4
Métodos de Estimação da Entropia e da Informação Mútua

A estimação da entropia e da informação mútua, envolvendo apenas variáveis aleatórias discretas, é simples, com aplicação direta da definição de Shannon. No entanto, quando uma das variáveis envolvidas é contínua, torna-se necessária a aplicação de um método de estimação de densidade. Um dos mais simples e amplamente utlizado é o Histograma. Este método posteriormente descrito será, doravante, referenciado por Método Shannon/Histograma. O segundo método, apresentado como alternativa, está baseado na entropia quadrática de Rényi, aliada ao famoso método de estimação de densidade Janela de Parzen, e na definição da Informação Mútua de Cauchy-Schwartz, tornando os cálculos diretos, sem necessidade de um passo de pré-processamento. Este método será, doravante, referenciado por Método Cauchy-Schwartz/Parzen.

Em ambos os métodos, as variáveis contínuas são normalizadas no intervalo [-1, 1], por razões numéricas na aplicação de Redes Neurais (com fins, neste caso, de classificação). O algoritmo de normalização aplicado é o seguinte:

$$vn = \frac{2\left(v - \min v\right)}{\max v - \min v} - 1 \tag{4.1}$$

onde *v* representa a variável original e *vn*, a normalizada.

Os tipos de cálculo necessários, genericamente, ao algoritmo MIFS-U são os seguintes:

- Entropia de uma variável discreta (EntD),
- Entropia de uma variável contínua (EntC),
- Informação mútua entre variáveis discretas (IM-DD),
- Informação mútua entre variáveis contínuas (IM-CC), e
- Informação mútua entre uma variável discreta e outra contínua (IM-DC).

4.1
Método Shannon / Histograma

No caso de variáveis contínuas, para evitar adotar um modelo paramétrico para a densidade desconhecida, um recurso freqüente é a aplicação de métodos não-paramétricos de estimação de densidade. O mais antigo e amplamente utilizado é o histograma (Silverman, 1986). Neste estudo, considera-se, verdadeiramente, o histograma de freqüências relativas, não o histograma de densidade em si, em que a única diferença é que o segundo integra a 1 (Scott, 1992).

Como todas as variáveis contínuas são normalizadas no intervalo [-1, 1], o intervalo é simplesmente dividido em 20 subintervalos de igual amplitude (h = 0,1). Cada subintervalo é interpretado como uma classe, e a respectiva freqüência relativa apurada é, então, tomada como a probabilidade. Ou seja, é realizada uma discretização – a variável contínua passa a ser tratada como discreta – não oferecendo, pois, mais nenhum óbice ao cálculo das informações necessárias, podendo-se, facilmente, aplicar a definição da entropia de Shannon, de uso dominante na literatura

Assim, a função de probabilidade da variável contínua discretizada é dada por

$$\hat{f}_x(c_j) = \frac{1}{n}\left(qtde\ de\ amostras\ \ pertencentes\ à\ classe\ c_j\right) \quad , j = 1,...,20. \tag{4.2}$$

Com intuito de manter uma nomenclatura harmônica, representar-se-á, aqui, simplesmente por x uma determinada classe de uma variável contínua discretizada ou um valor (distinto) de uma variável discreta (e por $\mathbb{X}$ o conjunto de tais valores ou classes), não sendo, pois, mais necessária tal distinção.

Assim, pode-se escrever

$$\hat{f}_x(x) = \frac{1}{n}\sum_{i=1}^{n}\xi(x_i, x) \quad , \forall x \in \mathbb{X}.$$

onde $\xi_x(x_i)$ é a Função Indicadora, ou seja,

$$\xi(x_i, x) = \begin{cases} 1 & , se\ x_i \in x\left(classe\right) \\ 0 & , caso\ contrário \end{cases} \qquad ou \qquad \xi(x_i, x) = \begin{cases} 1 & , se\ x_i = x\left(valor\right) \\ 0 & , caso\ contrário \end{cases} \tag{4.3}$$

(Variável Contínua Discretizada) (Variável Discreta)

Para a distribuição conjunta de duas variáveis (discretas ou contínuas discretizadas), tem-se o seguinte:

$$\hat{f}_{xy}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \xi_x(x_i)\xi_y(y_i) \quad , \forall x \in \mathbb{X}, \forall y \in \mathbb{Y}. \tag{4.4}$$

Portanto, considerando a discretização das variáveis contínuas, como uma passo de pré-processamento dos dados, os cálculos necessários ao algoritmo MIFS-U, passam a ser tão-somente os seguintes:

- Entropia de uma variável discreta (EntD),
- Informação mútua entre variáveis discretas (IM-DD),

