Aprendizado Baseado na Teoria da Informação

1. Visão geral

O campo de pesquisa conhecido como aprendizado baseado na teoria da informação (ITL, do inglês *information-theoretic learning*) (PRINCIPE, 2010) oferece um conjunto de critérios e algoritmos de adaptação de modelos para a solução de tarefas de aprendizado supervisionado e não-supervisionado, os quais são capazes de prover uma extração mais efetiva do conteúdo estatístico dos dados ao explorarem medidas derivadas da teoria da informação, como entropia e informação mútua (COVER & THOMAS, 2006).

2. Fundamentos de ITL

De modo geral, as medidas de informação exploradas em ITL dependem das funções densidade de probabilidade (PDFs, do inglês *probability density functions*) dos sinais disponíveis.

No entanto, uma vez que, em muitos cenários, não se tem conhecimento sobre estas PDFs, é necessário lançar mão de técnicas que realizem uma estimação diretamente a partir de dados amostrados. Neste contexto, a fim de não impor *a priori* um modelo para a PDF, a preferência em ITL é pelo uso de métodos não-paramétricos de estimação. Em especial, destacamos o método da janela de Parzen.

2.1. Estimador não-paramétrico de densidade

Considere que temos à disposição N amostras independentes e identicamente distribuídas (i.i.d.) $\{x_i\}_{i=0}^{N-1}$. A estimativa de Parzen da PDF da variável aleatória X, utilizando uma função *kernel* arbitrária, é dada por:

$$\hat{p}_X(x) = \frac{1}{N}\sum_{i=0}^{N-1} \kappa_{\sigma_\kappa}(x - x_i) = \frac{1}{N\sigma_\kappa}\sum_{i=0}^{N-1} \kappa\left(\frac{x - x_i}{\sigma_\kappa}\right), \quad (1)$$

onde $\kappa_{\sigma_\kappa}(x) = \frac{1}{\sigma_\kappa}\kappa\left(\frac{x}{\sigma_\kappa}\right)$, $\kappa(x)$ é uma função *kernel* (contínua e simétrica) e σ_κ é um parâmetro de suavização conhecido como largura do *kernel* (em inglês, *kernel size* ou *bandwidth*).

A função $\kappa(x)$ deve satisfazer às condições definidas pelo teorema de Mercer e, além disso, obedecer às seguintes propriedades para que $\hat{p}_X(x)$ seja uma PDF válida:

1) $\kappa(x) \geq 0$.

2) $\int_{\mathbb{R}} \kappa(x)dx = 1$.

3) $\lim_{x\to\infty}|x\kappa(x)| = 0$.

Na literatura, existem vários exemplos de funções *kernel*, tais como: uniforme, triangular, Epanechnikov e Gaussiana.

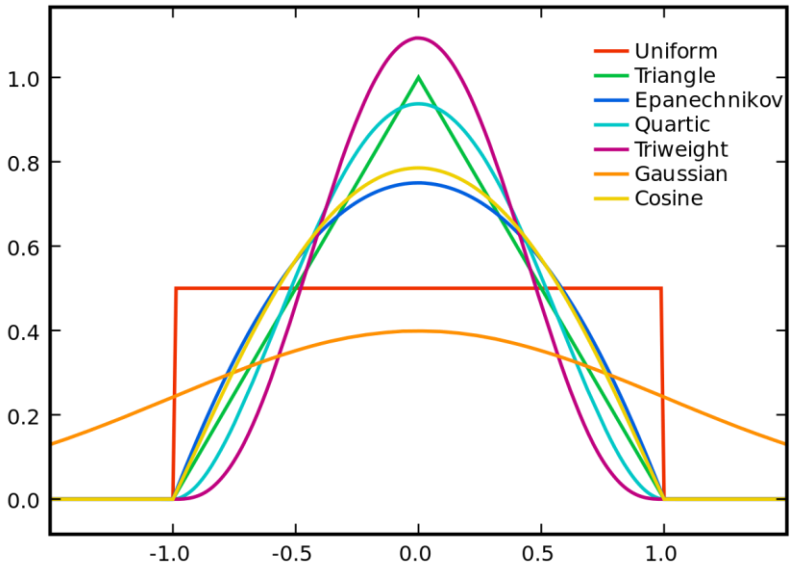


Figura 1. Exemplos de diferentes funções *kernel*.

