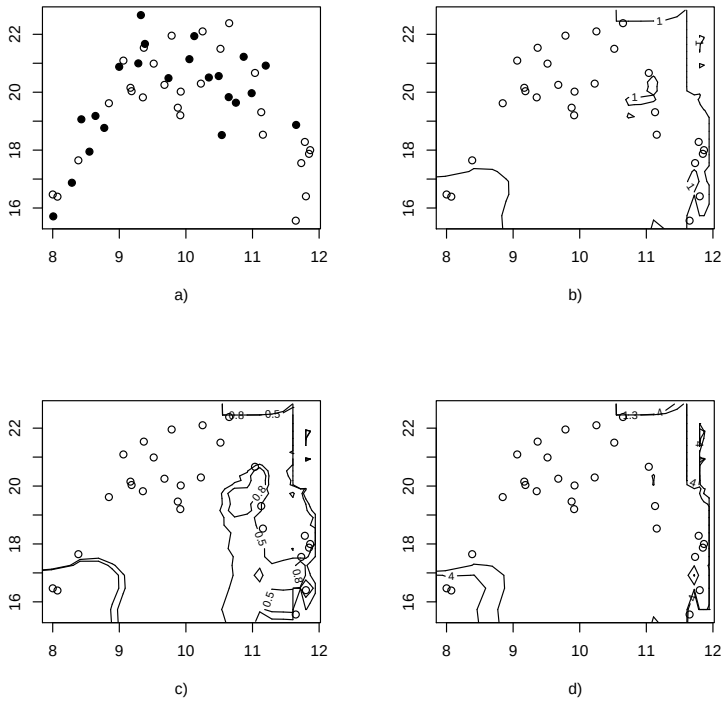

Figura 7.8: Resultados do Exemplo 7.7.2 com tamanho 50 e 15% de censura em cada variável.

Enquanto na Figura 7.5, a maioria dos pontos se localizam claramente em duas regiões de dependência local positiva e negativa, na Figura 7.6 a grande maioria dos pontos se localizam na região de dependência local negativa, especialmente por efeito da censura do ponto $(9.0, 18.1)$. Enquanto isso, os coeficientes de correlação linear r dos pares que não contém censura correspondentes são 0.29 e 0.27.

Na Figura 7.7, seis dos pontos de menores valores de x apresentam correlação local positiva, e a grande maioria dos pontos restantes apresentam correlação local negativa. O coeficiente de correlação linear r dos pares que não contém censura é de 0.00.

Na Figura 7.8, pela censura em alguns dos pares de menores valores de x, apenas os dois pontos de menores valores de x apresentam correlação local positiva, e quase todos os restantes apresentam correlação local negativa. O coeficiente de correlação linear r dos pares que não contém censura é de -0.16.

Nas Figuras 7.5, 7.6, 7.7 e 7.8 a maioria dos pontos apresenta correlação local negativa por causa da maior dispersão dos valores (x, y) gerados com correlação negativa.

Exemplo 7.7.3 Para uma amostra de pares gerados de tamanho $n = 20$, onde os dados são gerados com $X_i \sim N(5, 0.6)$ e $Y_i \sim N(5, 0.6)$, $i = 1, \ldots, 5$, $X_i \sim N(10, 0.6)$ e $Y_i \sim N(5, 0.6)$, $i = 6, \ldots, 10$, $X_i \sim N(5, 0.6)$ e $Y_i \sim N(10, 0.6)$, $i = 11, \ldots, 15$ e $X_i \sim N(10, 0.6)$ e $Y_i \sim N(10, 0.6)$, $i = 16, \ldots, 20$ e, atribuindo censuras uniformes de 10% em cada variável, o procedimento de estimação de θ resulta nos gráficos da Figura 7.9. Para a mesma amostra é atribuído censuras uniformes de 15% em cada variável e o procedimento de estimação de θ resulta nos gráficos da Figura 7.10. O mesmo procedimento é aplicado em uma amostra de tamanho $n = 50$, com censuras de 10% e 15% em cada variável, no qual os resultados se encontram nas Figuras 7.11 e 7.12, respectivamente.

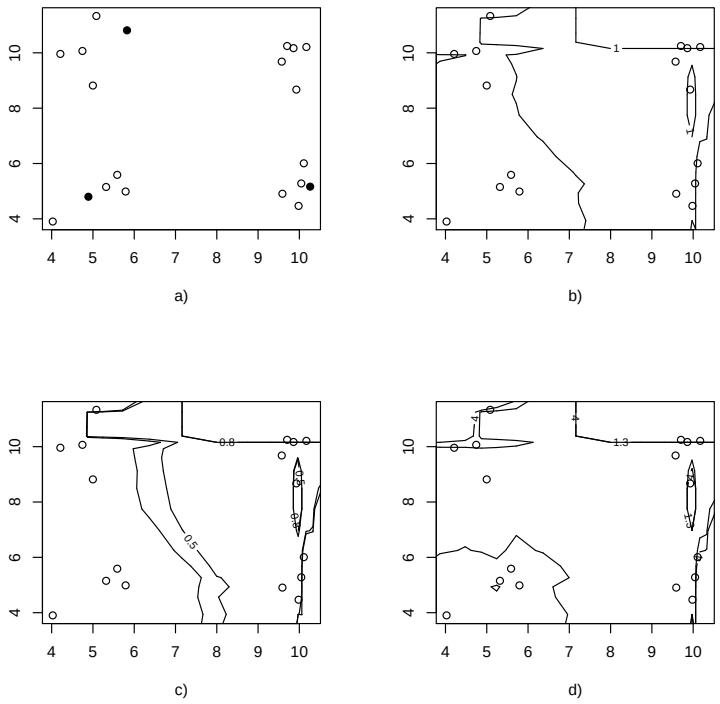

Figura 7.9: Resultados do Exemplo 7.7.3 com tamanho 20 e 10% de censura em cada variável.