EntD – Entropia de Shannon de uma variável discreta

$$\hat{H}(X) = -\sum_{x \in \mathbb{X}} \hat{f}_x(x) \log \hat{f}_x(x) \tag{4.5}$$

IM-DD – Informação Mútua (de Shannon) entre duas variáveis discretas

$$\hat{I}(X;Y) = \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \hat{f}_{xy}(x,y) \log \frac{\hat{f}_{xy}(x,y)}{\hat{f}_x(x)\hat{f}_y(y)} \tag{4.6}$$

4.2 Método Cauchy-Schwartz / Parzen-Rosenblatt

No contexto de seleção de variáveis de sistemas não-lineares, estimar a informação mútua entre variáveis diretamente dos dados, onde pelo menos uma delas é contínua, sem a necessidade de fazer hipóteses acerca da distribuição a *priori* dos dados, tem vital importância prática. Isto pode ser alcançado com a utilização da divergência de Cauchy-Schwartz, a qual é uma substituta da divergência de Kullback-Leibler, integrada à Janela de Parzen.

A divergência de Kullback-Leibler, baseada no conceito de entropia de Shannon, é, por sua simplicidade, uma medida usual de informação mútua entre duas variáveis aleatórias. Entretanto, nem esta, nem a equivalente para a entropia de Rényi podem ser integradas à Janela de Parzen (Príncipe et al.,1998).

Xu et al. (1998) apresentaram um método que integra a Janela de Parzen à Divergência de Cauchy-Schwartz, para estimar a informação mútua diretamente dos dados.

Portanto, no caso de variáveis contínuas, é utilizado, aqui, o método não-paramétrico de estimação de densidade de Parzen-Rosenblatt, ou, tão-simplesmente, Janela de Parzen, a seguir apresentado.

4.2.1
Estimador de Densidade Janela de Parzen

Segundo Scott (1992), dado um conjunto de amostras de f_x $\{x_1, x_2, ..., x_n\}$, o Estimador Núcleo – ou Estimador Janela de Parzen – pode ser escrito, compactamente como

$$\hat{f}_x(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i, h) \qquad (4.7)$$

onde $K_h(t) = K(t,h) = K(t/h)/h$ e $h = h(n) > 0$ é a largura da janela ou parâmetro de suavizamento.

Assim, $\hat{f}_x(x)$ pode ser vista como uma "média de curvas" centradas nas observações amostrais. Uma observação não deve representar apenas a si mesma, mas também sua vizinhança.

Comumente, define-se a Função Núcleo $K(\cdot)$ como não-negativa[3], com integral unitária, ou seja, como uma função densidade de probabilidade, e , ainda, simétrica e unimodal (Silverman, 1986).

Adotar-se-á a Função Núcleo Gaussiana, por uso posterior nesta dissertação, a seguir definida:

$$G(w,\phi) = (2\pi\phi)^{-\frac{1}{2}} \exp\left(-\frac{\omega^2}{2\phi}\right) \quad \text{, ou seja, } G(w,\phi) \sim N(0,\phi) \qquad (4.8)$$

onde $\phi = h^2 = \sigma^2$.

A escolha da largura da janela afeta muito mais o processo de estimação da densidade do que a escolha da Função Núcleo, não sendo esta, assim, crucialmente importante em questão de eficiência (Scott, 1992). Entretanto a

[3] Esta não é uma restrição necessária

Função Gaussiana tem uma propriedade que será extremamente vantajosa no contexto da dissertação.

Utilizando esta Função Núcleo, segue da definição que $\hat{f}_x$ será ela própria uma densidade de probabilidade. Além disso, $\hat{f}_x$ herda todas as propriedades de continuidade e diferenciabilidade de *G* (Silverman, 1986).

Mostra-se que o estimador de densidade Janela de Parzen é assintoticamente não-viesado e consistente, desde que a largura da janela *h*(*n*) tenda a zero a uma razão suficientemente pequena, quando *n* tende a infinito (Silverman, 1986; Scott, 1992; Wand & Jones, 1995).

A Figura 4.1 (Silverman, 1986) ilustra o efeito da largura da janela. O limite quando *h* tende a zero é (de certo modo) a soma de funções delta de Dirac posicionadas nas observações, já quando *h* torna-se grande, todo detalhe, espúrio ou não, é obscurecido.

