UFABC

DISSERTAÇÃO DE MESTRADO

Redes Adversárias Geradoras para Separação Semi-Cega de Fontes

Tito Caco Curimbaba Spadini

Orientador: **Dr. Ricardo Suyama**

Santo André

2019

Resumo

Abstract

Epígrafe

A juventude envelhece, a imaturidade é superada, a ignorância pode ser educada e a embriaguez passa, mas a estupidez é eterna.

ARISTÓFANES

Minha alma é uma orquestra oculta; não sei que instrumentos tangem e rangem, cordas e harpas, timbales e tambores, dentro de mim. Só me conheço como sinfonia.

FERNANDO PESSOA (BERNARDO SOARES)

Sometimes you need to sit lonely on the floor in a quiet room in order to hear your own voice and not let it drown in the noise of others.

CHARLOTTE ERIKSSON

One person's data is another person's noise.

K.C. COLE

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Abreviações

ANN	*Artificial Neural Networks*
BSS	*Blind Source Separation*
DNN	*Deep Neural Network*
DRNN	*Deep Recurrent Neural Networks*
DWT	*Discrete Wavelet Transform*
EM	*Expectation-Maximization*
FA	*Factorial Analysis*
FOBI	*Fourth Order Blind Identificiation*
HMM	*Hidden Markov Model*
ICA	*Independent Component Analysis*
LVA	*Latent Variable Analysis*
MI	*Mutual Information*
MLE	*Maximum Likelihood*
NMF	*Non-Negative Matrix Factorization*

PCA *Principal Component Analysis*

ReLU *Rectified Linear Unit*

SAR *Signal-to-Artifact Ration*

SBSS *Semi-Blind Source Separation*

SCA *Sparse Component Analysis*

SDR *Signal-to-Distortion Ratio*

SIR *Signal-to-Interference Ration*

SNR *Signal-to-Noise Ration*

STFT *Short-Time Foutier Transform*

CAPÍTULO

1

Introdução

Embora o propósito deste trabalho não seja pesquisar todos os detalhes de cada área envolvida, é interessante fazer um passeio por alguns dos pontos mais importantes que ajudam a entender este problema, incluindo suas principais dificuldades e a base de como cada parte funciona. Assim, ao longo deste capítulo serão apresentados e explicados os principais conceitos, técnicas e abordagens associadas ao tema deste trabalho; uma nova abordagem também será apresentada.

1.1 Separação Cega de Fontes

Onde quer que estejam, os seres humanos estão sujeitos à presença de vários tipos distintos de sinais analógicos, independentemente de serem capazes de serem percebidos sem o auxílio de algum tipo de equipamento. O canto de um pássaro, os aplausos de uma multidão e as badaladas de um sino são exemplos bastante comuns que qualquer indivíduo pode vivenciar em seu dia a dia. Talvez um olhar mais científico sobre tais ocorrências tão corriqueiras seja algo incomum para a maior parte da população, porém, ainda assim, em todos os casos mencionados há muitos sinais transitando por todos os lados.

Segundo o professor Bhagwandas Pannalal Lathi,

> "*A signal is a set of data or information.* "
> [Lathi, 2009]

Em uma tradução livre, segundo a citação do professor Lathi, um sinal é um conjunto de dados ou de informação. Pode parecer uma frase simples – e, de fato, é –, mas carrega em si um significado enorme, sobretudo se forem consideradas algumas de suas implicações; trata-se, afinal, de uma definição bastante abrangente, mas bastante importante para tudo o que viria a ser futuramente desenvolvido a partir de então.

Sinais carregam informação, mas, para que um sinal seja chamado de "informação" usando esta nomenclatura em particular, uma dependência específica deve ser satisfeita: sua utilidade; caso contrário, poderia ser simplesmente considerado um ruído. Assim, do ponto de vista da engenharia de informação, o que permite distinguir linguisticamente um sinal puramente ruidoso de um sinal puramente informativo é a sua utilidade.

Devido à sua suscetibilidade a interferências causadas por sinais indesejados, que podem ser chamados de ruídos em alguns casos, qualquer sinal originalmente coletado da natureza traz consigo algum ruído. A origem deste ruído é irrelevante neste momento, mas suas características e a maneira como ele interfere no sinal original de informação pura são altamente relevantes, porque influencia a escolha das técnicas necessárias para filtrar o ruído [Tuzlukov, 2002] e um apropriado tratamento de ruído é muito importante [Boll, 1979; Donoho, 1995; Widrow *et al.*, 1975; Lee, 1980]. Tratamentos ineficazes podem resultar em danos aos dados, além de gastos de recursos valiosos, como tempo e energia, podendo até mesmo introduzir novos ruídos, o que só intensificaria ainda mais a complexidade do problema.

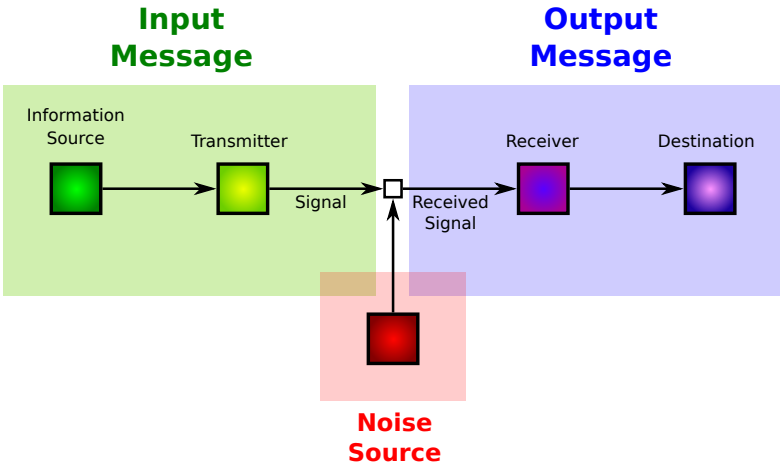

Figura 1.1: Diagrama esquemático representando um sistema de comunicação, inspirado em [Shannon, 1948].

Para visualizar melhor o processo de interferência causado pelo ruído sobre o sinal de informação, veja a Figura 1.1. Na verdade, a versão original dessa figura foi elaborada e utilizada pelo professor Claude Elwood Shannon em um cenário focado em comunicação, mas continua sendo uma boa ilustração da ideia por detrás do problema em sua forma mais ampla.

O primeiro bloco, a Fonte de Informação (do inglês, *Information Source*), representa a parte útil do sinal, o sinal original; o segundo, o Transmissor (do inglês, *Transmitter*), é responsável por enviar a mensagem através do meio (canal), representado pelo pequeno quadrado branco no centro do diagrama, que pode sofrer os efeitos de múltiplos tipos de ruído; o Receptor (do inglês, *Receiver*) interpreta o sinal recebido e fornece ao destino final (do inglês, *Destination*) as informações obtidas. Nesta mesma figura há um bloco intitulado *Noise Source*, que representa, de modo geral, todas as formas possíveis de ruído que sejam, de alguma maneira, capazes de interferir nocivamente no sinal assim que trafega ao longo do canal.

Por outro lado, há casos em que o objeto de interesse depende da separação dos sinais que estão envolvidos na composição, ou o interesse está no processo separação em si. Na literatura, há muitos bons exemplos de trabalhos que corroboram a importância de realizar uma separação de fontes, como [Belouchrani *et al.*, 1997; Nugraha *et al.*, 2016].

A tarefa a ser realizada para efetuar o processo de separação é conhecida como Separação de Fontes. Essa separação pode ser feita utilizando-se diversas técnicas distintas, que podem oferecer desempenhos melhores ou piores, depen-

dendo da situação, por isso a escolha da ferramenta deve ser feita cautelosamente. Todo problema a ser atacado envolve um cenário em particular, que é um dos principais fatores a serem considerados.

De modo geral, há cinco elementos a serem considerados ao todo o processo. São eles: as fontes de origem de cada sinal, o processo de mistura que gera sinais resultantes, os sensores que registraram os sinais misturados, o processo de separação que separa os sinais misturados, e os sinais independentes já separados. Alterações, ainda que sutis, em qualquer um dos cinco elementos mencionados já pode resultar em uma mudança de cenário, o que, por sua vez, pode implicar a demanda por uma mudança de técnica para que um bom desempenho na tarefa de separação seja obtido.

Casos que envolvem fontes de sinais e processos de mistura com características inicialmente desconhecidas são tratados como pertencentes a uma categoria conhecida como "cega", referindo-se à falta de informações relevantes sobre as duas partes mencionadas; por isso o nome Separação Cega de Fontes (**BSS**, do inglês *Blind Source Separation*). Também é comum ser utilizado o termo "não-supervisionado" para designar tal abordagem neste contexto. Quando se possui informações incompletas, parciais, sobre esses elementos, diz-se que se trata de uma categoria "semi-cega", por isso Separação Semi-Cega de Fontes (**SBSS**, do inglês *Semi-Blind Source Separation*).

Cabe aqui ressaltar o fato de que a separação cega de fontes é uma tarefa que não sofre limitações quanto às aplicações a que se destinam os sinais a serem separados, o que lhe confere um elevado grau de generalidade, permitindo que as ferramentas mais básicas exploradas com este mesmo objetivo possam ser utilizadas para aplicações em sinais cerebrais, música e telecomunicações, além de diversas outras áreas de aplicação.

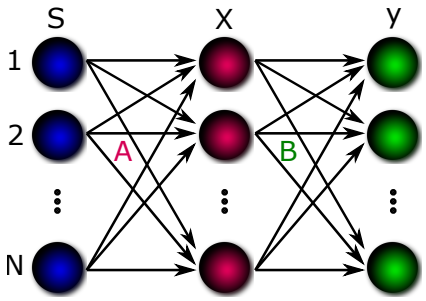

Figura 1.2: Uma representação genérica de um problema de BSS. S é o conjunto de fontes originais, X é o conjunto de sensores e y é o conjunto de saídas após todo o processo de separação de sinais. A representa a mistura do meio e do sensor em si; B, o processo de separação.

A Figura 1.2 ilustra um exemplo de problema geral de separação cega de fontes e pode ser usada para descrever esquematicamente uma ampla gama de diferentes aplicações. É importante lembrar que número de elementos em cada conjunto (S, X e y) pode ser diferente, ou seja, dado há casos que podem envolver um número de fontes superior ao número de sensores, assim como o contrário também pode ocorrer. Ainda que houvesse apenas um único sensor compondo X e um número enorme de fontes compondo S, por mais complexa que fosse a tarefa de separação, seria um caso possível e digno de ser interpretável como pertencente ao conjunto representado por tal ilustração.

Do ponto de vista da disponibilidade de informação, o cenário mais complexo trata-se do caso em que X é composto exclusivamente por um único sensor. Em um contexto de fontes de sinais de áudio, tal cenário é descrito como "monoaural" na literatura. Existem abordagens específicas para se trabalhar com tal cenário, mas esse não é o cenário a ser considerado neste trabalho.

Considere o cenário específico proposto na Figura 1.3, sobre uma situação usual de dois sinais sendo gravados por dois microfones. Em uma sala com os dois instrumentos musicais sendo tocados simultaneamente, ambos os microfones capturariam misturas resultantes compostas por variações de ambos os instrumentos musicais. Mas o resultado desejado seria um único sinal do violão e outro sinal único do piano. No cenário descrito, apesar de sua complexidade, quando a matriz da mistura é invertível, essa meta passa a ser plenamente alcançável.