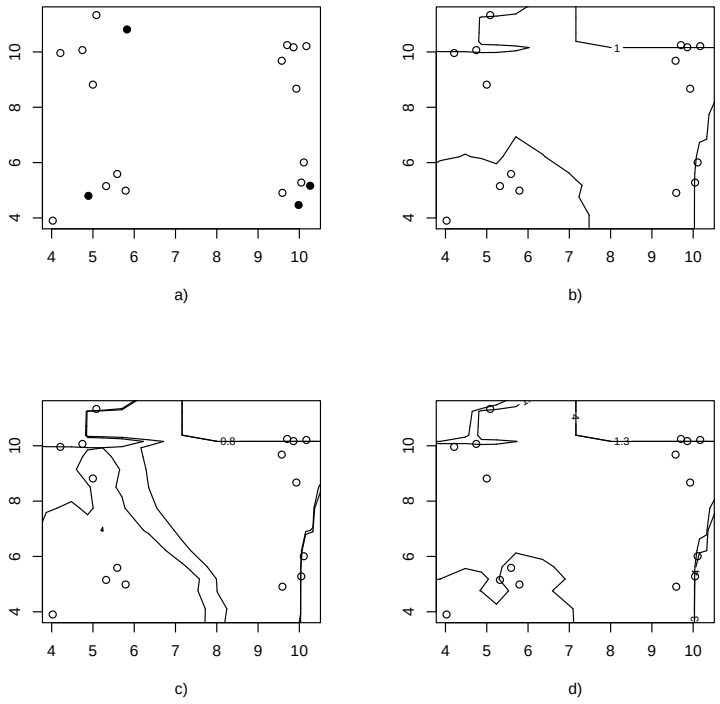

Figura 7.10: Resultados do Exemplo 7.7.3 com tamanho 20 e 15% de censura em cada variável.

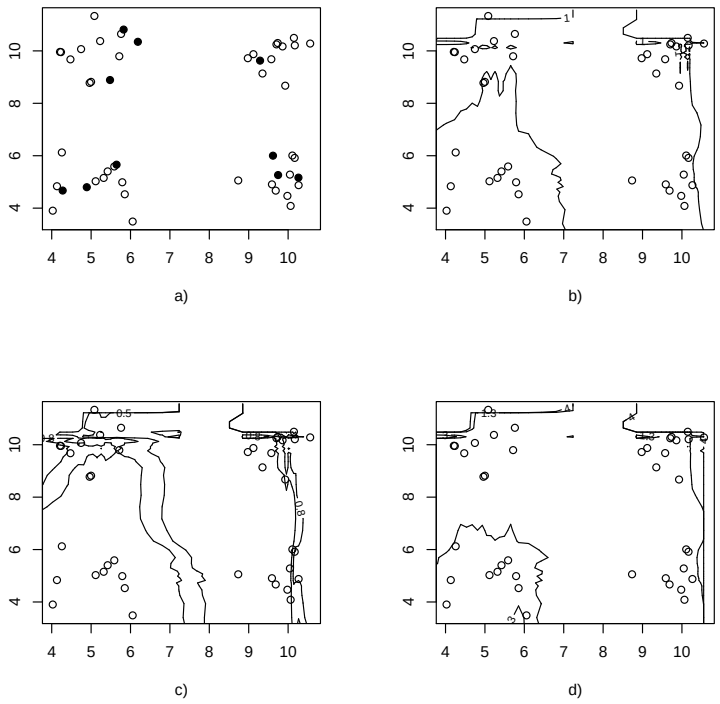

Figura 7.11: Resultados do Exemplo 7.7.3 com tamanho 50 e 10% de censura em cada variável.

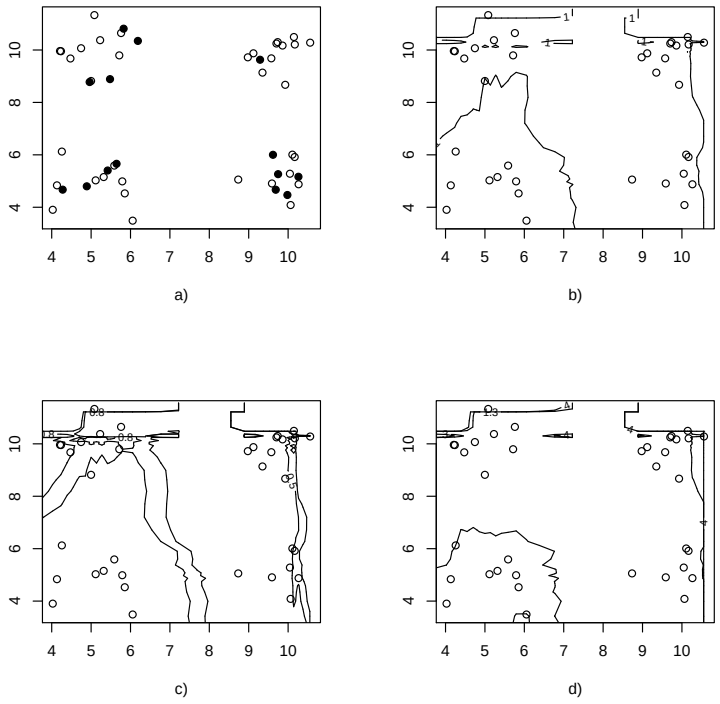

Figura 7.12: Resultados do Exemplo 7.7.3 com tamanho 50 e 15% de censura em cada variável.

Nas Figuras 7.9, 7.10, 7.11 e 7.12, apenas os pares de menores de valores (x, y) apresentam evidente correlação local positiva, para os pontos restantes pelas curvas de nível não se consegue estabelecer claramente o sinal da correlação local, o que ao final resulta coerente com a situação simulada. Ainda, os coeficiente de correlação linear r dos pares que não contém censura são muito pequenos.

Exemplo 7.7.4 Para os dados utilizados em Lin & Ying (1993) e Bae et al (2005), resumidos na Tabela 7.1, que consistem em 11 pares de valores da variáveis X: Tempo até a morte do enxerto de pele com acompanhamento rigoroso, e Y: Tempo até a morte do enxerto de pele com acompanhamento padrão. Na Figura 7.13 são apresentadas as curvas de nível de θ estimadas.

Tabela 7.1: Dados utilizados em Lin & Ying (1993).

i	1	2	3	4	5	6	7	8	9	10	11
x_i	37	19	57	93	16	22	20	18	63	29	60
δ_{1i}	1	1	0	1	1	1	1	1	1	1	0
y_i	29	13	15	26	11	17	26	21	43	15	40
δ_{2i}	1	1	1	1	1	1	1	1	1	1	1

Da Figura 7.13 observa-se que, exceto o ponto $(93, 26)$, da maior componente x, todos apresentam correlação local positiva, encontrando-se localizados acima da curva de nível de $\theta = 1$, ou seja, na direção de aumento da correlação positiva.