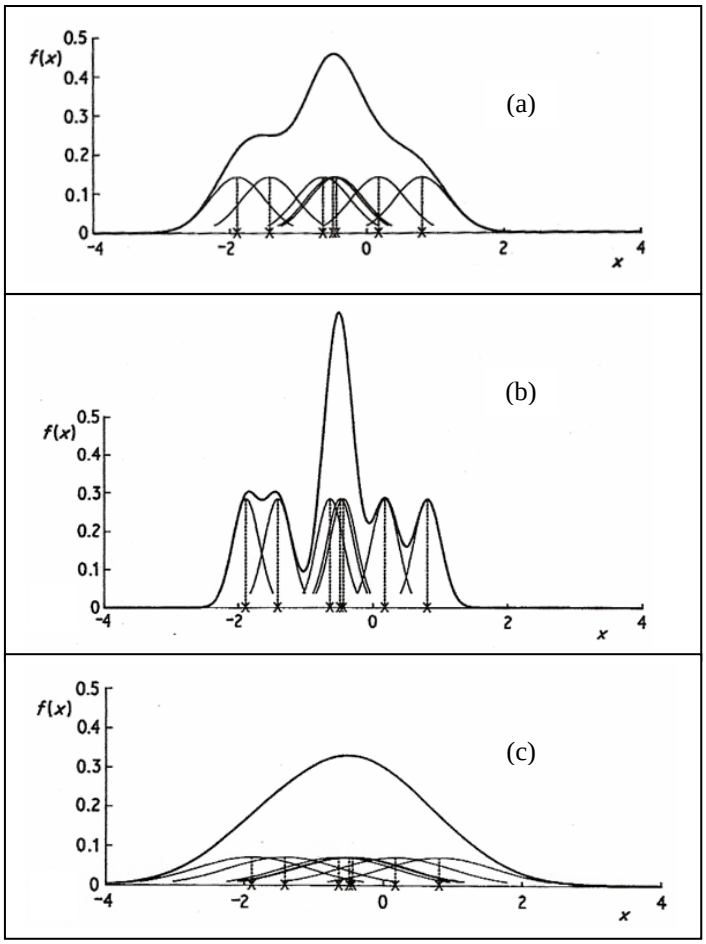

Figura 4.1 – Estimador Núcleo: efeito da largura da janela
(a) 0.4 (b) 0.2 (c) 0.8

No caso de amostra finita, a largura da janela tem de ser escolhida, através de um intercâmbio (*trade-off*) entre viés e variância, como será visto a seguir.

4.2.2
Escolha da Largura da Janela

Seja K uma função núcleo satisfazendo

- $\int K(t)dt = 1$
- $\int tK(t)dt = 0$ (4.9)
- $\int t^2 K(t)dt = m_2 \neq 0$

Segundo Silverman (1986), várias medidas têm sido estudadas quanto à proximidade do estimador de densidade $\hat{f}$ em relação à verdadeira densidade *f*. Quando se considera a estimação em um simples ponto, uma medida natural é o Erro Quadrático Médio (EQM) definido por

$$EQM\left(\hat{f}_x(x)\right) = E\left(\left(\hat{f}_x(x) - f_x(x)\right)^2\right) \tag{4.10}$$

O qual pode ser decomposto em

$$EQM\left(\hat{f}_x(x)\right) = \left[E\left(\hat{f}_x(x)\right) - f_x(x)\right]^2 + Var\left(\hat{f}_x(x)\right) \tag{4.11}$$

Entretanto a utilidade encontra-se em estimar f_x de forma global, não em um ponto fixo. Assim, a medida utilizada, que é , sem dúvida, a mais amplamente usada e de fácil trato, é o Erro Quadrático Integrado Médio (*EQIM*), o qual é expresso por

$$EQIM\left(\hat{f}_x(x)\right) = \int\left[E\left(\hat{f}_x(x)\right) - f_x(x)\right]^2 dx + \int Var\left(\hat{f}_x(x)\right)dx \tag{4.12}$$

O termo do viés e o termo da variância na Eq. (4.12) dependem da largura da janela de diferentes formas. Isto é melhor ilustrado analisando tal equação assintoticamente, isto é, para grandes amostras *n*.

Sob determinadas condições (veja, por exemplo, Silverman, 1986, seção 3.3), as quais fogem ao escopo desta dissertação, o *EQIM* é assintoticamente aproximado pelo Erro Quadrático Integrado Médio Assintótico (*EQIMA*) – bastante conhecido por sua sigla em inglês: *Asymptotic Mean Integrated Squared Error* (*AMISE*), que é dado por

$$EQIMA\left(\hat{f}(x)\right) = \frac{1}{nh}\int K^2(t)\,dt + \frac{1}{4}h^4 m_2^2 \int \left(f^{"}(x)\right)^2 dx \qquad (4.13)$$

Pode-se observar que a minimização do termo do viés, segunda parcela da Eq. (4.13), é obtida pela minimização de *h*; entretanto, o termo da variância, primeira parcela, é minimizado pela maximização de *h*. Existe, portanto, um inerente *trade-off* viés-variância associado com o método Janela de Parzen para estimação da densidade.