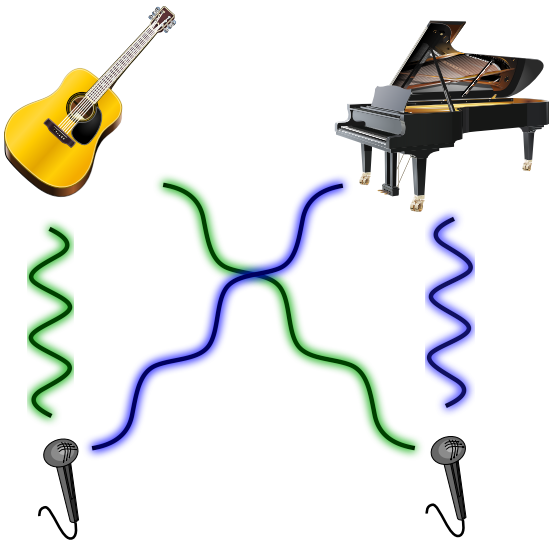

Figura 1.3: Um exemplo típico de um problema BASS com múltiplos sinais de interesse e múltiplos microfones. Note que ambos os microfones gravam os sinais acústicos de ambos os instrumentos musicais, mas, devido ao fato de não serem microfones perfeitos, além de registrarem os ruídos do ambiente, introduzirão à mistura o seu próprio ruído característico.

Existem várias técnicas diferentes que podem ser exploradas para alcançar um sinal resultante satisfatório para cada instrumento musical correspondente, por exemplo, o uso de informações tempo-frequenciais sobre cada instrumento musical envolvido e o uso de modelos probabilísticos para identificar as características desses sinais. Um bom trabalho explorando uma abordagem com múltiplos microfones foi feito em [Gannot *et al.*, 2017].

Se um dos microfones fosse removido, isso poderia ser interpretado erroneamente como uma variação simples do problema mostrado na Figura 1.3; a interpretação correta, na verdade, seria a de que trata-se de um problema totalmente diferente. Quando existem múltiplos microfones, um número maior de recursos está disponível, pois torna-se possível explorar características como distância relativa, o que contribui para facilitar o trabalho de separação de fontes.

É importante, no entanto, apontar para o fato de que esta seria uma perspectiva probabilística, ou seja, esta modelagem de problemas em particular exigiria uma análise não-determinística, o que geralmente requer um maior número de

considerações e implicações, bem como aumentar a complexidade do problema, apesar de permitir uma análise muito mais flexível; dois exemplos de trabalhos que usaram essa abordagem monoaural são [Huang *et al.*, 2015], que explorou Redes Neurais Recorrentes Profundas (**DRNN**, do inglês *Deep Recurrent Neural Networks*), e [Chien & Yang, 2016], que usou a Fatoração de Matriz Não Negativa (**NMF**, do inglês *Non-Negative Matrix Factorization*) Bayesiana.

A seguir serão trazidas algumas abordagens existentes sobre o uso de redes neurais e afins em um contexto de separação de fontes, oferecendo uma perspectiva histórica e exemplos de trabalhos marcantes da área.

1.2 Redes Neurais no contexto de BSS

Existem diversas abordagens distintas para se realizar a tarefa de separação cega de fontes. Algumas delas exploram o uso de redes neurais, sobretudo quando há maior disponibilidade de poder computacional, mas a utilização de redes neurais em dois trabalhos não implica atacar o problema da mesma maneira em ambos os trabalhos. É importante que sejam consideradas as informações *a priori*, a arquitetura da rede, as características exploradas, a presença ou ausência de rótulos para comparação, as técnicas de extração e seleção de características, as técnicas de validação cruzada, as métricas para aferição de desempenho etc.

Devido às suas propriedades generalistas, possivelmente interpretadas como um caráter positivo de ampla flexibilidade, as Redes Neurais Artificiais (**ANN**, do inglês *Artificial Neural Networks*), que serão explicadas mais detalhadamente no Capítulo 3, podem oferecer um alto nível de efetividade na tarefa de separação cega de fontes, explicada na Seção 1.1 e mais detalhada no Capítulo 2. Tais características conferem às ANNs uma imensa abrangência quanto à sua gama de aplicações e a tarefa de separação de fontes pode ser muito beneficiada de tais qualidades; alguns exemplos interessantes serão discutidos ainda nesta seção.

Um dos mais importantes trabalhos da área é [Hérault *et al.*, 1985], que trouxe um conceito chamado *computação de arquitetura neuromimética*[1], que foi uma das essências do uso de redes neurais com a finalidade de separar sinais com base em uma abordagem de aprendizado não supervisionado.

Mais tarde, outro importante trabalho sobre separação de fontes foi publicado

[1] "Neuromimética" pode ser interpretado como uma imitação do sistema (ou de partes do sistema) nervoso.

[Cardoso, 1989], mas com enfoque na utilização de momentos de altas ordens[2] para isso. O trabalho reforçou a ideia de que a independência estatística dos sinais tratava-se de uma propriedade muito mais forte do que a sua mera descorrelação e, também, desde que fosse constatada uma diferença de distribuição entre os sinais envolvidos, seria possível identificar as assinaturas das fontes sem a dependência de quaisquer modelos *a priori*, sendo que as tais assinaturas são autovetores da matriz de covariância após ter sido realizado um processo de ortonormalização e ponderação não linear.

O Algoritmo 1 foi proposto para realizar a tarefa chamava-se FOBI (do inglês, ***F****ourth-****O****rder* ***B****lind* ***I****dentification*)

[2]Em estatística, é comum entender como ordens superiores as ordens de grau igual ou maior que 3.

Algoritmo 1 FOBI (Fourth-Order Blind Identification) [Hérault *et al.*, 1985].

1: Calcular a covariância:

$$R_X \leftarrow E\left(XX^T\right) , \tag{1.1}$$

2: Fatorar a covariância:

$$R_X = CC^T , \tag{1.2}$$

3: Ortonormalizar os dados:

$$Y \leftarrow C^{-1}X , \tag{1.3}$$

4: Formar a covariância ponderada:

$$\tilde{R}_Y \leftarrow E\left(|Y|^2 Y Y^T\right) , \tag{1.4}$$

5: Extrair os autovetores da matriz de covariância:

$$\tilde{R}_Y = \sum_{i=1}^{N} (\mu_i + N - 1)\, Y_i Y_i^T , \tag{1.5}$$

6: Extrair as mensagens[a]:

$$\alpha_i \leftarrow Y_i^T Y , \tag{1.6}$$

7: Identificar as assinaturas[b]:

$$X_i \leftarrow C Y_i . \tag{1.7}$$

[a] Uma sequência de valores (dados) que pode carregar em si alguma informação, em um contexto de comunicação, geralmente é chamada de "mensagem".

[b] Assinatura pode ser entendida como o sinal original de informação advinda de uma dada fonte de comunicação.

1.2.1 Aplicações Contemporâneas

O trabalho [Uhlich *et al.*, 2015] utilizou-se de uma Rede Neural Profunda (**DNN**, do inglês *Deep Neural Network*) para realizar um processo de extração de instrumentos musicais presentes em sinais resultantes de misturas musicais. Neste trabalho, a informação *a priori* apenas assumiu conhecer os instrumentos que compunham as misturas, o que permitiu que um conjunto de dados de misturas para treinamento fosse construído a partir de trechos de áudio de sinais contendo exclusivamente instrumentos individuais, ou seja, misturas artificialmente produzidas. Fizeram parte das composições os seguintes instrumentos musicais:

fagote, violoncelo, clarinete, chifre, piano, saxofone, trompete, viola e violino.

A arquitetura explorada, que utilizou Unidade Linear Retificada (**ReLU**, do inglês *Rectified Linear Unit*) [Nair & Hinton, 2010] como sua função de ativação, era composta por camadas intermediárias contendo o mesmo número de neurônios presentes na camada de saída, e a inicialização foi feita com base no algoritmo de mínimos quadrados. Foram consideradas como métricas a Relação Sinal-Distorção (**SDR**, do inglês *Signal-to-Distortion Ratio*), a Relação Sinal-Interferência (**SIR**, do inglês *Signal-to-Interference Ratio*), e a Relação Sinal-Artefato (**SAR**, do inglês *Signal-to-Artifact Ratio*).

Este outro trabalho [Li & Zhang, 2012] utilizou-se de uma combinação entre Análise de Componentes Independentes (**ICA**, do inglês *Independent Component Analysis*) e redes neurais para realizar a separação cega de fontes ruidosas de sinais de voz em múltiplos canais, fazendo importantes considerações quanto aos ruídos associados durante todo o processo, inclusive pelo uso de redes neurais especificamente capazes de fazer tratamentos de remoção de ruídos dos sinais. A métrica de Relação Sinal-Ruído (**SNR**, do inglês *Signal-to-Noise Ratio*) foi utilizada para se analisar o desempenho deste trabalho. É importante mencionar também a realização de duas etapas altamente influentes no desempenho obtido por este trabalho, que são: o branqueamento dos dados e a técnica *Windage wipe off*.

1.3 Desafios

Assim como ocorre em qualquer outra área do conhecimento, a área de separação de fontes possui os seus desafios. Conforme pesquisas na matemática, na ciência e na engenharia avançam, cada vez mais formas eficazes e eficientes são desenvolvidas para que esse objetivo possa ser atingido. Contudo, uma característica intrínseca da ciência é a sua incessante busca por respostas; e sabe-se que ao longo da busca pela resposta de uma pergunta, inevitavelmente, novas dúvidas serão encontradas, sendo que estas requererão novas buscas por respostas, preservando-se, então, tal apaixonante comportamento *Ad infinitum*.

Novas dificuldades, novas barreiras, novos problemas e novas dúvidas sempre existirão. Por mais que se consiga vencer, transcender, resolver e responder, sempre haverá muito espaço novo a ser explorado por quem estiver interessado e for apto a desbravar-se nesse meio. As subseções a seguir trarão explanações

acerca de alguns dos desafios mais comuns desta área.

1.3.1 Subparametrizado

Um caso bastante corriqueiro desta área de pesquisa é conhecido como subparametrizado (ou, para fontes em inglês, *Undetermined*), que consiste em um cenário cujo número de variáveis envolvidas é superior ao número de equações, impossibilitando que abordagens mais simples e diretas atinjam a eficácia desejada, pois, como se sabe, trata-se de um sistema com número infinito de soluções possíveis, o que está em desacordo com o que se almeja obter.

Para os fins deste trabalho, dado que as aplicações finais serão todas focadas em sinais de áudio, pode-se adaptar o uso de tais termos para uma linguagem mais próxima a tal área, de modo que a interpretação do problema se faça mais clara a partir de então. Assim sendo, o caso subparametrizado pode ser interpretado como sendo aquele em que o número de fontes de áudio é maior que o número de microfones.

Apenas para fins de exemplificação, poderia-se imaginar uma performance musical realizada por uma orquestra sendo gravada com um único microfone; a tarefa de separação neste caso seria subparametrizada, dado que seriam muitos instrumentos musicais a serem separados com base em um único sinal de uma mistura resultante do único microfone utilizado.

