UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E COMPUTAÇÃO

Extensões Multidimensionais para Correntropia e suas Aplicações em Estimativas Robustas.

Joilson Batista de Almeida Rêgo

Orientador: Prof. Dr. Allan de Medeiros Martins

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação da UFRN (área de concentração: Automação e Sistemas) como parte dos requisitos para obtenção do título de Doutor em Ciências.

Número de ordem PPgEEC: D123
NATAL, RN – Agosto de 2014

UFRN / Biblioteca Central Zila Mamede
Catalogação da Publicação na Fonte

Rêgo, Joilson Batista de Almeida.
Extensões multidimensionais para correntropia e suas aplicações em estimativas robustas / Joilson Batista de Almeida Rêgo. – Natal, RN, 2014.
94 f. : il.

Orientador: Prof. Dr. Allan de Medeiros Martins.

Tese (Doutorado) – Universidade Federal do Rio Grande do Norte. Centro de Tecnologia. Programa de Pós-Graduação em Engenharia Elétrica e de Computação.

1. Função densidade de probabilidade – Tese. 2. Entropia – Tese. 3. Potencial de informação – Tese. 4. Correntropia – Tese. I. Martins, Allan de Medeiros. II. Universidade Federal do Rio Grande do Norte. III. Título.

RN/UF/BCZM CDU 621.3

Extensões Multidimensionais para Correntropia e suas Aplicações em Estimativas Robustas.

Joilson Batista de Almeida Rêgo

Tese de Doutorado aprovada em 08 de agosto de 2014 pela banca examinadora composta pelos seguintes membros:

Prof. Dr Allan de Medeiros Martins (Orientador) .....DEE/UFRN

Prof. Dr Aarão Lyra (Examinador Externo).....UnP/RN

Prof. Dr Evandro de Barros Costa (Examinador Externo).....IC/UFAL

Prof. Dr Guilherme de Alencar Barreto (Examinador Externo).....CT/UFC

Prof. Dr Adrião Duarte Dória Neto (Examinador Interno).....DCA/UFRN

Information is the resolution of uncertainty.

– Claude Shannon.

Agradecimentos

A Deus e a Imaculada que favoreceram benignamente os meus trabalhos.

A meu orientador prof. Dr. Allan de Medeiros Martins, por seu apoio e por tudo o que ele representa como Educador, Pesquisador, Professor e Amigo.

Ao Prof. Dr. José Carlos Príncipe e os pesquisadores do CNEL por abrirem as portas do maravilhoso mundo da *ITL*.

Aos Profs. Que compõem o LACI, pela amizade, estímulo e incentivo durante esta jornada.

Aos professores que contribuíram direta ou indiretamente para a realização deste trabalho através da transmissão do conhecimento que auxiliou na construção desta pesquisa.

Aos meus queridos familiares, em especial meus pais que não mediram esforços para promover educação aos filhos, a minha esposa e aos meus queridos filhos, fonte de inspirações e motivo de tanto esforço e dedicação.

Não posso deixar de dizer o quão é difícil nominar apenas algumas pessoas a quem tenho que agradecer, afinal são tantos que se fosse escrever não caberia nas páginas subsequentes, então fica aqui um abraço a todos que de alguma forma contribuíram para a realização deste.

Aos demais professores, funcionários e amigos da UFAL e da UFRN.

Resumo

O presente trabalho usa uma medida de similaridade denominada correntropia no desenvolvimento de um novo método para estimar uma relação linear entre as variáveis e suas amostras. O objetivo é estender o conceito de correntropia a partir de duas variáveis para quaisquer dois vetores (mesmo com diferentes dimensões), utilizando conceitos estatísticos. Através das extensões multidimensionais que serão apresentadas, o problema de regressão ou identificação de sistemas pode ser formulado de uma maneira diferente, buscando retas e hiperplanos que possuem a máxima densidade de probabilidade dos dados desejados. Experimentos mostraram que o novo algoritmo tem uma boa atualização de ponto fixo e robustez a ruídos impulsivos.

Palavras-Chave: Função densidade de probabilidade - *pdf*, Entropia, Potencial de Informação, correntropia.

Abstract

This present work uses a generalized similarity measure called correntropy to develop a new method to estimate a linear relation between variables given their samples. Towards this goal, the concept of correntropy is extended from two variables to any two vectors (even with different dimensions) using a statistical framework. With this multidimensionals extensions of Correntropy the regression problem can be formulated in a different manner by seeking the hyperplane that has maximum probability density with the target data. Experiments show that the new algorithm has a nice fixed point update for the parameters and robust performs in the presence of outlier noise.

Keywords: probability density functions – *pdf*, Entropy, Potential Information, Correntropy.

Sumário

Lista de Figuras..... 10

1 - Introdução..... 13

1.1 – Motivação e contribuições 16

2 –Teoria da Informação na Aprendizagem de Sistemas..... 18

2.1 – Alguns Conceitos Básicos e Definições..... 18

2.1.1 – Reconhecimento de padrões..... 20

2.1.2 – Regressão..... 20

2.1.3 – Estimação de função densidade de probabilidade - Métodos não paramétricos 21

2.2 – *Information Theoretic Learning*..... 26

2.3 – Entropia..... 27

2.3.1 – Entropia de uma variável aleatória contínua..... 29

2.3.2 – Princípio da Máxima Entropia..... 29

2.4 – Informação Mútua..... 31

2.4.1 – Propriedades..... 32

2.5 – Divergente de Kullback – Leibler..... 33

3 – Entropia de Rényi, Potencial de Informação e Correntropia..... 35

3.1 – Entropia de Rényi..... 35

3.1.1 – Caracterizando a Entropia de Shannon e definindo a entropia de Rényi. 35

3.1.2 – Entropia quadrática de Rényi..... 38

3.2 – Estimador quadrático de Rényi..... 38

3.3 – Potencial de Informação – *IP*..... 40

3.4 – Correntropia..... 42

3.4.1 – Definições e propriedades da correntropia..... 42

3.4.1.1 – Propriedades..... 45

4 – Extensões Multidimensionais para Correntropia..... 50

4.1 –Correntropia para a linha $x_1 = x_2 = \cdots = x_L$ (Extensão 01)..... 51

4.1.1 – Experimento 01: Variação na largura do kernel, pouca variância no ruído e sem a presença de outliers..... 58

4.1.2 – Experimento 02: Variação no ruído..... 60

4.1.3 – Experimento 03: percentagem e bias nos outliers..... 61

4.1.4 – Um algoritmo em ponto fixo para a extensão 01..... 62

4.2 – Qualquer possível k combinação em linha (Extensão 02)..... 64

4.3 – Qualquer possível subespaço interno linear e ortogonal (Extensão 03)..... 66

4.3.1 – Um algoritmo em ponto fixo para a extensão 03..... 69

4.4 – Qualquer possível combinação linear de k componentes (Extensão 04)..... 71

4.4.1 - Aplicação na identificação de um sistema MIMO..... 74

5 – Conclusões..... 78

Referências..... 80

Anexo A : ***Reproducing Kernel Hilbert Space - RKHS***..... 84

A.1 – Definição do *Reproducing Kernel Hilbert Space – RKHS*..... 85

A.2 – *RKHS* em Inferência Estatística..... 86

A.2.1 – Representação de uma função aleatória definida em um intervalo finito. 87

A.2.2 – Teoria de aprendizagem estatística..... 90

Lista de Figuras

Figura 01 – Ilustração do mecanismo de aprendizagem por adaptação... 19
Figura 02 – exemplo da estimação de Parzen utilizando o conceito de janela, utilizando um *kernel* gaussiano com variação em *h* e 2000 amostras de treinamento..... 24
Figura 03 – exemplo da estimação de Parzen utilizando o conceito de janela. utilizando um *kernel* gaussiano com *h* fixo e variação no número de amostras de treinamento..... 25
Figura 04 – Relação entre a entropia e a informação mútua..... 32
Figura 05 – Aprendizagem baseada num funcional de custo utilizando ITL..... 34
Figura 06 – Correntropia como a integral no espaço gaussiano ao longo da reta $x = y$..... 44
Figura 07 – reta $x = y$ e $z = 0$ 53
Figura 08 – gráfico comparativo entre a estimação da probabilidade utilizando o MCC, a extensão 01 da correntropia com σ = 0,005 e o EMQ com pouca variância no ruído e sem a presença de *outliers*..... 58
Figura 09 – gráfico comparativo entre a estimação utilizando o EMQ, MCC e a extensão 01 da correntropia com σ = 0,05 com pouca variância no ruído e sem a presença de *outliers*..... 59
Figura 10 – gráfico comparativo entre a estimação utilizando o EMQ, MCC e a extensão 01 da correntropia com σ = 0,05 na presença de um ruído intenso..... 60
Figura 11 – gráfico comparativo entre a estimação utilizando o EMQ, MCC e a extensão 01 da correntropia com σ = 0,05 na presença de *outliers* e um *bias*..... 61
Figura 12 – reta $x = y$,$z = z_0$ e $z = z_0, x = x_0$..... 64

Figura 13 – Plano $z=z_0$ 67
Figura 14 –Estimativa do sistema utilizando a extensão 03 ($\sigma^2 = 1{,}8$) e a estimativa do sistema utilizando MMQP..... 70
Figura 15 – Variância no erro entre o MMQP e a Extensão 03 da correntropia..... 71
Figura 16 – Plano $z = x$ e y 72

Lista de Abreviaturas

EMQ	*Erro Médio Quadrático*
RNA	*Redes Neurais Artificiais*
BSS	*Blind Source Separation*
ICA	*Independent Components Analysis*
CEE	*Critério Baseado na Entropia do Erro*
ITL	*Information Theoretic Learning*
MCC	*Critério da Máxima Correntropia*
IP	*Information Potential*
pdf	*Função distribuição de probabilidade*
IT	*Information Theory*
pmf	*Função massa de probabilidade*
RKHS	*Reproducing Kernel Hilbert Space*

Capítulo 1

Introdução

Um problema bastante comum na Engenharia é como extrair o máximo de informações relevantes presentes nos dados. Atualmente, nos deparamos com uma quantidade enorme de dados, dos quais muitos deles são irrelevantes ou redundantes. Para conseguirmos extrair informações precisas, necessitamos filtrar ou trabalhar tais dados, possivelmente com sistemas adaptativos que são utilizados para fornecer um modelo estimado que melhor represente os dados, desse modo necessitamos utilizar um critério para avaliar a diferença entre os valores estimados e os reais (erro). Uma forma é utilizar a média do quadrado do erro (erro médio quadrático – EMQ) como um critério para selecionar um estimador adequado [1].

Norbert Wiener utilizou o conceito do EMQ, em problemas de filtragem adaptativa, com o objetivo de reduzir a quantidade de ruído presente nos dados baseando-se em uma abordagem estatística [2]. A partir do trabalho de Wiener foram desenvolvidos vários outros sistemas adaptativos incluindo Redes Neurais Artificiais (RNA), que se concentraram na maioria em torno do erro médio quadrático resultando em superfícies quadráticas de forma que expressões analíticas podem ser facilmente encontradas e analisadas.

Essa prática foi estendida ao treinamento de inúmeros sistemas adaptativos, gerando um entendimento de que o momento de segunda ordem (variância com relação à média) [3] é suficiente na determinação de soluções para quase todos os problemas de ordem prática na Engenharia, onde a distribuição gaussiana emerge naturalmente como uma distribuição para os processos envolvidos, como consequência do Teorema Central do Limite – TLC [4]. Nas últimas

décadas inúmeras pesquisas seguiram nesta linha, adotando o erro médio quadrático e/ou a variância no erro na solução de tais problemas (supervisionado ou não supervisionado).

No entanto, novos problemas surgiram que não podiam mais serem resolvidos através de métodos que utilizavam estatística de segunda ordem, e que exigiam novas técnicas utilizando estatística de ordem superior nos processos envolvidos, tais como: Separação Cega de Fontes (BSS - *Blind Source Separation*), Análise de Componente Principal (ICA - *Independent Components Analysis*), dentre outros [5] [6]. No entanto, enquanto as pesquisas avançavam na área de sistemas adaptativos, paralelamente na área de comunicações avançava e se consolidava o conceito da teoria da informação. Shannon utilizou o conceito de entropia como uma medida de incerteza e o transformou numa medida de quantidade de informação, sendo inicialmente designada de Teoria Matemática da Informação. Motivado pelo problema das comunicações com segurança, Shannon estabeleceu um modelo de sistemas de comunicação genérico e formalizou os conceitos de medida de informação, de capacidade de transferência de informação sobre um canal e de codificação [7]. A teoria da informação, é uma teoria matemática consistente, que teve significativa importância em áreas como a teoria das probabilidades, estatística, ciência da computação, física, economia, biologia e química [8].

Outra contribuição nesta área foi apresentada pelo Matemático Húngaro Alfred Rényi que apresentou uma família paramétrica de entropias sendo a entropia de Shannon um caso particular da entropia de Rényi [9]. O trabalho de Rényi não foi inicialmente reconhecido como uma ferramenta a ser utilizada na Engenharia, até que na década de 90 passou a ser utilizada em diferentes campos, dentre eles o de processamento de imagens, onde foi apresentada uma técnica geral para limiarização de imagens digitais com base na entropia de Rényi [10] mais recentemente temos aplicações em Estimação espectral e reconhecimento de padrões [11], em algoritmos de aprendizagem de máquinas [12] e em aplicações biomédicas [13], dentre outras. Ainda na década de 90, foi proposto um método para estimar a entropia de dados utilizando a entropia de Shannon e a lei dos grandes números, como um estimador baseado na média das amostras, porém não muito eficiente [14].

Os algoritmos de adaptação que utilizam o EMQ consideram apenas os momentos de segunda ordem da distribuição do erro, bem eficiente se levarmos em conta que o erro possui uma distribuição gaussiana, no caso em que a distribuição do erro é não gaussiana, faz sentido utilizarmos funções de custo alternativas para a adaptação.

Pensando nisso, Princípe e Erdogmus [6] propuseram uma abordagem diferente baseada na Entropia do Erro (CEE), cujo objetivo é remover a incerteza tanto quanto possível a partir do sinal de erro. Eles desenvolveram um estimador de entropia não paramétrico baseado na entropia de Rényi, que pode ser aplicado diretamente aos dados obtidos a partir de experimentos, manipulando diretamente a informação contida nesses, desta forma foi aplicado com sucesso a entropia de Rényi e outros critérios de otimização na resolução de problemas, tais como: separação cega de fontes, redução de dimensionalidade, extração de características, dentre outros. Princípe foi o primeiro a utilizar o termo *Information Theoretic Learning* – ITL, na literatura de sistemas adaptativos [1].

Em ITL o principal objetivo é encontrar funções de custo que manipulem diretamente a informação presente nos dados, a questão fundamental em ITL é como estimar os dois principais descritores estatísticos propostos pela teoria da informação, a entropia e divergentes diretamente das amostras. ITL é independente da arquitetura de aprendizagem de máquina, e exige apenas a disponibilidade dos dados sem a necessidade de nenhum conhecimento a *priori* sobre a distribuição dos mesmos [15], pode-se neste caso pensar funções de custo baseadas no erro como uma medida a ser minimizada para alcançar a máxima entropia. ITL quantifica as propriedades gerais dos dados, mas podemos aplicar tais conceitos para medir similaridade entre variáveis aleatórias que é geralmente expressa como a correlação.

No entanto, a correlação trata da informação a partir dos momentos estatísticos de segunda ordem levando em conta que a *pdf* dos dados é gaussiana, em casos em que a *pdf* é não gaussiana ou que se necessita generalizar a similaridade incluindo a informação presente nos momentos de alta ordem, faz-se necessário generalizar a correlação. Recentemente, o conceito de uma nova generalização de funções de correlação foi apresentado, denominada de *correntropia* [16]. O

nome indica uma forte relação com a correlação, mas possuem diferenças bem significativas, a *Correntropia* é uma função definida positiva que mede a similaridade linear ou não linear entre variáveis aleatórias localmente e envolve estatística de alta ordem dos dados de entrada, enquanto que a correlação é quadraticamente dependente das distâncias entre amostras nos espaço conjunto [16]. A *correntropia* vem sendo utilizada como uma nova função de custo para sistemas adaptativos com um critério denominado de critério da máxima correntropia (MCC) [17]. O MCC é um critério local de medida de semelhança muito útil quando o ruído presente nos dados é não gaussiano, não possui média zero e possuem muitos ruídos impulsivos (*outliers*) [17].

Neste trabalho, iremos apresentar uma generalização para *correntropia* no desenvolvimento de um método para a estimação de uma relação entre os dados e suas amostras. Para atingir esse objetivo iremos expandir o conceito de *correntropia* de duas variáveis para dois vetores (mesmo de dimensões diferentes) utilizando ferramentas estatísticas.

1.1 – Motivação e Contribuições

Motivados pelos recentes avanços científicos na área de ITL, iremos neste trabalho estender aplicações da *correntropia* para estimadores lineares. Os conceitos desta nova definição de correlação generalizada são estendidos a variáveis multidimensionais, em vez de apenas comparar duas variáveis como apresentado na definição original. A correntropia apresenta várias propriedades interessantes que conecta áreas como ITL, métodos de *Kernel* e estimadores robustos.

A organização da tese é a seguinte. Após um breve histórico sobre teoria da informação, ITL e métodos de *Kernel* de modo a apresentar as terminologias, em seguida no capítulo 2, abordaremos alguns conceitos básicos e definições da teoria de aprendizagem em sistemas e dos momentos da teoria de informação inseridos no contexto de ITL. No capítulo 03, apresentamos a definição da entropia e da entropia quadrática de Rényi bem como a conceituação de potencial de informação – IP, apresentamos também uma introdução à unificação entre ITL e métodos de *Kernel* ou aprendizagem estatística através

de uma nova medida de correlação generalizada. No capítulo 04, os conceitos dessa correlação generalizada são estendidos a espaços multidimensionais, bem como alguns exemplos são apresentados para corroborar nosso entendimento e que possamos vislumbrar aplicações em diferentes campos. Finalmente, apresentamos as conclusões e possíveis futuras linhas de pesquisa.

Capítulo 2

Teoria da Informação na Aprendizagem de Sistemas.

O Erro médio quadrático (EMQ) é um critério de desempenho utilizado na aprendizagem supervisionada dada à sua simplicidade matemática que permite uma análise teórica simples. Ele provém meios suficientes para explorar a estatística de segunda ordem de um sistema com distribuições gaussianas [2]. No entanto, recentemente, novos problemas surgiram que não poderiam mais serem resolvidos através de métodos que utilizam estatística de segunda ordem, de modo que o conceito de aprendizagem passou a ser visto a partir da minimização da entropia do erro utilizando-se da entropia quadrática de Rényi com técnicas de janelamento de Parzen para a solução de tais problemas.
A técnica de janelamento de Parzen é utilizada para obter uma estimativa da função densidade de probabilidade do erro (*pdf* do erro), preservando o mínimo global, desde que certas restrições sejam satisfeitas. A aplicação do critério da entropia do erro em problemas práticos de aprendizagem supervisionada é conceitualmente simples e será apresentado, ou seja, dado um determinado sinal produzido por um sistema desconhecido (dados de treinamento), estima-se a entropia do erro do conjunto de dados [18].

2.1 – Alguns Conceitos Básicos e Definições.

Um passo importante na extração de conhecimento a partir de um conjunto de dados é a modelagem. Modelar é encontrar maneiras de desenvolver e implementar modelos matemáticos a partir de dados reais, no entanto para saber se o modelo representa bem os dados necessitamos de métricas, dentre as quais citamos o EMQ.

Vamos assumir que o nosso problema é encontrar um modelo estatístico que encontre uma relação entre duas variáveis aleatórias $\{x, d\}$ a partir de um conjunto de medidas obtidas de uma fonte de sinal desconhecida.

A seguinte figura ilustra o processo.

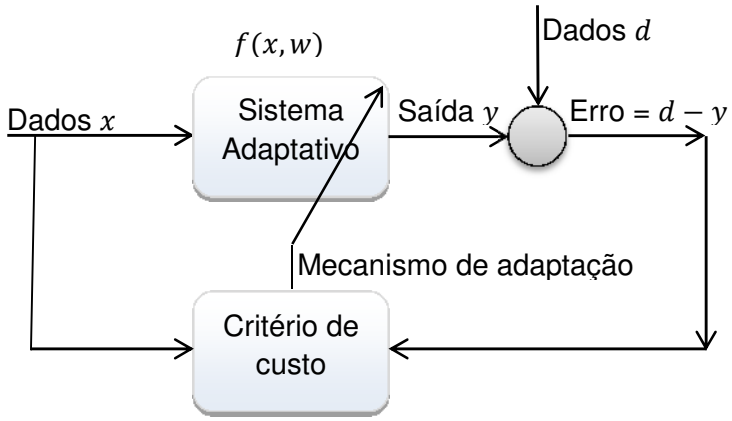

Figura 01 – Ilustração do mecanismo de aprendizagem por adaptação.

O cenário apresentado na figura 01 é geralmente descrito como aprendizagem de máquina ou sistema adaptativo supervisionado, cujo objetivo principal é aprender a partir dos dados apresentados e melhorar o seu desempenho através do processo, de forma que podemos entender como aprendizagem o processo pelo qual os parâmetros livres w do sistema são adaptados através da estimulação do mesmo produzindo uma resposta y. Como resultado o sistema sofre modificações nos seus parâmetros livres e responde de uma maneira nova devido a tais modificações a partir do sinal de erro.

Um conjunto de regras bem definidas para a solução do problema de aprendizagem é denominado de mecanismo de adaptação ou algoritmo de aprendizagem, no processo de aprendizagem o sinal de saída representando a resposta do sistema adaptativo é comparado com uma resposta desejada representada por d. Consequentemente, é produzido um sinal de erro ($Erro = d - y$). Este sinal de erro aciona um mecanismo de adaptação, cujo propósito é aplicar uma correção aos parâmetros w, tais ajustes têm como objetivo aproximar o sinal de saída y da resposta desejada d. Este objetivo é alcançado minimizando-se um critério ou funcional de custo, definido em função do sinal de erro. O processo de ajuste dos parâmetros livres continua até o sistema atingir um estado estável, onde o processo é então encerrado. O processo de

aprendizagem descrito é denominado na literatura como *aprendizagem por correção de erro* [19].