Existem vários métodos de seleção da largura da janela *h*, cada um tendo suas propriedades (Wand & Jones, 1995). Um desses métodos utiliza a Eq. (4.13) para seleção da largura da janela, obtendo-se uma expressão para o *h* ótimo, diferenciando-se tal equação e igualando-se a zero.

$$h_{ot} = k_2^{-2/5}\left(\int K^2(t)\,dt\right)^{1/5}\left(\int\left(f^{"}(x)\right)^2 dx\right)^{-1/5} n^{-1/5} \qquad (4.14)$$

Uma forma direta de se estimar $\int\left(f^{"}(x)\right)^2 dx$ é assumir que a densidade desconhecida é uma densidade normal com variância σ_x^2 (Silverman, 1986). Seja φ tal densidade.

$$\int\left(f^{"}(x)\right)^2 dx = \sigma^{-5}\int\left(\varphi^{"}(x)\right)^2 dx = \frac{3}{8}\pi^{-1/2}\sigma^{-5} \approx 0,212\,\sigma^{-5} \qquad (4.15)$$

Assim, se uma Função Núcleo Gaussiana é utilizada, então a largura da janela obtida da eq. (4.14), substituindo o valor (4.15), será dada por

$$h_{ot} = \left(4\pi\right)^{-1/10} \left(\tfrac{3}{8}\pi^{-1/2}\right)^{-1/5} \sigma \; n^{-1/5} = \left(\tfrac{4}{3}\right)^{1/5} \sigma \; n^{-1/5} = 1,06\,\sigma \; n^{-1/5} \qquad (4.16)$$

Este método é conhecido como “Regra de Referência Normal”.

O desvio padrão σ pode ser estimado, a partir dos dados, pelo desvio padrão amostral s, ou utilizar uma medida robusta, tal como o intervalo interquartil, e, neste caso, a expressão de h_{ot}, torna-se então

$$h_{ot} = 0,79\, I_Q\, n^{-1/5} \;\; , \text{ onde } \;\; I_Q = Q_3 - Q_1 \qquad (4.17)$$

Nesta dissertação, utilizou-se uma solução de compromisso entre os dois estimadores de dispersão, semelhante à forma apresentada por Silverman (1986):

$$\hat{h}_{ot} = 0,9 \min(s, \hat{I}_q)\, n^{-1/5} \qquad (4.18)$$

No caso multivariado, a Regra de Referência Normal, torna-se (Silverman, 1986; Scott, 1992):

$$h_{ot}^{(i)} = \left(\frac{4}{d+2}\right)^{1/(d+4)} \sigma_i \, n^{-1/(d+4)} \text{ , onde } d = \text{dimensão e } i = 1,...,d. \qquad (4.19)$$

Enfocando-se apenas o caso bivariado, de interesse desta dissertação, e adotando-se uma única largura de janela para ambas as variáveis, tem-se

$$h_{ot} = \sigma\, n^{-1/6} \text{ , onde } \sigma = \left(\frac{\sigma_1 + \sigma_2}{2}\right)^{1/2} \qquad (4.20)$$

Reiterando as mesmas considerações quanto à estimação do desvio padrão no caso univariado, nesta dissertação, adotou-se o seguinte:

$$h_{ot} \approx 0,85 \min\left(\left(\frac{s_1^2 + s_2^2}{2} \right)^{-\frac{1}{2}}, \frac{\hat{I}_q^{(1)} + \hat{I}_q^{(2)}}{2} \right) n^{-\frac{1}{6}} \tag{4.21}$$

4.2.3
Cálculos Necessários ao MIFS-U

Os tipos de cálculo necessários ao algoritmo MIFS-U são apresentados a seguir, utilizando-se, indistintamente, a notação h_{R_2} na representação da entropia de Rényi, seja ela diferencial ou não.

EntD – Entropia de Rényi de uma variável discreta

$$\hat{h}_{R_2}(X) = -\log \sum_{x \in \mathbb{X}} \hat{f}_x^2(x) \tag{4.22}$$

EntC – Entropia de Rényi de uma variável contínua

Estimação da densidade através da Função Núcleo Gaussiana

$$\hat{f}_x(x) = \frac{1}{n}\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2) \right) \tag{4.23}$$

Logo,

$$\begin{aligned} \hat{f}_x^2(x) &= \left[\frac{1}{n}\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2) \right) \right]^2 = \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \sum_{j=1}^{n} G(x - x_i, \sigma^2) G(x - x_j, \sigma^2) \right) \end{aligned} \tag{4.24}$$

E, aplicando-se a propriedade da integração do produto de núcleos Gaussianos (vide anexo), tem-se

$$\int_{\mathbb{X}} \hat{f}_x^2(x)dx = \frac{1}{n^2}\int_{\mathbb{X}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} G(x-x_i,\sigma^2)\,G(x-x_j,\sigma^2)\right)dx =$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{X}} G(x-x_i,\sigma^2)\,G(x-x_j,\sigma^2)\,dx = \quad (4.25)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i-x_j,2\sigma^2)$$

Esta propriedade facilita avaliar a entropia quadrática de Rényi, a qual é uma função do quadrado da função de densidade de probabilidade.