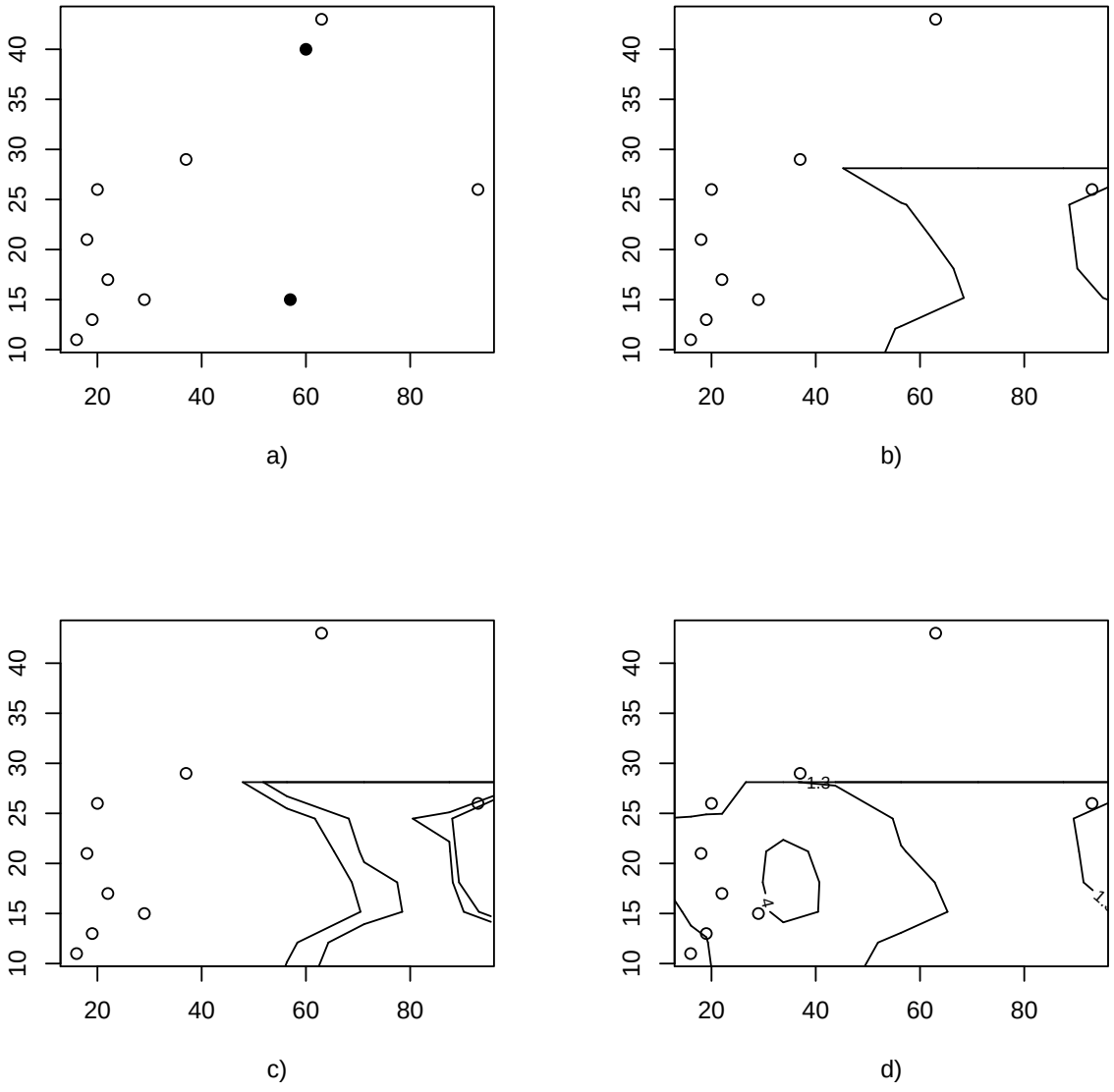

Figura 7.13: Estimativas do valor de θ da tendência da relação.

7.8 Conclusões

Se conhece pouca literatura sobre a análise de dependência, em dados bivariados com presença de censura. No entanto, a partir do coeficiente θ de dependência desenvolvido por Clayton, D. G. (1978), apresentamos aqui uma metodologia nova que permite analisar dependência local, pela análise das curvas de nível da superfície que se obtém para θ, utilizando métodos não paramétricos de suavização das funções da equação 7.2.7.

A determinação da correlação local, pela análise das curvas de nível de θ, se mostra coerente nos diversos exemplos apresentados neste capítulo, o que sugere a boa coerência dos procedimentos necessários para a estimação de θ e da metodologia aqui proposta.

Na pratica, os gráficos das curvas de nível mostram a correlação local coerentemente, quando ocorre o aumento do grau de censura de 10% para 15% dos casos em análise.

Com esta nova metodologia de análise de dependência, e com as informações adicionais de outras variáveis sobre os pacientes, podemos, por exemplo: i) determinar causas ou características comuns em indivíduos que pertençam à uma mesma região de dependência local, ii) determinar características comuns em indivíduos que apresentem valores semelhantes de θ e iii) separar grupos de indivíduos que apresentem valores semelhantes de θ, com vistas à outras análises estatísticas adicionais.

8 Considerações Finais

Três situações não tão comuns de dependência bivariada serviram como motivação inicial desta dissertação, as quais foram denominadas no Capítulo 5 como "Quadrática", "Exponencial" e "Normais". Para estas situações os coeficiente de correlação usuais τ de Kendall, ρ de Spearman e linear r de Pearson não são suficientes para descrever as possíveis estruturas de dependência geral, e menos ainda as estruturas de dependência local existentes.

Um dos procedimentos mais utilizados para análise de situações de dependência bivariada mais complexas, é a análise de Cópulas, a qual foi descrita no Capítulo 2 com abordagem não paramétrica e ilustrada, no Capítulo 5, com as situações de dependência que denominamos "Quadrática", "Exponenciais" e "Normais". Por ser um procedimento não paramétrico, a análise de dependência recai sobre as curvas de nível obtidas. Verificou-se da análise das curvas de nível das cópulas em alguns casos, como no exemplo das "Normais", que não identificam os agrupamentos existentes ou não apontam a dependência claramente, como no caso da "Quadrática".

Do estudo da Função de Sibuya, nos Capítulos 3 e 5, observamos que as curvas de nível da Função de Sibuya estimada são de fácil interpretação, permitindo a identificação das regiões de dependência local positiva ou negativa. Entretanto, o procedimento não têm coerência quando se trata de quantificar a magnitude da dependência, com base no distanciamento do valor da função de Sibuya em relação a 1.0: o valor da função de Sibuya para situação de independência entre as variáveis X e Y. Também verificamos incoerência do teste de Gonçalves (2008) por apresentar tendência de julgar amostras de variáveis independentes como sendo dependentes, mas observamos que o teste pode ser utilizado como ponto de partida na análise de dependência, para posteriormente classificar as regiões de acordo com as curvas de nível obtidas por meio de suavização.