Exemplo:

Considere uma variável aleatória $X \sim N(0,1)$. Foram obtidas $N = 40$ amostras (observações) desta variável, a partir das quais realizamos a estimação da PDF por meio do método da janela de Parzen.

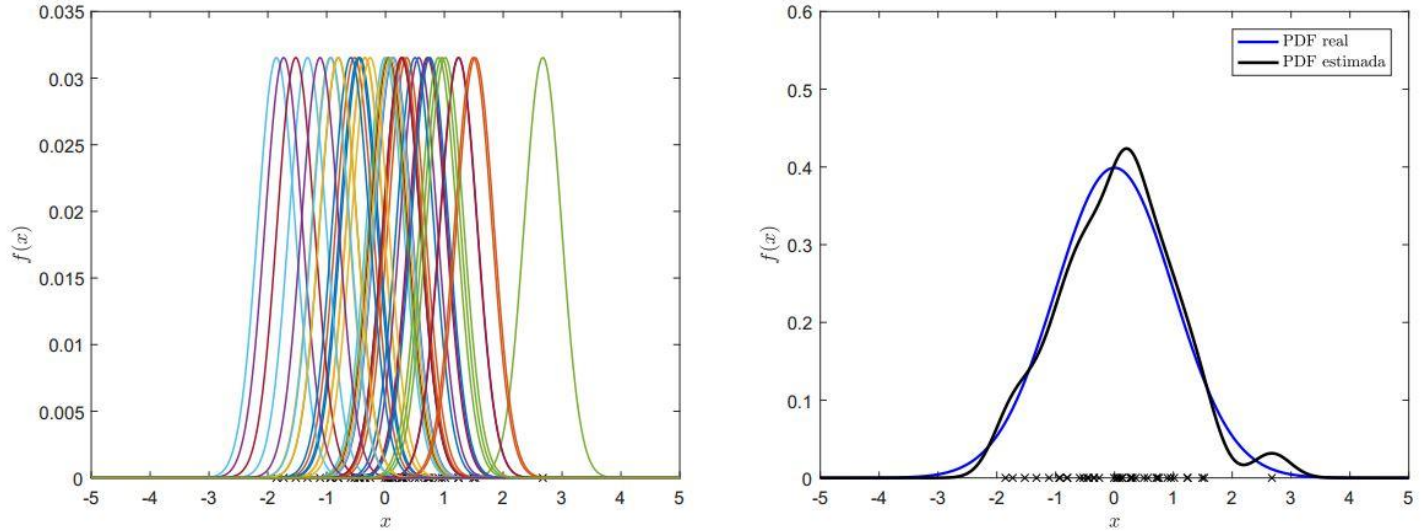


Figura 2. Componentes Gaussianas e PDF estimada utilizando $\sigma_\kappa = \sqrt{0{,}1}$.