Assim, a estimação da entropia, é expressa por

$$\hat{h}_{R_2}(X) = -\log\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i-x_j,2\sigma^2)\right] \quad (4.26)$$

Portanto, a entropia quadrática de Rényi pode ser avaliada como uma soma de interações locais definidas pelo núcleo sobre todos os pares de amostras.

IM-DD – Informação Mútua (de Cauchy-Schwartz) entre duas variáveis discretas

$$I_{CS}(X;Y) = h_{R_2}(f_{XY}\times f_X\, f_Y) - \frac{1}{2}h_{R_2}(f_{XY}) - \frac{1}{2}h_{R_2}(f_X\, f_Y) \quad (4.27)$$

onde

$$\hat{h}_{R_2}(\hat{f}_{xy}\times\hat{f}_x\,\hat{f}_y) = -\log\sum_{\mathbb{Y}}\sum_{\mathbb{X}}\hat{f}_{xy}(x,y)\,\hat{f}_x(x)\,\hat{f}_y(y) \quad (4.28)$$

$$\hat{h}_{R_2}(\hat{f}_{xy}) = -\log\sum_{\mathbb{Y}}\sum_{\mathbb{X}}\hat{f}_{xy}^2(x,y) \quad (4.29)$$

$$\hat{h}_{R_2}(\hat{f}_x\,\hat{f}_y) = -\log\sum_{\mathbb{Y}}\sum_{\mathbb{X}}\hat{f}_x^2(x)\,\hat{f}_y^2(y) \quad (4.30)$$

IM-CC – Informação Mútua (de Cauchy-Schwartz) entre duas variáveis contínuas

Como visto, a entropia de uma única variável é facilmente avaliada como interações entre pares de amostras. Este conceito será agora estendido à informação mútua entre variáveis.

Estimando-se as densidades através da Função Núcleo Gaussiana, tem-se

$$\hat{f}_x(x) = \frac{1}{n}\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2)\right) \tag{4.31}$$

$$\hat{f}_y(y) = \frac{1}{n}\left(\sum_{k=1}^{n} G(y - y_k, \sigma^2)\right) \tag{4.32}$$

$$\begin{aligned} \hat{f}_{xy}(x,y) &= \frac{1}{n}\sum_{i=1}^{n} G((x,y) - (x_i, y_i), \sigma^2 I_2) = \\ &= \frac{1}{n}\sum_{i=1}^{n} G(x - x_i, \sigma^2) G(y - y_i, \sigma^2) \end{aligned} \tag{4.33}$$

onde I_2 é a matriz identidade de ordem 2.

Logo,

$$\begin{aligned} \hat{f}_x^2(x) &= \left[\frac{1}{n}\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2)\right)\right]^2 = \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} G(x - x_i, \sigma^2) G(x - x_j, \sigma^2)\right) \end{aligned} \tag{4.34}$$

$$\begin{aligned} \hat{f}_y^2(y) &= \left[\frac{1}{n}\left(\sum_{j=1}^{n} G(y - y_k, \sigma^2)\right)\right]^2 = \\ &= \frac{1}{n^2}\left(\sum_{k=1}^{n}\sum_{l=1}^{n} G(y - y_k, \sigma^2) G(y - y_l, \sigma^2)\right) \end{aligned} \tag{4.35}$$

$$\hat{f}_{xy}^{2}(x,y) = \left[\frac{1}{n}\sum_{i=1}^{n} G(x-x_i,\sigma^2)\,G(y-y_i,\sigma^2)\right]^2 = \\ = \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} G(x-x_i,\sigma^2)\,G(y-y_i,\sigma^2)\,G(x-x_j,\sigma^2)\;G(y-y_j,\sigma^2)\right) \tag{4.36}$$