A formulação de problemas de aprendizagem é bastante ampla e engloba muitos problemas específicos, no entanto vamos exemplificar apenas três destes problemas que são: reconhecimento de padrões, regressão e estimação da função densidade de probabilidade.

2.1.1 – Reconhecimento de padrões.

O reconhecimento de padrões tem como objetivo a classificação de objetos em um número de categorias ou classes, é parte integrante da maioria dos sistemas de aprendizagem para a tomada de decisões. A abordagem baseia-se em argumentos probabilísticos decorrentes da natureza estatística nos dados. Adotando esse raciocínio, podemos projetar sistemas que classificam ou rotulam padrões desconhecidos em classes.

Seja um sistema cuja tarefa seja classificar n classes distintas $C_i, i = 1, \cdots, n$, a um padrão desconhecido apresentado e sendo representado por um vetor de características $\boldsymbol{x}$, com probabilidade condicional $P(C_i|\boldsymbol{x})$, representando a probabilidade do parâmetro desconhecido pertencer a uma respectiva classe C_i dado o vetor correspondente $\boldsymbol{x}$. Os classificadores geralmente calculam o máximo da função probabilidade definido pelas classes de modo que, o padrão desconhecido é então designado ou rotulado para a classe correspondente a esse valor máximo ou, se utilizar a probabilidade do erro o problema então consistirá na minimização da probabilidade do erro de classificação [20].

2.1.2 – Regressão.

Um dos principais métodos estatístico cujo objetivo é estimar uma ou mais variáveis (dependente) em função de outra(s) (independente). Por exemplo, Se y deve ser estimada em função de x por meio de uma equação, tal equação é denominada *equação de regressão* de y sobre x e a curva correspondente é a *curva de regressão* de y sobre x. De modo geral, pode-se ajustar mais de uma curva a determinado conjunto de dados.

A dispersão dos dados em relação a uma curva de regressão indica que, para determinado valor de x, há efetivamente valores distintos de y distribuídos em torno da reta ou curva de regressão. Esta ideia de distribuição conduz naturalmente à conclusão de que existe uma relação entre o problema de ajustar a curva e a teoria de probabilidade.

Esse relacionamento é visualizado a partir da analogia utilizando-se as variáveis aleatórias X e Y que podem assumir os diversos valores amostrais x e y, respectivamente. Supõe-se que as variáveis aleatórias tenham uma função distribuição de probabilidade conjunto, ou função de densidade conjunta, $f(x,y)$, conforme se trate de variáveis discretas ou contínuas [21].

Dada a função de densidade, ou função de densidade conjunta $f(x,y)$ de duas variáveis aleatórias de X e Y, é natural, de acordo com as observações apresentadas, perguntar se existe uma função $g(X)$ tal que,

$$\mathbb{E}[(Y - g(X))^2] \to 0. \qquad (2.1)$$

Uma determinada curva com equação $y = g(X)$ de acordo com a propriedade (2.1) é denominada *curva de regressão de erro médio quadrático* ou de *mínimos quadrados* [21]. Quando não conhecemos $f(x,y)$, podemos utilizar o critério (2.1) para obter curvas de regressão aproximadas através da estimação.

2.1.3 – Estimação da função densidade de probabilidade – Métodos não paramétrico.

Na década de 60 alguns pesquisadores sugeriram vários métodos para estimar *pdf's*, então chamados de não paramétricos, que são basicamente variações das aproximações executadas com o histograma de uma *pdf* desconhecida, bastante comum a partir dos conceitos básicos estatísticos utilizados na estimação de funções densidade de probabilidade. O objetivo desses métodos consiste em estimar a densidade a partir de uma gama de funções que não são restritas a um conjunto paramétrico de funções. Dentre os nãos paramétricos destacaremos o método de janelamento de Parzen.

A ideia básica em estimar a densidade é dada por,

$$\hat{p}(x) \approx \frac{k}{nV}$$

Onde V é o volume de uma pequena região em x com k valores em n amostras dos dados. Inicialmente no método de janelamento de Parzen, iremos supor uma determinada região como sendo um hipercubo L – dimensional, se h for o comprimento de uma aresta do referido hipercubo, então o seu volume é dado por [22],

$$V = h^L. \tag{2.2}$$

Podemos obter uma expressão analítica em k, definindo uma função janela $\Phi: \mathbb{R}^L \to \mathbb{R}$ tal que:

$$\Phi(u) = \begin{cases} 1 & |u_i| \leq {}^1/_2\,,\ i = 1, \dots, n \\ 0\,, & caso\ contrário \end{cases}. \tag{2.3}$$

Que define um hipercubo unitário em $\mathbb{R}^L$ centrado na origem. Da analogia, segue que $\Phi\big((x - x_i)/h\big)$ é igual à unidade quando x_i estiver contido no hipercubo de volume V centrado em x, e nulo caso contrário. Portanto, o número de dados contidos no hipercubo de lado h centrado em x é,

$$k = \sum_{i=1}^{n} \Phi\left(\frac{x - x_i}{h}\right) \tag{2.4}$$

Como sabemos que o hipercubo de lado h em $\mathbb{R}^L$ possui volume dado pela equação (2.2). Daí, podemos escrever nosso estimador da *pdf* como:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^L} \Phi\left(\frac{x - x_i}{h}\right) = \frac{1}{nh^L}\sum_{i=1}^{n} \Phi\left(\frac{x - x_i}{h}\right). \tag{2.5}$$

Conhecido como método de janelamento de Parzen [22]. De tal modo que a expressão (2.5) pode ser vista como uma média de funções de x e das amostras x_i. De forma resumida, temos a função janela sendo utilizada para interpolar de modo que cada amostra contribua para a estimativa de acordo com a sua distância a partir de x.

A função Φ deve satisfazer as seguintes condições,

$$\Phi(x) \geq 0, \forall x \text{ e } \int \Phi(u)du = 1$$

E se mantivermos a relação $V = h^L$, então segue que $\hat{p}(x)$ também satisfaz essas condições. Ou seja,

$$\hat{p}(x) \geq 0, \forall x$$

$$\int \hat{p}(x)dx = \int \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^L}\Phi\left(\frac{x - x_i}{h}\right)dx = \frac{1}{nh^L}\sum_{i=1}^{n}\int \Phi\left(\frac{x - x_i}{h}\right)dx$$

$$= \frac{1}{nh^L}\sum_{i=1}^{n} h^L = 1.$$

No entanto, a função definida em (2.5) apresenta algumas desvantagens, tais como: estimativas de densidade que possuem descontinuidades e igual ponderação de todos os pontos x_i independentemente da sua distância até o ponto de estimativa x, por esta razão, a janela de parzen é geralmente substituída por uma função *kernel* suave, sendo a mais comumente utilizada, a função Gaussiana ou distribuição normal [22], de forma que podemos escrever a estimativa da *pdf* como,

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h^L} exp\left(-\frac{(x - x_i)^T(x - x_i)}{2h^2}\right) \quad (2.6)$$

Desta forma a *pdf* desconhecida pode ser aproximada por uma média de *n* gaussianas centradas em amostras distintas do conjunto de treinamento, e a influência de cada gaussiana é mais localizada no espaço de características em torno da área de seu valor médio. A expansão de uma *pdf* é a soma das gaussianas, e o parâmetro desconhecido h denominado por largura do *kernel*, que afeta diretamente a largura e a amplitude das gaussianas, é escolhido pelo usuário. Para ilustrar tais conceitos iremos apresentar como exemplo um conjunto de amostras com $x_i \epsilon \mathbb{R}$, distribuídos de acordo com a seguinte *pdf*,

$$p(x) = 0{,}33\mathcal{N}(0; 0{,}2) + 0{,}66\mathcal{N}(2; 0{,}2)$$

A distribuição de probabilidade escolhida como exemplo é uma soma de duas distribuições gaussianas (normal) com médias 0 e 2 e desvios padrão de 0,2, respectivamente. Desta forma, iremos estimar a *pdf* utilizando uma aproximação através do janelamento de parzen utilizando *kernel* gaussiano. Primeiramente fixamos o número de amostras ($n = 2000$) e variamos a largura do *kernel* ($h = 0{,}01; 0{,}05; 0{,}1\ e\ 1$), obtendo os seguintes resultados:

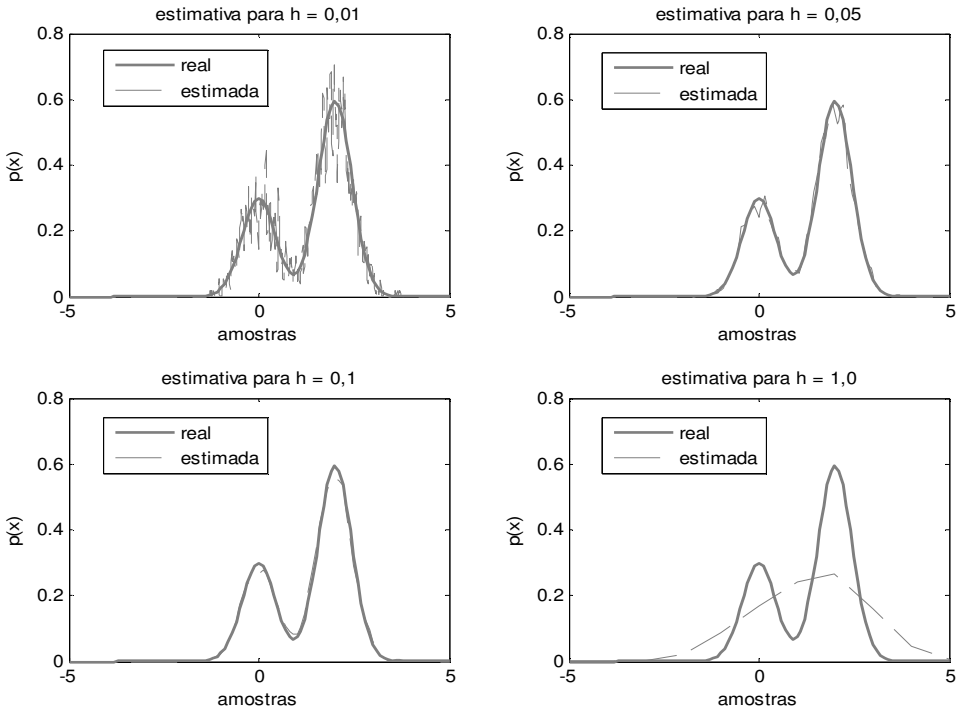

Figura 02 – pdf estimada (linha tracejada) via janelamento de Parzen da *pdf* real (linha sólida), utilizando um *kernel* Gaussiano com variação em h e um tamanho fixo de 2000 amostras de treinamento.

Podemos observar a partir da figura 02 que para uma quantidade fixa de amostras de treinamento, um valor pequeno na largura do *kernel* representa uma grande variância nos dados estimados, enquanto que valores maiores desta representam uma suavização na variação da densidade. Iremos em seguida, fixar o valor da largura do *kernel* (h = 0,1) e variarmos o número de amostras de treinamento ($n = 500;\ 1000;\ 3000\ e\ 10000$) obtendo os seguintes resultados:

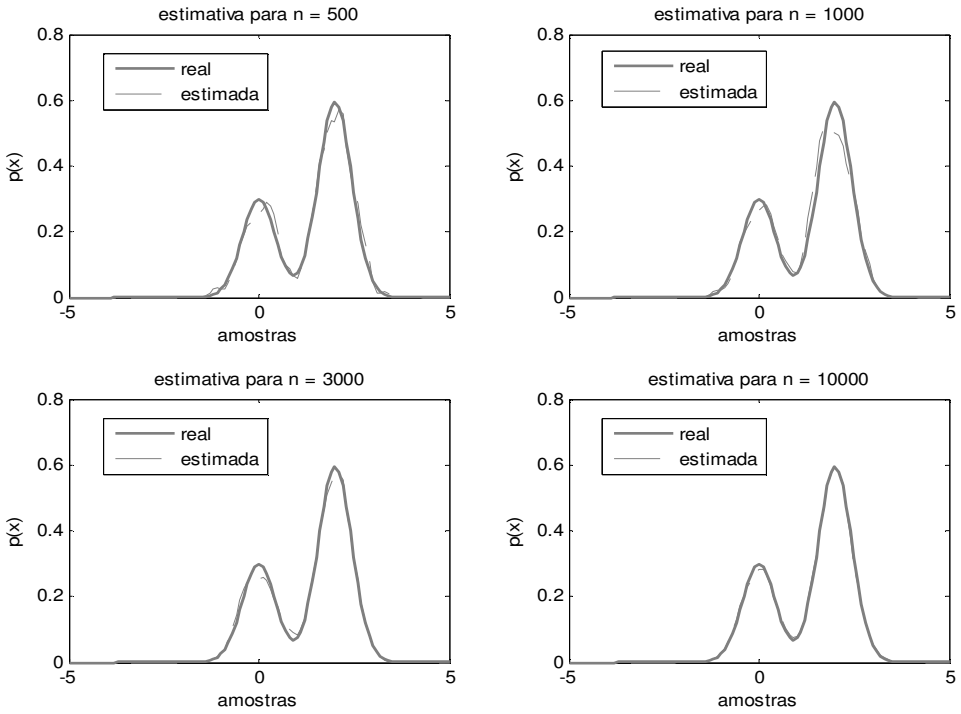

Figura 03 – *pdf* estimada(linha tracejada) via janelamento de Parzen da *pdf* real (linha sólida), utilizando um *kernel* Gaussiano com h fixo e variação no número de amostras de treinamento.

Podemos observar na figura 03 que para uma largura de *kernel* fixa, a variância dos dados estimados decresce à medida que aumentamos o número de amostras de treinamento. Desta forma, podemos concluir que a estimativa da *pdf* utilizando o método de janelamento de Parzen deve levar em conta tanto a largura do kernel como a quantidade de amostras utilizadas. A escolha da largura do *kernel* é crucial, e várias abordagens na escolha de tal parâmetro pode ser encontrada na literatura, por exemplo, Silverman [23]. Normalmente, um grande número de amostras é necessário para um bom desempenho, no entanto devemos observar que este número cresce exponencialmente com a dimensionalidade dos dados (maldição da dimensionalidade).

2.2 – *Information Theoretic Learning*

Recentemente, por outro lado, especialmente nas áreas de processamento de sinais e comunicação, novos problemas encontrados foram exigindo o uso de propriedades estatísticas de ordem superior nos processos envolvidos. Nesses problemas, a escolha de um critério baseado em medidas estatística da informação depende de outros requisitos e não mais do EQM. Pensando nisso, Príncipe e outros propuseram a combinação de um funcional de mapeamento com uma função de custo baseado na modelagem dos dados com o objetivo de obter uma eficiência computacional [6]. A modelagem lida com a construção de descritores escalares das *pdf* 's, que caracteriza a estrutura dos dados. A vantagem destes descritores é que eles resolvem o processamento dos dados com simplicidade computacional.

No entanto, o critério baseado na medida de informação não sofre a limitação da gaussianidade inerente às funções de custo baseadas em momentos de segunda ordem (EQM). Isto é conseguido com a utilização de descritores tais como entropia, divergentes e informação mútua (conceitos provenientes da *Teoria da Informação – IT*), combinados com estimadores não paramétricos das *pdf*'s, que trazem robustez e generalização para o funcional de custo e melhoram o desempenho em muitos cenários realistas, tais como processamento estatístico de sinais e aprendizagem de máquina. No entanto, existem diferenças significativas entre a aplicação da teoria da informação para sistemas de comunicação apresentada por Shannon e a realidade em processamento de sinal adaptativo e aprendizagem de máquina. Tal combinação foi denominada de *Information Theoretic Learning – ITL* [1].

Uma grande vantagem da ITL é que com pequenas modificações, ela pode ser utilizada em: aprendizagem convencional, métodos de filtragem adaptativa, aprendizagem de máquina e métodos de *kernel*. Vamos agora analisar o mecanismo de adaptação da figura 01 a partir do ponto de vista da *Teoria da Informação – IT*.

As informações contidas na *pdf* conjunta $p(x, d)$ devem ser transferidas da melhor forma possível aos parâmetros livres w do sistema. Portanto, deseja-se a extração do máximo de informação possível da *pdf* do erro $p(e)$, ajustando w

de modo a aproximar o máximo possível $y = f(x, w)$ de d em um sentido de informação. Neste caso, a entropia pode ser utilizada como uma medida da incerteza presente no erro, de modo que o funcional de custo pode ser utilizado para minimizar a entropia do erro [1].

2.3 – Entropia

O termo *entropia* como um conceito científico foi inicialmente utilizado na termodinâmica por Clausius em 1850. Sua interpretação probabilística no contexto estatístico é atribuída a Boltzmann (1877). No entanto, a relação explícita entre entropia e probabilidade foi registrada depois de vários anos por Planck (1906). Em 1928 Hartley fundamentou que quando um símbolo é escolhido a partir de um conjunto finito de símbolos, então o número de escolhas pode ser considerado como uma medida de informação ou como ele denominou *quantidade de informação*. Shannon, em um artigo publicado (1949) utilizou tal conceito para dar uma descrição simplificada a propriedades de uma longa sequência de símbolos, e aplicou os resultados a um número de problemas básicos na teoria de codificação e transmissão de dados. Suas notáveis contribuições formam a base da moderna Teoria da Informação – *IT* [7]. Jaynes (1957) aplicou o princípio da máxima entropia a uma variedade de problemas envolvendo a determinação de parâmetros desconhecidos em dados incompletos. No nosso contexto, a entropia é definida como a incerteza presente em uma variável aleatória, e será adotada como critério para aplicações onde a manipulação do conteúdo presente na informação em sinais é desejado ou necessário [1].

Inicialmente vamos considerar uma variável aleatória X, onde cada evento pode ser considerado como uma mensagem. Então, considerando uma variável aleatória discreta X, modelada como $X = \{x_k | \, k = 0, \pm 1, \dots, \pm n\}$, onde os valores das amostras x_i é um número discreto e $(2k + 1)$ é o número total de níveis discreto. Para completar o modelo, o evento $X = x_k$ ocorre com probabilidade $p_k = P(X = x_k)$ onde é requerido que

$$0 \leq p_k \leq 1 \; e \; \sum_{k=-n}^{n} p_k = 1 \tag{2.7}$$

Suponha que o evento $X = x_k$ ocorre com probabilidade $p_k = 1$. Em tal situação, temos certeza da ocorrência do evento e a informação contida na mensagem é nula. No entanto, se a probabilidade p_k possuir um valor baixo então a mensagem é incerta e o conteúdo da informação é elevado, temos a máxima incerteza se $p_k = 0{,}5$. Então os conceitos de incerteza e informação são relacionados entre si e a quantidade de informação é relacionada como o inverso da probabilidade de ocorrência de um evento [1].

Definimos a quantidade de informação ganha após a ocorrência do evento $X = x_k$ com probabilidade p_k como uma função logarítmica

$$I(x_k) = log\left(\frac{1}{p_k}\right) = -\log p_k \tag{2.8}$$

onde a base do logaritmo é arbitrária, em qualquer caso a informação dada na equação (2.8) apresenta as seguintes propriedades:

$$I(x_k) = 0 \quad para \quad p_k = 1 \tag{2.9}$$

obviamente, se temos certeza absoluta na realização de um evento, então não há informação ganha pela sua ocorrência.

$$I(x_k) \geq 0 \quad para \quad 0 \leq p_k \leq 1 \tag{2.10}$$

Isto é, a ocorrência do evento $X = x_k$ ou fornece alguma ou nenhuma informação, mas nunca provoca uma perda de informação. Ou seja,

$$I(x_k) > I(x_i) \quad para \quad p_k < p_i \tag{2.11}$$

de modo que o evento mais improvável é o que obtemos mais informação através de sua ocorrência.

A quantidade de informação $I(x_k)$ é uma variável aleatória discreta com probalilidade p_k. O valor médio de $I(x_k)$ dentro do range completo de valores discretos é dado por

$$H(X) = \mathbb{E}[I(x_k)] = \sum_{k=-n}^{n} p_k I(x_k) = -\sum_{k=-n}^{n} p_k \log p_k. \tag{2.12}$$

A quantidade $H(X)$ é denominada de *entropia* de uma variável aleatória X num conjunto finito e de valores discreto. A entropia $H(X)$ pode ser vista como a *quantidade média de informação transmitida por mensagem*. Podemos observar então que a entropia $H(X)$ é limitada por

$$0 \leq H(X) \leq \log(2k+1) \tag{2.13}$$

onde $2k+1$ é o número total de níveis discreto. Daí podemos concluir que $H(X) = 0$ se, e somente se, a probabilidade $p_k = 1$ para algum k e as probabilidades restantes são todas nulas; este valor na entropia nos diz que não há incertezas. $H(X) = \log(2k+1)$ se, e somente se, $p_k = {}^{1}/_{(2k+1)}$ para todo *k*; este valor na entropia corresponde a máxima incerteza [1].

2.3.1 – Entropia em uma variável aleatória contínua.

Considere uma variável aleatória contínua X com função densidade de probabilidade $p_X(x)$. Por analogia com o conceito de entropia da variável aleatória discreta, podemos definir

$$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log p_X dx = -\mathbb{E}[\log p_X(x)] \tag{2.14}$$

Se tivermos um vetor contínuo aleatório $\boldsymbol{X}$ consistindo de n variáveis aleatórias $X_1, X_2, \ldots, X_n$. Podemos definir a entropia de $\boldsymbol{X}$ como

$$h(\boldsymbol{X}) = -\int_{-\infty}^{\infty} p_{\boldsymbol{X}}(\boldsymbol{x}) \log p_{\boldsymbol{X}} d\boldsymbol{x} = -\mathbb{E}[\log p_{\boldsymbol{X}}(\boldsymbol{x})] \tag{2.15}$$

onde $p_{\boldsymbol{X}}(\boldsymbol{x})$ é a função de densidade de probabilidade de $\boldsymbol{X}$ e $\boldsymbol{x}$ é o valor da amostra de $\boldsymbol{X}$.