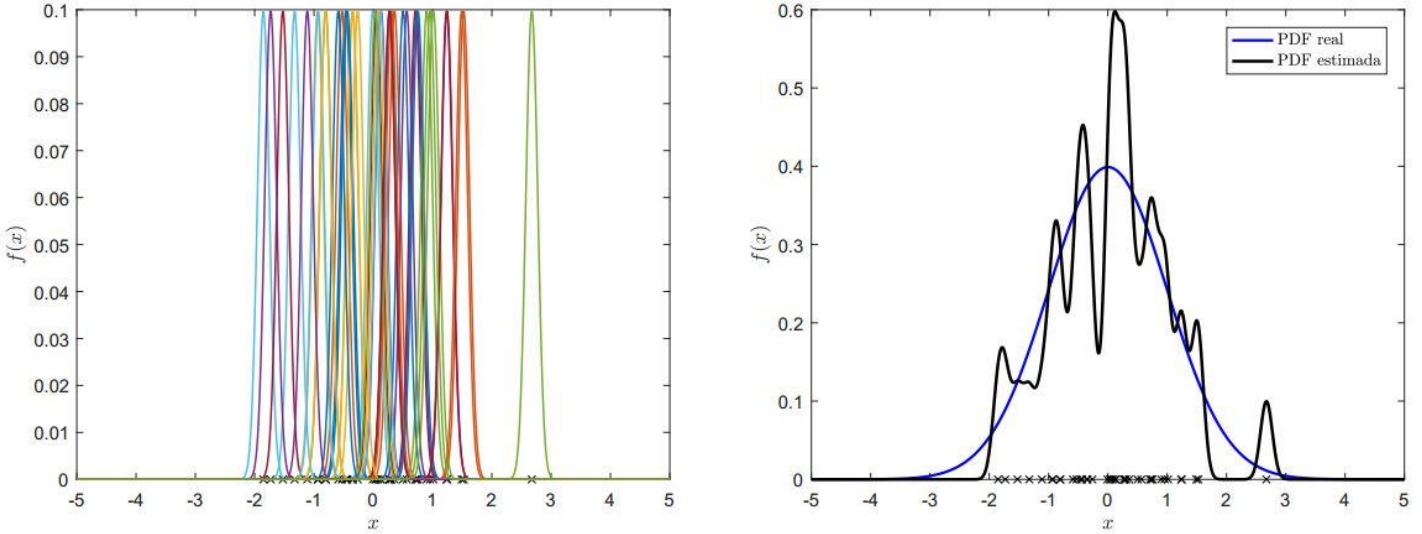


Figura 3. Componentes Gaussianas e PDF estimada utilizando $\sigma_\kappa = \sqrt{0{,}01}$.

2.2. Entropia de Rényi

A definição que Alfred Rényi dá ao conceito de entropia é particularmente interessante em virtude da maior facilidade de se estimar seu valor com o auxílio de métodos de estimação de PDFs.

Seja $p_X(x)$ a PDF de uma variável aleatória contínua X. A entropia de Rényi para esta variável, denotada por $H_\alpha(X)$, é definida como (RÉNYI, 1961):

$$H_\alpha(X) \triangleq \frac{1}{1-\alpha} \log \int p_X^\alpha(x) dx, \tag{2}$$

onde $\alpha \geq 0$. No caso limite em que $\alpha \to 1$, a entropia de Rényi se torna equivalente à entropia de Shannon (COVER & THOMAS, 2006). O argumento do logaritmo na definição da entropia de Rényi é chamado de potencial de informação (IP, do inglês *information potential*).

2.2.1. Entropia quadrática

Um caso de particular interesse da entropia de Rényi ocorre para $\alpha = 2$, quando temos a chamada entropia quadrática:

$$H_2(X) = -\log \int p_X^2(x)dx. \tag{3}$$

Neste caso, o valor do potencial de informação é dado por:

$$V_2(X) = \int p_X^2(x)dx = E_X\{p_X(x)\}, \tag{4}$$

e corresponde ao valor esperado (média) da PDF de X.

Se substituirmos $p_X(x)$ por sua estimativa dada pela janela de Parzen, conforme indicado em (1), obtemos um estimador (amostral) para a entropia quadrática. No caso em que a função *kernel* é gaussiana, temos que:

$$\hat{H}_2(X) = -\log \int \left(\frac{1}{N} \sum_{i=0}^{N-1} G_{\sigma_\kappa}(x - x_i) \right)^2 dx$$

$$= -\log \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \int G_{\sigma_\kappa}(x - x_i) G_{\sigma_\kappa}(x - x_j) dx, \quad (5)$$

em que $G_{\sigma_\kappa}(x - x_i) = \frac{1}{\sqrt{2\pi}\sigma_\kappa} \exp\left(\frac{-(x-x_i)^2}{2{\sigma_\kappa}^2}\right)$.

Propriedade: $\int G_{\sigma_\kappa}(x - x_i) G_{\sigma_\kappa}(x - x_j) dx = G_{\sigma_\kappa\sqrt{2}}(x_i - x_j)$.

Assim, chegamos à expressão final do estimador da entropia quadrática:

$$\hat{H}_2(X) = -\log \left(\frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} G_{\sigma_\kappa\sqrt{2}}(x_i - x_j) \right). \quad (6)$$

Observações:

- O argumento do logaritmo, que corresponde ao potencial de informação, é um valor escalar que pode ser diretamente estimado a partir das amostras, à semelhança dos estimadores de média e variância (PRINCIPE, 2010). Isto significa

que é possível evitar o cálculo explícito da entropia de Rényi, o qual exigiria a estimação da PDF e uma integração numérica no domínio da variável aleatória, através da estimação do IP.