1.3.2 Convolutivo

Apesar de existir todo um conjunto de cenários mais simples com misturas de características lineares, modelos projetados para cenários com tais características tendem a ser menos flexíveis, mostrando-se aceitáveis quase que exclusivamente quando há um rigor menos intenso na demanda por qualidade

1.3.3 Não-Linear

1.4 Nova Abordagem – Deep Learning e GAN

1.5 Objetivos

Os principais objetivos perseguidos por este trabalho em particular são: a realização da tarefa de separação cega de fontes de áudio usando redes adversárias geradoras e o desenvolvimento de uma nova topologia baseada em seu modelo original.

1.6 Estrutura da Dissertação

O Capítulo 2, intitulado "Separação Cega de Fontes", explicará alguns diferentes cenários, bem como aspectos e influências de linearidade e memória; também serão expostos problemas relacionados e técnicas clássicas, que já são amplamente conhecidas e tipicamente utilizadas para realizar tarefas relacionadas ao problema de separação cega de fontes.

O Capítulo 3 é reservado ao fornecimento de informações importantes sobre toda a base do que será explorado sobre Redes Neurais Artificiais ao longo deste trabalho, partindo de elementos históricos, passando pelas explicações sobre *Perceptrons*, arquiteturas e propriedades, explanando a essência do *Backpropagation* –um dos algoritmos mais importantes desta área– e, finalmente, chegando às redes profundas, que são as que serão, de fato, utilizadas nos algoritmos deste trabalho.

O Capítulo 4 explicará diversos aspectos sobre as Redes Adversárias Geradoras, desde sua origem, as bases de inspiração apoiadas em conceitos de teoria dos jogos, elementos estruturais, características intrínsecas desse modelo, alternativas para contornar problemas constatados e, por fim, aplicações em imagens, que, até então, é o tipo de mídia mais amplamente explorado quando tal ferramenta é utilizada.

O Capítulo 5 entrará no que, efetivamente, é o objetivo principal deste trabalho, i.e., a utilização de redes adversárias geradoras para a realização da tarefa de separação cega de fontes, que neste caso são sinais de áudio. Haverá uma seção destinada à explicação sobre como utilizar tal recurso. Também serão feitas diversas considerações relevantes, bem como uma ampla revisão bibliográfica contemporânea. Ao fim, serão discutidas as limitações identificadas e serão trazidas algumas propostas.

O Capítulo 6 trará uma série de simulações para diferentes cenários com suas

respectivas metodologias explicadas, bem como resultados devidamente acompanhados de suas discussões.

O Capítulo 7 encerrará o trabalho com as conclusões obtidas ao longo de todo este trabalho, bem como perspectivas futuras para possíveis caminhos a serem doravante trilhados.

CAPÍTULO

2

Separação Cega de Fontes

Antes de iniciar as discussões acerca da tarefa de separação cega de fontes propriamente dita, pode ser interessante realizar uma explicação sobre um cenário básico e generalista.

Assim sendo, imagine que haja três diferentes sinais de áudio originais, $s_1(t)$, $s_2(t)$ e $s_3(t)$, que tenham sido emitidos por diferentes fontes de origem, e que, após terem passado por processos de mistura, foram registrados como sinais misturados resultantes por três microfones. Desprezando-se quaisquer outros elementos que possam ser capazes de influenciar diretamente nas misturas, pode-se entender que os sinais resultantes dos processos de mistura são passíveis de serem representados por somas ponderadas dos sinais originais, como mostrado na Equação (2.1),

$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\ x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\ x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \end{cases}, \tag{2.1}$$

sendo que x_i representa o sinal resultante do processo de mistura, também conhecido como *observação*; a_{ij} compreende o peso da ponderação atribuído a cada elemento da mistura, que é um dado inicialmente desconhecido; e s_i trata-se do sinal original, que é desconhecido e que é justamente o que se deseja obter ao final de todo o processo e, como de costume, cada sinal é tipicamente compreen-

dido como um processo estocástico estacionário de média zero com valores reais [Papoulis, 1984].

Uma versão matricial da Equação (2.1) pode ser vista na Equação (2.2):

$$\underbrace{\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{bmatrix}}_{\mathbf{s}}. \tag{2.2}$$

Isolando-se a matriz de sinais originais, $\mathbf{s}$, será necessário obter a inversa da matriz $\mathbf{A}$, que será tratada como $\mathbf{W}$ neste trabalho, ou seja, $\mathbf{W} = \mathbf{A}^{-1}$, com $\mathbf{W}$ sendo

$$\mathcal{W} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}, \tag{2.3}$$

então, pode-se apresentar uma nova formulação que permitirá a obtenção dos sinais originais, desde que algumas condições sejam satisfeitas, que são: a necessidade de o número de observações (sinais resultantes após o processo de mistura) ser superior ao número de fontes (sinais originais); a constatação de uma independência estatística das fontes; a não estacionariedade dos processos de geração dos sinais envolvidos; e o respeito ao limite superior de uma única fonte de distribuição Gaussiana [Comon, 1994; Romano *et al.*, 2010].

Há também a necessidade de a matriz $\mathbf{A}$ ser, de fato, inversível[1] e, também, a necessidade de a matriz $\mathbf{A}$ ser plenamente conhecida, ou seja, que todos os valores de a_{ij} sejam conhecidos. Em tal situação, pode-se representar o problema da seguinte maneira:

$$\underbrace{\begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}}_{\mathbf{W}} \underbrace{\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix}}_{\mathbf{x}}, \tag{2.4}$$

[1] Uma forma relativamente prática de verificar tal condição depende apenas do valor do determinante da matriz em questão; caso seja igual a zero, a matriz não é inversível.

ou ainda, novamente na forma de um sistema de equações,

$$\begin{cases} y_1(t) = w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \\ y_2(t) = w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t) \\ y_3(t) = w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t) \end{cases} . \tag{2.5}$$

Mas esta é apenas uma visão teórica de uma parte–digamos–mais controlada do problema. O que se tem na prática, efetivamente, é que a tarefa de se obter cada componente em questão pode ser algo impraticável ou, no mínimo, inviável. Assim, antes de partir para a tarefa de identificação dos componentes em si, deve-se realizar um pré-processamento, a ser melhor explicado na Seção 2.1.

2.1 Pré-processamento

Independentemente de qual seja o processo a ser utilizado sobre um certo conjunto de dados, deve-se sempre procurar garantir que tais dados estejam adequadamente preparados para que possam ser devidamente trabalhados. Para que isso seja possível, há uma série de tarefas anteriores às etapas associadas ao grupo de processamento; a esse grupo de tarefas anteriores dá-se o nome de pré-processamento. Uma possível interpretação para tal etapa é a de que trata-se, simplesmente, de toda a preparação dos dados para que, só então, eles venham a ser, de fato, processados (ou, se preferir, trabalhados).

Antes de processar as observações dos sinais misturados, então, deve-se realizar uma tarefa conhecida como Branqueamento[2] (do inglês *Whitening*) [Bell & Sejnowski, 1997; Koivunen & Kostinski, 1999], que garantirá que, para um dado vetor composto por valores aleatórios com média zero, a variância será igual a 1 e – mais importante que isso – seus componentes serão descorrelacionados. A Figura 2.1 ilustra esquematicamente o que é almejado pelo processo de branqueamento; e tal processo pode ser realizado utilizando-se os passos representados pelo Algoritmo 2.

[2] Além do termo *Whitening* designado a este processo em literaturas de língua inglesa, pode-se encontrar também o termo *Sphereing*, referindo-se precisamente ao mesmo processo. A Figura 2.1 ajuda a compreender o motivo dessa outra nomenclatura.

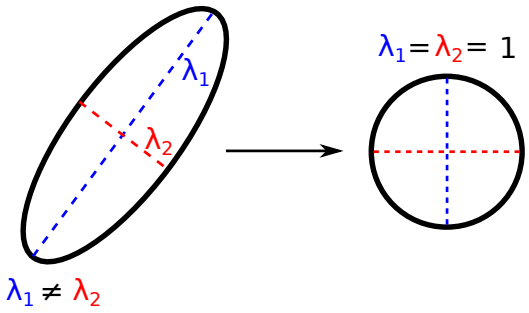

Figura 2.1: Simplificação esquemática, a partir de uma perspectiva geométrica (inspirada em [Zafeiriou, 2015]), dos objetivos do branqueamento utilizando o PCA.

Algoritmo 2 Branqueamento de dados por PCA [Zafeiriou, 2015].

1: Calcular o produto escalar matricial:

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \boldsymbol{\mu}) \tag{2.6}$$

2: Efetuar a auto-análise:

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T \tag{2.7}$$

3: Calcular os autovetores:

$$\mathbf{U} \leftarrow \mathbf{X}\mathbf{V}\boldsymbol{\Lambda}^{-\frac{1}{2}} \tag{2.8}$$

4: Calcular as d características:

$$\mathbf{Y} \leftarrow \mathbf{U_d}^T\mathbf{X} \tag{2.9}$$

5: Rearranjar a matriz de covariância de $\mathbf{Y}$:

$$\mathbf{Y}\mathbf{Y}^T = \mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U} = \boldsymbol{\Lambda} \tag{2.10}$$

6: Obter a matriz de projeção normalizadora:

$$\mathbf{W} \leftarrow \mathbf{U}\boldsymbol{\Lambda}^{-\frac{1}{2}}. \tag{2.11}$$

Vale ressaltar o fato de que o procedimento adotado pelo Algoritmo 2 carece de um único passo intermediário, que é o de manter os n primeiros componentes

de $\mathbf{U_d}$, com $n \leq d$, para ser interpretado no sentido mais amplo como, basicamente, o próprio algoritmo de Análise de Componentes Principais (**PCA**, do inglês *Principal Component Analysis*) [Pearson, 1901].

Para fins meramente ilustrativos, uma simulação foi elaborada para mostrar em um gráfico os efeitos do processo de branqueamento por PCA. Para esta simulação foram utilizados dados artificialmente gerados por meio de duas variáveis aleatórias de mesma distribuição uniforme. Os dados, compostos por 1000 amostras para cada uma das variáveis envolvidas, representados no gráfico da Figura 2.2 foram posicionados com base nos valores das variáveis aleatórias X_1 e X_2, que podiam variar entre -1.75 e 1.75, que foram arbitrariamente escolhidos.

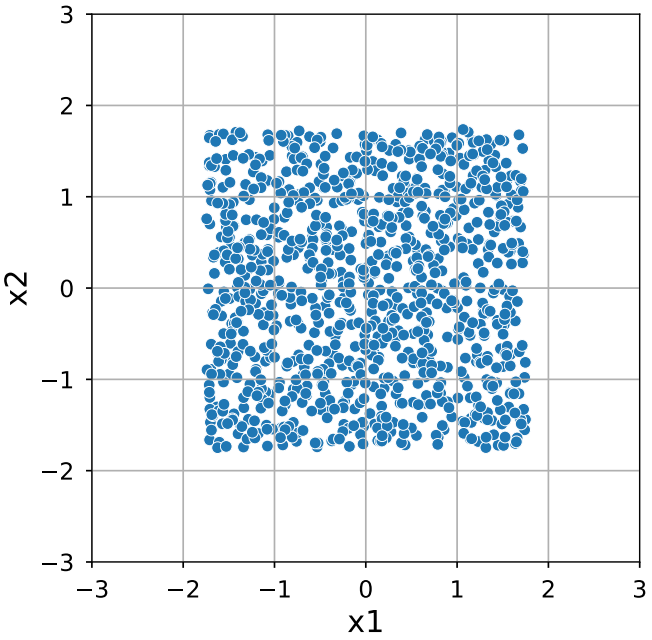

Figura 2.2: Distribuição de fonte conjunta para duas variáveis aleatórias de distribuição uniforme.