2.3.2 – Princípio da Máxima Entropia

Supondo que temos um processo com um conjunto de estados conhecidos, mas com probabilidades desconhecidas, e que de alguma forma aprendemos

algumas restrições com relação à distribuição de probabilidade dos estados, cujas restrições podem ser um conjunto de valores médios ou limitados. O problema consiste em escolher um modelo probabilístico que seja ótimo em algum senso, dado um conhecimento a priori acerca deste modelo. Usualmente encontramos uma quantidade muito grande de modelos que satisfazem as restrições. A questão é: que modelo escolher?

A resposta a esta questão foi apresentada por Jaynes em 1957 com o princípio da máxima entropia que afirma: "*quando uma inferência é feita com base numa informação incompleta, deve-se utilizar uma distribuição de probabilidades que maximiza a entropia, sujeito a restrições na distribuição*." De fato, a noção de entropia define um tipo de medida sobre o espaço de distribuição de probabilidade de tal forma que as distribuições com elevada entropia sejam favorecidas. Analisando esta informação podemos concluir que o princípio da máxima entropia trata de um problema de otimização com restrições cuja solução pode ser apresentada a partir da maximização da entropia

$$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log p_X dx \qquad (2.16)$$

em todas as funções distribuições de probabilidade $p_X(x)$ de uma variável aleatória X, sujeita as seguintes restrições:

1. $p_X(x) \geq 0$,
2. $\int_{-\infty}^{\infty} p_X(x)dx = 1$
3. $\int_{-\infty}^{\infty} p_X(x)g_i(x)dx = a_i\,,\ para\ i = 1,2,\dots,m.$

Onde $g_i(x)$ é uma função de x. Podemos observar que as duas primeiras restrições descrevem duas propriedades das funções densidade de probabilidade, enquanto que a restrição 3 define os momentos da variável aleatória X, dependendo de como a função $g_i(x)$ é formulada. Na resolução deste problema de otimização com restrições, utilizamos o *método dos multiplicadores de Lagrange*, especificamente através da função Lagrangiana.

$$J(p) = \int_{-\infty}^{\infty} \left[- p_X(x) \log p_X(x) + \alpha_0 p_X(x) + \sum_{i=1}^{m} \alpha_i g_i\,(x) p_X(x) \right] dx \qquad (2.17)$$

onde $\alpha_1, \alpha_2, \ldots, \alpha_m$ são os *multiplicadores de Lagrange*. Derivando o integrando da equação (2.17) com relação a $p_X(x)$ e igualando o resultado a zero, temos:

$$-\left[\left(1\log p_X(x) + p_X(x)\frac{1}{p_X(x)}\right) + \alpha_0 + \sum_{i=1}^{m}\alpha_i g_i\,(x)\right] = 0$$

$$-1 - \log p_X(x) + \alpha_0 + \sum_{i=1}^{m}\alpha_i g_i\,(x) = 0$$

Resolvendo a equação para $p_X(x)$, temos:

$$\exp\left(-1 - \log p_X(x) + \alpha_0 + \sum_{i=1}^{m}\alpha_i g_i\,(x)\right) = \exp(0)$$

$$p_X(x) = \exp\left(-1 + \alpha_0 + \sum_{i=1}^{m}\alpha_i g_i\,(x)\right) \tag{2.18}$$

Os multiplicadores de lagrange da equação (2.18) são escolhidos a partir das restrições 2 e 3. A equação (2.18) define a distribuição de máxima entropia para o problema [1].

2.4 – Informação Mútua.

Considere duas variáveis aleatórias X e Y com distribuição conjunta $p(x,y)$ e distribuições marginais $p(x)$ e $p(y)$. A informação mútua $I(X;Y)$ é a entropia entre a distribuição conjunta e o produto das marginais.

$$I(X;Y) = \mathbb{E}[I(x_k,y_i)] = \sum_i\sum_k p(x_k,y_i)\log\frac{p(x_k|y_i)}{p(x_k)}$$

$$= \sum_i\sum_k p(x_k,y_i)\log\frac{p(x_k,y_i)}{p(x_k)p(y_i)} \tag{2.19}$$

Ou no caso contínuo

$$I(X;Y) = \mathbb{E}[I(x_k,y_i)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right)dxdy \tag{2.20}$$

2.4.1 – Propriedades

1. Não negatividade: A informação mútua $I(X;Y)$ é sempre não negativa, ou seja, $I(X;Y) \geq 0$

A Informação mútua só é nula se, e somente se, as variáveis aleatórias X e Y foram estatísticamente independentes.

2. $I(X;Y) = I(Y;X)$;
3. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(X|Y)$;
4. $I(X;Y) = H(X) + H(Y) - H(X|Y)$.

Podemos agora começar a definir a entropia conjunta de um par de variáveis aleatórias X e Y como

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log p(x,y) = -\mathbb{E}_{X,Y}\left[\log p(X,Y)\right] \qquad (2.21)$$

Do mesmo modo podemos definir a entropia condicional de X e Y como

$$H(X|Y) = -\sum_{x}\sum_{y} p(x,y)\log p(y|x) = -\mathbb{E}_{X,Y}\left[\log p(y|x)\right] \qquad (2.22)$$

e a relação entre informação mútua e suas propriedades pode ser visualizada através do diagrama de venn, apresentado na seguinte figura.

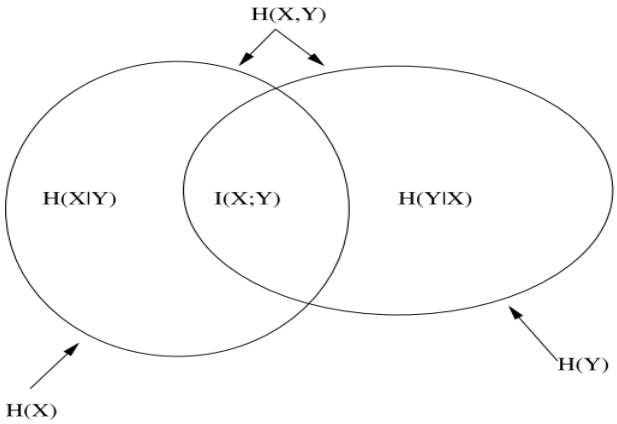

Figura 04 – Relação entre a entropia e a informação mútua.

2.5 – Divergente de Kullback – Leibler

Até agora, estamos relacionando o conceito de entropia com incerteza. Parece razoável que a diminuição da entropia deve resultar em um ganho de informação, mas não temos um tratamento sistemático deste aspecto. O divergente *KL* (Kullback – Leibler) está associado com o conceito de ganho de informação, o qual é considerado por muitos como o conceito fundamental na *IT*.

O divergente de Kullback – Leibler, determina a quantidade de informação inserida em um sistema descrito por uma probabilidade $p(x)$ quando atribui-se uma outra medida $q(x)$ a este sistema [24]. Representa a "distância" entre duas distribuições de probabilidade e é definida como

$$D_{KL}(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right]. \tag{2.23}$$

O divergente *KL* é efetivamente uma medida de dissimilaridade entre duas distribuições de probabilidade. No entanto, o termo divergente se dá devido ao fato dele somente obedecer a um dos postulados da distância, ou seja o divergente é não negativo, mas não é simétrico e não obedece a desigualdade triangular.

A informação mútua é um caso particular do divergente *KL*, que é obtida quando $p(x)$ é uma *pdf* conjunta de x e y, e $q(x)$ é o produto das probabilidades marginais [24]:

$$D_{KL}(p(X,Y)||q(X,Y)) = I(X,Y). \tag{2.24}$$

O divergente *KL* para uma variável contínua é dado por

$$D_{KL}(p(x)||q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right]. \tag{2.25}$$

A partir desta breve explanação dos conceitos iniciais podemos repensar o modelo da figura 01 para melhor adequá-lo a um problema de identificação, filtragem ou regressão do ponto de vista de um funcional de custo baseado em *ITL*.

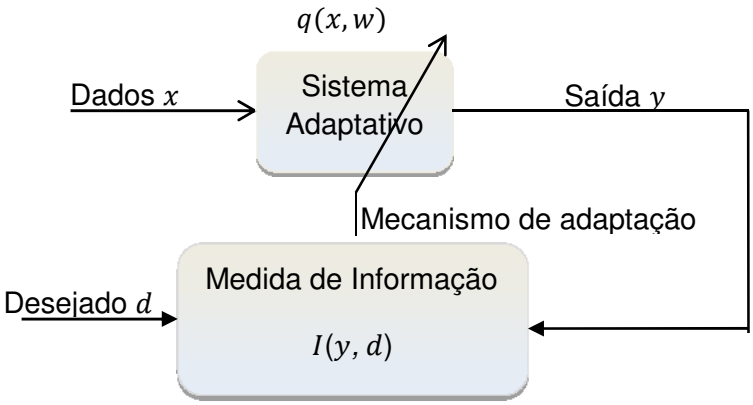

Figura 05 – Aprendizagem baseada num funcional de custo utilizando *ITL*.

Ao analisarmos a figura 05, podemos ver que o funcional de custo recebe duas fontes de informação, a saída do sistema (y) e a resposta desejada (d), então o objetivo é minimizar o divergente (ou maximizar a informação) de forma a fazer a melhor aproximação possível entre a saída do sistema e a resposta desejada, no entanto no lugar de utilizarmos o critério do EQM iremos utilizar o critério da entropia do erro – EEC. Ou seja,

$$\begin{aligned} \min_{\mathbf{W}} D_{KL}(y|z) &= \int p(y)\log\frac{p(y)}{p(z)}dy \\ \max_{\mathbf{W}} I(y,z) &= H(y) - H(y|z) \\ \min_{\mathbf{W}} H(e) &= \int p(e)\log p(e)de. \end{aligned} \tag{2.26}$$

Podemos concluir que as informações contidas na *pdf* conjunta $p(y, d)$ devem ser transferidas da forma mais eficiente possível ao mecanismo de adaptação. Portanto, é de se esperar a máxima extração de informação da *pdf* do erro, adaptando o sistema de modo a aproximar a saída do sistema à resposta desejada em um sentido de informação. Neste caso a entropia pode ser utilizada como uma medida de incerteza do erro, de forma que a função custo pode ser utilizada para minimizar a entropia do erro [1].

No próximo capítulo iremos apresentar o conceito da entropia de Rényi, bem como o conceito de potencial de informação e a ponte entre potencial de informação e Janelamento de Parzen, estendendo o conceito de entropia para espaços funcionais.

Capítulo 3

Entropia de Rényi, Potencial de Informação e Correntropia.

No capítulo anterior vimos que *ITL* é uma ferramenta para não parametrização de sistemas adaptativos baseada no conceito de entropia e divergente. Vimos também que a entropia de Shannon e o divergente de Kullback – Leibler são talvez as duas medidas mais importantes na *Teoria da Informação – IT* e em suas aplicações. Neste capítulo apresentaremos uma generalização do conceito da entropia de Shannon e divergentes como um funcional de custo em adaptação a aprendizagem, bem como os conceitos de uma função que relaciona diretamente a entropia de Rényi e a estimativa dos dados utilizando janelamento de Parzen, envolvendo os conceitos de correlação e entropia.

3.1 – Entropia de Rényi.

Alfréd Rényi, matemático Húngaro, em 1950 estendeu o conceito da entropia de Shannon apresentando uma família paramétrica de medidas de entropia como uma generalização matemática da entropia de Shannon.

A partir da equação funcional de Cauchy, Rényi buscou uma definição geral para medidas de informação que preservassem os axiomas da probabilidade de Kolmogorov e a aditividade de eventos independentes [9].

3.1.1 – Caracterizando a Entropia de Shannon e definindo a entropia de Rényi.

Em 1949, Shannon apresentou uma maneira de medir a quantidade de informação presente em uma mensagem transmitida, ele propôs o uso da

entropia como uma medida da incerteza contida em uma distribuição de probabilidade como uma definição da informação presente nos dados.

Se considerarmos X como uma variável aleatória contínua com distribuição densidade de probabilidade $p(x)$, a entropia de Shannon pode ser definida como

$$H(X) = -\int p(x) \log p(x) dx \tag{3.1}$$

Onde a entropia é um funcional da *pdf* $p(x)$ e pode ser denotada como $H(p)$. Se considerarmos X como sendo uma variável aleatória discreta com função massa de probabilidade $P\{X = x_k\} = p_k,\ (k = 0,1, \dots, n)$ a entropia (discreta) de Shannon pode ser definida como

$$H(X) = \sum_{k=1}^{n} p_k \log p_k \tag{3.2}$$

Alguns postulados foram apresentados para caracterizar a medida de entropia de Shannon, dentre eles podemos citar:

1. $H(P) = H(p_k),\ (k = 0,1, \dots, n)$ é uma função simétrica dos elementos p_k;
2. $H(p, 1-p)$ é uma função contínua em p para $0 \leq p \leq 1$;
3. $H\left({}^{1}/_{2}, {}^{1}/_{2}\right) = 1$;
4. $H[tp_1, (1-t)p_1, p_2, \dots, p_n] = H(p_1, p_2, \dots, p_n) + p_1 H(t, 1-t)$ para qualquer distribuição $P = p_k,\ (k = 0,1, \dots, n)$ e para $0 \leq t \leq 1$.

A prova destes postulados caracteriza a equação (3.2). Inicialmente vamos considerar duas distribuições de probabilidade $P = p_k,\ (k = 0,1, \dots, n)$ e $Q = q_j,\ (j = 0,1, \dots, m)$ em que $(P.Q)$ é o produto direto das distribuições P e Q consistindo nos números $p_k.q_j$. Então, temos a equação (3.2) na forma

$$H[P.Q] = H[P] + H[Q], \tag{3.3}$$

que denota uma das mais importantes propriedade da entropia, denominada de *aditividade*, ou seja, a entropia de um evento combinando dois experimentos independentes é igual a soma das entropias de cada experimento. No entanto, a equação (3.3) não pode substituir o postulado (4), e que existem muitas outras

medidas de entropia que atendem aos postulados (1), (2) e (3) e a equação (3.3), dentre estas medidas Rényi apresentou [9]

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\int p^\alpha(x)dx, \tag{3.4}$$

Para o caso contínuo e,

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_{k=1}^{n} p_k^\alpha\right), \tag{3.5}$$

Para o caso discreto, onde $\alpha > 0$ e $\alpha \neq 1$. A medida $H_\alpha(X)$ é denominada de entropia de Rényi ou medida de informação de Rényi de ordem α. Vale citar que $\lim_{\alpha\to 1} H_\alpha(X) = H(X)$, equivale a medida da entropia de Shannon, desta forma podemos ver que a entropia de Shannon é uma caso particular da entropia de Rényi. Conforme pode ser demonstrado

$$\lim_{\alpha\to 1} H_\alpha(X) = \lim_{\alpha\to 1}\frac{1}{1-\alpha}\log\int p^\alpha(x)dx$$

$$= \frac{\lim_{\alpha\to 1}\frac{1}{\int p^\alpha(x)dx}\int logp(x).p^\alpha(x)dx}{\lim_{\alpha\to 1} -1} = -\int p(x)\log p(x)dx = H(X),$$

desta forma fica demonstrado que a entropia de Shannon é um caso particular da entropia de Rényi quando $\alpha \to 1$. Se observarmos podemos concluir que na entropia de Shannon a função probabilidade é ponderada pelo termo do logaritmo, enquanto que na entropia de Rényi o logaritmo aparece fora do termo que envolve a potência α da função probabilidade. Para não confundir os momentos da função distribuição de probabilidade e da função massa de probabilidade com os momentos dos dados, Príncipe em [1] nominou o termo do logaritmo como potencial de informação α (IP_α), ou seja:

$$V_\alpha(X) \triangleq \int p^\alpha(x)dx, \tag{3.6}$$

ou

$$V_\alpha(X) \triangleq \sum_{k=1}^{n} p_k^\alpha, \tag{3.7}$$

de modo que podemos reescrever a entropia de Rényi como:

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\big(V_\alpha(X)\big) = -\log\left(\sqrt[\alpha-1]{V_\alpha(X)}\right), \tag{3.8}$$

Desta forma podemos observar que a entropia de Rényi é muito mais flexível devido ao parâmetro α, derivando assim uma família paramétrica de entropias permitindo várias medidas de incertezas (ou dissimilaridades) dentro de uma dada distribuição [1].

3.1.2 – Entropia quadrática de Rényi.

Se considerarmos $\alpha = 2$ na equação (3.4) ou (3.5) podemos observar um caso particular da entropia de Rényi, ou seja teremos uma função monotônica e decrescente do potencial de informação da função distribuição de probabilidade – *pdf* ou da função massa de probabilidade – *pmf* $p(x)$. H_2 implicitamente utiliza a distância euclidiana do ponto $p(x)$ no simplex para a origem do espaço [1].

$$H_2(X) = -\log\int p^2(x)dx, \tag{3.9}$$

para o caso contínuo e,

$$H_2(X) = -\log\left(\sum_{k=1}^{n} p_k^2\right) \tag{3.10}$$

para o caso discreto.

3.2 – Estimador quadrático de Rényi.

O problema de estimar entropia aparece em muitas áreas do conhecimento, embora existam muitas abordagens que permite estimar a entropia de uma variável aleatória, geralmente recorre-se a uma estimativa não paramétrica, desde que as amostras não pertençam a uma família paramétrica de *pdf* ‘s conhecidas, caso contrário estima-se a *pdf* e em seguida calcula-se a entropia [6]. O Estimador quadrático de Rényi propõe uma abordagem direta de

estimativa da entropia a partir das amostras, estimando $\mathbb{E}[\hat{p}(X)]$, como um escalar [1].

Supondo que temos *n* amostras independentes e identicamente distribuídas $\{x_1, x_2, \dots, x_n\}$, a partir dessas variáveis aleatórias, podemos estimar a *pdf* utilizando o método de estimativa de Parzen [25], utilizando uma função *kernel* $k(.)$ ou seja,

$$\hat{p}_X(x) = \frac{1}{N\sigma}\sum_{i=1}^{n} k\left(\frac{x - x_i}{\sigma}\right), \tag{3.11}$$

onde σ é o tamanho, largura do *kernel* ou parâmetro de suavização. A função *kernel* deve satisfazer a algumas condições, tais como:

1. $\int_{-\infty}^{\infty} k(x)dx = 1$
2. $k(x) \geq 0$
3. $\lim_{x\to\infty} |xk(x)| = 0.$

Geralmente, mas nem sempre, $k(.)$ será uma função densidade de probabilidade simétrica e a qualidade do estimador é normalmente quantificado pelo *bias* e a variância, de modo que o melhor tamanho de *kernel* é uma relação entre o *bias* e a variância do estimador [25]. No entanto, para o potencial de informação $V_2(X)$, estamos apenas interessados em estimar um escalar $\mathbb{E}[\hat{p}(X)]$. Embora existam diferentes *kernels* que atendem ao requisito, a função *kernel* gaussiana é a mais utilizada, de modo que

$$\mathbb{G}_\sigma(x, x_i) = \frac{1}{(2\pi\sigma^2)^{d/2}} exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right), \tag{3.12}$$

onde d representa a dimensão dos dados, σ a largura do *kernel*.

Para um tamanho de *kernel* fixo, temos que o valor esperado da estimativa da *pdf* é a convolução da função *kernel* com a *pdf* dos dados quando $n \to \infty$. Ou seja,

$$\mathbb{E}[\hat{p}(x)] = \lim_{n\to\infty} \hat{p}(x) = p(x) * \mathbb{G}_\sigma(x, x_i).$$

Assumindo uma função *kernel* do tipo gaussiana na estimativa da *pdf* utilizando o método de Parzen e substituindo na expressão da entropia quadrática de Rényi (equação (3.9)), obtemos o seguinte estimador

$$\widehat{H}_2(X) = -\log \int_{-\infty}^{\infty} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{G}_{\sigma}(x - x_i) \right)^2 dx$$

$$= -\log \frac{1}{n^2} \int_{-\infty}^{\infty} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{G}_{\sigma}(x - x_j) \mathbb{G}_{\sigma}(x - x_i) \right) dx$$

$$= -\log \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{-\infty}^{\infty} \mathbb{G}_{\sigma}(x - x_j) \mathbb{G}_{\sigma}(x - x_i)\, dx$$

$$= -\log \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{G}_{\sigma\sqrt{2}}(x_j - x_i) \right). \qquad (3.13)$$

A escolha do *kernel* gaussiano se deve ao fato de que outros *kernels* não apresentam uma resolução conveniente da integral, visto que a gaussiana mantém a forma funcional sob convolução. No entanto, qualquer função definida positiva com picos na origem (a maioria dos *kernels*) pode ser utilizada na estimativa, mas as expressões ficam um pouco mais complicadas. Lembrando que o argumento do logaritmo já foi denominado anteriormente de potencial de informação (*IP*), ou seja, o estimador quadrático.

3.3 – Potencial de Informação - *IP*.

Lembrando a definição de entropia quadrática utilizada na equação (3.9). E, sabendo que o logaritmo é uma função monotonicamente crescente, a maximização / minimização da entropia quadrática pode ser equivalente à minimização / maximização do argumento do logaritmo, que foi definido como sendo o potencial de informação – *IP* que pode ser estimado diretamente dos dados através da seguinte expressão [6]:

$$\hat{V}_{2,\sigma}(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{G}_{\sigma\sqrt{2}}(x_j - x_i) = \mathbb{E}[\hat{p}(x)] \tag{3.14}$$

onde $\hat{V}_{2,\sigma}(X)$ é o potencial de informação que depende de σ. Podemos observar também que,

$$\hat{V}_{2,\sigma}(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{G}_{\sigma\sqrt{2}}(x_j - x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\Phi(x_j), \Phi(x_i)\rangle$$

$$\hat{V}_{2,\sigma}(X) = \langle\frac{1}{n}\sum_{j=1}^{n}\Phi(x_j), \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\rangle \tag{3.15}$$

onde $\hat{V}_{2,\sigma}(X)$ é o vetor média dos dados mapeados / transformados pela função Φ, ou seja, o potencial de informação pode ser expresso como a norma quadrática do vetor média dos dados no espaço de características de Mercer.