- O desvio padrão da função *kernel* Gaussiana, denotado por σ_κ, é um parâmetro livre que afeta o processo implícito de estimação da PDF (SILVERMAN, 1986) e, consequentemente, a estimação da própria entropia de Rényi.

O estimador mostrado em (6) apresenta algumas propriedades interessantes (ERDOGMUS & PRINCIPE, 2002; PRINCIPE, 2010):

- Invariância com respeito à média da densidade subjacente às amostras;
- Valor mínimo ocorre quando todas as amostras da variável aleatória são iguais;
- O mínimo global é suave, i.e., tem gradiente nulo e matriz hessiana semi-definida positiva.

À luz destas características, bem como da possibilidade de implicitamente explorar o conteúdo estatístico da própria PDF, surge a ideia de se utilizar o estimador da entropia de Rényi referente ao sinal de erro entre a saída gerada por um modelo e um sinal de referência como um critério alternativo de aprendizado.

2.3. Medidas de divergência e informação mútua

A divergência de ordem α entre duas PDFs $p(x)$ e $q(x)$ é definida como (RÉNYI, 1970):

$$D_\alpha(p;q) = \frac{1}{\alpha - 1} \log \int p(x) \left(\frac{p(x)}{q(x)} \right)^{\alpha - 1} dx.$$

Lembrando que a informação mútua é um caso especial da divergência de Küllback-Leibler, quando se considera a "distância" entre a PDF conjunta e o produto das marginais, temos que a informação mútua de ordem α entre duas variáveis aleatórias X e Y pode ser escrita como:

$$I_\alpha(X;Y) = \frac{1}{\alpha - 1} \log \int \int p_{X,Y}(x,y) \left(\frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \right)^{\alpha - 1} dxdy.$$

3. Principais critérios de aprendizado

3.1. Entropia do erro

Explorando os conceitos de entropia de Rényi e estimação de PDF baseada em funções *kernel*, o critério de mínima entropia do erro (MEE, do inglês *minimum error entropy*) é definido como (ERDOGMUS & PRINCIPE, 2002; PRINCIPE, 2010):

$$\min \widehat{H}\big(e(n)\big) \qquad (7)$$
$$\text{s.a.} \quad E\{e(n)\} = 0$$

O objetivo do critério MEE é remover o máximo possível de incerteza do sinal de erro. Neste sentido, a PDF do erro se tornaria, idealmente, uma função delta de Dirac, o que indicaria que toda a informação contida nos dados disponíveis $\{\mathbf{x}(i); y(i)\}_{i=0}^{N-1}$ foi assimilada pelo modelo em seus parâmetros livres.

3.2. Correntropia

Outra entidade de grande relevância em ITL é a medida conhecida como correntropia (SANTAMARIA, POHKAREL & PRINCIPE, 2006; LIU, POHKAREL & PRINCIPE, 2007; PRINCIPE, 2010), a qual pode ser vista como uma generalização da noção de correlação, capaz de quantificar a similaridade entre duas variáveis aleatórias.

Definição:

$$c(X;Y) = E_{XY}\{\kappa_{\sigma_\kappa}(X,Y)\} = \int\int \kappa_{\sigma_\kappa}(x,y)p_{X,Y}(x,y)dxdy, \qquad (8)$$

em que $\kappa_{\sigma_\kappa}(\cdot)$ denota uma função *kernel* com parâmetro σ_κ e $p_{X,Y}(x,y)$ representa a função densidade de probabilidade conjunta das variáveis X e Y.

Em cenários práticos, nos quais a PDF conjunta é desconhecida, uma média amostral pode ser empregada para aproximar o cálculo da esperança, resultando em:

$$\hat{c}(X;Y) = \frac{1}{N}\sum_{i=0}^{N-1} \kappa_{\sigma_\kappa}(x_i, y_i). \tag{9}$$

Considerando o uso de *kernels* Gaussianos, o estimador da correntropia é dado por:

$$\hat{c}(X;Y) = \frac{1}{N}\sum_{i=0}^{N-1} G_{\sigma_\kappa}(x_i - y_i). \tag{10}$$

Seja $E = X - Y$ a variável aleatória de erro entre X e Y. Sua PDF pode ser estimada com o auxílio do método da janela de Parzen:

$$\hat{p}_E(e) = \frac{1}{N}\sum_{i=0}^{N-1} G_{\sigma_\kappa}(e - e_i). \tag{11}$$