E, aplicando-se a propriedade da integração do produto de núcleos Gaussianos (vide anexo), tem-se

$$\int_{\mathbb{X}} \hat{f}_x^2(x)dx = \frac{1}{n^2}\int_{\mathbb{X}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} G(x-x_i,\sigma^2)\,G(x-x_j,\sigma^2)\right)dx = \\ = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{X}} G(x-x_i,\sigma^2)\,G(x-x_j,\sigma^2)\,dx = \\ = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i-x_j,2\sigma^2) \tag{4.37}$$

$$\int_{\mathbb{Y}} \hat{f}_y^2(y)dy = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} G(y_k-y_l,2\sigma^2) \quad \text{(analogamente)} \tag{4.38}$$

$$\int_{\mathbb{Y}}\int_{\mathbb{X}} \hat{f}_{xy}^2(x,y)\,dx\,dy = \\ = \frac{1}{n^2}\int_{\mathbb{Y}}\int_{\mathbb{X}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} G(x-x_i,\sigma^2)\,G(y-y_i,\sigma^2)\,G(x-x_j,\sigma^2)\;G(y-y_j,\sigma^2)\right)dx\,dy \\ = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\int_{\mathbb{X}} G(x-x_i,\sigma^2)\,G(x-x_j,\sigma^2)\,dx\right)\left(\int_{\mathbb{Y}} G(y-y_i,\sigma^2)\,G(y-y_j,\sigma^2)\,dy\right) = \\ = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i-x_j,2\sigma^2)\,G(y_i-y_j,2\sigma^2) \tag{4.39}$$

Assim, como já visto,

$$I_{CS}(X;Y) = h_{R_2}(f_{XY}\times f_X\,f_Y) - \frac{1}{2}h_{R_2}(f_{XY}) - \frac{1}{2}h_{R_2}(f_X\,f_Y) \tag{4.40}$$

onde

$$\hat{h}_{R_2}(\hat{f}_{xy} \times \hat{f}_x \hat{f}_y) = -\log \int_{\mathbb{Y}} \int_{\mathbb{X}} \hat{f}_{xy}(x,y)\, \hat{f}_x(x)\, \hat{f}_y(y) dx\, dy =$$

$$= -\log \int_{\mathbb{Y}} \int_{\mathbb{X}} \hat{f}_{xy}(x,y) \hat{g}_{xy}(x,y) dx\, dy =$$

$$= -\log \left[\int_{\mathbb{Y}} \int_{\mathbb{X}} \left(\frac{1}{n} \sum_{i=1}^{n} G(x - x_i, \sigma^2) G(y - y_i, \sigma^2) \right) \right.$$

$$\left. \left(\frac{1}{n} \left(\sum_{j=1}^{n} G(x - x_j, \sigma^2) \right) \right) \left(\frac{1}{n} \left(\sum_{l=1}^{n} G(y - y_l, \sigma^2) \right) \right) dx\, dy \right] =$$

$$= -\log \left[\frac{1}{n^3} \int_{\mathbb{Y}} \int_{\mathbb{X}} \left(\sum_{i=1}^{n} G(x - x_i, \sigma^2) G(y - y_i, \sigma^2) \right) \right.$$

$$\left. \left(\sum_{j=1}^{n} G(x - x_j, \sigma^2) \right) \left(\sum_{l=1}^{n} G(y - y_l, \sigma^2) \right) dx\, dy \right] =$$

$$= -\log \left\{ \frac{1}{n^3} \int_{\mathbb{Y}} \int_{\mathbb{X}} \sum_{i=1}^{n} \left[G(x - x_i, \sigma^2) \left(\sum_{j=1}^{n} G(x - x_j, \sigma^2) \right) \right. \right.$$

$$\left. \left. G(y - y_i, \sigma^2) \left(\sum_{l=1}^{n} G(y - y_l, \sigma^2) \right) \right] dx\, dy \right\} =$$

$$= -\log \left\{ \frac{1}{n^3} \int_{\mathbb{Y}} \int_{\mathbb{X}} \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{n} G(x - x_i, \sigma^2) G(x - x_j, \sigma^2) \right) \right. \right.$$

$$\left. \left. \left(\sum_{l=1}^{n} G(y - y_i, \sigma^2) G(y - y_l, \sigma^2) \right) \right] dx\, dy \right\} =$$

$$= -\log \left\{ \frac{1}{n^3} \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{n} \int_{\mathbb{X}} G(x - x_i, \sigma^2) G(x - x_j, \sigma^2) dx \right) \right. \right.$$

$$\left. \left. \left(\sum_{l=1}^{n} \int_{\mathbb{Y}} G(y - y_i, \sigma^2) G(y - y_l, \sigma^2) dy \right) \right] \right. =$$

$$= -\log \left\{ \frac{1}{n^3} \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2) \right) \left(\sum_{l=1}^{n} G(y_i - y_l, 2\sigma^2) \right) \right] \right.$$

$$\hat{h}_{R_2}(\hat{f}_{xy} \times \hat{f}_x \hat{f}_y) = -\log \int_{\mathbb{Y}} \int_{\mathbb{X}} \hat{f}_{xy}(x,y)\, \hat{f}_x(x)\, \hat{f}_y(y) dx\, dy =$$

$$= -\log \left\{ \frac{1}{n^3} \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2) \right) \left(\sum_{l=1}^{n} G(y_i - y_l, 2\sigma^2) \right) \right] \right\} \tag{4.41}$$