Desta forma podemos concluir que a equação (3.14) é um dos principais resultados obtido em ITL, ou seja, o potencial de informação, um escalar, pode ser estimado diretamente das amostras sem a necessidade de resolver integrais, utilizando *kernels* gaussianos, o *IP* é semelhante aos estimadores convencionais de média e variância que trabalham diretamente com as amostras, independente da *pdf*, mas infelizmente este estimador possui um parâmetro livre que deve ser escolhido (largura do *kernel*). Portanto, quando o *IP* é estimado, os valores resultantes da entropia dependem do tamanho do kernel selecionado [1]. Silverman em [23] apresenta algumas sugestões de escolha deste parâmetro livre.

O Potencial de informação criou uma relação direta entre teoria da informação e um espaço de Hilbert associado a um *kernel* de tal forma que reproduz (através do produto interno) cada função no espaço. Este espaço é denominado de RKHS - *Reproducing Kernel Hilbert Space*. Desta forma as funções de custo baseadas em ITL, quando estimadas utilizando o método de janelamento de Parzen, também podem ser expressas utilizando o produto interno num espaço de características, que pode ser definido pela função *kernel* de Parzen, sugerindo uma estreita relação entre ITL e os métodos de *kernel* (aprendizagem estatística). Ou seja, se tivermos um conjunto de dados, e o correspondente

conjunto de dados mapeados (transformados), verifica-se que a média quadrática dos vetores mapeados (transformados), é igual ao potencial de informação definido na equação (3.14) para uma função *kernel* baseada no teorema de Mercer [26].

Em 2006 I. Santamaria e outros, inspirados nos conceitos de ITL apresentados, propuseram uma função correlação generalizada. Esta função está diretamente relacionada com a entropia quadrática de Rényi e a estimativa de dados utilizando o método de Parzen e, foi denominada de *Correntropia* [16].

3.4 – Correntropia

Esta nova medida de correlação generalizada inclui tanto as informações da distribuição dos dados quanto da estrutura temporal de um processo. Esta medida pode ser interpretada como um método de *kernel* com o ponto de vista da ITL demonstrando algumas propriedades bem relevantes. Ou seja, define uma função de correlação generalizada em termos de um produto interno vetorial num espaço de características definido pelo *kernel*, sendo o produto interno uma medida de similaridade no espaço de Hilbert, então essa função mede o efeito da interação entre variáveis aleatórias.

Por outro lado do ponto de vista da ITL, esta nova medida quantifica a forma e o tamanho de um conjunto de dados no espaço característica, o que dá uma informação da distribuição estatística no espaço de entrada, pois a função distribuição de probabilidade estimada com *kernels* de Parzen pode ser vista como a definição de um campo potencial de informação sobre um espaço de amostras, então, é interessante utilizar o IP para definir medidas de similaridade neste espaço e que não possuem a limitação dos momentos convencionais [1]. Pensando nisso, Santamaria e outros, propuseram em [16] uma nova medida de similaridade, denominada *correntropia*.

3.4.1 – Definição e propriedades da correntropia.

A correntropia pode ser definida a partir de uma regra básica no estudo de funções aleatórias de segunda ordem e interpretado pelo seu *kernel*. Parzen em [27] definiu o *kernel* covariância R de uma função aleatória de segunda ordem

$\{X(t), t \in T\}$ como uma função no espaço produto $T\ x\ T$, com valor, em cada t_1 e t_2 na variável T, dado por

$$R(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)]. \tag{3.16}$$

Parzen também demonstrou que as propriedades de continuidade, diferenciabilidade e integrabilidade do *kernel* levam as mesmas propriedades das funções aleatórias. Podemos utilizar essa definição para variáveis e vetores aleatórios. A transformação não linear induzida pelo mapeamento através do *kernel* de Mercer para o espaço de características, onde a função correntropia $V(X,Y)$ pode ser definida como uma função em $\mathbb{R}$ e dada por [16]:

$$V(X,Y) = \mathbb{E}[k_\sigma(X,Y)] = \iint k_\sigma(x,y)\, p_{x,y}(x,y)dxdy \tag{3.17}$$

onde $\mathbb{E}[\cdot]$ denota o operador matemático do valor esperado sobre as variáveis aleatórias X, Y e k_σ uma função *kernel* definida positiva com largura de banda σ. Iremos considerar esse *kernel* como sendo Gaussiano definido na equação (3.12) que é simétrico e invariante a translação.

Na prática, a *pdf* conjunta é desconhecida e somente dispomos de um conjunto finito de dados $\{(x_i, y_i),\ i = 1,2,\cdots N\}$ distribuídos de acordo com alguma distribuição de probabilidade que pode ser estimada através de estimadores de Parzen. Utilizando o estimador de Parzen com kernel gaussiano simétrico e bidimensional para estimar a pdf conjunta, temos:

$$\hat{p}_{X,Y,\sigma}(x,y) = \frac{1}{N}\sum_{i=1}^{n} \mathbb{G}_\sigma\left(\begin{bmatrix}x\\y\end{bmatrix} - \begin{bmatrix}x_i\\y_i\end{bmatrix}\right), \tag{3.18}$$

onde

$$\mathbb{G}_\sigma\left(\begin{bmatrix}x\\y\end{bmatrix}\right) = \frac{1}{2\pi\sqrt{|\sum|}} exp\left(-\frac{1}{2}\left(\begin{bmatrix}x\\y\end{bmatrix}^T \sum^{-1} \begin{bmatrix}x\\y\end{bmatrix}\right)\right), \sum = \begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}$$

de modo que,

$$\mathbb{G}_\sigma\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \mathbb{G}_\sigma(x)\mathbb{G}_\sigma(y),$$

então,

$$p_{X,Y}(x,y) \approx \hat{p}_{X,Y,\sigma}(x,y) = \frac{1}{N}\sum_{i=1}^{n} \mathbb{G}_\sigma\,(x - x_i)\mathbb{G}_\sigma(y - y_i), \tag{3.19}$$

quando a largura do kernel tende a zero e o produto $N\sigma$ tende ao infinito, de acordo com a condição do método de Parzen, a igualdade vale para a equação (3.19) [1]. Integrando tal equação ao longo da reta $x = y$, temos:

$$\int \hat{p}_{X,Y,\sigma}(x,y) = \int \frac{1}{N}\sum_{i=1}^{N} \mathbb{G}_\sigma\,(x - x_i)\mathbb{G}_\sigma(y - y_i)\,dxdy$$

$$\int \hat{p}_{X,Y,\sigma}(x,y)\big|_{x=y=u} = \frac{1}{N}\sum_{i=1}^{N} \int \mathbb{G}_\sigma(u - x_i)\mathbb{G}_\sigma(u - y_i)\,du$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{G}_{\sqrt{2}\sigma}\,(x_i - y_i) = \hat{V}_\sigma(X,Y) \tag{3.20}$$

ou seja, a integral da estimativa de Parzen com *kernel* Gaussiano, é exatamente a estimação da correntropia com largura de *kernel* $\sqrt{2}\sigma$ [1]. A seguinte figura ilustra intuitivamente como a correntropia fornece a pdf do evento $p(X = Y)$.

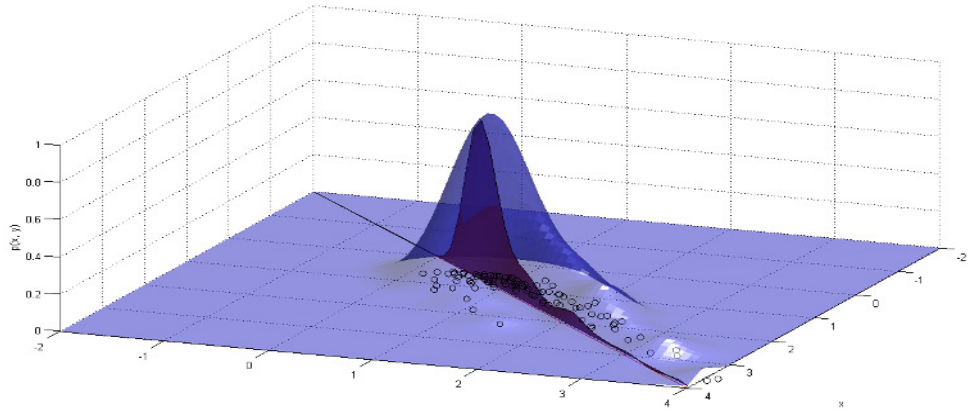

Figura 06 – Correntropia como a integral no espaço gaussiano ao longo da reta $x = y$.

Se utilizarmos a expansão em série de Taylor para o *kernel* Gaussiano, temos que a correntropia pode ser reescrita como:

$$V(X,Y) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\, 2^n \sigma^{2n}} \mathbb{E}[\|X-Y\|^{2n}] \tag{3.20}$$

que envolve todos os momentos de ordem par da variável aleatória $\|X-Y\|$. Especificamente, o termo correspondente para $n=1$ na equação (3.20) é proporcional a [16]:

$$\begin{aligned} &\mathbb{E}[\|X\|^2] + \mathbb{E}[\|Y\|^2] - 2\mathbb{E}[\langle X,Y\rangle] \\ &\quad = \sigma_X^2 + \sigma_Y^2 - 2\,R(X,Y) \end{aligned} \tag{3.21}$$

analisando a equação (3.21) podemos perceber que a função covariância convencional é um caso particular da correntropia, e que a informação contida nessa é incluída na nova função [16]. Desta forma vimos claramente a correntropia como uma generalização da correlação e como uma estimativa do *IP* para variáveis aleatórias a partir do método de Parzen.

3.4.1.1 – Propriedades.

Algumas importantes propriedades da correntropia definida em (3.17) são apresentadas a seguir [16]:

1 - Para um *kernel* simétrico definido positivo a correntropia é um RKHS (definido em A.1).

Demonstração: seja $\mathbb{H}$ um espaço de Hilbert de funções no conjunto X, e $k(X,Y)$ um kernel simétrico definido positivo no $\mathbb{R}^2$. Então a função $V(t_1,t_2)$ é dita simétrica e definida positiva se para um conjunto finito de pontos $\{x_1, x_2, \dots, x_n\} \in X$ e para qualquer conjunto de números reais correspondente, nem todos nulos $\{\alpha_1, \alpha_2, \dots, \alpha_n\} \in \mathbb{R}$, temos que:

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, k(X_i, X_j)\right] > 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\,\mathbb{E}\big[k(X_i,X_j)\big] = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\,V(i,j)\ > 0.$$

Como demonstrado a correntropia é um operador simétrico e definido positivo e de acordo com o teorema de Moore – Aronszajn [28] todo *kernel* real simétrico e definido positivo de duas variáveis, existe um único RKHS com um *reproducing kernel* k. De modo, a concluir que a correntropia é um *reproducing kernel*.

2 - A correntropia é positiva e limitada, e para um kernel gaussiano temos que $0 < V(X,Y) < {}^{1}/_{\sigma\sqrt{2\pi}}$. Então, a correntropia é máxima se, e somente se, $X = Y$.

3 - Para um *kernel* gaussiano a correntropia é uma soma ponderada de todos os momentos de ordem par da variável aleatória $X - Y$. Analisando a equação (3.20), vemos que a largura do *kernel* aparece como um termo de ponderação do momento de segunda ordem $(n = 1)$ e os momentos de ordem mais elevada. Com $\sigma > 1$, os momentos de alta ordem somem mais rapidamente devido ao denominador da equação (3.20), de modo que o momento de segunda ordem tende a dominar a abordagem. Então, o problema da largura do *kernel* em correntropia é diferente do problema de estimativa da pdf [1].

4 - Quando a largura do *kernel* tende a zero, o valor da correntropia se aproxima do valor da probabilidade $p(X = Y)$; que é dado por:

$$\begin{aligned}\lim_{\sigma\to 0} V(X,Y) &= \lim_{\sigma\to 0}\iint \mathbb{G}_\sigma(x-y)\,p_{X,Y}(x,y)dxdy \\ &= \iint \delta\,(x-y)p_{X,Y}(x,y)dxdy = \iint p_{X,Y}(x,y)dx \end{aligned} \qquad (3.22)$$

5 - Então, seja uma pdf conjunta $p_{X,Y}(x,y)$, e $\hat{p}_{X,Y,\sigma}(x,y)$ sua estimativa de Parzen com largura de *kernel* σ, de modo que a Correntropia estimada com largura de *kernel* $\sqrt{2}\sigma$ é a integral da estimativa de Parzen ao longo da reta $x = y$.

6 – Se *X* e *Y* são variáveis aleatórias independentes com distribuições de probabilidade $p_X(x)$ e $p_Y(y)$ de modo que a distribuição de probabilidade conjunta seja dada por:

$$p_{X,Y}(x,y) = p_X(x)p_y(Y), \tag{3.23}$$

utilizando o método de estimativa de Parzen com *kernel* gaussiano para estimar essas pdf's e integrando a equação (4.10) ao longo da reta $x = y$, temos:

$$\int \hat{p}_{X,Y,\sigma}(x,y)\big|_{x=y} = \int \frac{1}{N}\sum_{i=1}^{N} \mathbb{G}_\sigma\,(x - x_i)\mathbb{G}_\sigma(x - x_i)\,dxdx$$

$$\int \hat{p}_{X,Y,\sigma}(x,y)\big|_{x=y} = \int \left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{G}_\sigma\,(x - x_i)\right)^2 dx$$

$$\int \hat{p}_{X,Y,\sigma}(x,y)\big|_{x=y} = \frac{1}{N^2}\int \left(\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{G}_\sigma(x - x_i)\mathbb{G}_\sigma(x - x_i)\right)dx$$

$$\int \hat{p}_{X,Y,\sigma}(x,y)\big|_{x=y} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int \mathbb{G}_\sigma(x - x_i)\mathbb{G}_\sigma(x - x_i)\,dxdx$$

$$\hat{p}_{X,Y,\sigma}(x,y) = \hat{V}_\sigma(X,Y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{G}_{\sqrt{2}\sigma}\big(x_i - x_j\big) = \hat{V}_{2,\sigma}(X), \tag{3.24}$$

que é uma estimativa do potencial de informação (*IP*) definido na equação (3.14), então podemos concluir que uma estimativa do *IP* para variáveis aleatórias independentes é na verdade uma estimativa da correntropia. Então do ponto de vista dos métodos de *kernel*, $p_X(\cdot) = \mathbb{E}[\varphi(X)], p_Y(\cdot) = \mathbb{E}[\varphi(Y)]$ são dois pontos no RKHS, e o IP é o produto interno entre dois vetores criados por essas duas pdf's [1].

Assim, o estimador baseado no método de estimação de Parzen pode ser expresso em termos de um produto interno no espaço definido por Mercer [24]. De modo que podemos reescrever a equação (3.14) para obter:

(3.25)

$$\hat{V}_{2,\sigma}(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{G}_{\sqrt{2}\sigma}(x_i - x_j) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle\varphi(x_i), \varphi(x_j)\rangle$$

$$\hat{V}_{2,\sigma}(X) = \langle\frac{1}{N}\sum_{j=1}^{N}\varphi(x_i), \frac{1}{N}\sum_{j=1}^{N}\varphi(x_i)\rangle = \|\boldsymbol{m}\|^2.$$

Ou seja, podemos observar que a estimativa do potencial de informação pode ser expressa como a norma quadrática de um vetor de dados no RKHS definido pelo *kernel* [1]. A partir deste conceito podemos definir a próxima propriedade da correntropia.

7 – Se *X* e *Y* são variáveis aleatórias estatísticamente independentes, então:

$$V(X, Y) = \langle\mathbb{E}(X), \mathbb{E}(Y)\rangle, \tag{3.26}$$

onde ⟨·,·⟩ representa o produto interno no RKHS definido pelo *kernel*.

8 – A correntropia representa uma simplificação da estatística de segunda ordem dos dados projetados no RKHS definido por uma função *kernel*.

Demonstração: Pelo teorema de Mercer, um kernel positivo e simétrico pode ser decomposto como:

$$k(x, y) = \sum_{i=0}^{\infty}\lambda_i\varphi_i(x)\varphi_i(y) = \langle\Phi(x), \Phi(y)\rangle$$

$$\Phi: x \mapsto \sqrt{\lambda_i}\varphi_i(x), \qquad i = 1,2,\cdots,$$

onde $\{\varphi_i(x), i = 1,2,\cdots\}$ e $\{\lambda_i(x), i = 1,2,\cdots\}$ são sequências de autofunções e os correspondentes autovalores do *kernel*, respectivamente, e ⟨·,·⟩ representa o produto interno entre dois vetores de dimensão infinita $\Phi(x)$ *e* $\Phi(y)$. Pelo teorema de Moore – Aronszajn este kernel determina um RKHS de alta dimensionalidade, onde uma transformação não linear mapeia os dados originais na superfície de uma esfera no RKHS. Então, baseado em um kernel simétrico e definido positivo, a correntropia pode ser interpretada como [30]:

$$V(X, Y) = \mathbb{E}[\langle\Phi(x), \Phi(y)\rangle] = \mathbb{E}[\Phi(x)^T\Phi(y)] \tag{3.27}$$

admitindo que a dimensão do espaço de características é *M* (infinito se o *kernel* for Gaussiano) e que temos um mapeamento dos dados dado pelo *kernel*, ou seja, o vetor $\Phi(X) = [\varphi_1(X)\ \varphi_2(X) \cdots \varphi_M(X)]^T$. Então a estatística de segunda ordem entre os dois vetores $\Phi(X)$ e $\Phi(Y)$ pode ser expresso pela seguinte matriz de correlação [1]:

$$\begin{aligned} R_{XY} &= \mathbb{E}[\Phi(X)\Phi(Y)^T] \\ &= \begin{bmatrix} \mathbb{E}[\varphi_1(X)\varphi_1(Y)] & \cdots & \mathbb{E}[\varphi_1(X)\varphi_M(Y)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\varphi_M(X)\varphi_1(Y)] & \cdots & \mathbb{E}[\varphi_M(X)\varphi_M(Y)] \end{bmatrix} \end{aligned} \tag{3.28}$$

no entanto, a partir da equação (3.28) podemos concluir que

$$V(X,Y) = \mathbb{E}[\Phi(x)^T\Phi(y)] = traço\ (R_{XY}) \tag{3.29}$$

ou seja, se o *kernel* é definido positivo então a matriz é definida positiva e o traço da matriz de correlação é igual a soma dos autovalores, que demonstra que a correntropia representa uma simplificação da estatística de segunda ordem no espaço de características [1].

Como comentado no capítulo, em 2006 foi proposta uma nova medida de correlação generalizada que está diretamente relacionada com a entropia quadrática de Rényi e a estimativa de dados utilizando janelamento de Parzen. Iremos estender no próximo capítulo o conceito da correntropia entre duas variáveis aleatórias para vetores aleatórios *L* – dimensionais.

Capítulo 4

Extensões Multidimensionais para Correntropia.

O escopo deste capítulo segue a linha de raciocínio do capítulo anterior, considerando o caso em que duas ou mais variáveis aleatórias são consideradas em conjunto. Ou seja, estávamos colocando ao longo do capítulo anterior a medida de similaridade apenas entre duas variáveis aleatórias considerando o $\mathbb{R}^2$, no entanto é possível generalizar esse conceito não somente considerando a densidade ao longo de retas em duas dimensões (como *X* = *Y*), mas densidades ao longo de linhas e hiperplanos em espaços *L* – dimensionais utilizando o estimador de Parzen para funções densidade conjunta multidimensional.

Supondo que dispomos de um vetor *L* – dimensional aleatório, cujos componentes sejam variáveis aleatórias. Então supondo *N* amostras, iremos considerar o estimador de Parzen da *pdf* (utilizando um *kernel* gaussiano simétrico e radial) como:

$$\hat{p}(x_1, x_2, \cdots, x_n) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{L/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L}\left(x_j - x_j^i\right)^2\right)\right) \qquad (4.1)$$

vamos nos referir aos valores x_j^i como sendo o *j – ésimo* componente (de dimensão *L*) do dado *i*. A seguir analisaremos a correntropia para diferentes configurações na distribuição das variáveis aleatórias, sendo que cada caso considerado consistirá em um caso particular de um espaço amostral *L* – dimensional.