Calculando o valor da PDF de E na origem ($e = 0$) e explorando a simetria da função gaussiana, *viz.*, $G_{\sigma_\kappa}(e) = G_{\sigma_\kappa}(-e)$, obtemos:

$$\hat{p}_E(0) = \frac{1}{N}\sum_{i=0}^{N-1} G_{\sigma_\kappa}(e_i). \tag{12}$$

Comparando (10) com (12), podemos constatar que a correntropia (cruzada) entre as variáveis X e Y corresponde ao valor estimado da PDF do erro $E = X - Y$ na origem.

3.2.1. Propriedades

- Usando a expansão em série de Taylor do *kernel* Gaussiano em torno da origem, a correntropia pode ser reescrita como:

$$\hat{c}(X;Y) = \frac{1}{\sqrt{2\pi}\sigma_\kappa} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n {\sigma_\kappa}^{2n} n!} E\{\|x_i - y_i\|^{2n}\}. \qquad (13)$$

  Esta expressão revela que, implicitamente, a correntropia engloba todos os momentos estatísticos de ordem par da variável aleatória $E = X - Y$.

- Para *kernels* simétricos, a correntropia $c(X;Y)$ é uma função simétrica.
- Adotando o *kernel* Gaussiano, a correntropia é positiva e limitada, $0 < c(X;Y) < \frac{1}{\sqrt{2\pi}\sigma_\kappa}$, sendo que o máximo valor ocorre quando $X = Y$.

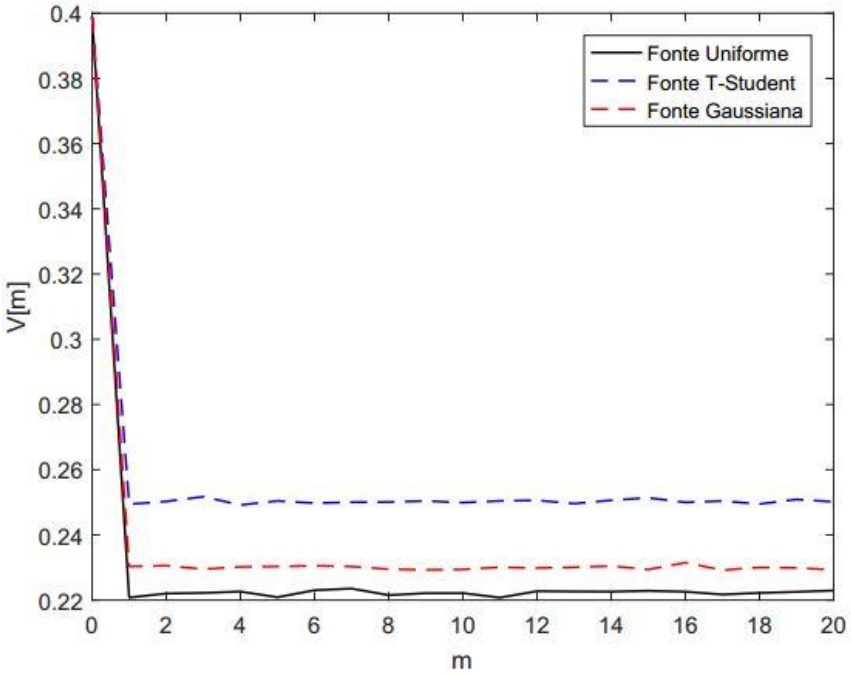
- Quando consideramos que as variáveis aleatórias estão associadas a um mesmo processo aleatório, porém em instantes de tempo diferentes, temos a chamada função de autocorrentropia*:

$$c(m) = E_{Y_n Y_{n-m}}\{\kappa_{\sigma_\kappa}(Y_n, Y_{n-m})\} = \int\int \kappa_{\sigma_\kappa}(y_n, y_{n-m}) p_{Y_n, Y_{n-m}}(y_n, y_{n-m}) dy_n dy_{n-m}, \quad (14)$$