$$\hat{h}_{R_2}(\hat{f}_{xy}) = -\log \int_{\mathbb{Y}} \int_{\mathbb{X}} \hat{f}_{xy}^2(x,y) dx\, dy =$$

$$= -\log \left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2) G(y_i - y_j, 2\sigma^2) \right] \tag{4.42}$$

$$\hat{h}_{R_2}(\hat{f}_x \hat{f}_y) = -\log \int_{\mathbb{Y}} \int_{\mathbb{X}} \hat{f}_x^2(x)\, \hat{f}_y^2(y) dx\, dy = -\log\left[\left(\int_{\mathbb{X}} \hat{f}_x^2(x) dx\right)\left(\int_{\mathbb{Y}} \hat{f}_y^2(y)\, dy\right)\right] =$$

$$= -\log\left[\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2)\right)\left(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} G(y_k - y_l, 2\sigma^2)\right)\right] =$$

$$= -\log\left\{\frac{1}{n^4}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2)\right]\left[\sum_{k=1}^{n}\sum_{l=1}^{n} G(y_k - y_l, 2\sigma^2)\right]\right\}$$

$$\hat{h}_{R_2}(\hat{f}_x \hat{f}_y) = -\log \int_{\mathbb{Y}} \int_{\mathbb{X}} \hat{f}_x^2(x)\, \hat{f}_y^2(y) dx\, dy =$$

$$= -\log\left\{\frac{1}{n^4}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2)\right]\left[\sum_{k=1}^{n}\sum_{l=1}^{n} G(y_k - y_l, 2\sigma^2)\right]\right\} \quad (4.43)$$

IM-DC – Informação Mútua (de Cauchy-Schwartz)
entre uma variável discreta (*Y*) e outra contínua contínua (*X*)

Sejam

v = número de valores distintos da v.a.d. Y na amostra.

y_p = p-ésimo valor distinto da da v.a.d. Y na amostra.

n_p = número de amostras de X relacionadas ao valor y_p da v.a.d Y.

Aqui são utilizadas duas notações diferentes para as amostras de *X*. Uma amostra é identificada com um único subscrito x_i $(1 \le i \le n)$, quando a determinação do valor de *Y* a ela relacionado é irrelevante. Quando for relevante, x_{ps} indica a amostra de *X*, com índice $1 \le s \le n_p$, relacionada ao valor y_p $(1 \le p \le v)$ de *Y*.

Estimando-se as densidades, através da Função Núcleo Gaussiana, no caso da variável contínua

$$\hat{f}_y(y_p) = \frac{n_p}{n} \qquad \sum_{p=1}^{v} n_p = n \quad (4.44)$$

$$\hat{f}_x(x) = \frac{1}{n}\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2)\right) \tag{4.45}$$

$$\hat{f}_{x|y}(x \mid y_p) = \frac{1}{n_p}\sum_{s=1}^{n_p} G(x - x_{ps}, \sigma^2)\,,\ p = 1,...,v \tag{4.46}$$

$$\begin{aligned}\hat{f}_{xy}(x, y_p) &= \hat{f}_y(y_p)\,\hat{f}_{x|y}(x \mid y_p) \\ &= \left(\frac{n_p}{n}\right)\left(\frac{1}{n_p}\sum_{s=1}^{n_p} G(x - x_{ps}, \sigma^2)\right) = \\ &= \frac{1}{n}\sum_{s=1}^{n_p} G(x - x_{ps}, \sigma^2)\,, \quad p = 1,...,v\end{aligned} \tag{4.47}$$

E a estimação das entropias, empregando a propriedade da integração do produto de núcleos Gaussianos (vide anexo), são expressas por

$$\begin{aligned}\hat{h}_{R_2}(\hat{f}_{xy} \times \hat{f}_x\,\hat{f}_y) &= -\log\sum_{p=1}^{v}\int_{\mathbb{X}} \hat{f}_{xy}(x, y_p)\,\hat{f}_x(x)\,\hat{f}_y(y_p)\,dx = \\ &= -\log\sum_{p=1}^{v}\left[\,\hat{f}_y(y_p)\int_{\mathbb{X}} \hat{f}_{xy}(x, y_p)\,\hat{f}_x(x)\,dx\right] = \\ &= -\log\sum_{p=1}^{v}\left[\left(\frac{n_p}{n}\right)\int_{\mathbb{X}}\left(\frac{1}{n}\sum_{s=1}^{n_p} G(x - x_{ps}, \sigma^2)\right)\left(\frac{1}{n}\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2)\right)\right)dx\right] = \\ &= -\log\frac{1}{n^3}\sum_{p=1}^{v}\left[n_p\int_{\mathbb{X}}\left(\sum_{s=1}^{n_p} G(x - x_{ps}, \sigma^2)\right)\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2)\right)dx\right] = \\ &= -\log\frac{1}{n^3}\sum_{p=1}^{v}\left[n_p\sum_{s=1}^{n_p}\sum_{i=1}^{n}\int_{\mathbb{X}} G(x - x_{ps}, \sigma^2)\,G(x - x_i, \sigma^2)dx\right] = \\ &= -\log\frac{1}{n^3}\sum_{p=1}^{v}\left[n_p\sum_{s=1}^{n_p}\sum_{i=1}^{n} G(x_{ps} - x_i, 2\sigma^2)\right]\end{aligned}$$