Na implementação dos cálculos para obtenção da Cópula e da Função de Sibuya suavizadas, é utilizado a propriedade de que transformações crescentes em ambas as variáveis preservam a estrutura da Cópula e da Função de Sibuya, afim de obter um procedimento que não seja necessário a escolha de uma função kernel com características específicas.

Do estudo do Chi-plot, nos Capítulos 4 e 6, destacamos a riqueza de informação, assim como a facilidade de interpretação do gráfico dos pares (χ_i, λ_i), permitindo a identificação de regiões de dependência local, além de permitir a identificação de estruturas de dependência complexas. O valor de cada χ_i, que mede a dependência local, é a principal

fonte de informação à respeito da dependência local e está associado ao par (x_i, y_i).

Do estudo do Chi-plot, nos Capítulos 4 e 6, verificamos que as características esperadas do Chi-plot **C1-C5** sob independência das variáveis X e Y não são coerentes. Entretanto, após estabelecermos a distribuição assintótica de χ, $N(0, 1/(n-1))$, para a verificação de independência entre X e Y, pelo critério C1 reformulado e a avaliação da normalidade de χ por algum dos testes considerados, observamos maior coerência.

Desenvolvemos e apresentamos uma forma gráfica muito eficiente para identificar a dependência local, distinguindo cada ponto no gráfico de dispersão, segundo a posição do seu valor χ_i no Chi-plot.

No Capítulo 7 desenvolvemos e apresentamos um procedimento de análise de dependência bivariada com presença de censuras, que permite identificar regiões de dependência local, a partir da medida de correlação geral de Gumbel (1960). Observamos resultados coerentes sobre as distintas situações analisadas e a possibilidade de utilização na identificação de grupos com semelhanças quanto à magnitude da correlação e/ou sinal.

Para os estudos ao longo desta dissertação, sobre a análise de dependência bivariada por meio dos procedimentos de Cópulas, da função de Sibuya, do Chi-plot, e para dados com presença de censura, elaboramos e implementamos procedimentos específicos para cálculo e suavização de funções de densidade e de distribuição marginais e conjuntas, que apresentaram excelente desempenho.

9 Apêndice A

Nesta seção são apresentados os resultados de validação da análise por simulação dos critérios de validação dos gráficos Chi-plot para os casos de dependência linear, quadrática, exponencial e por agrupamento(normais) apresentados no Capítulo 5.

Tabela 9.1: Amostras que satisfazem os testes para C2 à C5 do Capítulo 5 Entre 300 Amostras

	Testes com α=5%				
	n	25% cada quadrante	50% de pontos $\lambda > 0$	50% de pontos $\chi > 0$	Uniformidade de λ
Linear	20	0	57	0	65
	50	0	1	0	1
	100	0	0	0	0
	200	0	0	0	0
Quadrática	20	236	259	225	274
	50	242	267	243	272
	100	247	269	238	280
	200	237	250	241	264
Normais	20	262	300	199	300
	50	206	300	178	300
	100	136	300	138	300
	200	88	300	97	300
Exponenciais	20	200	270	156	289
	50	77	254	71	262
	100	31	240	36	253
	200	7	185	5	206

A Tabela 9.2 classifica como independente as amostras cujos estratégias são feitas pelos testes determinados por:

A: O resultado em pelo menos 1 dos testes C2, C3 ou C4 é independente;

B: Os resultados dos testes C2, C3 ou C4 dão independência conjuntamente

C: Uniformidade de λ ou o resultado em pelo menos 1 dos testes C2, C3 ou C4 é independente,

onde os testes C2, C3 e C4 são denominados como no Capítulo 5, ou seja, pela referência:
C2: Cada um dos quatro quadrantes do chi-plot, determinados por $\chi = 0$ e $\lambda = 0$ contém

25% dos pares da amostra;

C3: 50% dos valores chi_i são maiores do que zero;

C4: 50% dos valores λ_i são maiores que zero;

Tabela 9.2: Classificados como Independentes dos Testes A,B e C

	n	A	B	C
Linear	20	37	0	32
	50	1	0	1
	100	0	0	0
	200	0	0	0
Quadrática	20	272	203	261
	50	279	220	268
	100	283	220	272
	200	272	218	257
Normais	20	300	161	300
	50	300	169	300
	100	300	114	300
	200	300	72	300
Exponenciais	20	300	197	300
	50	256	58	249
	100	242	26	234
	200	185	5	176

Tabela 9.3: Amostras Classificadas como Independentes Através dos IC das Equações (4.2.7) e (6.2.6)

	n	$c_p = 1,54$	$c_p = 1,78$	$c_p = 2.18$	$z_{1-\alpha/2} = 1,64$	$z_{1-\alpha/2} = 1,96$	$z_{1-\alpha/2} = 2,57$
Linear	20	0	0	0	0	0	1
	50	0	0	0	0	0	0
	100	0	0	0	0	0	0
	200	0	0	0	0	0	0
Quadrática	20	10	4	24	11	10	85
	50	0	0	0	0	0	0
	100	0	0	0	0	0	0
	200	0	0	0	0	0	0
Normais	20	178	177	224	222	209	275
	50	185	191	148	219	237	242
	100	192	167	113	226	219	204
	200	176	160	96	226	228	219
Exponenciais	20	116	107	179	164	141	248
	50	91	97	91	111	144	164
	100	67	63	40	92	97	109
	200	20	18	11	28	37	54

Tabela 9.4: Amostras Classificadas como Independentes Através dos IC das Equações (4.2.7) (6.2.6) e C4

	n	$c_p = 1,54$	$c_p = 1,78$	$c_p = 2,18$	$z_{1-\alpha/2} = 1,64$	$z_{1-\alpha/2} = 1,96$	$z_{1-\alpha/2} = 2,57$
Linear	20	0	0	0	0	0	1
	50	0	0	0	0	0	0
	100	0	0	0	0	0	0
	200	0	0	0	0	0	0
Quadrática	20	10	4	24	11	10	82
	50	0	0	0	0	0	0
	100	0	0	0	0	0	0
	200	0	0	0	0	0	0
Normais	20	178	177	224	222	209	275
	50	185	191	148	219	237	242
	100	192	167	113	226	219	204
	200	176	160	96	226	228	219
Exponenciais	20	116	106	178	163	140	242
	50	88	95	88	108	141	158
	100	67	63	40	91	96	107
	200	20	18	11	28	37	52