4.1 – *Correntropia para a linha* $x_1 = x_2 = \cdots = x_L$ (*Extensão 01*)

Esta é a mais simples extensão e consiste no cálculo de uma integral de linha definida por um vetor de componentes idênticos (o bissetor do primeiro quadrante no espaço *L* – dimensional). Ou seja, consiste na resolução da seguinte integral:

$$V(x_1, x_2, \cdots, x_L) = \int \cdots \int \hat{p}(x_1, x_2, \cdots, x_L)\delta(x_1 = x_2 = \cdots x_L)dx_1 \cdots dx_L \,. \tag{4.2}$$

A notação delta utilizada pode ser entendida como um delta multidimensional que está localizado no espaço em que as condições em seu interior são verdadeiras. Podemos escrever essa integral como uma integral simples em que todos os parâmetros da *pdf* conjunta são iguais, ou seja:

$$V_X = \int \hat{p}(x, x, \cdots, x)dx$$

isto é facilmente estimado a partir da equação (4.1) de modo que,

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L}\left(x_j - x_j^i\right)^2\right)\right)dx$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L}\left(x^2 - 2xx_j^i + {x_j^i}^2\right)\right)\right)dx \tag{4.3}$$

o argumento do polinômio envolve o somatório dos termos de segunda ordem que podem ser escritos como um produto de um quadrado perfeito de termos polinomiais, ou seja:

$$\sum_{j=1}^{L}\left(x^2 - 2xx_j^i + {x_j^i}^2\right) = \sum_{j=1}^{L}x^2 - 2x\sum_{j=1}^{L}x_j^i + \sum_{j=1}^{L}{x_j^i}^2 = (x-\gamma)^2 + c^2$$

$$\sum_{j=1}^{L}\left(x^2 - 2xx_j^i + {x_j^i}^2\right) = Lx^2 - 2x\sum_{j=1}^{L}{x_j^i}^2 + \sum_{j=1}^{L}{x_j^i}^2 \quad (\div L)$$

$$x^2 - \frac{2}{L}x\sum_{j=1}^{L}{x_j^i}^2 + \frac{1}{L}\sum_{j=1}^{L}{x_j^i}^2 = x^2 - 2x\gamma + \gamma^2 + c^2$$

sendo,

$$\gamma = \frac{1}{L}\sum_{j=1}^{L}{x_j^i}^2$$

e

$$\gamma^2 + c^2 = \frac{1}{L}\sum_{j=1}^{L}{x_j^i}^2 \rightarrow c^2 = \frac{1}{L}\sum_{j=1}^{L}{x_j^i}^2 - \left(\frac{1}{L}\sum_{j=1}^{L}{x_j^i}^2\right)^2$$

onde $\gamma_i = \sum_{j=1}^{L} {x_j^i}/{L}$ e $C_i = \sqrt{\sum_{j=1}^{L} {\left(x_j^i\right)^2}/{L} - (\gamma_i)^2}$ são os fatores polinomiais. γ é o termo que completa os quadrados e C é a constante remanescente. Substituindo na equação (4.3). temos,

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{L}{2\sigma^2}(x^2 - 2x\gamma_i + \gamma_i^2) + c^2\right)dx$$

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{L}{2\sigma^2}(x^2 - 2x\gamma_i + \gamma_i^2)\right)\exp\left(-\frac{L}{2\sigma^2}C_i^2\right)dx \tag{4.4}$$

seja $\sigma' = {\sigma}/{\sqrt{L}}$ podemos reescrever a expressão como:

$$\hat{V}_X = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L^{L/2}}\frac{1}{(2\pi\sigma'^2)^{(\mathrm{L}-2)/2}}\frac{1}{(2\pi\sigma'^2)^{1/2}}\exp\left(-\frac{L}{2\sigma'^2}(\mathrm{x} - \gamma_\mathrm{i})^2\right)$$

$$\frac{1}{(2\pi\sigma'^2)^{1/2}} exp\left(-\frac{L}{2\sigma'^2}C^2\right) dx$$

como a integral é uma gaussiana completa, então a convolução entre duas gaussianas é uma gaussiana, de modo que,

$$\hat{V}_X = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L^{L/2}}\frac{1}{(2\pi\sigma'^2)^{(L-2)/2}}\mathbb{G}_{\sigma'}(\mathrm{C_i^2})$$

e finalmente, substituindo o valor de $\mathrm{C_i^2}$ na expressão anterior e fatorando, temos:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\frac{1}{(2\pi\sigma^2)^{(L-2)/2}}\mathbb{G}_{\sigma'}\left(\frac{1}{L}\sum_{j=1}^{L}x_j^{i\,2} - \left(\frac{1}{L}\sum_{j=1}^{L}x_j^{i}\right)^2\right). \qquad (4.5)$$

que representa o cálculo da correntropia em um vetor com *L* componentes idênticos em *N* amostras. Na definição original da correntropia o objetivo é maximizar a probabilidade $P(x = y)$, aqui o nosso principal objetivo é maximixar a probabilidade $P(x_1 = \cdots = x_L)$. A seguinte figura apresenta um exemplo ilustrativo.

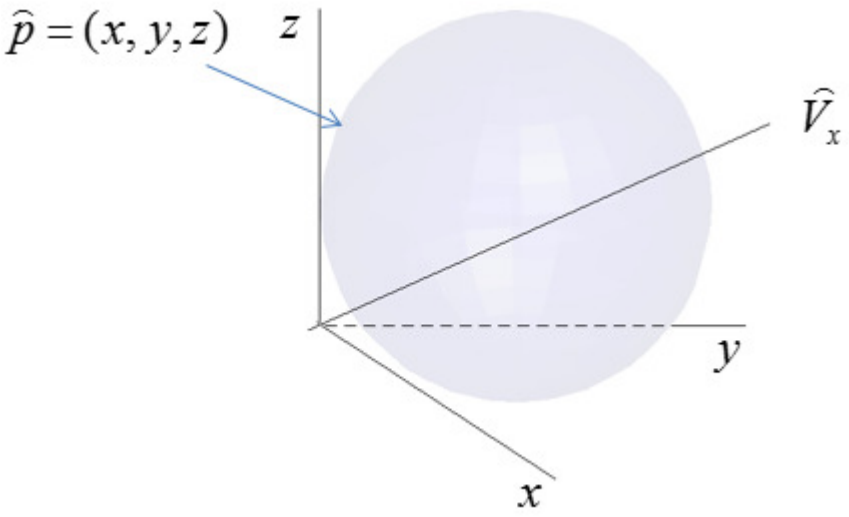

Figura 07 – reta $x = y, z = 0$.

Agora é possível estender o conceito da correntropia para retas definidas por vetores de componentes idênticas (o bissetor ao primeiro quadrante no espaço *L*-dimensional). Ao analisarmos a expressão (4.5) vemos o cálculo da

correntropia da variância entre os dados, que representa uma medida da dispersão desses em torno das médias considerando todos os valores da distribuição. Desta forma, podemos, por exemplo, utilizar essa extensão para estimar a correntropia que torne uma igualdade do tipo $aX_1 = aX_2 = \cdots = Y$ verdadeira.

Para validar o modelo iremos utilizar essa ferramenta em um problema clássico de regressão (calibração de sensores). Num problema de regressão temos uma relação $Y = f(X) + \varepsilon$, onde f representa uma função desconhecida, ε o ruído inerente ao processo e Y às observações ou medidas. O objetivo é encontrar uma função aproximada que minimize ao máximo o efeito do ruído tornando a igualdade a mais próxima possível. No problema que iremos abordar, dispomos de um conjunto de sensores onde estes efetuam medições em volts e em seguida essas medidas são convertidas em graus centígrados através de uma constante de conversão.

O nosso objetivo é encontrar a constante de conversão que melhor represente o conjunto de medições a uma temperatura tida como referência ou padrão. Neste caso, dispomos de N medidas em L sensores. Embora pareça um problema simples, pois dispomos de um número finito de sensores com a mesma taxa de conversão, uma solução natural seria a divisão da temperatura pela tensão em volts, no entanto, podemos observar que temos medidas distintas o que acarretaria diferentes valores da constante de conversão, tornando a solução inviável visto que, devemos atender para cada temperatura a seguinte igualdade:

$$aV_1(i) = \cdots = aV_j(i) = \cdots = aV_L(i) = T_i$$

onde a representa a constante de conversão de volts em graus centígrados, $V_j(i)$ a medida da temperatura i ($i = 1,..., N$) no sensor j ($j = 1,..., L$) e T_i a temperatura i do sensor padrão utilizado como referência na calibração. Como teremos que efetuar a calibração de forma que os sensores ofereçam uma resposta satisfatória a diferentes temperaturas, a formulação do problema é dada por:

$$aV_L(i) = T_i \tag{4.6}$$

neste caso podemos observar *L* equações e apenas uma variável a ser estimada e cuja solução será abordada de três maneiras distintas: Erro médio quadrático, correntropia e a extensão 01 da correntropia.

Neste problema consideraremos as grandezas aV e T sendo V, a tensão em Volts, $\underline{a}$ a constante de conversão, T a temperatura e o erro como a diferença entre ambas, ou seja, $e = aV - T$. O Erro Médio Quadrático (EMQ) entre as grandezas é definido como,

$$EMQ(aV, T) \approx \mathbb{E}\left[\left(aV_j - T_i\right)^2\right] = \frac{1}{NL}\sum_i \sum_j \left(aV_j - T_i\right)^2$$

O valor esperado entre a estimativa na conversão e a temperatura de referência. Neste caso, a similaridade entre as grandezas pode ser vista como o quão diferente aV é de T, esta ideia intuitiva nos diz que o EMQ é uma medida de similaridade entre grandezas. No entanto, o termo quadrático aumenta a contribuição de possíveis amostras que estejam distantes do valor médio da distribuição do erro, de modo que o EMQ não seja um método indicado quando a distribuição do erro possuir ruídos impulsivos (*outliers*: um determinado dado ou observação é denominado de *outlier*, se ele se afasta do padrão linear definido pelos outros dados) [31] ou média diferente de zero. No entanto, sabemos que os algoritmos utilizados na estimação de variáveis com EMQ, levam em conta os momentos de segunda ordem da distribuição do erro, que proporciona bons resultados se o erro possuir uma função distribuição de probabilidade (*pdf*) do tipo gaussiana. Entretanto, se a *pdf* do erro for de natureza não gaussiana, faz sentido utilizarmos funções de custo alternativas para a adaptação.

A Correntropia pode ser vista como uma alternativa ao EMQ, e tem sido utilizada com sucesso em diversos problemas na Engenharia. Para o nosso problema em que dispomos de duas grandezas de natureza aleatória, podemos escrever a correntropia a partir da equação (3.20) como,

$$V_a = \frac{1}{N}\sum_{i=1}^{N} exp\left[-\frac{1}{\sigma^2}\left(a\sum_{j=1}^{L} V_j - T_i\right)^2\right] \tag{4.7}$$

desta forma, podemos utilizar a correntropia para estimar a variável que torne a igualdade $aV = T$ verdadeira. Ou seja, o nosso objetivo é maximixar a probabilidade $p_{V,T}(aV = T)$. No entanto, a estimativa utilizando a correntropia apresenta um parâmetro livre (largura do *kernel*) que deve ser escolhido pelo usuário, geralmente utilizando conceitos tais como a regra de Silverman, máxima verossimilhança ou validação cruzada [23] e para sanar essa deficiência iremos apresentar a extensão 01 da correntropia na solução do citado problema. Inicialmente o nosso problema foi apresentado na equação (4.6) de modo que o nosso objetivo agora é maximizar a probabilidade para um conjunto de componentes idênticos utilizando a extensão 01 da correntropia, através do cálculo da integral de Parzen para tais componentes:

$$p_{V,T}(aV_1(i) = aV_2(i) = aV_3(i) = \cdots = aV_L(i), T_i)$$

de modo que a equação (4.5) pode ser escrita como:

$$\alpha = \frac{1}{L+1}\left(a\sum_{j=1}^{L} V_j + T_i\right)$$

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N} \exp\left\{-\frac{1}{\sigma^2}[(\alpha^2) - (\alpha)^2]\right\}. \tag{4.8}$$

Desta forma o nosso principal objetivo agora é maximizar a estimativa da distribuição de probabilidade através do *kernel* da variância. No entanto durante os experimentos notamos uma instabilidade numérica na forma apresentada, sendo assim reescrevemos a equação (4.6) como:

$$aV_1(i) = T_i = aV_2(i) = T_i = aV_3(i) = \cdots = aV_L(i) = T_i$$

o que não altera o problema original. Neste caso podemos utilizar a extensão para estimar a variável *a*, de forma que a expressão (4.8) é dada por,

$$\hat{V}_x = \sum_{i=1}^{N} \exp\left[-\frac{1}{\sigma}\left(\frac{1}{(L+1)}\left(T_i^2 + a^2\sum_{j=1}^{L} {v_j^i}^2\right) - \left(\frac{1}{(L+1)}\left(T_i + a\sum_{j=1}^{L} v_j^i\right)\right)^2\right)\right]$$

durante a simulação utilizando essa estimativa notamos uma grande sensibilidade do estimador com relação à variância nas medidas, quanto maior a variância mais o escalar ***a*** funciona como um fator de amplificação não somente nas medições como também ao ruído, pois se observarmos o termo do *kernel* na expressão (4.5), veremos que a extensão trabalha com a variância nos dados, considerando o problema proposto iremos fazer uma mudança de variável para evitarmos o efeito apresentado na expressão anterior, ou seja:

$$b + V_1^{'}(i) = T_i^{'} = b + V_2^{'}(i) = T_i^{'} = b + V_3^{'}(i) = \cdots = b + V_L^{'}(i) = T_i^{'}$$

onde $b + V_L^{'}(i) = aV_L(i)$ de modo que $\log\left(aV_L(i)\right) = \log(a) + \log(V_L(i))$ onde $\log(a) = b$ e $\log(V_L(i)) = V_L^{'}(i)$, reescrevendo o conjunto de equações, temos:

$$\log(a) + \log(V_1(i)) = log(T_i) = \cdots = \log(a) + \log(V_L(i)) = log(T_i)$$

o objetivo é otimizar o estimador para a constante ***a*** de modo que estamos escolhendo um valor tal que a probabilidade de $aV_L(i) = T_i$ seja a máxima independentemente de $aV_1(i), aV_2(i), \cdots, aV_L(i)$. Desta forma podemos reescrever a extensão 01 da correntropia como,

$$\beta = \frac{1}{2L}\left(b\sum_{j=1}^{L} V_j' + log(T_i)\right)$$

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N} \exp\left\{-\frac{1}{\sigma^2}[(\beta^2) - (\beta)^2]\right\} \qquad (4.9)$$

neste caso conseguimos tornar a extensão 01 robusta à variância nas medidas. Uma vez apresentada as equações envolvidas na solução, iremos testar os algoritmos com um conjunto de dados do problema proposto.

4.1.1 – Experimento 01: variação na largura do kernel, pouca variância no ruído e sem a presença de outliers

Primeiramente, mostraremos o desempenho dos algoritmos a diferentes valores do parâmetro σ - largura do *kernel*, o objetivo aqui é analisar o comportamento dos mesmos quando variamos o σ utilizando *kernels* gaussianos em uma faixa de valores. Este parâmetro controla as propriedades da correntropia, ou seja, controla a ênfase dada aos momentos de ordem superior sobre os momentos de segunda ordem, portanto uma escolha apropriada deste parâmetro é importante. Se, por exemplo, este parâmetro for de valor pequeno e os dados forem muito ruidosos, a correntropia não será capaz de distinguir o sinal do ruído, e se for muito grande os momentos de alta ordem da correntropia serão ignorados [1]. Nas seguintes figuras serão apresentadas estimações da correntropia (equação 4.7), a extensão 01 (equações 4.8 e 4.9) com a variação da largura do kernel, através da maximização da probabilidade sobre a constante de conversão e a minimização do erro médio quadrático. Iniciamos os testes com o parâmetro a ser estimado (*a* = 10, por exemplo), com a presença de ruídos nas medições e variamos apenas a largura do *kernel*, a seguinte figura apresenta o resultado para um σ = 0.0005.

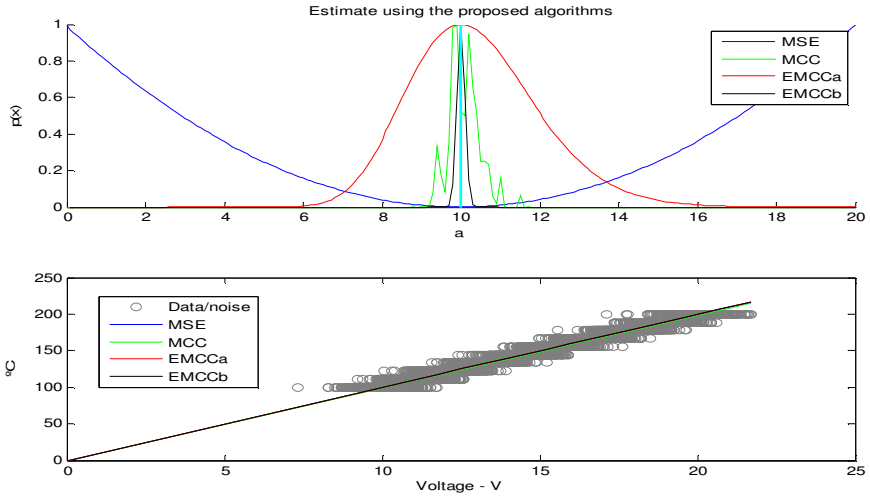

Figura 08 – gráfico comparativo entre a estimação da probabilidade utilizando o MCC, a extensão 01 da correntropia com σ = 0,0005 e o EMQ com pouca variância no ruído e sem a presença de *outliers*.

E os algoritmos forneceram os seguintes valores para a constante de conversão:

Algoritmo	Valor estimado para *a*
EMQ	10
CC	9.9
Ext01_a	10
Ext01_b	10

Ao analisarmos a figura 08 vemos claramente a maximixação da probabilidade para os algoritmos que envolvem a correntropia e uma minimização quadrática no gráfico do EMQ em torno da constante de conversão. Vemos também uma leve sensibilidade da correntropia ao parâmetro σ. Ou seja, Como estamos trabalhando na estimação da média do erro, em dados com pouca variância no ruído presente nas medições e sem a presença de *outliers*, os estimadores apresentaram bons resultados. Na figura 09 pode ser observado que uma vez sendo ajustado corretamente o parâmetro σ a correntropia apresenta uma resposta satisfatória, ou seja, ao analisarmos as figuras podemos ver claramente que a extensão 01 da correntropia é mais suave ao parâmetro de ajuste *σ* (largura do *kernel*), suprindo uma deficiência no algoritmo de estimação utilizando a correntropia convencional.

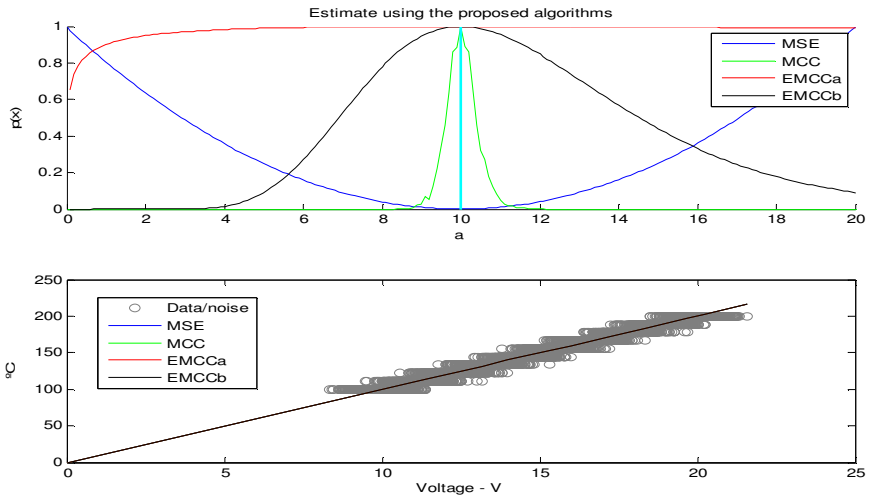

Figura 09 – gráfico comparativo entre a estimação utilizando o EMQ, MCC e a extensão 01 da correntropia com σ = 0,05 com pouca variância no ruído e sem a presença de *outliers*.

e os algoritmos forneceram os seguintes valores para a constante de conversão:

Algoritmo	Valor estimado para *a*
EMQ	10
CC	10
Ext01_a	10
Ext01_b	10

4.1.2 – Experimento 02: variação no ruído

Iremos agora analisar os algoritmos na presença de um ruído de natureza gaussiana acrescido aos dados de modo que a variância pode criar *outliers*. A largura do *kernel* para o MCC, Ext01_a e Ext01_b foram ajustados para σ = 0.05 e com uma variância no ruído de 5.0, obtendo como resultado a seguinte figura,

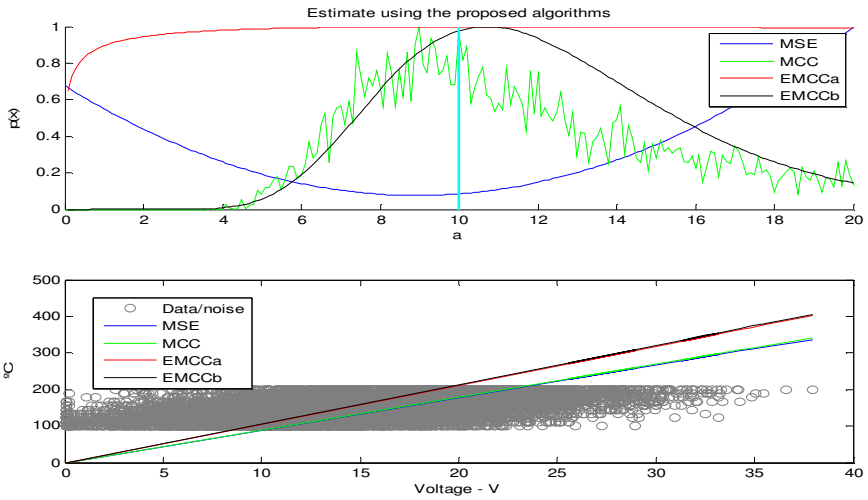

Figura 10 – gráfico comparativo entre a estimação utilizando o EMQ, MCC e a extensão 01 da correntropia com σ = 0,05 na presença de um ruído intenso.

Na figura 10 podemos observar uma significante diferença no desempenho dos algoritmos, fica claro tanto a suavidade do estimador utilizando a extensão da correntropia quanto a sua insensibilidade a grandes variações no ruído presente, sendo os valores obtidos na estimação apresentados na seguinte tabela.

Os algoritmos forneceram os seguintes valores para a constante de conversão:

Algoritmo	Valor estimado para *a*
EMQ	8
CC	8.9
Ext01_a	10.7
Ext01_b	10.5

Claramente, a extensão da correntropia é mais suave e diminui consideravelmente a potência do ruído presente nos dados.