A Figura 2.3 mostra a distribuição conjunta dos mesmos dados exibidos pela Figura 2.2, mas agora em uma versão que passou pelo processo de branqueamento antes de ter sido plotada.

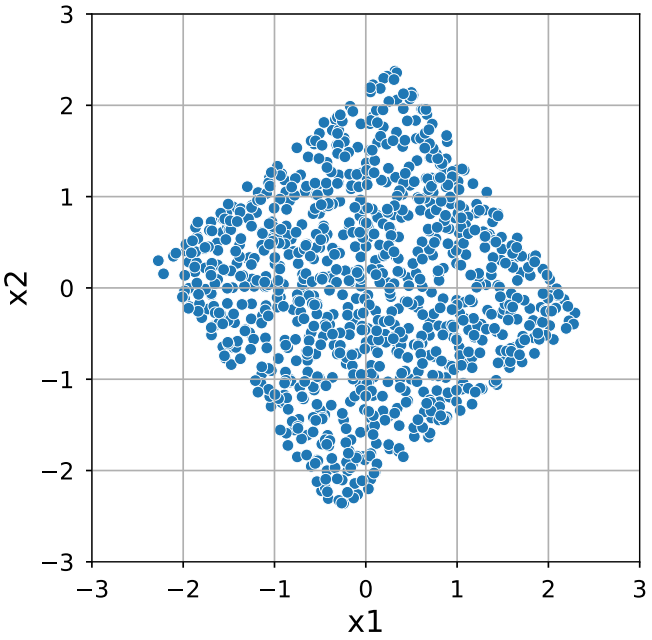

Figura 2.3: Distribuição conjunta de dados branqueados para duas variáveis aleatórias de distribuição uniforme.

Uma forma de se obter os valores de w_{ij} é a partir da independência estatística das misturas. A constatação de distinções que preservem uma elevada não-gaussianidade nas distribuições envolvidas pode ser explorada para efetuar eficazmente a separação dos sinais misturados com base nessa independência; a partir disso foi elaborado o método que ficou conhecido como Análise de Componentes Independentes, que será melhor explicado na Seção 2.4.

2.2 Características dos Modelos

O problema de separação de sinais não é algo simples de ser resolvido; para compreendê-lo melhor, pode ser muito convidativa a utilização de formas esquemáticas ilustrativas que abordem o problema de um modo mais genérico. Contudo, a utilização de formas generalistas em demasia pode transmitir a errônea impressão de que as técnicas gerais podem ser aplicadas em múltiplos cenários consideravelmente distintos com grau de eficácia equiparável, o que não reflete a

realidade. Para se atingir maiores níveis de eficácia e de eficiência, pode ser muito vantajoso compreender especificamente o cenário em que se realizará a separação de fontes e, a partir de modelos que considerem as características inerentes a tal cenário, seja feita a separação. Características relacionadas à linearidade e à memória serão melhor tratadas a seguir.

2.2.1 Linearidade

De acordo com o Princípio da Superposição, sistemas lineares compostos pela soma de múltiplos estímulos terão como respostas a soma das respostas individuais para cada entrada individual no sistema, assim como a proporção escalar da entrada resultará em igual proporção escalar da saída. Em outras palavras, a soma das entradas em um determinado sistema linear resultará na soma das saídas desse mesmo sistema, respeitando-se as suas respectivas proporções. Vale lembrar que o Princípio da Superposição pode ser matematicamente descrito a partir de duas propriedades conhecidas como Aditividade e Homogeneidade:

Definição 2.2.1 Princípio da Superposição

Sejam $\boldsymbol{F}(\cdot)$ uma função linear, $\boldsymbol{x}$ uma variável de entrada e α um escalar. Assim, a partir das propriedades de Aditividade e Homogeneidade

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{x}_1 + \boldsymbol{x}_2) &= \boldsymbol{F}(\boldsymbol{x}_1) + \boldsymbol{F}(\boldsymbol{x}_2) && \textbf{Aditividade} \\ \boldsymbol{F}(\alpha\boldsymbol{x}) &= \alpha\boldsymbol{F}(\boldsymbol{x}) && \textbf{Homogeneidade} \end{aligned}$$

pode-se chegar ao **Princípio da Superposição**:

$$\begin{aligned} \boldsymbol{F}(\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2) &= \boldsymbol{F}(\alpha_1\boldsymbol{x}_1) + \boldsymbol{F}(\alpha_2\boldsymbol{x}_2) \\ &= \alpha_1\boldsymbol{F}(\boldsymbol{x}_1) + \alpha_2\boldsymbol{F}(\boldsymbol{x}_2) \end{aligned} \tag{2.12}$$

Assim sendo, a partir da definição do princípio da superposição, é possível considerar uma analogia para sinais misturados por um processo de mistura como um sistema linear, representado aqui por $\boldsymbol{F}(\cdot)$. Considerando, então, os sinais $\boldsymbol{s}_1[n]$ e $\boldsymbol{s}_2[n]$ e os escalares α_1 e α_2, o modelo linear pode ser representado por

$$\boldsymbol{F}(\alpha_1\boldsymbol{s}_1[n] + \alpha_2\boldsymbol{s}_2[n]) = \alpha_1\boldsymbol{F}(\boldsymbol{s}_1[n]) + \alpha_2\boldsymbol{F}(\boldsymbol{s}_2[n])\,. \tag{2.13}$$

Todos os casos que não possam ser corretamente representados por formas análogas advindas de uma generalização feita a partir da Equação (2.13) são chamados *não-lineares*. Os casos não-lineares comumente são mais complexos e, assim, demandam o uso de técnicas mais avançadas, além de nem sempre permitirem que resultados com níveis satisfatórios de qualidade sejam alcançáveis por meio das técnicas que geralmente costumam ser exploradas para tais fins. Algumas das técnicas exploradas nesses casos são: ICA Não-Linear, Minimização de Informação Mútua e Gassianização.

2.2.2 Memória

Uma forma bastante usual de se analisar a questão da memória é separando-a em sistemas sem memória e sistemas com memória, sendo que neste contexto é comum referir-se aos sistemas com memória como sendo sistemas convolutivos, dado que tais sistemas são dependentes de memória, além de serem tipicamente lineares na maior parte dos casos. Sistemas com memória tendem a depender de tratativas importantes para que seja possível realizar implementações dependentes de processamento em tempo real, devido ao fato de serem mais suscetíveis a problemas provocados por atrasos, sobretudo quando os atrasos sofrem variações em um largo intervalo.

Definição 2.2.2 Convolução

Sejam $s_1[n]$ e $s_2[n]$ dois sinais de tempo discreto; a convolução entre eles é definida por:

$$\begin{aligned} y[n] &= s_1[n] * s_2[n] \\ &= \sum_{i=-\infty}^{+\infty} s_1[i] \cdot s_2[n-i] \end{aligned}. \tag{2.14}$$

Sendo que a convolução possui as propriedades de Associatividade, Comutatividade e Distributividade, apresentadas a seguir:

$$(s_1[n] * s_2[n]) * s_3[n] = s_1[n] * (s_2[n] * s_3[n]) \quad \textbf{Associatividade}$$

$$s_1[n] * s_2[n] = s_2[n] * s_1[n] \quad \textbf{Comutatividade}$$

$$s_1[n] * (s_2[n] + s_3[n]) = s_1[n] * s_2[n] + s_1[n] * s_3[n] \quad \textbf{Distributividade}$$

2.3 Problemas Relacionados

2.3.1 Extração Cega de Fontes

2.3.2 Desconvolução

2.3.3 Denoising

2.4 Análise de Componentes Independentes

O método de Análise de Componentes Independentes (**ICA**, do inglês *Independent Component Analysis*) é um dos mais importantes e consagrados métodos de separação de sinais já desenvolvidos até então. É, provavelmente, o método mais amplamente utilizado para as mais diversas aplicações dependentes de um processo de separação de fontes, sobretudo para BSS. O método de ICA pode ser interpretado como uma extensão agregadora do PCA [Comon, 1994], explorando independência estatística em vez de apenas considerar correlação e covariância.

Apenas como uma questão de curiosidade, é interessante lembrar que uma das conferências internacionais de maior importância na área de separação de sinais trazia os termos "Independent Component Analysis" em seu título, certamente que por motivos mais do que justos na época, o que se manteve até o ano de 2009; a partir de então, em vez de disso, por uma questão de generalidade e abrangência, dadas as evoluções na área, compreendeu-se que seria mais conveniente classificar o ICA como um dos possíveis métodos de algo ainda maior, chamado de Análise de Variáveis Latentes[3] (**LVA**, do inglês *Latent Variable Analysis*), então o termo que assumiu tal lugar no nome desta importante conferência internacional foi justamente "Latent Variable Analysis".

Além do ICA, outros métodos que também se encaixam na categoria de LVA são: o Modelo Oculto de Markov (**HMM**, do inglês *Hidden Markov Model*) [Ra-

[3]De modo praticamente oposto às variáveis observáveis, as variáveis latentes são as que não podem ser diretamente aferidas, lidas ou constatadas; em vez disso, são variáveis que dependem de algum tipo de transformação, adaptação ou inferência para que, só então possam ser obtidas e, enfim, exploradas.

biner, 1989], PCA, a Análise Fatorial (**FA**, do inglês *Factorial Analysis*)[4] [Vincent, 1953; Anderson & Gerbing, 1988; Kaiser, 1960], e o algoritmo de Maximização de Expectativas (**EM**, do inglês *Expectation-Maximization*) [Dempster *et al.*, 1977].

Partindo agora para a definição do que vem a ser efetivamente o método de ICA, pode-se dizer que, segundo [Hyvärinen *et al.*, 2004], possuindo-se diretamente apenas os sinais resultantes (misturados), $x_i(t)$, o ICA se trata da estimação de dois específicos elementos, que são: a matriz ponderadora **A**, que é responsável pelos pesos das combinações lineares que neste contexto são convenientemente compreendidas como misturas; e os sinais originais independentes, $s_i(t)$, ou seja,

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{bmatrix}, \tag{2.15}$$

que, assim como havia sido mostrado no início deste capítulo, também poderiam ser compreendidas na forma de um sistema de equações lineares, tal como

$$\begin{cases} s_1(t) = & w_{11}x_1(t) + w_{12}x_2(t) + \cdots + w_{1n}x_3(t) \\ s_2(t) = & w_{21}x_1(t) + w_{22}x_2(t) + \cdots + w_{2n}x_3(t) \\ \quad \vdots & \\ s_n(t) = & w_{n1}x_1(t) + w_{n2}x_2(t) + \cdots + w_{nn}x_3(t) \end{cases} \tag{2.16}$$

porém, apesar de em alguns cenários a visualização do problema na forma de um sistema de equações, tal como pode ser observado em (2.16), ser mais humanamente conveniente – talvez por uma questão de hábito –, as implementações computacionais são dependentes de uma linguagem que esteja de acordo com representações matriciais, por isso a forma observada na Equação (2.15) acaba sendo bem mais conveniente para trabalhos como este.