Tabela 9.5: Amostras Independentes por Algum Teste de $N(0, 1/(n-1))$

	n	10%	5%	1%
Linear	20	0	0	1
	50	0	0	0
	100	0	0	0
	200	0	0	0
Quadrática	20	164	197	237
	50	3	5	17
	100	0	0	0
	200	0	0	0
Normais	20	279	285	291
	50	246	259	276
	100	178	206	250
	200	153	185	215
Exponenciais	20	215	231	254
	50	130	140	166
	100	66	78	101
	200	15	18	23

Tabela 9.6: Amostras Independentes pelo IC das Equações (4.2.7) e (6.2.6) ou o Teste de $N(0, 1/(n-1))$

		Fisher & Switzer			Assintótico		
	n	10%	5%	1%	10%	5%	1%
Linear	20	0	0	1	0	0	1
	50	0	0	0	0	0	0
	100	0	0	0	0	0	0
	200	0	0	0	0	0	0
Quadrática	20	164	197	237	164	197	245
	50	3	5	17	3	5	17
	100	0	0	0	0	0	0
	200	0	0	0	0	0	0
Normais	20	281	287	293	282	290	295
	50	271	280	283	277	286	291
	100	248	248	262	256	261	274
	200	235	242	241	258	261	271
Exponenciais	20	222	237	260	231	239	277
	50	144	155	179	149	175	212
	100	92	97	105	106	116	138
	200	25	25	28	32	40	59

Tabela 9.7: Amostras Classificadas Como Independentes Pelos Testes Usuais

		p value $> 10\%$			$> 5\%$			$> 1\%$		
	n	$\tau = 0$	$\rho = 0$	$r = 0$	$\tau = 0$	$\rho = 0$	$r = 0$	$\tau = 0$	$\rho = 0$	$r = 0$
Linear	20	0	0	3	0	1	0	3	3	0
	50	0	0	0	0	0	0	0	0	0
	100	0	0	0	0	0	0	0	0	0
	200	0	0	0	0	0	0	0	0	0
Quadrática	20	247	241	290	267	260	256	293	290	289
	50	243	234	281	262	251	243	283	281	287
	100	253	243	282	268	264	261	290	282	294
	200	234	231	283	261	251	249	288	283	291
Normais	20	296	295	300	299	298	300	300	300	300
	50	295	295	300	300	300	300	300	300	300
	100	295	296	300	299	299	300	300	300	300
	200	299	299	300	299	299	300	300	300	300
Exponenciais	20	241	245	287	260	263	248	286	287	298
	50	189	189	273	226	222	222	272	273	289
	100	148	146	241	184	179	192	242	241	279
	200	79	77	186	112	112	143	183	186	246

10 Apêndice B

Neste Capítulo estão localizadas as rotinas de programação utilizadas para o Cálculo dos procedimentos mencionados neste trabalho.

10.1 Cópula

Função utilizada dentro da função copulasuavizada:

```
retvv=function(x,y=numeric(length(x)),z=numeric(length(x)),w){
tam=length(x)
rz=numeric(tam)
maisdeum=numeric(tam)
for(i in 1:tam){
for(j in 1:tam){if(x[i]==x[j] \& y[i]==y[j]) {rz[j]=rz[j]+z[i]
maisdeum[i]=maisdeum[i]+1}}}
for(i in 1:tam){rz[i]=rz[i]/maisdeum[i]}
i=1
while(i<=tam){
j=1
while(j<=tam){
if(i!=j) {if(rz[i]==rz[j]\&x[i]==x[j]\&y[i]==y[j]) { x=x[-j]
                                          y=y[-j]
                                          w=w[-i]
                                          rz=rz[-j]
                                         tam=length(x)
                                         } else j=j+1
} else j=j+1}
i=i+1}
tammatriz=length(x)
matriz=matrix(0,tammatriz,4)
matriz[,1]=x;
matriz[,2]=y;
matriz[,3]=rz;
matriz[,4]=w
return(matriz)
}
```

10.1.1 Função da Cópula suavizada

```
copulasuavizada = function(x,y){
n=length(x)
medx=mean(x)
medy=mean(y)
sdx=sd(x)
sdy=sd(y)
maxx=max(abs(x))
maxy=max(abs(y))
y=(y-medy)/sdy
x=(x-medx)/sdx
sortx=sort(x)
sorty=sort(y)
p=seq(0,1,0.025)
w=sortx[p*n]
w=append(w,c(sortx[1]-sdx/maxx,sortx[1]),0)
z=sorty[p*n]
z=append(z,c(sorty[1]-sdy/maxy,sorty[1]),0)
w=append(w,sortx[n]+sdx/maxx,length(w))
z=append(z,sorty[n]+sdy/maxy,length(z))
```

```
fx=numeric(n)
gy=numeric(n)
fazerw=length(w)
fazerz=length(z)
ux=numeric()
uy=numeric()
band=(sqrt(n))
for(j in 1:fazerw){for(i in 1:n){fx[i]=pnorm((w[j]-x[i])*band,0,0.3)}
ux[j]=(sum(fx)/n)}
for(k in 1:fazerz){
for(i in 1:n){gy[i]=pnorm((z[k]-y[i])*band,0,0.3)}
uy[k]= (sum(gy)/n)}
u=retvv(ux,w=w)
v=retvv(uy,w=z)
ux=u[,1]
uy=v[,1]
w=u[,4]
z=v[,4]
fazerw=length(ux)
fazerz=length(uy)
kxy=numeric(n)
fxy=matrix(0,fazerw,fazerz)
for(j in 1:fazerw){
for(k in 1:fazerz){
for(i in 1:n){kxy[i]=pnorm((w[j]-x[i])*band,0,0.3)*pnorm((z[k]-y[i])*band,0,0.3)}
fxy[j,k]= (sum(kxy)/n)}}
suppressWarnings(return("u"=ux,"v"=uy,"c"=fxy))
}
```