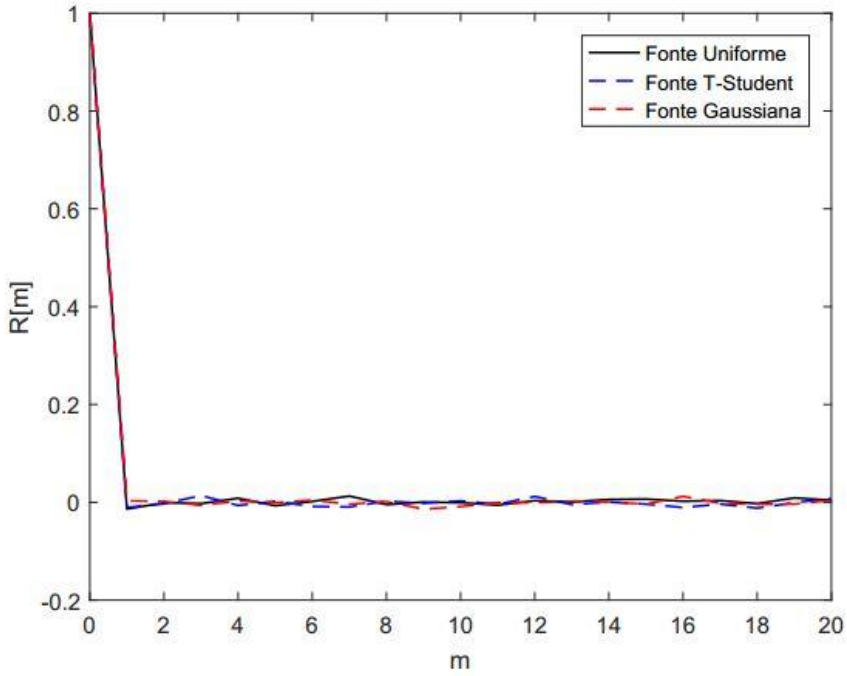
Nesta versão, a medida de correntropia incorpora não somente informações da distribuição estatística das variáveis, mas também a estrutura temporal do processo aleatório, o que pode ser particularmente útil quando se lida com sinais estatisticamente dependentes.

* Neste caso, estamos considerando um processo aleatório discreto no tempo e estacionário.

Exemplo:



(a) Autocorrentropia



(b) Autocorrelação

Figura 4. Perfis da função de autocorrentropia e de autocorrelação de três processos aleatórios brancos associados a diferentes PDFs, todos com média nula e variância unitária.

3.2.2. Critério baseado em correntropia

No âmbito de aprendizado supervisionado, foi proposto o critério da máxima correntropia (MC, do inglês *maximum correntropy*), o qual é definido da seguinte forma (PRINCIPE, 2010):

$$\max \hat{c}(y(n); \hat{y}(n)) \tag{15}$$

Neste caso, o objetivo do processo de aprendizado é maximizar o valor da PDF do erro entre a saída produzida pelo modelo ($\hat{y}(n)$) e o sinal de referência ($y(n)$) na origem, o que significa maximizar o número de amostras com pequenos desvios entre $\hat{y}(n)$ e $y(n)$.

3.3. Ilustração de alguns resultados

- Previsão da série de Mackey-Glass.