[4] Análise Fatorial, na verdade, trata-se de um conjunto de técnicas, não de uma única técnica *Stricto sensu*. Isso faz com que não haja um único trabalho específico que possa ser utilizado como a referência; o que há, porém, é uma vasta gama de trabalhos sobre as técnicas propriamente ditas, podendo ser sobre as técnicas em si ou aplicações –comumente, em pesquisas de Psicologia– que as explorem.

Ainda em [Hyvärinen *et al.*, 2004] é sugerido um método alternativo para implementar o ICA, que é encontrando-se uma transformação linear dada por uma matriz **W**, a mesma da Equação (2.4), de modo que as variáveis y_i, com $i = 1, \ldots, n$, sejam o mais independentes possível, porém o próprio autor faz a observação de que não chega a se tratar de um método realmente diferente do anterior, dado que a matriz **W** nada mais é do que a inversa da matriz **A**.

Dado que o ICA estima a transformação ortogonal que resta após ter sido feita a descorrelação[5], que pode ser realizada utilizando-se métodos já plenamente conhecidos e bastante consolidados a partir de qualquer mistura linear de componentes independentes, pode-se dizer que, ao menos em relação à tarefa de separação de fontes para cenários tais quais os descritos, o ICA está em uma posição vantajosa perante métodos clássicos, pois esses métodos mais clássicos são baseados na covariância, dada pela Equação (2.17), ou seja, ao final do processo, acabam explorando essencialmente a mesma informação de métodos que se baseiam puramente na descorrelação [Hyvärinen *et al.*, 2004].

Definição 2.4.1 Covariância

Sejam X e Y duas variáveis aleatórias. A covariância entre X e Y é definida por

$$\begin{aligned} cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned} \tag{2.17}$$

sendo $cov(X, Y)$ a covariância entre as variáveis aleatórias X e Y; e $E[\cdot]$ a esperança[a] de uma dada variável aleatória qualquer.

[a] Também conhecido como *valor esperado* ou *média* (estatística).

Assim sendo, considerando-se que a descorrelação entre as variáveis aleatórias X e Y, necessariamente, implica

$$E[XY] = E[X]E[Y], \tag{2.18}$$

[5] Tarefa de redução de autocorrelação em relação a um dado sinal ou, caso se trate de um conjunto de sinais, redução de correlação cruzada, porém, preservando-se os demais aspectos do sinal em questão.

então, pode-se concluir que, a partir das Equações (2.17) e (2.18), a covariância entre X e Y será zero, tal como demonstrado em (2.19)

$$\begin{aligned} cov(X, Y) &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] \\ &= 0 \,. \end{aligned} \tag{2.19}$$

Um dos princípios de estimação do ICA está associado à descorrelação não linear, obtendo-se uma matriz $\mathbf{W}$ capaz de reverter o processo de mistura, mas garantindo que todos os elementos que compõem a matriz de saída sejam, dois a dois, descorrelacionados entre si para todas as possíveis combinações, além de as transformações não lineares (cuidadosamente selecionadas), $g(y_i)$ e $h(y_j)$, também serem descorrelacionadas entre si.

Uma escolha apropriada das funções g e h é crucial para que se atinja tal objetivo, caso contrário, pode ser impraticável encontrar os componentes independentes. Assim sendo, a escolha não pode ser meramente arbitrária, aleatória ou mesmo influenciada por métodos nocivos. Existem, porém, dois métodos clássicos muito bem consolidados que ajudam a encontrar tais funções: a Estimação de Máxima Verossimilhança (**MLE**, do inglês *Maximum Likelihood Estimation*) [Fisher, 1922; Aldrich *et al.*, 1997], oriunda da teoria de estimação; e a Informação Mútua (**MI**, do inglês *Mutual Information*) [Cover & Thomas, 2012], advinda da teoria da informação.

Outro princípio da estimação do ICA trata-se da obtenção da máxima não-gaussianidade[6] possível, que pode ser estimada a partir da forma normalizada do quarto momento central da distribuição, a Curtose (do inglês *Kurtosis*). A curtose mede a forma de uma dada distribuição quanto à sua "cauda", sendo que seu valor será igual a zero exclusivamente quando se tratar de uma legítima distribuição Gaussiana. Os casos com decaimentos laterais monotônicos suaves e com pouca (ou nenhuma) região central de planalto aproximam-se mais da distribuição Laplaciana, decaindo-se suavemente ao longo do eixo das ordenadas enquanto se caminha longamente pelo eixo das abscissas; tal caso é conhecido como Platicúrtico (ou Super-Gaussiana). Por outro lado, casos que exibam abrupto decaimento de cauda nas laterais, geralmente acompanhados de uma região central similar a um planalto, tendem a se aproximar mais da distribuição uniforme; esse caso é conhecido como Leptocúrtico (ou Sub-Gaussiana).

[6] Pode-se compreender não-gaussianidade como independência estatística.

Definição 2.4.2 Curtose

Sejam o n-ésimo momento central de um sinal pode ser calculado por

$$\mu_n = E\left[(X - E[X])^n\right] \tag{2.20}$$

e a normalização padrão dada pela divisão do momento calculado pelo desvio padrão da mesma ordem n do momento em questão, ou seja,

$$\sigma^n = \left(\sqrt{E\left[(X - E[X])^2\right]}\right)^n . \tag{2.21}$$

O quarto momento central normalizado, também conhecido como Curtose, é definido por

$$k(X) = \frac{\mu_4}{\sigma^4} = \frac{E\left[(X - E[X])^4\right]}{\left(E\left[(X - E[X])^2\right]\right)^2} \tag{2.22}$$

Mas a curtose não é a única maneira de se mensurar a não-gaussianidade; na verdade, nem mesmo é a melhor, dado que trata-se de uma métrica com elevada susceptibilidade a sofrer interferências indesejadas provocadas por *Outliers*[7], sobretudo por conta de seu termo de quarto grau, que faz com que até mesmo valores tipicamente considerados de magnitude não tão elevada sejam responsáveis por resultados indesejados. Isso não invalida a curtose como um recurso possível, mas o insere em um conjunto de recursos preferencialmente a serem evitados, pois suas características caminham em direção a uma instabilidade contornável.

Um dos conceitos mais importantes advindos da área de Teoria da Informação chama-se Entropia (do inglês *Entropy*) de Shannon, também conhecida como *informação média*, uma forma de se calcular a desordem em um contexto de teoria da informação.

[7]Dados que sejam considerados aberrantes ou atípicos, sendo, portanto, passíveis de serem considerados "fora da curva" para eventuais considerações em modelos. Modelos mais robustos tendem a identificar tais dados e efetuar um tratamento devido para evitar que danifiquem seu desempenho.

Definição 2.4.3 Entropia

Seja X uma variável aleatória. Então, a entropia de X, também entendida como a informação média de X, é dada por

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log_b (P(X = x_i)) \ , \tag{2.23}$$

sendo $P(X = x_i)$ a probabilidade de a variável aleatória X assumir o valor x_i; e, para os fins deste trabalho, $b = 2$.

A partir da entropia, pode-se chegar ao conceito de Negentropia[8] (do inglês *Negentropy*) [Schrödinger, 1944; Brillouin, 1953; Mahulikar & Herwig, 2009], que trata-se de uma alternativa à curtose, contudo, consideravelmente mais robusta.

Definição 2.4.4 Negentropia

A negentropia de uma variável aleatória X pode ser definida como

$$J(X) = H\left(X_{\text{gauss}}\right) - H(X) \ , \tag{2.24}$$

sendo que X_{gauss} trata-se de uma variável aleatória de distribuição gaussiana e cujas média e variância são as mesmas de X.

Tal como havia sido explicitado nesta mesma seção, a entropia pode ser utilizada como uma métrica da desordem de uma certa informação, o que permite dizer que o valor máximo de entropia caracteriza desordem máxima, ou seja, maior incerteza e, portanto, maior dificuldade de se obter uma dada informação; isso ocorre apenas no único e exclusivo caso em que a distribuição em questão se trata de uma Gaussiana. Por outro lado, entropia mínima implica maior certeza e maior facilidade para se obter a informação em questão. Efetuando-se uma análise, ainda que superficial, a respeito da Equação (2.23), é possível concluir que a entropia de Shannon jamais pode ser negativa. Assim, pode-se perceber que a negentropia também sempre será não negativa. Tais características mencionadas permitem compreender a negentropia como uma espécie de distância dada pela diferença entre as distribuições envolvidas, sendo ao menos uma delas sempre

[8] Também conhecida como *sintropia* ou *entropia negativa*.

uma distribuição Gaussiana [Suyama, 2007].

2.5 Análise de Componentes Esparsos

Além da técnica ICA, existem outras técnicas que também efetuam uma análise de componentes e que podem ser exploradas com o intuito de se realizar a tarefa de separação de sinais; esta seção abordará mais especificamente a Análise de Componentes Esparsos (**SCA**, do inglês *Sparse Component Analysis*). Contudo, antes mesmo de prosseguir para questões mais aprofundadas sobre tal técnica, pode ser importante começar explicando-se o significado de esparso. A ideia de esparsidade está intimamente relacionada à ideia de dispersão e distanciamento, podendo ser interpretada, portanto, como o oposto de densidade. A esparsidade pode, inclusive, ser quantificada, segundo o que se entende por grau de esparsidade de uma matriz.

Definição 2.5.1 Grau de Esparsidade

Seja uma matriz $\boldsymbol{M}$ definida por

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nn} \end{bmatrix}. \tag{2.25}$$

Considerando-se N como o número de elementos nulos da matriz e T como o número total de elementos da matriz, o Grau de Esparsidade, E, pode ser obtido a partir de

$$E = \frac{N}{T} \cdot \tag{2.26}$$

Pela definição, então, pode-se concluir que, em um cenário ideal, a matriz M seria composta por valores não nulos exclusivamente em sua diagonal principal, ou seja, em $m_{11}, m_{22}, \ldots, m_{nn}$; e, dado que uma das vantagens de se trabalhar com matrizes esparsas advém do fato de ser possível armazenar apenas os valores não nulos das matrizes, isso significa que, para este caso em particular, apenas os valores da diagonal principal seriam armazenados. Para que seja possível

trabalhar com matrizes em tal situação, são necessárias técnicas especiais de armazenamento e processamento, por exemplo, pelo uso de listas de coordenadas; tais técnicas computacionais, no entanto, não serão minuciosamente exploradas ou explanadas neste trabalho.

Em situações um pouco mais próximas ao que se encontra na realidade, as matrizes não são tão esparsas assim. Além de haver diversos elementos não nulos em posições que não sejam a diagonal principal, é possível que os próprios valores que devam ser considerados como nulos não sejam originalmente iguais a zero, o que dependeria de alguma técnica de pré-processamento que ajustasse tais valores com base em algum Limiar (do inglês *Threshold*), podendo resultar em um aumento considerável da esparsidade da matriz. As técnicas de análise de componentes esparsos possuem características geométricas intrínsecas que as favorecem em casos de cenários subparametrizados [Theis *et al.*, 2003], ou seja, conforme explicado na Seção 1.3.1, cenários em que o número de misturas observadas é inferior ao número de fontes.