10.2 Sibuya

```
sibuyasuavizada = function(x,y){
n=length(x)
medx=mean(x)
medy=mean(y)
sdx=sd(x)
sdy=sd(y)
maxx=max(abs(x))
maxy=max(abs(y))
y=(y-medy)/sdy
x=(x-medx)/sdx
tamanhow=(max(x)-min(x))/40;
w=seq(min(x)-sdx/maxx,max(x)+sdx/maxx,tamanhow)
tamanhoz=(max(y)-min(y))/40;
z=seq(min(y)-sdy/maxy,max(y)+sdy/maxy,tamanhoz)
fxy=matrix(0,length(w),length(z))
band=sqrt(n)
kxy=numeric(n)
fx=numeric(n)
gy=numeric(n)
fazerw=length(w)
fazerz=length(z)
for(k in 1:fazerz){
for(j in 1:fazerw){
for(i in 1:n){
kxy[i]=pnorm((w[j]-x[i])*band,0,0.3)*pnorm((z[k]-y[i])*band,0,0.3);
fx[i]=pnorm((w[j]-x[i])*band,0,0.3)
gy[i]=pnorm((z[k]-y[i])*band,0,0.3)              }
fxy[j,k]=fxy[j,k] + (sum(kxy)/n)/(sum(fx)*sum(gy)/n^2)
minimo=max((sum(fx)/n+sum(gy)/n-1)/(sum(fx)*sum(gy)/n^2),0)
maximo=min(1/(sum(fx)/n),1/(sum(gy)/n))
if(fxy[j,k]<minimo) fxy[j,k]=minimo
if(fxy[j,k]>maximo) fxy[j,k]=maximo
kxy=numeric(n)
fx=numeric(n)
gy=numeric(n)
```

```
}}
w=sdx*w+medx
z=sdy*z+medy
x=sdx*x+medx
y=sdy*y+medy
return("w"=w,"z"=z,"s"=fxy)
}
```

10.3 Teste sibuya

```
sibuyateste=function(x,y){
n=length(x);
f=numeric(n)
g=numeric(n)
for(j in 1:n){
for(i in 1:n){
   if (x[i]<=x[j])  f[j]=(f[j]+1);
   if (y[i]<=y[j])  g[j]=(g[j]+1);
}}
hfg=numeric(n)
for(k in 1:n){
for(j in 1:n){
if (x[j]<=x[k]\&y[j]<=y[k]) hfg[k]= hfg[k]+1;
}}
f=f/(n);
g=g/(n);
hfg=hfg/(n);
simpirica=numeric(n)
for(i in 1:n){
simpirica[i]=hfg[i]/(f[i]*g[i])
}
mediasoma=sum(simpirica)/(n);
xsquare=sum((simpirica-mediasoma)^ 2)/((n-1));
zn=(mediasoma-1)*sqrt(n)/sqrt(xsquare);
intervalo=numeric(2);
intervalo[1]=mediasoma-1.96*sqrt(xsquare)/sqrt(n);
intervalo[2]=mediasoma+1.96*sqrt(xsquare)/sqrt(n);
print("O intervalo de Confiança para 95\% é:");
print(intervalo);
print("valor z  supostamente de uma N(0,1)")
print(zn);
print("Valor médio de Sibuya")
print(mediasoma)
}
```

10.4 Chi-plot

```
chiplot=function(x,y){
n=length(x)
f1=numeric(n)
g1=numeric(n)
ch=numeric(n)
chi=numeric(n)
xsort=sort(x)
ysort=sort(y)
yysort=y[order(x)]
for(i in 1:n){
j=1
k=1
cont1=-1
cont2=-1
menorx=T
menory=T
while(menorx){ \
```

```
if(xsort[j]<=x[i]){ cont1=cont1+1
j=j+1
if(j>n) menorx=F } else menorx=F
           }
while(menory){
if(ysort[k]<=y[i]){ cont2=cont2+1
k=k+1
if(k>n) menory=F} else menory=F
           }
f1[i]=cont1
g1[i]=cont2
j=1
k=1
cont=-1
menor2=T
while(menor2){
if(xsort[k]<=x[i]){k=k+1
if(k>n) menor2=F} else menor2=F
           }
k=k-1
menor3=j<=k
while(menor3){ if(yysort[j]<=y[i]) {cont=cont+1}
                                    j=j+1
                                    if(j>k) menor3=F}
ch[i]=cont}
f1=f1/(n-1)
g1=g1/(n-1)
hfg=ch/(n-1)
f2=numeric()
g2=numeric()
for(i in 1:n){
f2[i]=f1[i]-0.5
g2[i]=g1[i]-0.5
chi[i]=(hfg[i]-f1[i]*g1[i])/(sqrt(f1[i]*(1-f1[i])*g1[i]*(1-g1[i])))
               }
lambda=numeric(n)
sinal=numeric(n)
maxi=numeric(n)
for(i in 1:n){
if (f2[i]*g2[i]<0) sinal[i]=-1 else sinal[i]=1
if ((f2[i])^2<(g2[i])^2) maxi[i]=(g2[i])^2 else maxi[i]=(f2[i])^2
lambda[i]=4*sinal[i]*maxi[i]
}
return("x"=x,"y"=y,"l"=lambda,"chi"=chi)
}
```