4.1.3 – Experimento 03: percentagem e bias nos outliers

Neste experimento o principal objetivo é mostrar a robustez dos estimadores baseados na correntropia através do comportamento de tais algoritmos quando dispomos não somente de um percentual de *outliers* presentes nas medições, mas também um *bias* nestes. Primeiramente utilizamos um percentual de 10 por cento de *outliers* em um deslocamento (*bias*) de 10 unidades, obtendo o resultado visto na seguinte figura,

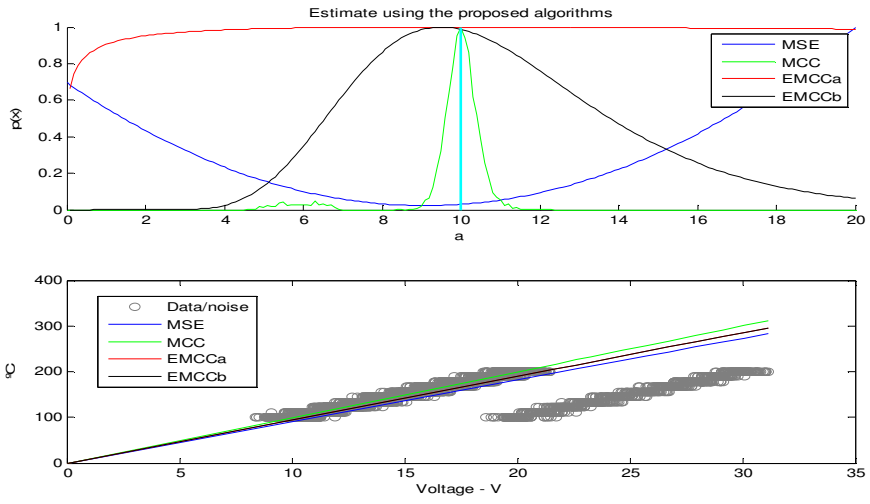

Figura 11 – gráfico comparativo entre a estimação utilizando o EMQ, MCC e a extensão 01 da correntropia com σ = 0,05 na presença de *outliers* e um *bias*.

e os algoritmos forneceram os seguintes valores para a constante de conversão:

Algoritmo	Valor estimado para *a*
EMQ	9
CC	10
Ext01_a	9.5
Ext01_b	9.5

Ao analisarmos a figura 11 podemos observar que a extensão da correntropia caracteriza-se pelas vantagens de ser insensível a largura do *kernel*, tal como o EMQ e de filtrar melhor os resultados na presença de ruídos gaussianos, desta forma concluímos que a extensão 01 da correntropia corresponde a um estimador de variância mínima sendo caracterizado pela suavidade do EMQ e a robustez da correntropia a ruídos impulsivos.

Mostramos nessa aplicação a primeira extensão multidimensional da correntropia num clássico problema de regressão (calibração de sensores), a correntropia foi utilizada com sucesso em muitos problemas práticos da Engenharia principalmente os não - lineares, de natureza não gaussiana e com ruído impulsivo, estendemos esse conceito a mais uma aplicação utilizando uma extensão multidimensional da correntropia.

4.1.4 – Um algoritmo em ponto fixo para a extensão 01:

Nesta seção apresentaremos uma solução recursiva para uma função de custo a partir da extensão 01 da correntropia, essencialmente baseada na teoria de ponto fixo e pode ser utilizada para a obtenção de estimadores ou pesos em algoritmos de métodos de *kernel.* Os algoritmos em ponto fixo possuem uma convergência de segunda ordem para uma solução ótima com poucas iterações [32].

Para obter uma solução em ponto fixo, inicialmente iremos considerar um conjunto de dados no espaço conjunto de probabilidade composto pelas entradas e uma resposta desejada, ou seja, a saída é composta da combinação linear das entradas. Tal como,

$$y = \sum_{j=1}^{L} a_j x_j^i$$

a correntropia convencional calcula a probabilidade de $\boldsymbol{a}^T\boldsymbol{x} = y$, de modo que a probabilidade será a máxima quando $\boldsymbol{a}^T\boldsymbol{x} - y \to 0$ para todas as entradas. Ou seja, podemos reescrever a correntropia da expressão (4.7) como,

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L} exp\left[-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L} a_j x_j^i - d_i\right)^2\right]$$

considerando $\alpha = \frac{1}{2\sigma^2}\left(\sum_{j=1}^{L} a_j x_j^i - d_i\right)^2$, d_i a resposta desejada do sistema, derivando a expressão com relação ao parâmetro a_j e igualando a zero, temos:

$$\sum_{i=1}^{N}\frac{1}{L} exp[\alpha]\left[-\frac{1}{\sigma^2}\left(\sum_{k=1}^{L-1} a_k x_k^i - d_i\right)\sum_{k=1}^{L-1} x_k^i\right] = 0$$

$$-\sum_{i=1}^{N} exp[\alpha]\sum_{k=1}^{L-1} a_k x_k^i \sum_{k=1}^{L-1} x_k^i + \sum_{i=1}^{N}\frac{1}{L} exp[\alpha] d_i \sum_{k=1}^{L-1} x_k^i = 0$$

$$a_j \sum_{i=1}^{N} exp[\alpha]\sum_{k=1}^{L-1}\left(x_k^i\right)^2 = \sum_{i=1}^{N}\frac{1}{L} exp[\alpha] d_i \sum_{k=1}^{L-1} x_k^i$$

$$a_j = \frac{\sum_{i=1}^{N}\frac{1}{L} exp\left[-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L} a_j x_j^i - d_i\right)^2\right] d_i \sum_{k=1}^{L-1} x_k^i}{\sum_{i=1}^{N} exp\left[-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L} a_j x_j^i - d_i\right)^2\right]\sum_{k=1}^{L-1}\left(x_k^i\right)^2}$$

que representa uma expressão em ponto fixo para o cálculo dos parâmetros de estimação para um problema de regressão utilizando a correntropia a partir de um conjunto de amostras, e para a expressão (4.8), temos:

$$a_j = \frac{\sum_{i=1}^{N} exp\left[-\frac{1}{\sigma}\left(\frac{1}{(L+1)}\left(d_i^2 + a^2\sum_{j=1}^{L} x_j^{i^2}\right) - \left(\frac{1}{(L+1)}\left(d_i + a\sum_{j=1}^{L} x_j^i\right)\right)^2\right)\right] d_i \sum_{k=1}^{L-1} x_k^i}{\sum_{i=1}^{N} exp\left[-\frac{1}{\sigma}\left(\frac{1}{(L+1)}\left(d_i^2 + a^2\sum_{j=1}^{L} x_j^{i^2}\right) - \left(\frac{1}{(L+1)}\left(d_i + a\sum_{j=1}^{L} x_j^i\right)\right)^2\right)\right]\left[\sum_{j=1}^{L} x_j^{i^2} - \left(\frac{1}{L}\sum_{j=1}^{L} x_j^{i^2}\right)^2\right]}$$

que representa uma expressão em ponto fixo para o cálculo dos parâmetros citados utilizando a extensão 01 da correntropia. A prova de convergência de uma equação em ponto fixo tal como as equações apresentadas, é conhecida como o teorema da contração de Banach uma importante ferramenta na teoria de espaços métricos e que garante a existência e unicidade de pontos fixos de certos espaços métricos. [33].

4.2 – *Qualquer possível k combinação em linha* (*Extensão 02*).

Aqui, consideraremos a geração possível de qualquer linha através da combinação de hiperplanos de dimensão *L* – *K*. Ou seja, consiste na resolução da seguinte integral (com *k* constantes $X_2, X_5, \cdots, X_9$ por exemplo) no espaço conjunto *L* - dimensional.

$$V_X = \int \cdots \int \hat{p}(x_1, x_2, \cdots, x_L)\delta(x_1 = x_7 = \cdots = x_L, x_2 = X_2, \cdots, x_9 = X_9)dx_1 \ldots dx_L$$

que pode novamente ser reescrita como a integral

$$V_X = \int \cdots \int \hat{p}(x, X_2, \cdots, X_5, X_6, x, x, X_9, x, \cdots, x)dx$$

onde os parâmetros que possuem índice $(X_2, X_5, X_7, \cdots)$ são tratados como constantes. A seguinte figura ilustra exemplos possíveis da extensão 02.

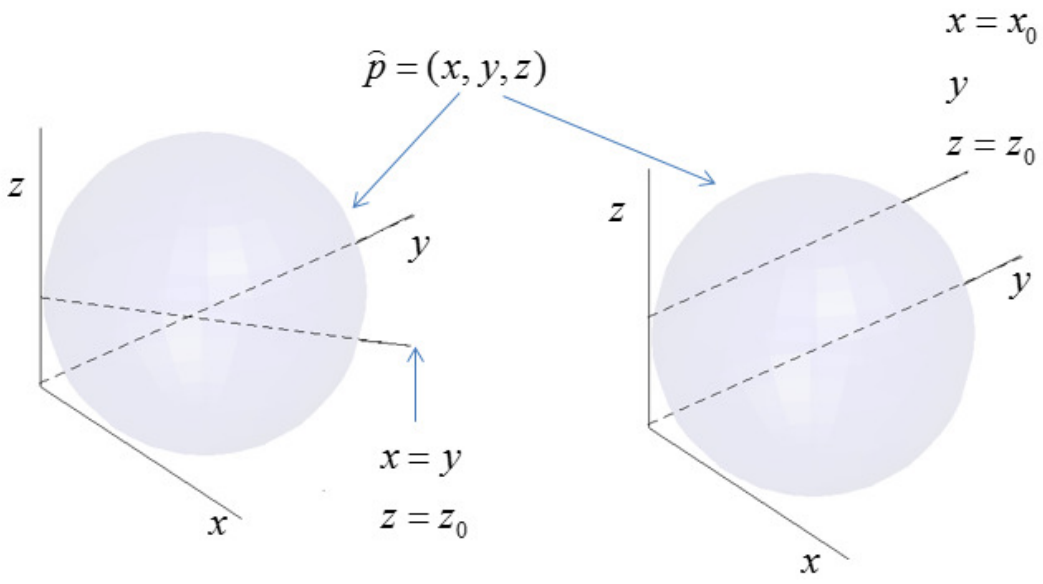

Figura 12 – reta $x = y, z = z_0$ e $z = z_0, x = x_0$.

Substituindo a equação (4.3) na integral anterior, temos:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L}\left(x'_j - x_j^i\right)^2\right)\right)dx \qquad (4.10)$$

onde x_j' ou representa uma constante ou uma variável definida pelo conjunto I dado por:

$$x'_j = \begin{cases} x_j, & j \in I \\ x_j, & j \notin I \end{cases}$$

$$I = \{i_1, i_2, i_3, \cdots, i_n\}, n \leq L$$

agora, desenvolvendo a expressão (4.10), temos:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j\in I}\left(x_j - x_j^i\right)^2 + \sum_{j\notin I}\left(x_j - x_j^i\right)^2\right)\right)dx$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{(2\pi\sigma^2)^{L/2}} exp\left(-\frac{1}{2\sigma^2}\sum_{j\in I}\left(x_j - x_j^i\right)^2\right) exp\left(-\frac{1}{2\sigma^2}\sum_{j\notin I}\left(x_j - x_j^i\right)^2\right)dx$$

utilizando os resultados obtidos na extensão 01 e separando as exponenciais, temos:

$$= \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{n}\frac{1}{(2\pi\sigma^2)^{(n-2)/2}}\mathbb{G}_{\sigma'}\left(\frac{1}{n}\sum_{j\in I}x_j^{i\,2} - \left(\frac{1}{n}\sum_{j\in I}x_j^i\right)^2\right)$$

$$\frac{1}{(2\pi\sigma^2)^{(L-n-1)/2}}\mathbb{G}_{\sigma'}\left(\sum_{j\notin I}\left(x_j - x_j^i\right)^2\right).$$

E finalmente,

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\frac{1}{(2\pi\sigma^2)^{(L-3)/2}}\mathbb{G}_{\sigma'}\left(\frac{1}{n}\sum_{j\in I}x_j^{i\,2} - \left(\frac{1}{n}\sum_{j\in I}x_j^i\right)^2\right)\mathbb{G}_{\sigma'}\left(\sum_{j\notin I}\left(x_j - x_j^i\right)^2\right) \qquad (4.11)$$

Se considerarmos um novo conjunto de dados, dado por:

$$D = \left\{ \sum_j a_j x_j^i, \cdots \sum_l a_l x_l^i, [(y_1 = y_2 = \cdots = y_n) - d_i] \right\}$$

O objetivo é otimizar a integral para todos os $a's$ de modo que estamos escolhendo valores tal que a probabilidade de $d_i = y_i$ seja a máxima independentemente de $x_1^i,...,x_L^i$. Para aplicações práticas, será proposto um algoritmo em ponto fixo para determinar os parâmetros $a's$ ideais. Derivando a expressão (4.11) com relação à a_j e igualando a zero, teremos a seguinte expressão em ponto fixo para atualização de a_j.

$$a_j = -\frac{\sum_i exp\left(-\frac{1}{2\sigma^2}\left(\sum_j a_j x_j^i - d_i\right)^2\right)\left(\sum_k a_k x_k^i - d_i\right)x_j^i}{\sum_i exp\left(-\frac{1}{2\sigma^2}\left(\sum_j a_j x_j^i - d_i\right)^2\right)x_j^{i^2}}$$

A extensão 02 aborda sistemas com múltiplas entradas e saídas idênticas**.**

4.3 – *Qualquer possível subespaço interno linear e ortogonal* (*Extensão 03*).

Vamos agora estender os conceitos apresentados não somente a linhas, mas a planos ou hiperplanos graças às definições de subespaço da álgebra linear. A limitação aqui é que os hiperplanos devem ser ortogonais aos eixos (combinação dos eixos coordenados). A integral é dada por:

$$V_X = \int \cdots \int \hat{p}(x_1, x_2, \cdots, x_L)\delta(x_2 = X_2, \cdots, x_9 = X_9)dx_1 \cdots dx_L$$

agora temos uma integral de dimensão (*L* – *K*) e pode ser escrita como:

$$V_X = \int \cdots \int \hat{p}(x'_1, X_2, \cdots, X_5, X_6, x'_7, x'_8, \cdots, x'_L)dx'_1 \ldots dx'_L$$

onde o conjunto com os sobrescritos são tratados como variáveis de integração e o conjunto de índices como constantes. A seguinte figura ilustra um exemplo dessa integral.

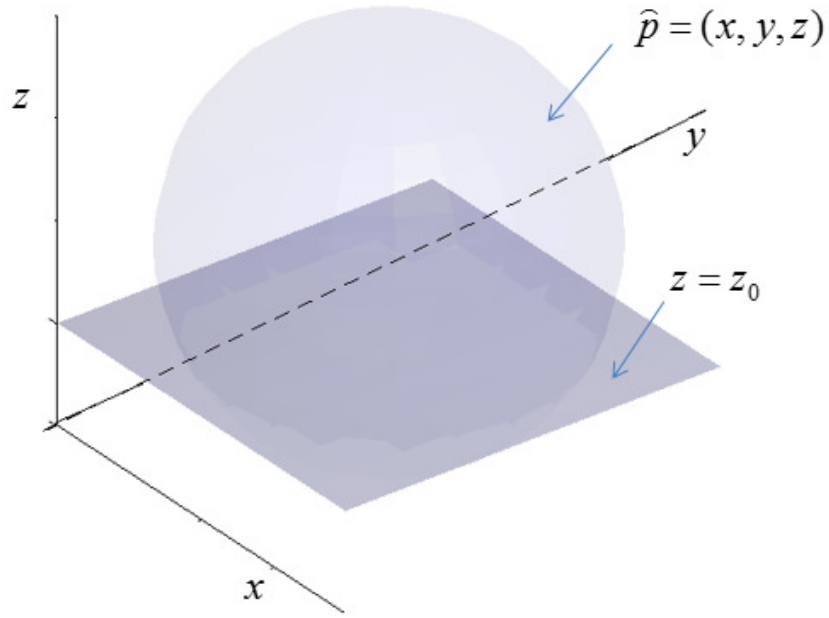

Figura 13 – Plano $z{=}z_0$

Substituindo a equação (4.1) na integral anterior, temos:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int\cdots\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{L}\left(x_j'' - x_j^i\right)^2\right)\right)\prod_{j\in I} dx'_j \tag{4.12}$$

onde x_j'' ou representa uma constante ou uma variável definida pelo conjunto I dado por:

$$x_j'' = \begin{cases} x'_j, & j \in I \\ x_j, & j \notin I \end{cases}$$

$$I = \{i_1, i_2, i_3, \cdots, i_n\}, n \leq L$$

agora, desenvolvendo a expressão (4.12), temos:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int\cdots\int \frac{1}{(2\pi\sigma^2)^{L/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j\in I}\left(x'_j - x_j^i\right)^2 + \sum_{j\notin I}\left(x_j - x_j^i\right)^2\right)\right)\prod_{j\in I} dx'_j$$

utilizando os resultados obtidos na extensão 01 e separando as exponenciais, temos:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi\sigma^2)^{(L-n-1)/2}}\mathbb{G}_{\sigma'}\left(\sum_{j\notin I}\left(x_j - x_j^i\right)^2\right)\int\cdots\int\frac{1}{(2\pi\sigma^2)^{n/2}}$$

$$\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j\in I}\left(x'_j - x_j^i\right)^2\right)\right)\prod_{j\in I}dx'_j$$

e finalmente,

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi\sigma^2)^{(L-n-1)/2}}\mathbb{G}_{\sigma'}\left(\sum_{j\notin I}\left(x_j - x_j^i\right)^2\right) \tag{4.13}$$

a extensão 03 trata-se da marginalização convencional presente na estatística, em que a dimensão de cada probabilidade marginal não é restrita a ser unidimensional.

Para demonstrar esses resultados iremos considerar essa formulação multidimensional da correntropia como uma integral ao longo de hiperplanos em um problema de regressão robusta, as técnicas de regressão robusta são um importante complemento ao método clássico que utiliza os mínimos quadrados, de modo que pequenas alterações na distribuição da amostra produza pequenas alterações nas estimativas limitando a influência dos *outliers* [31].

O uso da extensão 03 justifica-se por ser esta considerada uma técnica robusta não somente com relação aos *outliers*, mas a não linearidade presente nos dados. Nesse caso, utilizaremos a extensão 03 para o seguinte conjunto de dados composto por dados no espaço (*L+1*) – dimensional $\left(x_1^i,\cdots,x_L^i,d_i\right)$ onde d_i é uma função de $x_1^i,\cdots,x_L^i$ adicionada de ruído.

$$D = \left\{x_{i,1},\cdots,x_{i,L},a_0 + \sum_{j=1}^{L}a_j x_j^i - d_i\right\}_{i=1,\cdots,N} \tag{4.14}$$

a integral no plano $z = 0$ calcula a probabilidade de $z = 0$ para todo $x_1^i,\cdots,x_L^i$, e otimizar a integral para todos os a_j significa que estamos escolhendo valores tal

que a probabilidade de $a_0 + \sum_{j=1}^{L} a_j x_j^i - d_i$ tenda a zero independentemente de $x_1^i, \cdots, x_L^i$. Isso nos leva a um estimador robusto aos *outliers* em espaços *L* – dimensionais, além de extrair informações a partir do sinal de erro.

4.3.1 – Um algoritmo em ponto fixo para a extensão 03:

Para aplicações práticas iremos desenvolver um algoritmo em ponto fixo para encontrar os valores ótimos dos $a_j's$. Inserindo as amostras apresentadas em (4.14) na expressão da extensão 03 (4.13) utilizando um *kernel* gaussiano, obteremos o valor da probabilidade de $a_0 + \sum_{j=1}^{L} a_j x_j^i - d_i$ como uma função de a_j.

$$V_x = \frac{1}{N}\sum_{i=1}^{N} exp\left[-\frac{1}{2\sigma^2}\left(a_0 + \sum_{j=1}^{L} a_j x_j^i - d_i\right)^2\right] \tag{4.15}$$

observa-se que o somatório em *j* na variável interna ao *kernel* é somente sobre o valor de *z* e o valor ideal de *z* é que seja nulo. Ou seja,

$$z(i) = a_0 + \sum_{j=1}^{L} a_j x_j^i - d_i \to 0$$

Derivando (4.15) com relação a a_j e igualando-se a zero, temos a seguinte expressão em ponto fixo para os valores de a_j,

$$a_j = -\frac{\sum_{i=1}^{N} x_j^i \left(\sum_{k=1}^{L-1} a_k x_k^i - d_i\right) exp\left[-\frac{1}{2\sigma^2}\left(a_0 + \sum_{k=1}^{L-1} a_k x_k^i - d_i\right)^2\right]}{\sum_{i=1}^{N}\left(x_j^i\right)^2 exp\left[-\frac{1}{2\sigma^2}\left(a_0 + \sum_{k=1}^{L-1} a_k x_k^i - d_i\right)^2\right]} \tag{4.16}$$

o raciocínio aqui é pensar o problema de regressão linear como um ajuste de hiperplanos de acordo com alguma das formulações apresentada e otimizar a integral com relação ao modelo.

Para validar o algoritmo iremos considerar um caso onde $x_{i,0} = 1$ para todo i e o sistema é dado por,

$$z = 2{x_{i,3}}^3 - 2{x_{i,2}}^2 + 2x_{i,1} + x_{i,0} + \varepsilon_i$$

Onde ε_i é a componente de um ruído gaussiano. A seguinte figura mostra o resultado da estimação regressiva na presença de ruído impulsivo (*outliers*). Pode-se observar que a extensão 03 conseguiu estimar o sistema perfeitamente, enquanto o MMQP (Mínimos Quadrados Ponderado) não teve boa aproximação. Os *outliers*, na extensão 03, foram ponderados pela largura do *kernel* gaussiano minimizando sua influência. Esse parâmetro controla as propriedades da correntropia, ou seja, controla a ênfase dada aos momentos de ordem superior sobre os momentos de segunda ordem, portanto uma escolha apropriada deste parâmetro é importante, se, por exemplo, este parâmetro for de valor pequeno e os dados forem muito ruidosos a correntropia não será capaz de distinguir o sinal do ruído, e se for muito grande os momentos de alta ordem da correntropia serão ignorados [34], Já no MMPQ o termo quadrático aumenta a contribuição de possíveis amostras que estejam distantes do valor médio da distribuição do erro, de modo que o MMPQ não seja um método indicado quando a distribuição do erro possuir *outliers* ou média diferente de zero [31].

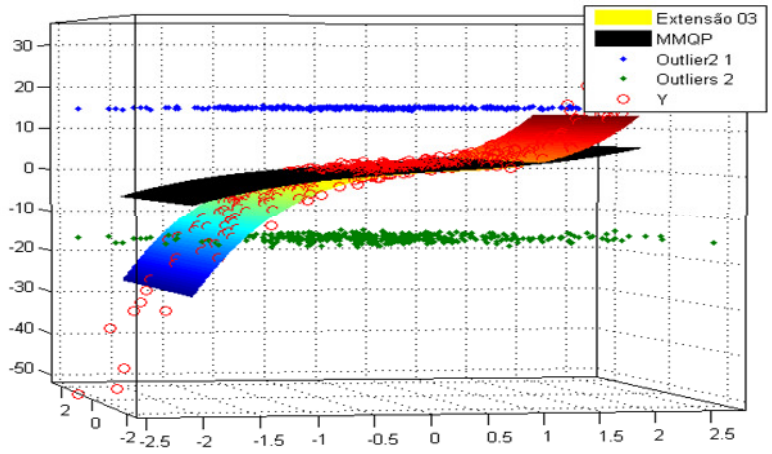

Figure 14: os círculos vermelhos caracterizam o conjunto de dados acrescido de *outliers* (pontos azuis e verdes). A superfície colorida é a estimativa do sistema utilizando a extensão 03 ($\sigma^2 = 1{,}8$) e a superfície preta é a estimativa do sistema utilizando MMQP.