O Algoritmo 3 é uma versão bastante simplificada que explica sucintamente os quatro passos básicos essenciais de uma técnica de SCA. A versão simplificada pode ser útil para oferecer uma visão superficial do caminho a ser trilhado. Contudo, a explicação acerca de tal algoritmo ainda será um pouco mais aprofundada ainda nesta mesma seção; não haverá um detalhamento matemático muito rigoroso, mas serão fornecidas fontes de referência que poderão ser consultadas para o caso de os objetivos do leitor dependerem de tal abordagem.

Algoritmo 3 Versão simplificada da Análise de Componentes Esparsos (SCA) [Gribonval & Lesage, 2006].

1: Aplicar transformação linear esparsificante à mistura.
2: Estimar $\boldsymbol{A}$ a partir da dispersão de $C_{\boldsymbol{x}}$.
3: Estimar as representações da fonte.
4: Reconstruir as fontes a partir da inversa de uma transformação esparsificante.

Dando continuidade às explicações um pouco mais minuciosas sobre o SCA, tomando como base ainda o Algoritmo 3, pode ser interessante começar pelo fato de que a análise será feita individualmente, uma a uma, para cada mistura observada; ou seja, caso haja três sinais observados resultantes de misturas, serão necessárias três execuções completas do algoritmo para que seja possível realizar a tarefa almejada de separação de sinais. Observe que a primeira etapa do

algoritmo se refere à aplicação de uma transformação linear esparsificante à mistura observada, que é a entrada do algoritmo. É muito comum que a escolha dessa transformação esparsificante seja feita entre Transformada Discreta Wavelet (**DWT**, do inglês *Discrete Wavelet Transform*) e a Transformada de Fourier de Tempo Curto (**STFT**, do inglês *Short-Time Fourier Transform*).

2.6 Fatoração de Matrizes Não-Negativas

Diferentemente de como ocorre com outros casos mencionados neste trabalho, a Fatoração de Matrizes Não Negativas (**NMF**, do inglês *Non-Negative Matrix Factorization*) não se trata de um algoritmo em si; em vez disso, trata-se de uma técnica que pode ser compreendida de maneira quase autoexplicativa a partir de seu nome, ou seja, independentemente de qual seja o algoritmo para se atingir tal objetivo, a ideia é a de se obter duas matrizes de valores não negativos que, multiplicando-se uma pela outra, obtém-se a matriz resultante que se deseja representar ao fim.

O objetivo de se efetuar tal processo de decomposição matricial é, principalmente, o de facilitar a análise dos resultados obtidos, mas cabe aqui ressaltar o fato de que nem todos os casos são plenamente solucionáveis. Pode ser preciso trabalhar com aproximações, mas isto não significa que os resultados sejam necessariamente imprecisos e inacurados.

CAPÍTULO

3

Redes Neurais Artificiais

A história de **Redes Neurais Artificiais** começou em 1943 com a fundação do modelo matemático proposto em [McCulloch & Pitts, 1943], que em 1951 permitiu abordagens inspiradas na biologia cerebral e por aplicações em inteligência artificial [Kleene, 1951], como no caso da criação de **Perceptron**[1] [Rosenblatt, 1958], um algoritmo de classificação binária baseado no aprendizado supervisionado, um dos desenvolvimentos mais importantes no campo, publicado em 1958. Outro grande passo foi dado em 1967, quando foi publicado um trabalho sobre redes envolvendo múltiplas camadas [Ivakhnenko & Lapa, 1967]. A Seção 3.1 esclarecerá o funcionamento do Perceptron, então a Seção 3.2 trará explicações sobre arquitetura e propriedades das redes neurais.

Alguns anos mais tarde, mais precisamente em 1972, dois grandes problemas foram identificados nos modelos de redes neurais então conhecidos: a incapacidade dos perceptrons básicos para processar operações de "ou exclusivo" (XOR) e a escassez de poder computacional exigido para se trabalhar com redes neurais de muitas camadas [Minsky & Papert, 1972]. Estes problemas dificultaram muito para que houvesse qualquer avanço relevante nesta área durante vários anos, mas em 1975 foi proposto o **Backpropagation** [Werbos, 1975], um novo algoritmo

[1] A menor unidade de processamento de uma RNA, que pode fazer pequenos trabalhos de processamento por si só, mas pode cooperar com outros para alcançar uma convergência, mesmo para um trabalho enorme distribuído por um grande número de neurônios trabalhando em uma rede.

que faria esta área de pesquisa continuar a ser interessante, uma vez que resolveria o problema de operação XOR e aceleraria o processamento em redes multicamadas ajustando os pesos das camadas por meio de uma distribuição de erro. Uma explicação detalhada sobre Backpropagation será oferecida na Seção 3.3.

Considerando a alta demanda de poder computacional que tais modelos trouxeram consigo, pode-se dizer que muitos avanços na área foram conquistados nos anos seguintes, também por conta de trabalhos envolvendo paralelismo, por exemplo. [Rumelhart *et al.*, 1986], de 1986. Em 1993 algumas melhorias foram alcançadas em redes neurais [Hush & Horne, 1993]; e, em 2010, o uso de Unidades de Processamento Gráfico na paralelização desta tarefa [Scherer *et al.*, 2010] foi extremamente importante.

Também em 2006, um modelo de representações de alto nível foi proposto usando camadas sucessivas de variáveis latentes com máquinas Boltzmann [Hinton *et al.*, 2006]. E, ainda mais recentemente, em 2013, foi introduzida uma rede suficientemente avançada para reconhecer conceitos avançados, como gatos em vídeos do YouTube [Le, 2013] de maneira não supervisionada. Essas implementações de um grande número de camadas passaram a incluir em sua nomenclatura o termo "Deep" e, portanto, o termo **Deep Learning** tornou-se popular quando se refere a redes neurais artificiais profundas, também chamadas **Deep Neural Networks**.

Atualmente, existem inúmeras aplicações de Deep Learning para as mais diversas áreas de atividade, como neurociências [Varatharajan *et al.*, 2018], IoT (Internet das Coisas) [Molanes *et al.*, 2018], segurança redes de computadores [Abeshu & Chilamkurti, 2018], reconhecimento facial [Ranjan *et al.*, 2018], instrumentação [Lei *et al.*, 2018], telecomunicações [Challita *et al.*, 2018], imageamento médico [Iakovidis *et al.*, 2018], detecção de objetos [Han *et al.*, 2018] tratamento de distúrbios da voz [Muhammad *et al.*, 2018], mobilidade [Herrera-Quintero *et al.*, 2018] e tantos outros.

3.1 Perceptron

O perceptron é compreendido como a menor unidade neural capaz de realizar a tarefa de classificar dados linearmente separáveis, separando-os através de uma

região fronteiriça denominada hiperplano[2].

3.2 Arquiteturas e Propriedades

Uma RNA é uma estrutura composta por perceptrons interconectados distribuídos entre múltiplas camadas. Sua origem vem da ideia de imitar a função cerebral. Uma das características mais importantes das RNAs é a capacidade de aprender com ou sem supervisores (professores), ou seja, as RNAs podem adotar uma abordagem de aprendizado supervisionado ou uma abordagem de aprendizado não supervisionada [Haykin, 1999], dependendo da implementação do algoritmo de acordo com os objetivos do projeto.

Independentemente do tipo de rede neural artificial, os princípios básicos permanecem os mesmos; existem camadas de entrada, camadas ocultas e camadas de saída. Cada uma delas será explicada em mais detalhes posteriormente. Camadas ocultas são as principais responsáveis pelo aumento da demanda por poder de processamento, pois é onde residem os neurônios que realizam todo o processamento principal da rede.

3.3 Backpropagation

3.4 Aplicação aos Problemas Mencionados

3.5 Redes Profundas

3.5.1 Neurônios

3.5.2 Treinamento

[2]Hiperplano nada mais é do que a generalização de um plano para dimensões superiores; trata-se de uma forma geométrica com uma dimensão a menos que o hiperespaço, que é o caso que utiliza-se de todas as dimensões disponíveis

3.5.3 Inicialização

CAPÍTULO

4

Redes Adversárias Geradoras

4.1 Redes Neurais como Modelos Geradores

4.2 Redes Adversárias

4.3 Teoria dos Jogos

Teoria dos Jogos é um enorme campo de estudos, geralmente mais comumente explorado pelos matemáticos, que tem um escopo bastante amplo e utiliza modelos matemáticos para oferecer visões importantes sobre cenários competitivos ou cooperativos que envolvam indivíduos que possam ter objetivos ou preferências diferentes [Myerson, 1997]. Este campo de estudo tem uma ampla gama de aplicações em telecomunicações [Han *et al.*, 2012], biologia [Smith, 1974], economia [Friedman, 1998; Kreps, 1990; Gibbons, 1992], setor jurídico [Baird *et al.*, 1998], administração [Sanfey, 2007; Camerer, 2011], computação [Abraham *et al.*, 2006] etc.

RAND Corporation[1], uma organização sem fins lucrativos independente, foi uma das instituições que mais estimulou o uso de cientistas e matemáticos para fins aplicados logo após a Segunda Guerra Mundial. Entre tantos outros matemáticos e cientistas brilhantes que participaram dos projetos da RAND estão John von Neumann, um pioneiro do computador digital moderno, e John Forbes Nash, que ganhou o Prêmio Nobel de Economia em 1994.

Segundo John von Neumann e Oskar Morgenstern:

> "*The game is simply the totality of the rules which describe it.* "
> [von Neumann *et al.*, 1944]

Ou seja, efetuando-se uma tradução livre, pode-se compreender que, segundo os professores Neumann e Morgenstern, o jogo é simplesmente a totalidade das regras que o descrevem.

Apesar do nome popularmente convidativo – talvez por estar erroneamente associado a elementos meramente lúdicos – não pode haver confusão quanto à importância dessa área, assim como o rigor dos modelos matemáticos envolvidos; tais confusões podem ocorrer devido à alta ambiguidade da linguagem coloquial [von Neumann *et al.*, 1944].

Na Subseção 4.3.1 será discutido mais sobre o **Dilema do Prisioneiro**, um exemplo de cenário que pode envolver cooperação e competição.

4.3.1 Dilema dos Prisioneiro

Originalmente, esse modelo havia sido proposto ao longo de 1950, na RAND, por Merrill Meeks Flood e Melvin Dresher; mais tarde, Albert William Tucker foi responsável por uma excelente interpretação do problema e pela atribuição do nome pelo qual é conhecido agora o modelo matemático, o **Dilema do Prisioneiro** [Poundstone, 1992].

O cenário Dilema do Prisioneiro é composto por dois membros de uma gangue criminosa que são presos e mantidos separados um do outro e submetidos a um interrogatório, sem que haja qualquer tipo de comunicação entre eles durante todo esse processo. *A priori*, o promotor tem provas suficientes para garantir a

[1] O nome "RAND" vem de uma contração do termo, em inglês, "research and development" (**R**esearch **AN**d **D**evelopment).

condenação de ambos os criminosos, mas apenas para crimes menores, o que não é suficiente para o promotor. Ambos os criminosos têm o direito de tomar a decisão de permanecer em silêncio ou de trair seu comparsa, delatando-o.

É importante enfatizar o fato de que, embora ambos os prisioneiros estejam cientes das possíveis consequências de cada situação, não se sabe sobre o que o outro decidiu fazer. Se o Prisioneiro A escolher cooperar com o Prisioneiro B, dependendo da estratégia adotada pelo Prisioneiro B, as consequências podem ser de 1 ano de prisão tanto para A quanto para B se o Prisioneiro B também cooperar com A; ou, se B denunciar A, 4 anos de prisão para A e imediata liberdade para B. Ao inverter as estratégias entre A e B, as penas são as mesmas, mudando apenas os presos. Porém, se ambos os presos escolherem denunciar seu parceiro, ambos serão sentenciados a três anos de prisão.