10.5 Caso com Censura

```
#Utilizados na Construção da densidade bivariada
N=function(s,t,xa,ya){
n=length(xa)
cont=0
for(i in 1:n) if(xa[i]>s&ya[i]>t) cont=cont+1
return(cont)
}

gama=function(s,t,xa,ya,cy,j){ if(xa[j]>s&ya[j]<=t&cy[j]==0) return(1) else return(0) }

G=function(pf,sf,xa,ya,cx,cy){
n=length(xa)
f1=1
f2=1
for(i in 1:n) f1=f1*(N(xa[i],0,xa,ya)/(N(xa[i],0,xa,ya)+1))^(1-cx[i])
for(i in 1:n) f2=f2*(N(sf,ya[i],xa,ya)/(N(sf,ya[i],xa,ya)+1))^(gama(pf,sf,xa,ya,cy,i))
return(f1*f2)
}
```

```
I=function(pt,st){ if(pt>st) return(1) else return(0) }

####densidade bivariada
densidade.censurada=function(xa,ya,cx,cy,s,t){
n=length(xa)

fxy=matrix(0,length(s),length(t))
band1=1/bw.nrd0(xa)
band2=1/bw.nrd0(ya)

AG1=matrix(,n,n)
for(l in 1:n){
for(i in 1:n){if(i==1) AG1[i,l]=0 else AG1[i,l]= G(xa[i-1],ya[l],xa,ya,cx,cy) }}

AG2=matrix(0,n,n)
for(l in 1:n){
for(i in 1:n){AG2[i,l]= G(xa[i],ya[l],xa,ya,cx,cy) }}

kxy=matrix(0,2,n)
fazers=length(s)
fazert=length(t)
for(k in 1:fazert){
for(j in 1:fazers){
for(l in 1:n){
for(i in 1:n){
kxy[1,i]=dnorm((s[j]-xa[i])*band1)*dnorm((t[k]-ya[l])*band2)*cy[l]*(I(l,i-1)/(AG1[i,l]+1)-I(l,i)/(AG2[i,l]+1))
            }
fxy[j,k]=fxy[j,k]+(sum(kxy)/n)*(band1*band2)
            }
                 }
                 }
return(fxy)
}

#######contruçao densidades univariadas
KMu=function(xa,cx){
n=length(xa)
sobrev=numeric(n+1)

for(i in 2:n){
termo=1
cxorderx=cx[order(xa)]
for(j in 1:(i-1)){
termo=((n-(j))/(n-(j)+1))^(cxorderx[j])*termo
}
sobrev[i]=termo
}
sobrev[1]=1
s=numeric(n)
for(i in 1:(n-1)) s[i]=sobrev[i]-sobrev[i+1]
s[n]=sobrev[n]
suppressWarnings(return(s,"KM"=sobrev))
}

##Densidade suavizada univariada
dsu=function(x,ss=ss$s,s){
orderx=sort(x)
band=1/bw.nrd0(x)
n=length(orderx)
fxy=numeric(length(s))
kxy=numeric(n)
for(j in 1:length(s)){
for(i in 1:n){
kxy[i]=ss[i]*dnorm((s[j]-orderx[i])*band)
            }
```

```
fxy[j]=sum(kxy)*band
            }
return(fxy)
}


###Kapla Meier Bivariado
Sobrev=function(s,t,xa,ya,cx,cy){
n=length(xa)
cont1=0
for(i in 1:n) if(xa[i]>=s&ya[i]>=t) cont1=cont1+1
cont1=cont1/n
nmax=numeric(n)

for(i in 1:n){
cont2=0
for(j in 1:n){ if (max(xa[j],ya[j])>=max(xa[i],ya[i])) cont2=cont2+1
nmax[i]=cont2+1
            }
            }
prod1=1
for(i in 1:n) if (max(xa[i],ya[i])<=max(s,t)) prod1=prod1*(1-(1-cx[i]*cy[i])/nmax[i])
res=cont1/prod1
return(res)
            }

KMvalor=function(x,xa,cx){
xortxa=sort(xa)
nx=length(xa)
pos=1
for(j in 1:(nx-1)) if (x>xortxa[j]&x<=xortxa[j+1]) pos=j+1
if(x>xortxa[nx]) pos=nx+1
valor=KMu(xa,cx)$KM[pos]
return(valor)
}


EB=function(xa,ya,cx,cy){

n=length(xa)
xas=xa
yas=ya
cxs=cx
cys=cy

tcensurado=T
k=1
while(tcensurado){
                      if(cx[k]==0||cy[k]==0) { xa=xa[-k]
                                              ya=ya[-k]
                                              cx=cx[-k]
                                              cy=cy[-k]}  else k=k+1
                if(k>=length(xa)) tcensurado=F
                }
cy=cy[order(ya)]
ya=ya[order(ya)]

cx=cx[order(xa)]
xa=xa[order(xa)]

xa=append(xa,0,0)
xa=append(xa,(1+1/n)*max(xa),length(xa))
nx=length(xa)

ya=append(ya,0,0)
ya=append(ya,(1+1/n)*max(ya),length(ya))
```

```
ny=length(ya)

cx=append(cx,1,0)
cx=append(cx,1,length(cx))
cy=append(cy,1,0)
cy=append(cy,1,length(cy))

nx=length(xa)
ny=length(ya)

rsobrev=matrix(,nx,ny)
for(i in 0:(nx-1)){
for(j in 0:(ny-1)){
rsobrev[i+1,j+1]=Sobrev(xa[i+1],ya[j+1],xas,yas,cxs,cys)
                  }
                  }

su=s
tu=t

KMS=matrix(0,length(su),length(tu))
eixox=numeric(length(su))
eixoy=numeric(length(tu))

for(k in 1:length(su)){
soma1=0
for(i in 0:(nx-1)){
soma1=soma1+xa[i+1]*choose(nx-1,i)*su[k]^(i)*(1-su[k])^(nx-1-(i))
}
eixox[k]=soma1
}

for(k in 1:length(tu)){
soma2=0
for(i in 0:(ny-1)){
soma2=soma2+ya[i+1]*choose(ny-1,i)*tu[k]^(i)*(1-tu[k])^(ny-1-(i))
}
eixoy[k]=soma2
}

for(k in 1:length(su)){
for(m in 1:length(tu)){
soma=0

for(i in 0:(nx-1)){
for(j in 0:(ny-1)){
soma=soma+rsobrev[i+1,j+1]*choose(nx-1,i)*choose(ny-1,j)*su[k]^(i)*(1-su[k])^(nx-1-(i))*tu[m]^(j)*(1-tu[m])^(ny-1-(j))
}
}
KMS[k,m]=soma
}
}
KMS[1,1]=1
suppressWarnings(return("x"=eixox,"y"=eixoy,KMS))
}

#kapla meie suavizado com as escolhas nos pontos s
Kmsuave=function(s,xa,spulo){

ns=length(s)
nx=length(xa)

resultado=numeric(ns)
soma=numeric(nx)
band=1/bw.nrd0(xa)
for(i in 2:ns){
```

```
for(k in 1:nx){
soma[k]=spulo[k]*(pnorm((s[i]-xa[k])*band))
               }
resultado[i]=1-sum(soma)
               }
resultado[1]=1
return(resultado)
}

#função risco acumulado dado maior que t0
hazard.pt.dado.st.maior.st0=function(kmx,kmy,xa,ya,cx){
mdux=matrix(,length(kmy),length(kmx))
lengt=length(kmy)
for(k in 1:lengt){
xfazer=numeric()
cfazer=numeric()
for(j in 1:length(ya)) if(ya[j]>kmy[k]) {xfazer=append(xfazer,xa[j],0)
                                         cfazer=append(cfazer,cx[j],0)}
if(length(xfazer)>1) {k.m=KMu(xfazer,cfazer)
                      res=dsu(xfazer,k.m$s,kmx)
                      res2=Kmsuave(kmx,xfazer,k.m$s)}
minKMS=min(res2[res2!=0])/2
for(v in 1:length(res2)){
if(res2[v]==0)  res2[v]=minKMS }
mdux[k,]=res/res2
}
return(mdux)
}

#estimador de theta
est.theta=function(xa,cx,ya,cy,print=T){

KMS=EB(xa,ya,cx,cy)

fxy=densidade.censurada(xa,ya,cx,cy,KMS$x,KMS$y)

ht0=hazard.pt.dado.st.maior.st0(KMS$x,KMS$y,xa,ya,cx)
hs0=hazard.pt.dado.st.maior.st0(KMS$y,KMS$x,ya,xa,cy)

minKMS=min(KMS$KMS[KMS$KMS!=0])/5
if(minKMS>0.05) minKMS=0.0005
for(k in 1:length(KMS$x)){
for(j in 1:length(KMS$y)){
if(KMS$KMS[k,j]==0)  KMS$KMS[k,j]=minKMS }}
RESULTADO1=fxy/KMS$KMS

minhs0=min(hs0[hs0!=0])/5
if(minhs0>0.05) minhs0=0.0005
for(k in 1:length(KMS$x)){
for(j in 1:length(KMS$y)){if(hs0[k,j]==0)  hs0[k,j]=minhs0 }}

minht0=min(ht0[ht0!=0])/5
if(minht0>0.05) minht0=0.0005
for(k in 1:length(KMS$x)){
for(j in 1:length(KMS$y)){if(ht0[j,k]==0)  ht0[j,k]=minht0 }}

RESULTADO=RESULTADO1/(t(ht0)*hs0)

maxR=max(RESULTADO[RESULTADO!="Inf"])

for(k in 1:length(KMS$x)){
for(j in 1:length(KMS$y)){if(RESULTADO[k,j]=="Inf")  RESULTADO[k,j]=maxR }}

if(print){
par(mfrow=c(2,2))
plot(xa,ya,type="n",xlab="Scatter Plot",ylab="")
```

```
for(i in 1:length(cx)){
if(cx[i]==1&cy[i]==1) points(xa[i],ya[i]) else  points(xa[i],ya[i],pch=19)
}

contour(KMS$x,KMS$y,RESULTADO,levels=1,xlim=c(min(xa),max(xa)),ylim=c(min(ya),max(ya)),xlab="=1",ylab="")
for(i in 1:length(cx)){
if(cx[i]==1&cy[i]==1) points(xa[i],ya[i])
}

contour(KMS$x,KMS$y,RESULTADO,levels=c(0.8,0.5),xlim=c(min(xa),max(xa)),ylim=c(min(ya),max(ya)),xlab="=0.5, 0.8",ylab="")
for(i in 1:length(cx)){
if(cx[i]==1&cy[i]==1) points(xa[i],ya[i])
}

contour(KMS$x,KMS$y,RESULTADO,levels=c(1.3,4),xlim=c(min(xa),max(xa)),ylim=c(min(ya),max(ya)),xlab="=1.3, 4",ylab="")
for(i in 1:length(cx)){
if(cx[i]==1&cy[i]==1) points(xa[i],ya[i])
} }
suppressWarnings(return(KMS$x,KMS$y,RESULTADO))
}

#dados do exemplo real 7.7.4
xa=c(37, 19, 57, 93 ,16 ,22 ,20 ,18 ,63 ,29 ,60)
cx=c(1,1,0,1,1,1,1,1,1,1,0)
ya=c(29 ,13, 15, 26, 11, 17, 26, 21, 43, 15, 40)
cy=c(rep(1,11))
est.theta=(xa,cx,ya,cy)
```