$$\begin{aligned}\hat{h}_{R_2}(\hat{f}_{xy} \times \hat{f}_x\,\hat{f}_y) &= -\log\sum_{p=1}^{v}\int_{\mathbb{X}} \hat{f}_{xy}(x, y_p)\,\hat{f}_x(x)\,\hat{f}_y(y_p)\,dx = \\ &= -\log\frac{1}{n^3}\sum_{p=1}^{v}\left[n_p\sum_{s=1}^{n_p}\sum_{i=1}^{n} G(x_{ps} - x_i, 2\sigma^2)\right]\end{aligned} \tag{4.48}$$

$$\hat{h}_{R_2}(\hat{f}_{xy}) = -\log \sum_{p=1}^{v} \int_{\mathbb{Y}} \hat{f}_{xy}^2(x_i, y)\, dy =$$

$$= -\log \sum_{p=1}^{v} \int_{\mathbb{X}} \left(\frac{1}{n} \sum_{s=1}^{n_p} G(x - x_{ps}, \sigma^2) \right)^2 dx =$$

$$= -\log \frac{1}{n^2} \sum_{p=1}^{v} \int_{\mathbb{X}} \sum_{s=1}^{n_p} \sum_{t=1}^{n_p} G(x - x_{ps}, \sigma^2)\, G(x - x_{pt}, \sigma^2)\, dx =$$

$$= -\log \frac{1}{n^2} \sum_{p=1}^{v} \sum_{s=1}^{n_p} \sum_{t=1}^{n_p} \int_{\mathbb{X}} G(x - x_{ps}, \sigma^2)\, G(x - x_{pt}, \sigma^2) dx =$$

$$= -\log \frac{1}{n^2} \sum_{p=1}^{v} \sum_{s=1}^{n_p} \sum_{t=1}^{n_p} G(x_{ps} - x_{pt}, 2\sigma^2)$$

$$\hat{h}_{R_2}(\hat{f}_{xy}) = -\log \sum_{p=1}^{v} \int_{\mathbb{Y}} \hat{f}_{xy}^2(x_i, y)\, dy = \\ = -\log \frac{1}{n^2} \sum_{p=1}^{v} \sum_{s=1}^{n_p} \sum_{t=1}^{n_p} G(x_{ps} - x_{pt}, 2\sigma^2) \tag{4.49}$$

$$\hat{h}_{R_2}(\hat{f}_x \hat{f}_y) = -\log \sum_{p=1}^{v} \int_{\mathbb{X}} \hat{f}_x^2(x)\, \hat{f}_y^2(y_p)\, dx = -\log \left[\left(\sum_{p=1}^{v} \hat{f}_y^2(y_p) \right) \left(\int_{\mathbb{X}} \hat{f}_x^2(x)\, dx \right) \right] =$$

$$= -\log \left[\left(\sum_{p=1}^{v} \left(\frac{n_p}{n} \right)^2 \right) \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2) \right) \right] =$$

$$= -\log \left[\frac{1}{n^4} \left[\left(\sum_{p=1}^{v} n_p^2 \right) \left(\sum_{i=1}^{n} \sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2) \right) \right] \right]$$

$$\hat{h}_{R_2}(\hat{f}_x \hat{f}_y) = -\log \sum_{p=1}^{v} \int_{\mathbb{X}} \hat{f}_x^2(x)\, \hat{f}_y^2(y_p)\, dx = \\ = -\log \left[\frac{1}{n^4} \left[\left(\sum_{p=1}^{v} n_p^2 \right) \left(\sum_{i=1}^{n} \sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2) \right) \right] \right] \tag{4.50}$$

Apesar dos somatórios triplos nas três equações anteriores, estes são somente somas duplas sobre as amostras.

Por fim, basta levar os resultados na equação seguinte:

$$I_{CS}(X;Y) = h_{R_2}(f_{XY} \times f_X\, f_Y) - \frac{1}{2} h_{R_2}(f_{XY}) - \frac{1}{2} h_{R_2}(f_X\, f_Y) \tag{4.51}$$