A Figura 4.1 contém um exemplo hipotético para o Dilema do Prisioneiro.

A \ B	Cooperates	Defects
Cooperates	A: 1 year B: 1 year	A: 4 years B: free
Defects	A: free B: 4 years	A: 3 years B: 3 years

Figura 4.1: Estratégias e suas respectivas consequências para os prisioneiros A e B. A estratégia de cooperação significa que o prisioneiro ficará calado e aceitará as consequências; e a estratégia de desertar é trair seu parceiro denunciando-o pelo crime mais pesado.

O clássico Dilema do Prisioneiro, apresentado até agora, considera apenas uma ocorrência para ambos os prisioneiros envolvidos, mas não é suficiente para modelar adequadamente todas as situações possíveis; um algoritmo ainda mais flexível consideraria iterações [Press & Dyson, 2012].

A Subseção 4.3.2 tratá uma regra de decisão muito interessante que pode ajudar a escolher a estratégia que melhor se adapte a ambos os prisioneiros.

4.3.2 MiniMax

Esta subseção começa explicando alguns passos para se chegar à melhor decisão para o Dilema do Prisioneiro, seguindo o exemplo dado pela Figura 4.1.

Numa primeira tentativa ingênua, um prisioneiro pode pensar em ficar quieto, cooperando com seu cúmplice, mas como não é possível saber qual será a estratégia de seu cúmplice, este prisioneiro corre o risco de ser condenado a 4 anos de prisão. Isto é, uma simples mudança de estratégia egoísta por parte do outro preso, sobre o qual não há controle, pode causar uma grande alteração no resultado da penalidade do primeiro preso, indo do mínimo (1 ano) para o máximo (4 anos). Esta é, então, uma estratégia muito arriscada. Por outro lado, o prisioneiro pode escolher denunciar seu companheiro; isso certamente garantiria ao outro preso uma sentença mínima de 3 anos e um máximo de 4 anos, além de colocar o prisioneiro que tomou tal decisão em posição de poder sair em liberdade, desde que seu parceiro também não o entregue.

Para evitar as piores consequências possíveis para ambos os criminosos, eles têm que alcançar um forte *equilíbrio de Nash* [Nash, 1950], ou seja, chegar a uma posição cujo resultado para um deles não receberia uma melhoria significativa se mudasse sua estratégia unilateralmente. A ideia aqui, portanto, seria minimizar a sentença máxima; esta é uma regra de decisão chamada *MiniMax*[2] [v. Neumann, 1928; Blackwell *et al.*, 1956; Willem, 1997].

Assim, continuando com o exemplo exposto na Figura 4.1, o movimento mais seguro para ambos seria optar por delatar seu cúmplice, pois quem delatar terá sua pena superiormente limitada a 3 anos; Além disso, se o seu companheiro não fizer o mesmo, em vez de ser condenado à prisão, será libertado.

4.4 Redes Adversárias Geradoras

Explorando conceitos de Teoria dos Jogos e de Aprendizagem Profunda, foi possível formular um novo tipo de abordagem, utilizando duas redes neurais como jogadores adversários de um jogo competitivo; assim, foram propostas as Redes Adversárias Geradoras (**GAN**, do inglês *Generative Adversarial Networks*) [Goodfellow *et al.*, 2014].

[2] É possível encontrar variações deste termo, tais como *MinMax*, ou *Max Min*, ou MM, dependendo da referência utilizada.

Neste jogo competitivo para dois jogadores, existe um conjunto de dados já bem preparado, composto por amostras do mesmo tipo, escolhidas de forma adequada, mas com valores de atributos diferentes. O primeiro jogador, D, tem o objetivo de discriminar se uma amostra veio do conjunto de dados original ou não; o segundo, G, deve capturar a distribuição do conjunto de dados original e usá-la para gerar amostras completamente novas. Assim, enquanto um dos jogadores pretende gerar a imitação perfeita dos dados originais, o outro jogador tenta ser o melhor identificador de falsificações possível.

Como foi dito neste capítulo, ambos os jogadores deste jogo são redes neurais artificiais e, como há uma dependência de treinamento, este passo deve ocorrer gradualmente e concomitantemente para ambos os jogadores, caso contrário, dado que os jogadores estão jogando um contra o outro, caso contrário, pode ocorrer um desequilíbrio de evolução em favor de um dos atores, que tenderá a tornar o processo evolutivo cada vez mais desfavorável para um lado e, assim, em vez de conseguir uma boa evolução para ambos, apenas um deles evoluirá minimamente se comparado ao outro, o que sequer garante que o processo terá sido bom para ao menos um dos jogadores.

O modelo matemático baseado na regra de decisão Minimax, da Subseção (4.3.2), que é utilizado pelo algoritmo da GAN, pode ser visto na Equação (4.1).

$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})} [\log D(\boldsymbol{x})] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))], \quad (4.1)$$

onde V é a função de valor; D é o Discriminador, um perceptron multicamada que gera a probabilidade de $\boldsymbol{x}$ ter sido originado dos dados legítimos em vez da distribuição p_G; G é o Gerador (jogador); p_{data} é a distribuição de dados; e $p_{\boldsymbol{z}}$ é uma *a priori* nas variáveis de ruído de entrada.

O treinamento de Redes Adversariais Generativas usa um método baseado em Gradiente Descendente Estocástico, que pode ser visto no Algoritmo 4 [Goodfellow *et al.*, 2014].

4.5 Aplicação em Imagens

Algoritmo 4 Treinamento de Redes Adversárias Geradoras com Gradiente Descendente Estocástico por mini-lotes. O número de etapas a serem aplicadas ao Discriminador, k, é um hiper-parâmetro [Goodfellow *et al.*, 2014].

1: **for** número de iterações de treinamento **do**
2: **for** k passos **do**
3: • Amostrar lote de m amostras ruidosas $\left\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\right\}$ de ruído *a priori* $p_g(\boldsymbol{z})$.
4: • Amostrar lote de m exemplos $\left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\right\}$ da distribuição geradora de dados $p_{\text{data}}(\boldsymbol{x})$.
5: • Atualiza o discriminante aumentando seu gradiente estocástico:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right].$$

6: • Amostra lote de m amostras ruidosas $\left\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\right\}$ de ruído *a priori* $p_g(\boldsymbol{z})$.
7: • Atualiza o gerador diminuindo seu gradiente estocástico:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

As atualizações baseadas em gradiente podem utilizar qualquer regra de aprendizagem padrão baseada em gradiente.
Este algoritmo utilizou $k = 1$ e o utilizou momento como regra de aprendizagem padrão.

CAPÍTULO

5

GAN para BSS

5.1 Como Utilizar

5.2 Considerações

5.3 Revisão Bibliográfica

5.3.1 Bengio (ICA + GAN)

5.3.2 xxx (GAN + BSS)

5.4 Limitações

5.5 Propostas

CAPÍTULO

6

Simulações e Resultados

CAPÍTULO

7

Conclusões e Perspectivas

Referências Bibliográficas

[Abeshu & Chilamkurti, 2018] Abeshu, A., & Chilamkurti, N. 2018. **Deep learning: The frontier for distributed attack detection in fog-to-things computing**. *IEEE Communications Magazine*, 56(2), 169–175.

[Abraham *et al.*, 2006] Abraham, Ittai, Dolev, Danny, Gonen, Rica, & Halpern, Joe. 2006. **Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation**. *Pages 53–62 of: Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*. ACM.

[Aldrich *et al.*, 1997] Aldrich, John, *et al.* 1997. **Ra fisher and the making of maximum likelihood 1912-1922**. *Statistical science*, 12(3), 162–176.

[Anderson & Gerbing, 1988] Anderson, James, & Gerbing, David W. 1988. **Structural equation modeling in practice: A review and recommended two-step approach**. *Psychological Bulletin*, 103(3), 411–423.

[Baird *et al.*, 1998] Baird, Douglas G, Gertner, Robert H, & Picker, Randal C. 1998. ***Game theory and the law***. Harvard University Press.

[Bell & Sejnowski, 1997] Bell, Anthony J., & Sejnowski, Terrence J. 1997. **The "independent components" of natural scenes are edge filters**. *Vision Research*, 37(23), 3327 – 3338.

[Belouchrani *et al.*, 1997] Belouchrani, Adel, Abed-Meraim, Karim, Cardoso, J-F, & Moulines, Eric. 1997. **A blind source separation technique using second-order statistics**. *IEEE Transactions on signal processing*, 45(2), 434–444.

[Blackwell *et al.*, 1956] Blackwell, David, *et al.* 1956. **An analog of the minimax theorem for vector payoffs.** *Pacific Journal of Mathematics*, 6(1), 1–8.

[Boll, 1979] Boll, Steven. 1979. **Suppression of acoustic noise in speech using spectral subtraction**. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2), 113–120.

[Brillouin, 1953] Brillouin, Leon. 1953. **The negentropy principle of information**. *Journal of Applied Physics*, 24(9), 1152–1163.

[Camerer, 2011] Camerer, Colin F. 2011. ***Behavioral game theory: Experiments in strategic interaction***. Princeton University Press.

[Cardoso, 1989] Cardoso, J. . 1989 (May). **Source separation using higher order moments**. *Pages 2109–2112 vol.4 of: International Conference on Acoustics, Speech, and Signal Processing,*.

[Challita *et al.*, 2018] Challita, U., Dong, L., & Saad, W. 2018. **Proactive resource management for lte in unlicensed spectrum: A deep learning perspective**. *IEEE Transactions on Wireless Communications*, 17(7), 4674–4689.

[Chien & Yang, 2016] Chien, J., & Yang, P. 2016. **Bayesian factorization and learning for monaural source separation**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1), 185–195.

[Comon, 1994] Comon, Pierre. 1994. **Independent component analysis, a new concept?** *Signal Processing*, 36(3), 287 – 314. Higher Order Statistics.

[Cover & Thomas, 2012] Cover, T.M., & Thomas, J.A. 2012. ***Elements of Information Theory***. Wiley.

[Dempster *et al.*, 1977] Dempster, Arthur P, Laird, Nan M, & Rubin, Donald B. 1977. **Maximum likelihood from incomplete data via the em algorithm**. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

[Donoho, 1995] Donoho, D. L. 1995. **De-noising by soft-thresholding**. *IEEE Transactions on Information Theory*, 41(3), 613–627.

[Fisher, 1922] Fisher, R.A. 1922. **On the mathematical foundations of theoretical statistics**. *Phil. Trans. R. Soc. Lond. A*, 222(594-604), 309–368.

[Friedman, 1998] Friedman, Daniel. 1998. **On economic applications of evolutionary game theory**. *Journal of Evolutionary Economics*, 8(1), 15–43.

[Gannot *et al.*, 2017] Gannot, S., Vincent, E., Markovich-Golan, S., & Ozerov, A. 2017. **A consolidated perspective on multimicrophone speech enhancement and source separation**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4), 692–730.

[Gibbons, 1992] Gibbons, Robert. 1992. ***Game theory for applied economists***. Princeton University Press.