A seguinte figura ilustra uma comparação dos dois métodos extensão 03 e MMQP, para a regressão não linear. Observa-se que o erro do MMQP é bastante elevado, enquanto que na extensão 03 é aproximadamente nulo.

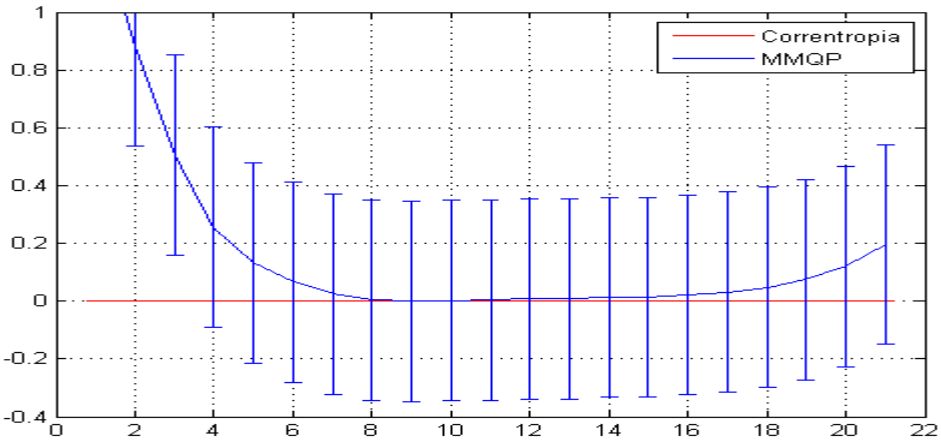

Figura 15 – Variância no erro entre o MMQP e a Extensão 03 da correntropia.

Como vimos nas figuras apresentadas, o modelo obtido com o MMQP é totalmente influenciado pelos *outliers*, enquanto que o modelo obtido através da extensão 03 é bem próximo do real. O modelo utilizando a extensão 03 foi obtido após 5 iterações em ponto fixo da expressão (4.16).

4.4 – *Qualquer possível k combinação linear de manifolds* (*Extensão 04*).

Finalmente, iremos considerar uma combinação de *k manifolds* para produzir hiperplanos que não são necessariamente ortogonais aos eixos coordenados. Um *manifold* é um espaço topológico que é localmente euclidiano (ou seja, em torno de cada ponto, existe uma vizinhança que é topologicamente a mesma que uma esfera unitária aberta no $\mathbb{R}^n$ [35]. A ideia básica é encontrar uma estrutura dimensional que está incorporada em um espaço de dimensão superior. No nosso caso (uma generalização da extensão 02 a hiperplanos) a integral é dada por:

$$V_X = \int \cdots \int \hat{p}(x_1, x_2, \cdots, x_L)\delta(x_1 = X_7, \cdots, x_9 = X_9)dx_1 \cdots dx_L$$

agora temos uma integral que pode ser escrita como:

$$V_X = \int \cdots \int \hat{p}(x'_1, X_2, \cdots, X_5, X_6, x'_7, x'_8, X_9, \cdots, x'_L)\, dx'_1 \ldots dx'_L$$

A seguinte figura ilustra uma integral do tipo,

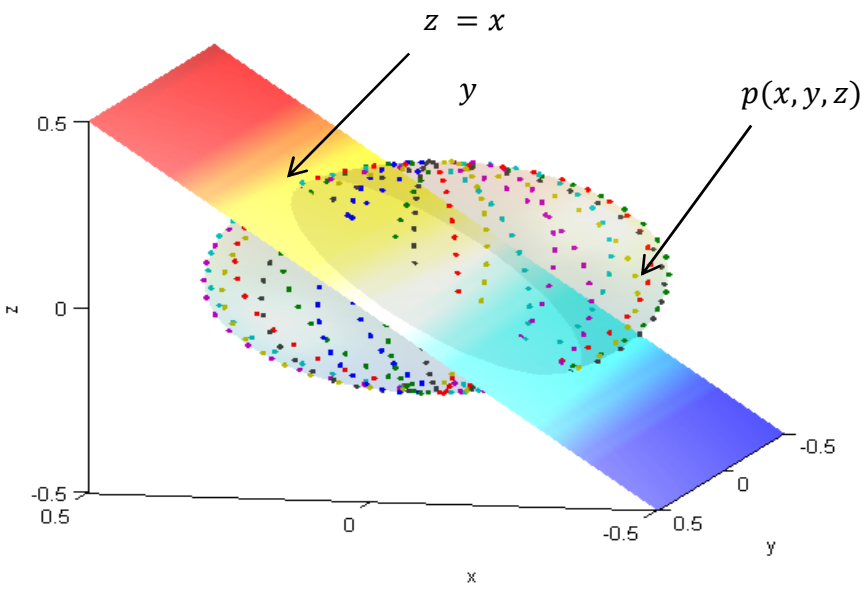

Figura 16 - Plano $z=x$ e y

assim como na extensão 03, as variáveis principais são as únicas a serem integráveis. Desta vez, a diferença básica é que as variáveis repetidas são permitidas, de modo que iremos considerar a integral:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N} \int \cdots \int \frac{1}{(2\pi\sigma^2)^{L/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_j \left(x''_j - x_j^i\right)^2\right)\right) \prod_j dx''_j$$

com

$$x''_j = \begin{cases} x'_{j,k}, & j \in I_{k=1,2,\ldots,s} \\ x_j, & j \notin I_{k=1,2,\ldots,s} \end{cases}$$

$$I = \{I_1, I_2, I_3, \cdots, I_s\}$$

$$I_k = \left\{I_1^k, I_2^k, I_3^k, \cdots, I_{n(k)}^k\right\}, \sum_{k=1}^{s} n(k) \leq L$$

No caso, dividimos o conjunto de *manifolds* I em subgrupos $\{I_1, I_2, I_3, \cdots, I_s\}$. Onde cada subgrupo k possui $n(k)$ elementos. A ideia básica é que todos os elementos

em cada subgrupo possua uma variável correspondente que tenha o índice do grupo $x_{i_1^k} = x_{i_2^k} = \cdots x_{i_{n(k)}^k} = x_k$ para que possam ser consideradas as variáveis repetidas. O número de variáveis no subgrupo pode ser menor que a dimensão L $(n(k) \leq L)$, então, qualquer termo que não pertença ao subgrupo pode ser considerado como uma constante.

Deste modo, podemos escrever a integral como:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N}\int \cdots \int \frac{1}{(2\pi\sigma^2)^{L/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j\in I_1}\left(x'_{j,1} - x_j^i\right)^2 + \cdots \right.\right.$$

$$\left.\left. + \sum_{j\in I_s}\left(x'_{j,s} - x_j^i\right)^2 + \sum_{j\notin I_{k=1,2,\ldots,s}}\left(x_j - x_j^i\right)^2\right)\right) \prod_{k=1,2,\ldots,s} dx'_{j,k}$$

onde a exponencial possui os elementos separados por subgrupos e o último termo é a constante do grupo.

Separando e desenvolvendo a expressão temos:

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{\left(L-\sum_{k=1}^{s} n(k)\right)/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j\notin I_{k=1,2,\ldots,s}}\left(x_j - x_j^i\right)^2\right)\right)$$

$$\prod_{k=1}^{s}\int \frac{1}{(2\pi\sigma^2)^{n(k)/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j\in I_k}\left(x'_{j,k} - x_j^i\right)^2\right)\right) dx'_{j,k}$$

que resulta em

$$\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{\left(L-1-\sum_{k=1}^{s} n(k)\right)/2}} \mathbb{G}_\sigma\left(\sum_{j\notin I_{k=1,2,\ldots,s}}\left(x_j - x_j^i\right)^2\right)$$

$$\prod_{k=1}^{s} \mathbb{G}_{\sigma'}\left(\frac{1}{n(k)}\sum_{j\in I_k} {x_j^i}^2 - \left(\frac{1}{n(k)}\sum_{j\in I_k} x_j^i\right)^2\right) \tag{4.17}$$

representando o caso mais geral das extensões propostas para a correntropia. De modo que nessa extensão podemos pensar diferentes combinações em *manifolds*.

4.4.1 - Aplicação na identificação de um sistema MIMO (*Multiple Inputs Multiple Outputs*).

Como exemplo, considere um sistema linear MIMO (mxn) tal que $\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{x}_i$ sendo $\boldsymbol{A}$ a matriz do sistema, $\boldsymbol{x}_i = \left[x_1^i, \cdots, x_m^i\right]^T$ um *m*-dimensional vetor de entradas e $\boldsymbol{y}_i = \left[y_1^i, \cdots, y_n^i\right]^T$ o correspondente vetor *n*-dimensional de saídas. Considerando o *L*-dimensional ($L = 2n$) espaço conjunto $\left(u_1^i, \cdots, u_n^i; y_1^i, \cdots, y_n^i\right)$. Então a extensão 04 será utilizada para o cálculo da probabilidade $u_j^i = y_j^i$ para todo *j*, podemos fazer para cada u_j^i uma combinação linear de todos $x_j^i (j = 1, \cdots, m)$ e otimizar a probabilidade com relação aos coeficientes. Essa probabilidade é a verossimilhança dos parâmetros na estimação da *pdf* conjunta (para os dados fornecidos). De fato, maximizar a correntropia neste caso corresponde a maximizar a verossimilhança. Este procedimento conduz a um estimador linear robusto (MIMO). Podemos também obter o sistema reverso utilizando a operação inversa. Como fizemos anteriormente para a regressão *L*-dimensional, considere o novo espaço conjunto dado pelos pontos

$$\left(\sum_j a_{1j} x_j^i, \cdots, \sum_j a_{nj} x_j^i; y_1^i, \cdots, y_n^i\right)$$

e a_{ij} os coeficientes do sistema a serem otimizado ou identificado.

Se substituirmos essa equação na expressão resultante da extensão 04 (4.17) poderemos obter a densidade de probabilidade de $\sum_j a_{ij} x_j^i$ sendo igual a y_j^i para todo *j*. Utilizando essa ideia iremos testar o estimador no caso mais simples (n = 2) (novamente, desconsiderando os termos constantes e utilizando um *kernel* gaussiano),

$$V_X = \sum_i \prod_{j=1}^{2} exp\left[-\frac{1}{\sigma^2}\left(\frac{1}{2}\left(\left(\sum_k a_{jk} x_k^i\right)^2 + \left(y_j^i\right)^2\right)\right) - \left(\frac{1}{2}\left(\sum_k a_{jk} x_k^i + y_j^i\right)\right)^2\right]$$

desta forma podemos escrever a extensão 04 como,

$$V_X = \sum_i exp\left\{-\frac{1}{\sigma^2}\sum_j\left[\left(\frac{1}{2}\left(\left(\sum_k a_{jk}x_k^i\right)^2 + \left(y_j^i\right)^2\right)\right) - \left(\frac{1}{2}\left(\sum_k a_{jk}x_k^i + y_j^i\right)\right)^2\right]\right\}$$

$$V_X = \sum_i exp\left\{-\frac{1}{\sigma^2}\sum_j\left[\left(\frac{1}{2}\left(\sum_k a_{jk}x_k^i\right)^2 + \frac{1}{2}\left(y_j^i\right)^2\right) - \frac{1}{4}\left(\left(\sum_k a_{jk}x_k^i\right)^2 + 2y_j^i\sum_k a_{jk}x_k^i + \left(y_j^i\right)^2\right)\right]\right\}$$

denominando o termo interno a exponencial de α,temos:

$$\alpha = -\frac{1}{\sigma^2}\sum_j\left[\left(\frac{1}{2}\left(\sum_k a_{jk}x_k^i\right)^2 + \frac{1}{2}\left(y_j^i\right)^2\right) - \frac{1}{4}\left(\left(\sum_k a_{jk}x_k^i\right)^2 + 2y_j^i\sum_k a_{jk}x_k^i + \left(y_j^i\right)^2\right)\right]$$

daí podemos escrever α como,

$$\alpha = -\frac{1}{\sigma^2}\sum_j\left[\frac{1}{2}\left(\sum_k a_{jk}x_k^i\right)^2 + \frac{1}{2}\left(y_j^i\right)^2 - \frac{1}{4}\left(\sum_k a_{jk}x_k^i\right)^2 - \frac{1}{2}y_j^i\sum_k a_{jk}x_k^i - \frac{1}{4}\left(y_j^i\right)^2\right]$$

simplificando a expressão temos,

$$\alpha = -\frac{1}{\sigma^2}\sum_j\left[\left(\frac{1}{4}\left(\sum_k a_{jk}x_k^i\right)^2 - \frac{1}{2}y_j^i\sum_k a_{jk}x_k^i + \frac{1}{4}\left(y_j^i\right)^2\right)\right]$$

$$\alpha = -\frac{1}{\sigma^2}\sum_j\left[\frac{1}{4}\left(\left(\sum_k a_{jk}x_k^i\right)^2 - 2y_j^i\sum_k a_{jk}x_k^i + \left(y_j^i\right)^2\right)\right]$$

$$\alpha = -\frac{1}{4\sigma^2}\sum_j\left(\left(\sum_k a_{jk}x_k^i\right) - y_j^i\right)^2$$

de modo que podemos reescrever a expressão (4.17) da extensão 04 como,

$$V_X = \sum_i exp\left[-\frac{1}{4\sigma^2}\sum_j\left(\left(\sum_k a_{jk}x_k^i\right) - y_j^i\right)^2\right]$$

derivando com relação a a_{ij} e igualando a zero teremos a seguinte atualização em ponto fixo

$$a_{lp} = \frac{\sum_i \exp\left(-\frac{1}{4\sigma^2}\sum_j\left(\sum_k a_{jk}x_k^i - y_j^i\right)^2\right)\left(y_l^i - \sum_{k\neq p} a_{lk}x_k^i\right)x_p^i}{\sum_i \exp\left(-\frac{1}{4\sigma^2}\sum_j\left(\sum_k a_{jk}x_k^i - y_j^i\right)^2\right){x_p^i}^2}$$

Testamos esse estimador com o sistema cuja matriz é dada por $\boldsymbol{A} = \begin{bmatrix} -0{,}2964 & 0{,}4867 \\ -2{,}8508 & -0{,}9583 \end{bmatrix}$ e as saídas são adicionadas de ruído tal como nos exemplos anteriores. Utilizamos 600 pontos aleatórios distribuídos uniformemente em uma grade de -1 a 1 (em ambos os eixos coordenados) como entradas. Utilizamos uma largura de *kernel* igual a 0.1 e o algoritmo convergiu com 5 iterações em ponto fixo, obtendo a seguinte matriz estimada $\widehat{\boldsymbol{A}} = \begin{bmatrix} -0{,}2402 & 0{,}4939 \\ -2{,}8314 & -0{,}9137 \end{bmatrix}$ e com um regressor linear clássico obtivemos $\widehat{\boldsymbol{A}} = \begin{bmatrix} -0{,}6641 & -3{,}2449 \\ 0{,}2801 & -1{,}1208 \end{bmatrix}$.

Para ilustrar o algoritmo iremos considerar as entradas,

$$\boldsymbol{x} = \begin{bmatrix} x_1^i \\ x_2^i \end{bmatrix}, \boldsymbol{A}_0 = \begin{bmatrix} a_{11}^0 & a_{12}^0 \\ a_{21}^0 & a_{22}^0 \end{bmatrix} \; e \; \boldsymbol{d}^i = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

e a matriz de saída,

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} \\ \hat{a}_{21} & \hat{a}_{22} \end{bmatrix}$$

Passo 01:

Cálculo do α para $j = 1$.

$$\alpha_1 = -\frac{1}{4\sigma^2}\left((a_{11}^0 x_1^i + a_{12}^0 x_2^i) - d_1^i\right)^2$$

Cálculo do α para $j = 2$.

$$\alpha_2 = -\frac{1}{4\sigma^2}\left((a_{21}^0 x_1^i + a_{22}^0 x_2^i) - d_2^i\right)^2$$

Passo 02:

Cálculo do $\hat{a}_{11}$:

$$\hat{a}_{11} = \frac{\sum_i \exp(\alpha_1)\left(a_{12}^0 x_2^i - d_1^i\right) x_1^i}{\sum_i \exp(\alpha_1)\left(x_1^i\right)^2}$$

Cálculo do $\hat{a}_{12}$:

$$\hat{a}_{12} = \frac{\sum_i \exp(\alpha_1)\left(a_{11}^0 x_2^i - d_1^i\right) x_1^i}{\sum_i \exp(\alpha_1)\left(x_1^i\right)^2}$$

Cálculo do $\hat{a}_{21}$:

$$\hat{a}_{21} = \frac{\sum_i \exp(\alpha_2)\left(a_{22}^0 x_2^i - d_2^i\right) x_2^i}{\sum_i \exp(\alpha_2)\left(x_2^i\right)^2}$$

Cálculo do $\hat{a}_{22}$:

$$\hat{a}_{22} = \frac{\sum_i \exp(\alpha_2)\left(a_{21}^0 x_2^i - d_2^i\right) x_2^i}{\sum_i \exp(\alpha_2)\left(x_2^i\right)^2}$$

Atualiza a matriz $\widehat{\boldsymbol{A}}$ e repetem-se os passos 01 e 02 até o algoritmo convergir.

Capítulo 5

Conclusões.

Esta tese apresentou um novo algoritmo que estende a aplicação de correntropia para qualquer dimensão, em vez de comparar apenas duas variáveis escalares como na definição original. Experimentos com regressão foram desenvolvidos com bons resultados, em particular, na presença de ruído impulsivo (*outliers*).

A extensão 01 estendeu o conceito da correntropia para retas definidas por vetores de componentes idênticas e consiste no cálculo de uma integral de linha definida pelo vetor. No experimento computacional foi observado que a extensão 01 é menos sensível ao parâmetro de ajuste do *kernel* suprindo uma deficiência no algoritmo de estimação utilizando a correntropia convencional. Concluimos também que a extensão 01 corresponde a um estimador de variância mínima sendo caracterizado pela suavidade do EMQ e a robustez da correntropia a ruídos impulsivos.

A Extensão 02 trata da qualquer combinação de retas através da combinação de hiperplanos e pode ser utilizada na análise de sistemas com múltiplas entradas e saídas idênticas. Na extensão 03, podemos estender o conceito da correntropia a planos ou hiperplanos ortogonais aos eixos. O uso da extensão 03 justifica-se por ser esta considerada uma técnica robusta não somente com relação aos *outliers*, mas a não linearidade presente nos dados.

A extensão 04 considera uma combinação de *manifolds* para produzir hiperplanos que não são necessariamente ortogonais aos eixos coordenados, é uma generalização da extensão 02 a hiperplanos, representado o caso mais geral entre as extensões propostas. Um exemplo de identificação de sistemas *MIMO* foi apresentado. No experimento da aplicação *MIMO*, a correntropia em um espaço de 4 dimensões corresponde a probabilidade de que cada uma das

duas entradas seja igual a sua saída correspondente. Além disso, a aplicação MIMO mostrou uma nova e interessante forma de visualizar a correntropia multidimensional. Os estimadores para as probabilidades são todos baseados na estimativa de Parzen da pdf conjunta dos dados, portanto um tamanho de *kernel* sub-ótimo deve ser estimado. Embora o exemplo *MIMO* tenho o mesmo número de entradas e saídas, a formulação pode ligar com dimensões arbitrárias. Isto exigiria uma solução semelhante à da extensão 01, onde para um sistema n x m deverá integrar $m + n$ dimensões.

Algoritmos em ponto fixo bem como resultados foram apresentados para mostrar o potencial em inúmeras aplicações tais como: Filtragem adaptativa, Identificação de sistemas, predição de séries temporais não lineares, classificação de padrões, etc. Este trabalho motiva futuras pesquisas na análise teórica da demonstração na convergências dos algoritmos, bem como a aplicação das extensões em diferentes problemas na Engenharia.

Referências.

[1] Jose C. Principe, *Information Theoretical Learning – Renyi's Entropy and Kernel Perspective*, Springer Science+Business Media,2010.

[2] S. Haykin, *Adaptive Filter Theory*, 4th Edition, Prentice Hall, 2002.

[3] A. McFarlane, F. A. Poch – *Introducción a la teoria de la Estadística*, 2ª Ed. Aguilar, 1960.

[4] A. Papoulis, S.U. Pillai – *Probability, Random Variables ans Stochastic Processes*, 4th Edition, McGraw Hill, 2002.

[5] S. Haykin, *Unsupervised Adaptive Filtering, Volume 1: Blind Source Separation*, John Wiley & sons, 2000.

[6] Erdogmus D., *Information theoretic learning: Renyi's entropy and its applications to adaptive systems training*, Ph.D. Dissertation, University of Florida,Gainesville, 2002.

[7] Shannon C., and Weaver W., *The mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

[8] S. Verdu. Fifty Years of Shannon Theory. In *Information Theory - 50 Years of Discovery*, pages 2178–2206, S. Verdu and S. McLaughlin (Eds.), IEEE Press, Piscataway, 2000.

[9] Renyi A., *On measures of entropy and information*, *Proc. of the 4th Berkeley Symp. Math. Statist. Prob. 1960*, vol. I, Berkeley University Press, pp. 457, 1961.

[10] P. Sahoo, C. Wilkins, and J. Yeager, "*Threshold Selection Using Renyi's Entropy," Pattern Recognition*, vol. 30, pp. 71-84, 1997.

[11] Beadle, E.; Schroeder, J.; Moran, B.; Suvorova, S., "An overview of Renyi Entropy and some potential applications," *Signals, Systems and Computers, 2008 42nd Asilomar Conference on* , vol., no., pp.1698,1704, 26-29 Oct. 2008.

