UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE COMUNICAÇÕES

Um Estudo sobre Separação Cega de Fontes e Contribuições ao Caso de Misturas Não-lineares

Autor
Leonardo Tomazeli Duarte

Orientador
Prof. Dr. João Marcos Travassos Romano

Co-orientador
Dr. Romis Ribeiro de Faissol Attux

Banca Examinadora:

- Prof. Dr. João Marcos Travassos Romano (FEEC/UNICAMP)
- Prof. Dr. Allan Kardec Duailibe Barros Filho (DEE/UFMA)
- Prof. Dr. Fernando José Von Zuben (FEEC/UNICAMP)
- Prof. Dr. João Bosco Ribeiro do Val (FEEC/UNICAMP)

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.

Campinas, 2 de Agosto de 2006

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

D85e

Duarte, Leonardo Tomazeli
Um estudo sobre separação cega de fontes e contribuições ao caso de misturas não-lineares / Leonardo Tomazeli Duarte. --Campinas, SP: [s.n.], 2006.

Orientadores: João Marcos Travassos Romano, Romis Ribeiro de Faissol Attux
Dissertação (Mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Processamento de sinais. 2. Entropia (Teoria da informação. 3. Teoria da informação. 4. Algoritmos genéticos. 5. Sistemas não-lineares. I. Romano, João Marcos Travassos. II. Attux, Romis Ribeiro de Faissol. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Título em Inglês: A Study on Blind Source Separation and Contributions to the Nonlinear Case
Palavras-chave em Inglês: Signal processing, Blind source separation, Independent component analysis, Nonlinear models, Evolutionary computation, Information theory
Área de concentração: Telecomunicações e Telemática
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: Allan Kardec Duailibe Barros Filho, Fernando José Von Zuben, João Bosco Ribeiro do Val
Data da defesa: 02/08/2006

Resumo

O presente trabalho tem como objetivo a realização de um estudo sobre o problema de separação cega de fontes. Em uma primeira parte, considera-se o caso clássico em que o sistema misturador é de natureza linear. Na seqüência, a extensão ao caso não-linear é tratada. Em particular, enfatizamos uma importante classe de modelos não-lineares, os modelos com não-linearidade posterior (PNL). Com o intuito de contornar uma dificuldade relacionada à convergência para mínimos locais no treinamento de sistemas separadores PNL, uma nova técnica é proposta. Tal solução se baseia no uso de um algoritmo evolutivo na etapa de treinamento e de um estimador de entropia baseado em estatísticas de ordem. A eficácia do algoritmo proposto é verificada através de simulações em diferentes cenários.

Abstract

The aim of this work is to study the problem of blind source separation (BSS). In a first part, the classical case in which the mixture system is of linear nature is considered. Afterwards, the nonlinear extension of the BSS problem is addressed. In special, an important class of nonlinear models, the post-nonlinear (PNL) models, is emphasized. In order to overcome a problem related to the convergence to local minima in the training of a PNL separating system, a novel technique is proposed. The bases of such solution are the application of an evolutionary algorithm in the training stage and the use of an entropy estimator based on order statistics. The efficacy of the proposal is attested by simulations conducted in different scenarios.

Agradecimentos

Agradeço

Aos meus pais, Sebastião e Aparecida, e ao meu irmão, Ronaldo. O amor e o apoio de vocês foram fundamentais para que esse trabalho se tornasse realidade.

À minha querida esposa, Camila, pelo amor, incentivo, apoio e também pela compreensão durante toda esta jornada.

Ao meu orientador, professor João Marcos Travassos Romano, pela orientação, apoio e confiança.

Ao meu co-orientador, professor Romis Ribeiro de Faissol Attux, pela orientação, pela amizade e pelo carinho desde a graduação.

Ao meu amigo, Ricardo Suyama, pelo apoio em todas as etapas desse projeto e, sobretudo, pela amizade e convivência nesses últimos anos.

Ao professor Fernando José Von Zuben, por suas valiosas sugestões e pelo apoio desde a graduação.

Ao professor João Bosco Ribeiro do Val, pela disponibilidade e pela revisão atenciosa deste trabalho.