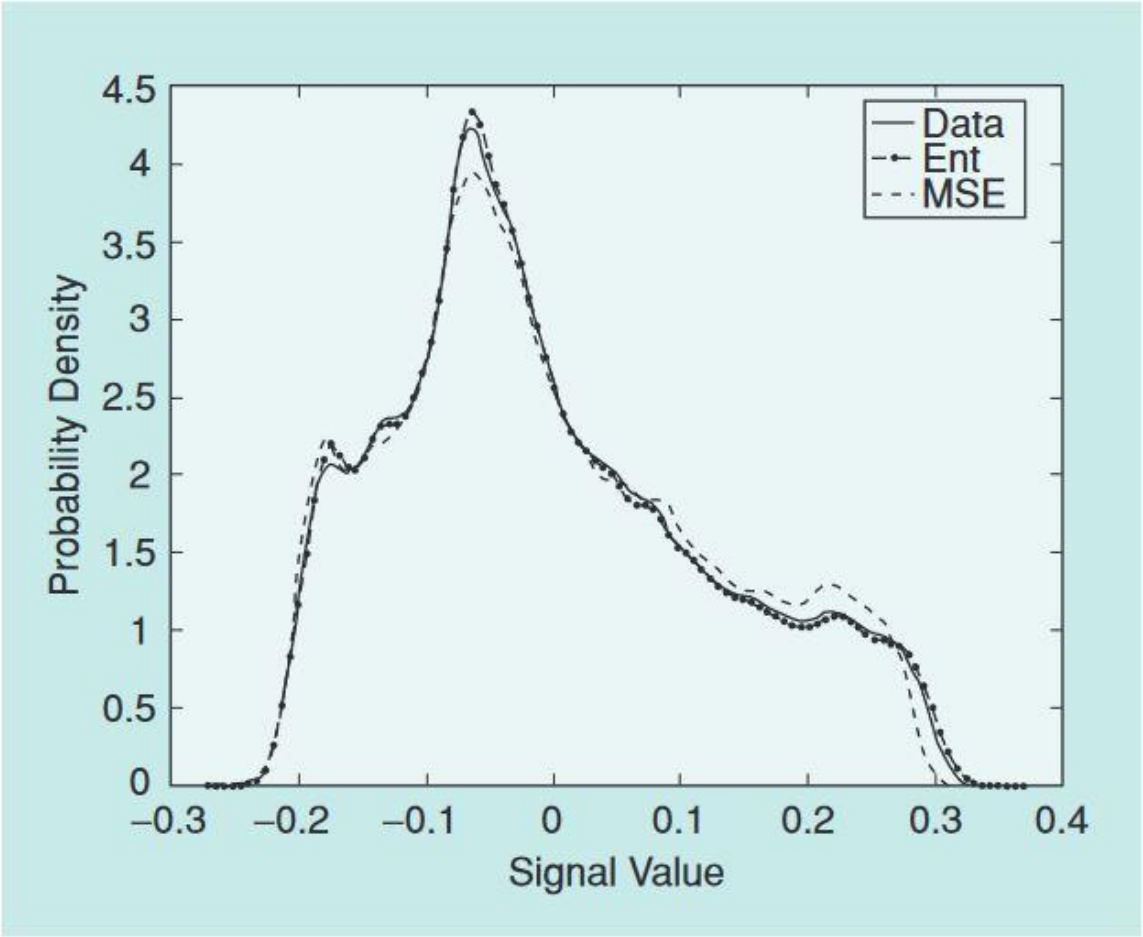


Figura 5. Densidade de probabilidade das amostras da série de Mackey-Glass (MG30) e as aproximações geradas por uma rede neural treinada com os critérios MEE e MSE (ERDOGMUS & PRINCIPE, 2006).

- Regressão não-linear com dados ruidosos e *outliers*.