[Goodfellow *et al.*, 2014] Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, & Bengio, Yoshua. 2014. **Generative adversarial nets**. *Pages 2672–2680 of:* Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., & Weinberger, K. Q. (eds), *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc.

[Gribonval & Lesage, 2006] Gribonval, Rémi, & Lesage, Sylvain. 2006. **A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges**. *Pages 323–330 of: ESANN'06 proceedings-14th European Symposium on Artificial Neural Networks*. d-side publi.

[Han *et al.*, 2018] Han, J., Zhang, D., Cheng, G., Liu, N., & Xu, D. 2018. **Advanced deep-learning techniques for salient and category-specific object detection: A survey**. *IEEE Signal Processing Magazine*, 35(1), 84–100.

[Han *et al.*, 2012] Han, Zhu, Niyato, Dusit, Saad, Walid, Başar, Tamer, & Hjørungnes, Are. 2012. ***Game theory in wireless and communication networks: theory, models, and applications***. Cambridge University Press.

[Haykin, 1999] Haykin, S.S. 1999. ***Neural Networks: A Comprehensive Foundation***. International edition. Prentice Hall.

[Hérault *et al.*, 1985] Hérault, Jeanny, Jutten, Christian, & Ans, Bernard. 1985. **Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé**. *In: 10 Colloque sur le traitement du signal et des images, FRA, 1985*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images.

[Herrera-Quintero *et al.*, 2018] Herrera-Quintero, L. F., Samper-Zapater, J. J., Svitek, M., & David, W. 2018. **Special section on its services to smart city context [guest editorial]**. *IEEE Intelligent Transportation Systems Magazine*, 10(2), 4–5.

[Hinton *et al.*, 2006] Hinton, Geoffrey E, Osindero, Simon, & Teh, Yee-Whye. 2006. **A fast learning algorithm for deep belief nets**. *Neural computation*, 18(7), 1527–1554.

[Huang *et al.*, 2015] Huang, P., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. 2015. **Joint optimization of masks and deep recurrent neural networks for monaural source separation**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12), 2136–2147.

[Hush & Horne, 1993] Hush, D. R., & Horne, B. G. 1993. **Progress in supervised neural networks**. *IEEE Signal Processing Magazine*, 10(1), 8–39.

[Hyvärinen *et al.*, 2004] Hyvärinen, A., Karhunen, J., & Oja, E. 2004. ***Independent Component Analysis***. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley.

[Iakovidis *et al.*, 2018] Iakovidis, D. K., Georgakopoulos, S. V., Vasilakakis, M., Koulaouzidis, A., & Plagianakos, V. P. 2018. **Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification**. *IEEE Transactions on Medical Imaging*, 1–1.

[Ivakhnenko & Lapa, 1967] Ivakhnenko, A.G., & Lapa, V.G. 1967. ***Cybernetics and forecasting techniques***. Modern analytic and computational methods in science and mathematics. American Elsevier Pub. Co.

[Kaiser, 1960] Kaiser, Henry F. 1960. **The application of electronic computers to factor analysis**. *Educational and Psychological Measurement*, 20(1), 141–151.

[Kleene, 1951] Kleene, Stephen Cole. 1951. ***Representation of events in nerve nets and finite automata***. Tech. rept. RAND PROJECT AIR FORCE SANTA MONICA CA.

[Koivunen & Kostinski, 1999] Koivunen, AC, & Kostinski, AB. 1999. **The feasibility of data whitening to improve performance of weather radar**. *Journal of Applied Meteorology*, 38(6), 741–749.

[Kreps, 1990] Kreps, David M. 1990. ***Game theory and economic modelling***. Oxford University Press.

[Lathi, 2009] Lathi, B.P. 2009. ***Linear Systems and Signals***. The Oxford series in electrical and computer engineering. Oxford University Press.

[Le, 2013] Le, Quoc V. 2013. **Building high-level features using large scale unsupervised learning**. *Pages 8595–8598 of: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE.

[Lee, 1980] Lee, J. 1980. **Digital image enhancement and noise filtering by use of local statistics**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(2), 165–168.

[Lei *et al.*, 2018] Lei, J., Liu, Q., & Wang, X. 2018. **Deep learning-based inversion method for imaging problems in electrical capacitance tomography**. *IEEE Transactions on Instrumentation and Measurement*, 1–12.

[Li & Zhang, 2012] Li, H., & Zhang, X. 2012 (July). **Blind separation of noisy mixed speech based on independent component analysis and neural network**. *Pages 105–108 of: 2012 International Conference on Computing, Measurement, Control and Sensor Network*.

[Mahulikar & Herwig, 2009] Mahulikar, Shripad P, & Herwig, Heinz. 2009. **Exact thermodynamic principles for dynamic order existence and evolution in chaos**. *Chaos, Solitons & Fractals*, 41(4), 1939–1948.

[McCulloch & Pitts, 1943] McCulloch, Warren S, & Pitts, Walter. 1943. **A logical calculus of the ideas immanent in nervous activity**. *The bulletin of mathematical biophysics*, 5(4), 115–133.

[Minsky & Papert, 1972] Minsky, M.L., & Papert, S. 1972. ***Perceptrons: An Introduction to Computational Geometry***. Mit Press.

[Molanes *et al.*, 2018] Molanes, R. Fernandez, Amarasinghe, K., Rodriguez-Andina, J., & Manic, M. 2018. **Deep learning and reconfigurable platforms in the internet of things: Challenges and opportunities in algorithms and hardware**. *IEEE Industrial Electronics Magazine*, 12(2), 36–49.

[Muhammad *et al.*, 2018] Muhammad, G., Alhamid, M. F., Alsulaiman, M., & Gupta, B. 2018. **Edge computing with cloud for voice disorder assessment and treatment**. *IEEE Communications Magazine*, 56(4), 60–65.

[Myerson, 1997] Myerson, R.B. 1997. ***Game Theory***. Harvard University Press.

[Nair & Hinton, 2010] Nair, Vinod, & Hinton, Geoffrey E. 2010. **Rectified linear units improve restricted boltzmann machines**. *Pages 807–814 of: Proceedings of the 27th international conference on machine learning (ICML-10)*.

[Nash, 1950] Nash, John F. 1950. **Equilibrium points in n-person games**. *Proceedings of the National Academy of Sciences*, 36(1), 48–49.

[Nugraha *et al.*, 2016] Nugraha, Aditya Arie, Liutkus, Antoine, & Vincent, Emmanuel. 2016. **Multichannel audio source separation with deep neural networks.** *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(9), 1652–1664.

[Papoulis, 1984] Papoulis, A. 1984. ***Probability, Random Variables, and Stochastic Processes***. McGraw-Hill Series in Electrical Engineering. McGraw-Hill.

[Pearson, 1901] Pearson, Karl. 1901. **Liii. on lines and planes of closest fit to systems of points in space**. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.

[Poundstone, 1992] Poundstone, W. 1992. ***Prisoner's Dilemma***. Doubleday.

[Press & Dyson, 2012] Press, William H, & Dyson, Freeman J. 2012. **Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent**. *Proceedings of the National Academy of Sciences*, 109(26), 10409–10413.

[Rabiner, 1989] Rabiner, L. R. 1989. **A tutorial on hidden markov models and selected applications in speech recognition**. *Proceedings of the IEEE*, 77(2), 257–286.

[Ranjan *et al.*, 2018] Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J. C., Patel, V. M., Castillo, C. D., & Chellappa, R. 2018. **Deep learning for understanding faces: Machines may be just as good, or better, than humans**. *IEEE Signal Processing Magazine*, 35(1), 66–83.

[Romano *et al.*, 2010] Romano, J.M.T., Attux, R., Cavalcante, C.C., & Suyama, R. 2010. ***Unsupervised Signal Processing: Channel Equalization and Source Separation***. CRC Press.

[Rosenblatt, 1958] Rosenblatt, Frank. 1958. **The perceptron: a probabilistic model for information storage and organization in the brain.** *Psychological review*, 65(6), 386.

[Rumelhart *et al.*, 1986] Rumelhart, D.E., McClelland, J.L., Group, PDP Research, & University of California, San Diego. PDP Research Group. 1986. ***Psychological and Biological Models***. A Bradford Book, no. v. 2. MIT Press.

[Sanfey, 2007] Sanfey, Alan G. 2007. **Social decision-making: insights from game theory and neuroscience**. *Science*, 318(5850), 598–602.

[Scherer *et al.*, 2010] Scherer, Dominik, Müller, Andreas, & Behnke, Sven. 2010. **Evaluation of pooling operations in convolutional architectures for object recognition**. *Pages 92–101 of: Artificial Neural Networks–ICANN 2010*. Springer.

[Schrödinger, 1944] Schrödinger, Erwin. 1944. ***What Is Life? the physical aspect of the living cell and mind***. Cambridge University Press, Cambridge.

[Shannon, 1948] Shannon, Claude Elwood. 1948. **A mathematical theory of communication**. *Bell system technical journal*, 27(3), 379–423.

[Smith, 1974] Smith, J Maynard. 1974. **The theory of games and the evolution of animal conflicts**. *Journal of theoretical biology*, 47(1), 209–221.

[Suyama, 2007] Suyama, Ricardo. 2007. ***Proposta de metodos de separação cega de fontes para misturas convolutivas e não-lineares***. Ph.D. thesis, Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

[Theis *et al.*, 2003] Theis, Fabian J, Jung, Andreas, Puntonet, Carlos G, & Lang, Elmar W. 2003. **Linear geometric ica: Fundamentals and algorithms**. *Neural computation*, 15(2), 419–439.

[Tuzlukov, 2002] Tuzlukov, V. 2002. ***Signal Processing Noise***. Electrical Engineering & Applied Signal Processing Series. CRC Press.

[Uhlich *et al.*, 2015] Uhlich, S., Giron, F., & Mitsufuji, Y. 2015 (April). **Deep neural network based instrument extraction from music**. *Pages 2135–2139 of: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[v. Neumann, 1928] v. Neumann, J. 1928. **Zur theorie der gesellschaftsspiele**. *Mathematische Annalen*, 100(1), 295–320.

[Varatharajan *et al.*, 2018] Varatharajan, R., Manogaran, Gunasekaran, & Priyan, M. K. 2018. **A big data classification approach using lda with an enhanced svm method for ecg signals in cloud computing**. *Multimedia Tools and Applications*, 77(8), 10195–10215.

[Vincent, 1953] Vincent, Douglas F. 1953. **The orgin and development of factor analysis**. *Applied statistics*, 107–117.

[von Neumann *et al.*, 1944] von Neumann, John, Morgenstern, Oskar, Kuhn, Harold W., & Rubinstein, Ariel. 1944. ***Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)***. Princeton University Press.

[Werbos, 1975] Werbos, P.J. 1975. ***Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences***. Harvard University.

[Widrow *et al.*, 1975] Widrow, B., Glover, J. R., McCool, J. M., Kaunitz, J., Williams, C. S., Hearn, R. H., Zeidler, J. R., Dong, J. Eugene, & Goodlin, R. C. 1975. **Adaptive noise cancelling: Principles and applications**. *Proceedings of the IEEE*, 63(12), 1692–1716.

[Willem, 1997] Willem, M. 1997. ***Minimax Theorems***. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Boston.

[Zafeiriou, 2015] Zafeiriou, Stefanos. 2015. ***Notes on Implementation of Component Analysis Techniques***.