11 Referências Bibliográficas

Bae W. Choi H. Park, B. U., Choongrak K. (2005). Smoothing techniques for the bivariate kaplan-meier estimator. *Communications in Statistics - Theory and Methods*, **34**(7), 1659-1674.

Berg A. & Politis D. N. (2007). Density estimation of censored data with infinite-order kernels. Preprint.

Berg A. & Politis D. N. (2009). Cdf and survival function estimation with infinite-order kernels. *Electronic Journal of Statistics*, **3**, 1436-1454.

Charpentier A., Fermanian J.-D. and Scaillet O. (2007). The estimation of copulas: theory and practice, in J. Rank (ed.), *Copulas: From theory to application in Finance*, Risk Publications, London, chapter Section 2.

Genest C. & Boies J. C. (2003). Detecting dependence with kendall plots. The American Statistician, 57(4), 275-284.

Clayton D. G. (1978). A model for Association in Bivariate Life Tables and Its Aplication in Epidemiological Studies of Family Tendency in Chronic Disease Incidence. *Biometrika*, **1**(65), 141-151.

Conover W. *Practical nonparametric statistics*. Wiley Series in Probability and Mathematical Statistics, second edition. 1980.

Fermanian J. D., Radulovic D., Wegkamp M. (2004). Weak convergence of empirical copula process. *Bernoulli*, **5**(10), 847- 860.

Fisher N. & Switzer P. (1985). Chi-Plots for assessing dependence. *Biometrika*, **2**(72), 253-265.

Fisher N. & Switzer P. (2001). Graphical Assessment of Dependence:Is a Picture Worth 100 Tests?. *The American Statistician*, 55, 233-239.

Gonçalves M. (2008). *Um Estudo Sobre Funções de Dependência e Medida de Risco*. Tese de doutorado, IME - Universidade de São Paulo, São Paulo.

Govindarajulu Z. *Nonparametric Inference*. University of Kentucky, USA. 2007.

Gumbel E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55, 698-707.

Heinz G., Peterson L. J., Johnson R. W., Kerk C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education*, **11**(2).

Kaplan E. L. & Meier, P. (1958). Nonparametric Estimation From Incomplete Observations. *Journal of the American Statistical Association*, 53, 457-481.

Kim C., Park B. U., Lim C. (2003). Bezier Curve Smoothing of The Kaplan-Meier Estimator. *Annals of the Institute of Statistical Mathematics*, **2**(55), 359-367.

Lin D. Y. & Ying Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, 80, 573-581.

Luc D. *Non-Uniform Random Variate Generation.* New York: Springer-Verlag, 1986, 27-82.

Morettin P. A., Toloi C. M. C., Chiann C. and de Miranda J. C. S. (2008). Wavelet Smoothed Empirical Copula Estimators. Preprint.

Memória J. M. P. *Breve História da Estatística.* Embrapa Informação Tecnológica. Brasília. 2004.

Nelsen R. B. *An Introduction to Copulas.* Springer: New York, second edition, 2006.

Nelsen R. B. Properties and applications of copulas: a brief survey. *Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance*, (Dhaene, J., Kolev, N., Morettin, P.A. (Eds.)), University Press USP: São Paulo, 10-28.

Splent P. & Smeeton N. C. (2001). *Applied Nonparametric Statistical Methods.* Chapman & Hall, third edition.

Ulisses U., Flavio H., Nikolai V. e Beatriz V. (2004). *Modelando Dependências via Cópulas.* Departamento de Estatística, IME - USP e IM - UFRJ.

Wells M. T. & Yeo K. P. (1996). Density Estimation With Bivariate Censored Data. *Journal of the American Statistical Association*, **91**(436), 1566-1574.