[12] Zhao Guan-hua; Hao Min, "Incremental learning algorithm of least squares support vector machines based on Renyi entropy," *Management Science and Engineering, 2009. ICMSE 2009. International Conference on* , vol., no., pp.95,100, 14-16 Sept. 2009.

[13] Jelinek, H.F.; Tarvainen, M.P.; Cornforth, D.J., "Renyi entropy in identification of cardiac autonomic neuropathy in diabetes," *Computing in Cardiology (CinC), 2012* , vol., no., pp.909,912, 9-12 Sept. 2012.

[14] Viola P. Schraudolph N., Sejnowski T., “Empirical entropy manipulation for real-world problems”, Proc. Neural Information Processing Systems NIPS 8, 851-857, 1995.

[15] Jose C. Principe, Dongxin Xu, JohnW. Fisher III, *Information-Theoretic Learning.* Chapter 7, 1999.

[16] I. Santamaria, P. P. Pokharel, and J. C. Principe, Generalized correlation function: Definition, properties and application to blind equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187 - 2197, June 2006.

[17] Liu Weifeng, P. P. Pokharel, and J. C. Principe, Correntropy: properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing,* 2006.

[18] Deniz Erdogmus and Jose C. Principe, *Comparison of Entropy and Mean Square Error Criteria*, Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ,2000, pages 75 – 80.

[19] S. Haykin, *Redes Neurais: princípios e práticas – 2ª Edição*. Porto Alegre: Bookman, 2001.

[20] Theodoridis S., K. Koutroumbas, *Pattern Recognition - 4ª Edição*, Academic Press, 2008.

[21] Neter, Kutner, Nachtsheim, and Wasserman, *Applied Linear Statistical Models 4ª Edição*, 1996.

[22] Duda R., Hart P., *Pattern Recognition and Scene Analysis*, Wiley, New York, 1973.

[23] Silverman B., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[24] S. Kullback, *Information Theory and Statistics – 2ed*. Washington: Dover Publications, Inc.,1968.

[25] Parzen E., *On the estimation of a probability density function and the mode*, *Ann. Math. Statist.*, 33:1065–1067, 1962.

[26] R. Jenssen, D. Erdogmus, J. C. Principe, and T. Eltoft, "*Toward a unification of information theoretic learning and kernel methods*," *IEEE Int. Workshop Machine Learning Signal Processing*, Sao Luis, Brazil, Sep. 2004, pp. 443–451.

[27] E. Parzen, *Statistical Inference on Time Series by Hilbert Space Methods*, I – Technical Report 23, Stanford University, 1959.

[28] N. Aronszajn, *Theory of Reproducing Kernels*. Trans. of the American Mathematical Society, Vol. 68, pag. 337 – 404,1950.

[29] J. Mercer, "*Functions of positive and negative type, and their connection with the theory of integral equations*," *Philosophical Transactions of the Royal Society of London*, vol. 209, pp. 415-446, 1909.

[30] Weifeng Liu, P. P. Pokharel, and J. C. Príncipe, Correntropy: *A Localized Similarity Measure*. International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 2006.

[31] M. G. Kendall e A. Stuart, *The Advanced Theory of Statistics – Vol 2: Inference and Relationship*, Charles Griffin & Company Limited, 1961

[32] A. Granas e J. Dugundji, Fixed Point Theory. Springer – Verlag, New York, 2003.

[33] J. K. Hunter e B. Nachtergaele, *Applied Analysis*, University of California at Davis, WorldScientific, 2000

[34] Jian-Wu Xu; Bakardjian, H.; Cichocki, A.; Principe, J.C., "A New Nonlinear Similarity Measure for Multichannel Biological Signals," *Neural Networks, 2007. IJCNN 2007. International Joint Conference on* , vol., no., pp.2046,2051, 12-17 Aug. 2007]

[35] Xiaoming Huo, Xuelei Ni, and Andre K Smith. A survey of manifold-based learning methods. In Recent Advances in Datamining of Enterprise Data Algorithms and Applications, pages 691 – 745. 2004.

Anexo A

Reproducing Kernel Hilbert Space - RKHS.

O *RKHS* foi utilizado Pela primeira vez no início do século 20 por S. Zaremba em seu trabalho sobre funções harmônicas e bi - harmônicas. Em 1907, ele foi o primeiro que introduziu, em um determinado caso, o *kernel* correspondente a uma classe de funções, e declarou sua propriedade de reproducing *kernel*, mas não desenvolveu a teoria e nem nominou o *kernel* utilizado.

Em 1909, Mercer analisou as funções que satisfaziam as propriedades de *reproducing kernel* na teoria de equações integrais desenvolvidas por Hilbert e denominou essas funções de *kernels* positivos definidos. Ele demonstrou que esses kernels positivos definidos possuem determinadas propriedades interessantes em relação aos *kernels* contínuos de equações integrais.

No entanto por um longo período esses resultados não foram investigados, então a ideia de *Reproducing Kernel Hilbert Space* - RKHS surgiu na dissertação de três matemáticos em Berlin, Szegö (1921), Bergman (1922) e Bochner (1922). Em particular, Bergman introduziu *RKHS* para a classe de funções harmônicas e analíticas que ele denominou de funções *kernel*.

Em 1935, Moore examinou os *kernels* definidos positivo em sua análise geral sob o nome de *matriz Hermitiana positiva*.

Mais tarde, a teoria de *RKHS* foi consolidada por Aronszajn em torno de 1948. E, Na teoria de probabilidades, a teoria de *kernels* definidos positivos foi utilizado por Kolmogorov, Parzen, dentre outros.

Aplicações do método *RKHS* em processos estocásticos de segunda ordem foram dadas por Loève (1948). O método RKHS foi utilizado com sucesso em séries temporais, detecção, filtragem e problemas de predição. (Parzen, 1960, 1961 e Kailath, 1967).

O *RKHS* pode ser utilizado em uma grande variedade de ajuste de curvas, estimação de funções, descrição e construção de modelos.

A.1 – Definição do *Reproducing Kernel Hilbert Space - RKHS*.

Um espaço de Hilbert $\mathbb{H}$ de funções num conjunto X é denominado de Espaço de Hilbert Reproduzido por *Kernel* – RKHS se for denotado pelo produto interno $\langle f, g\rangle$ e $\|f\| = \langle f, f\rangle^{1/2}$ for a norma em $\mathbb{H}$ para f e $g \in \mathbb{H}$. A função de valor complexo $k(y, x)$ de y e x em X é chamada de reproduzindo *kernel* de $\mathbb{H}$ se atender as seguintes propriedades:

1. Para cada x, $k_x(y) = k(y, x)$ como uma função de $y \in \mathbb{H}$.
2. Para cada $x \in X$ e $f \in \mathbb{H}$,

$$f(x) = \langle f, k_x\rangle. \qquad \text{(A.1)}$$

Então aplicando (A.1) na função k_x em y, temos $k_x(y) = \langle k_x, k_y\rangle$, para $x, y \in X$, e por (1), temos $k(y, x) = \langle k_x, k_y\rangle$, para $x, y \in X$. Pela relação acima, para $x \in X$ obtemos $\|k_x\| = \langle k_x, k_x\rangle^{1/2} = k(x, x)^{1/2}$.

Então, $\mathbb{H}$ é dito um Espaço de Hilbert Reproduzido por *Kernel* e as propriedades (1) e (2) são chamadas de propriedades de $k(x, y)$ em $\mathbb{H}$.

Teorema 3.1: Se um espaço de Hilbert $\mathbb{H}$ de funções num conjunto X admite um reproduzindo *kernel*, então o reproduzindo *kernel* $k(y, x)$ é unicamente determinado pelo espaço de Hilbert $\mathbb{H}$.

Demonstração: Seja $k(y, x)$ um reproduzindo kernel de $\mathbb{H}$. Suponha que existe um outro *kernel* $\tilde{k}(y, x)$ em $\mathbb{H}$. Então, para todo $y \in X$, aplicando a propriedade (2) para k e $\tilde{k}$, temos:

$$\begin{aligned}\left\|k_x - \tilde{k}_x\right\|^2 &= \langle k_x - \tilde{k}_x, k_x - \tilde{k}_x\rangle \\ &= \langle k_x - \tilde{k}_x, k_x\rangle - \langle k_x - \tilde{k}_x, k_x\rangle \\ &= \left(k_x - \tilde{k}_x\right)(x) - \left(k_x - \tilde{k}_x\right)(x) = 0\end{aligned}$$

Então, podemos concluir que $k_x = \tilde{k}_x$, ou seja, $k_x(y) = \tilde{k}_x(y), \forall\, y \in X$. De modo que $k(x,y) = \tilde{k}(x,y), \forall\, x,y \in X$.

Teorema 3.2: Para um espaço de Hilbert $\mathbb{H}$ de funções em X, existe um *reproduzindo* kernel k para $\mathbb{H}$ se, e somente se, para cada $x \in X$, a avaliação do funcional linear $\mathbb{H} \ni f \mapsto f(x)$ é um funcional linear limitado em $\mathbb{H}$.

Demonstração: Suponha que k é um reproduzindo *kernel* de $\mathbb{H}$. Pelas propriedades (1) e (2) e a desigualdade de Schwarz para o produto interno, para todo $x \in X$, temos:

$$|f(x)| = |\langle f, k_x\rangle| \leq \|f\|\|k_x\| = \|f\|\langle k_x, k_x\rangle^{1/2} = \|f\|k(x,x)^{1/2}$$

de modo que, a avaliação em x é um funcional linear limitado em $\mathbb{H}$ [23].

Definição 3.1: Seja $\mathbb{H}$ um espaço de Hilbert de funções no conjunto X, e $k(x,y)$ um kernel em X, tal que $K: X\,x\,X \to \mathbb{C}$. Então a função $k(x,y)$ é dita positiva definida se para um conjunto finito de pontos $\{x_1, x_2, \ldots, x_n\} \in X$ e para qualquer número complexo correspondente, nem todos nulos $\{\alpha_1, \alpha_2, \ldots, \alpha_n\} \in \mathbb{C}$, temos:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \bar{\alpha}_j\, k(x_i, x_j) > 0. \qquad \text{(A.2)}$$

E o kernel $k(x,y)$ é dito positivo definido e é um reproduzindo *kernel* porque atende as propriedades (1) e (2) [28].

A.2 – *RKHS* em Inferência Estatística.

Parzen num *tecnical report* [27] publicado em 1959, apresentou uma importante ferramenta para uso na representação de um processo estocástico com momento de segunda ordem finito por meio do RKHS.

Primeiramente, vamos considerar um processo estocástico $\{X(t), t \in T\}$, consistindo de variáveis aleatórias definidas no espaço probabilidade (Ω, *F*, *P*). Se cada uma dessas variáveis aleatórias $X(t)$ possui um momento finito de segunda ordem, então:

$$\mathbb{E}|X(t)|^2 = \int_{\Omega} |X(t)|^2 dP \ < \ \infty, \forall \ t \in T \qquad \text{(A.3)}$$

onde $\{X(t), t \in T\}$ é denominado de função aleatória ou função aleatória de segunda ordem. Podemos estender este conceito para definir um *espaço linear* $L_2(\Omega, F, P)$ como um conjunto de todas as variáveis aleatórias $X(t)$ que satisfazem a equação (A.3).

Podemos definir o produto interno entre duas variáveis aleatórias X e Y no $L_2(\Omega, F, P)$ por:

$$\langle X, Y \rangle = \mathbb{E} = [XY] \int_{\Omega} [XY] dP. \qquad \text{(A.4)}$$

Então $L_2(\Omega, F, P)$ é um espaço produto interno. As demonstrações podem ser vistas em [25]. Por isso, o espaço produto interno $L_2(\Omega, F, P)$ de todos os quadrados integráveis das variáveis aleatórias no espaço probabilidade (Ω, *F*, *P*) é um espaço de Hilbert. Então para cada $t \in T$, a variável aleatória $X(t)$ pode ser considerada como um ponto de dado no espaço de Hilbert $L_2(\Omega, F, P)$.

A.2.1 – Representação de uma função aleatória definida em um intervalo finito.

O método que iremos utilizar no estudo de funções aleatórias de segunda ordem consiste em analisar vários espaços concretos de Hilbert que são congruentes ao espaço de Hilbert gerado pela função aleatória.

Definição 3.2: Um espaço de Hilbert é dito ser a representação de uma função aleatória $\{X(t), t \in T\}$ se $\mathbb{H}$ é congruente a $L_2(X(t), t \in T)$.

Uma das fundamentações teóricas na utilização da abordagem utilizando RKHS no estudo de processos aleatórios de segunda ordem é que a função covariância dos processos induz um RKHS e existe um isomorfismo isométrico (identificação total entre dois espaços normados - congruentes) entre $L_2(X(t), t \in T)$ e o RKHS determinado pela sua função de covariância.

Definição 3.3: Uma família de vetores $\{\boldsymbol{X}(t), t \in T\}$ no espaço de Hilbert $\mathbb{H}$ é dita ser uma representação de uma função aleatória $\{X(t), t \in T\}$ se, para cada t_1, t_2 em T,

$$\langle \boldsymbol{X}(t_1), \boldsymbol{X}(t_2) \rangle = R(t_1, t_2) = \mathbb{E}[\boldsymbol{X}(t_1), \boldsymbol{X}(t_2)] \quad \text{(A.5)}$$

Iremos assumir que a função aleatória $X(t)$ é contínua com média quadrática. Sua função covariância *R* é então uma função contínua, simétrica e não negativa em $T\ x\ T$. Para tal função, existe uma expansão em série dada pelo teorema de Mercer [29].

Teorema de Mercer: Suponha $R(t_1, t_2)$ uma função contínua, simétrica e não negativa em um intervalo fechado e finito $T\ x\ T$. Denotada por $\{\lambda_k, k = 1, 2, \dots\}$ uma sequência de autovalores de $R(t_1, t_2)$ não negativos e por $\{\varphi_k(t), k = 1, 2, \dots\}$ uma sequência de autofunções normalizadas correspondente, ou seja, para todo inteiro t_1 e t_2,

$$\int_T R(t_1, t_2)\varphi_k(t_1)dt_1 = \lambda_k \varphi_k(t_2), \quad t_1, t_2 \in T\delta_{k,j} \quad \text{(A.6)}$$

$$\int_T \varphi_k(t)\varphi_j(t)dt = \delta_{k,j} \quad \text{(A.7)}$$

onde $\delta_{k,j}$ é a função delta de Kronecker, igual a 1 ou 0, de acordo com $k = j$ ou $k \neq j$. Então

$$R(t_1, t_2) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t_1)\varphi_k(t_2) \quad \text{(A.8)}$$

onde a série converge absolutamente e uniformemente em $T\ x\ T$. Comparando as equações (A.8) e (A.5), pode-se perceber que dá para obter um espaço de Hilbert que é uma representação de $\{X(t), t \in T\}$. Seja $\mathbb{H}$ o espaço de todas as sequências de valores reais $\{\alpha_k, k = 1, 2, \dots\}$ tal que

$$\sum_{k=1}^{\infty} \lambda_k \alpha_k^2 < \infty \quad \text{(A.9)}$$

O produto interno entre duas sequências

$\alpha = \{\alpha_k, k = 1,2, \dots\}$ e $\beta = \{\beta_k, k = 1,2, \dots\}$ é definido como:

$$\langle \alpha, \beta \rangle = \sum_{k=1}^{\infty} \lambda_k \alpha_k \beta_k \quad \text{(A.10)}$$

De (A.8) segue que para cada $t \in T$, a sequência

$$\varphi(t) = \{\varphi_k, k = 1,2, \dots\} \tag{A.11}$$

Pertence a $\mathbb{H}$. Além disso

$$R(t_1, t_2) = \langle \varphi(t_1), \varphi(t_2) \rangle \tag{A.12}$$

comparando (A.5) com (A.12) podemos ver que o espaço de Hilbert $L_2(X(t), t \in T)$ é congruente ao espaço de Hilbert $L_2(\varphi(t), t \in T)$. Assim através da definição de $\mathbb{H}$ obtivemos um RKHS de sequências, que é uma representação da função aleatória $\{X(t), t \in T\}$ [27].

Podemos também definir uma função em T da seguinte forma

$$f(t) = \sum_{k=1}^{\infty} \lambda_k \alpha_k \varphi_k(t) \tag{A.13}$$

onde a sequência $\{\alpha_k, k = 1, 2, \dots\}$ satisfaz a condição (A.9).

Se $\mathbb{H}$ é um conjunto composto de funções $f(\cdot)$ que podem ser representadas na forma (A.13) em termos das autofunções $\{\varphi_k(t), k = 1, 2, \dots\}$ e dos autovalores $\{\lambda_k, k = 1, 2, \dots\}$ da função covariância $R(t_1, t_2)$. E, se podemos definir um produto interno de duas funções em $\mathbb{H}$ como

$$\langle f, g \rangle = \sum_{k=1}^{\infty} \lambda_k \alpha_k \beta_k \tag{A.14}$$

onde f, g são da forma (A.13) e as sequências α_k e β_k satisfazem a condição (A.9). $\mathbb{H}$ possui duas importantes propriedades que o tornam um RKHS. Primeiramente, seja $R(t_1, \cdot)$ a função em T com valor em t_2 igual a $R(t_1, t_2)$, então pelo teorema de Mercer a auto expansão para a função covariância (A.8) é dada por

$$R(t_1, t_2) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t_1) \varphi_k(t_2) \tag{A.15}$$

portanto, $R(t_1, \cdot) \in \mathbb{H}$ para cada $t \in T$. Segundo, para cada função $f(\cdot) \in \mathbb{H}$ da forma dada por (A.13) e para cada $t \in T$,

$$\langle f, R(t_1, \cdot) \rangle = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t_1) \varphi_k(t_2) = f(t) \tag{A.16}$$

pelo teorema de Moore – Aronszajn, $\mathbb{H}$ é um RKHS com $R(t_1, t_2)$ como um reproduzindo *kernel*, daí segue que

$$\langle R(t_1,\cdot), R(t_2,\cdot)\rangle = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t_1)\varphi_k(t_2) = R(t_1, t_2) \qquad \text{(A.17)}$$

assim, podemos concluir que $\mathbb{H}$ é uma representação do processo estocástico $\{X(t), t \in T\}$ com função de covariância $R(t_1, t_2)$. Então vimos claramente duas formas de se obter o mesmo resultado para a função covariância.

A abordagem do método RKHS em processamento estatístico de sinais foi proposto por Parzen no final de 1950, que fornecia pela primeira vez uma analogia utilizando análise funcional para processos estocásticos definidos por momentos de segunda ordem (chamados de processos Gaussianos), porque eles podem ser abordados por métodos puramente geométricos quando estudados em termos de seus momentos de segunda ordem (função covariância). Embora envolvam variáveis aleatórias, alguns problemas de processamento estatístico de sinais podem ser resolvidos algebricamente no *RKHS* associado com suas funções de covariância com todas as vantagens geométricas do produto interno definido em tais espaços.

A ideia principal é que existe um mapeamento isomorfo isométrico entre o espaço de Hilbert de variáveis aleatórias gerado por um processo estocástico e sua função de covariância que determina um único RKHS.

A.2.2 – Teoria de aprendizagem estatística.

Outra abordagem é a utilização do *RKHS* na teoria de aprendizagem estatística que explora maneiras de estimar uma dependência funcional em uma determinada coleção de dados. Ou seja, podemos utilizar como exemplo o dispositivo apresentado na figura 01, em que temos um conjunto de dados (variáveis aleatórias independentes e igualmente distribuídas) $\boldsymbol{x}$ com distribuição de probabilidade desconhecida $p(\boldsymbol{x})$ na entrada, e um conjunto $\boldsymbol{y}$ de dados na saída (variáveis aleatórias) para cada entrada $\boldsymbol{x}$, de acordo com uma distribuição de probabilidade condicional desconhecida $p(\boldsymbol{y}|\boldsymbol{x})$, e um mecanismo de aprendizagem que pode implementar um conjunto de funções $f(\boldsymbol{x}, \boldsymbol{w})$. O

objetivo é encontrar a função $f(\boldsymbol{x}, \boldsymbol{w})$ que minimize o funcional de custo na situação em que a função distribuição de probabilidade conjunta é desconhecida e somente possuímos as informações presentes nos dados de treinamento $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$. Uma forma de solucionar o problema é a utilização de um funcional de risco ou medida de discrepância $L(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w}))$ entre a saída $\boldsymbol{y}$ dada a entrada $\boldsymbol{x}$ e a resposta $f(\boldsymbol{x}, \boldsymbol{w})$ do mecanismo de aprendizagem para selecionar a melhor função. Na teoria de aprendizagem estatística, desenvolvida por Vapnik e Chervoneskis em 1990, o funcional de risco é utilizado para procurar a melhor função baseado num parâmetro de regularização da função custo desenvolvido a partir de um espaço RKHS induzido por um *kernel* $k(\boldsymbol{x},\cdot)$. Schölkopf e Smola em [27] mostraram que o RKHS desempenha um importante papel nos algoritmos baseados em apredizagem por *kernel*.

A partir do teorema de Mercer [29] em que uma função kernel $k(x, y)$ simétrica e definida positiva pode ser escrita como o produto interno entre dois vetores num espaço de característica de alta dimensionalidade, ou seja:

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x) \varphi_k(y) \qquad \text{(A.18)}$$

então os algoritmos de aprendizagem estatística, utilizam o teorema de Mercer para mapear os dados de entrada a um espaço de características de alta dimensionalidade (geralmente infinita). Este espaço de características é um *RKHS*, onde o mapeamento não linear Φ constitui a base. Então, os algoritmos de aprendizagem podem ser implementados em termos de produto interno, pois este mapeamento não linear se torna bastante útil e interessante, uma vez que podemos utilizar o *truque do kernel* para calcular os produtos internos no espaço de características através das funções do kernel (Gaussianas, polinomiais, multiquadráticas, etc.) sem a necessidade de conhecer este mapeamento. Então, a elegância do método reside no fato de que o produto interno dos dados transformados pode ser implicitamente calculado no RKHS, sem a necessidade explícita de se conhecer o mapeamento não linear.