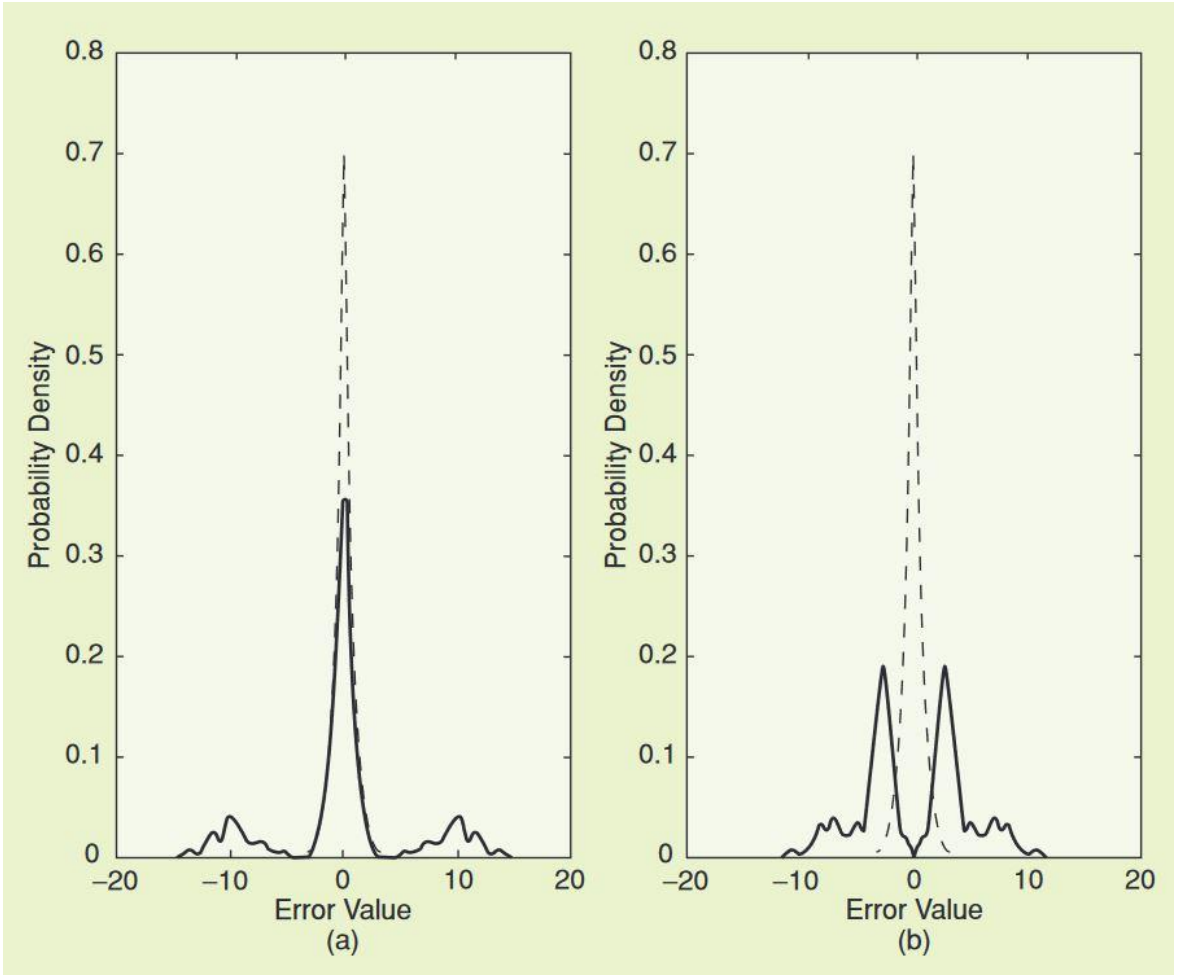


Figura 6. Distribuição do erro residual de regressão. Em (a), os parâmetros foram obtidos com base no critério MEE, enquanto em (b) foi utilizado o MSE. A linha pontilhada exibe a distribuição original das 100 amostras corrompidas com ruído Laplaciano. No conjunto de dados, também havia 40 *outliers* gerados a partir de uma distribuição $N(10,4)$.

4. Referências bibliográficas

BISHOP, C. M. **Pattern Recognition and Machine Learning**. Springer, 2006.

COVER, T. M.; THOMAS, J. A. **Elements of Information Theory**. Wiley-Interscience, 2nd ed., 2006.

ERDOGMUS, D.; PRINCIPE, J. C. An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems. *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1780-1786, 2002.

ERDOGMUS, D.; PRINCIPE, J. C. Generalized Information Potential for Adaptive Systems Training. *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1035-1044, 2002.

ERDOGMUS, D.; PRINCIPE, J. C. From Linear Adaptive Filtering to Nonlinear Information Processing – The Design and Analysis of Information Processing Systems. *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 14-33, 2006.

LIU, W.; POHKAREL, P. P.; PRINCIPE, J. C. Correntropy: Properties and Applications in Non-Gaussian Signal Processing. *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286-5298, 2007.

LUENBERGER, D.G. **Linear and Nonlinear Programming**. 2nd ed., Addison-Wesley Publishing Company, 1984.

PARZEN, E. On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, vol. 33, pp. 1065-1076, 1962.

PRINCIPE, J. C. **Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives**. Springer, 2010.

RÉNYI, A. On Measures of Information and Entropy. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pp. 547-561, 1961.

RÉNYI, A. **Probability Theory**. North Holland, Amsterdam, 1970.

ROSENBLATT, M. Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics*, vol. 27, pp. 832-837, 1956.

SANTAMARIA, I.; POHKAREL, P. P.; PRINCIPE, J. C. Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization, *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187-2197, 2006.

SHANNON, C. E. A Mathematical Theory of Communications. *Bell Systems Technical Journal*, vol. 27, no. 3, pp. 379-423, 1948.

SILVERMAN, B. **Density Estimation for Statistics and Data Analysis**. Chapman and Hall, 1986.