Ao professor Allan Kardec Duailibe Barros Filho, pelas sugestões técnicas e pela revisão criteriosa deste trabalho.

Ao meu amigo, Rafael Ferrari (Gremista), pelas importantes discussões durante este trabalho.

Aos meus amigos, Danilo e Murilo, pela ajuda com o LaTeX.

Ao meu amigo, professor Renato da Rocha Lopes, pelas valiosas discussões que resultaram no apêndice B deste trabalho.

Aos demais amigos do DSPCom: Aline, Alam, Charles, Cristiano Panazio, Cristina, Cynthia, Dayan, Fabiano, Fábio, Glauco, Gustavo (Parmera), Leandro, Mário, Moisés, Rafael (Krummen), Tarciano e Tiago; pelo companheirismo e pela ajuda.

A Celi, Amanda, Eloísa, Mazé, Noêmia, Lúcia, pela ajuda em diversas questões.

Aos professores e funcionários da FEEC, pelo apoio indispensável.

Aos meus tios, Luiz e Sueli, e aos meus primos, Alexandre e Bruno, pelo carinho e incentivo.

Aos meus sogros, Agenor e Yara, pelo apoio e pelo incentivo.

Aos meus amigos Diogo e Filipe, pela amizade e convivência durante a minha estadia em Campinas.

Aos meus colegas de graduação da turma EE00.

Aos meus colegas da pós-graduação.

A todos que compareceram na defesa.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro.

Sumário

Lista de Figuras

Lista de Tabelas

Abreviaturas

ACY:	Amari-Cichocki-Yang
BS:	Bell-Sejnowski
BSS:	*Blind Source Separation* – Separação Cega de Fontes
EASI:	*Equivariant Adaptive Source Separation*
ECG:	Eletrocardiograma
EEG:	Eletroencefalograma
EQM:	Erro Quadrático Médio
FA:	*Factor Analysis* – Análise de Fatores
fMRI:	*Functional Magnetic Resonance Imaging* – Ressonância Magnética Funcional
HOS:	*Higher Order Statistics* – Estatísticas (ou Momentos) de Ordem Elevada
JADE:	*Joint Approximate Diagonalization of Eigenmatrices*
MEG:	Magnetoencefalograma
MMSE:	*Minimum Mean Square Error* – Erro Quadrático Médio Mínimo
NBSS:	*Nonlinear Blind Source Separation* – Separação Cega de Fontes Não-linear
NICA:	*Nonlinear Independent Component Analysis* – Análise de Componentes Independentes Não-linear
NPCA:	*Nonlinear Principal Component Analysis* – Análise de Componentes Principais Não-linear
ICA:	*Independent Component Analysis* – Análise de Componentes Independentes
Infomax:	*Information Maximization*

PCA: *Principal Component Analysis* – Análise de Componentes Principais
PNL: *Post-Nonlinear*
SCA: *Sparse Component Analysis* – Análise de Componentes Esparsos
SNR: *Signal-to-Noise Ratio* – Relação Sinal-Ruído
TJ: Taleb-Jutten

Capítulo 1

Introdução

Um dos principais temas em processamento de sinais diz respeito à recuperação de sinais de interesse (fontes) através da observação de misturas deles. Nas estratégias clássicas, essa tarefa é realizada levando em conta características relevantes desses sinais fontes, ou, ainda, informações sobre o processo de mistura. Um exemplo típico ocorre nas situações em que as características espectrais dos sinais, misturados no tempo, são bem conhecidas. Neste caso, é possível separá-los a partir do emprego de filtros seletivos em freqüência, por exemplo. Evidentemente, há, nesta estratégia, uma hipótese tácita relativa à necessidade de que os sinais tenham espectros distintos.

Este exemplo simples ilustra bem algumas restrições presentes nas estratégias supervisionadas, ou seja, aquelas que pressupõem a existência de informações sobre os sinais e sistemas envolvidos no processo de mistura. Primeiramente, a necessidade deste conhecimento já é uma considerável limitação em si, dado que existem aplicações nas quais tal exigência pode ser inatingível. Um outro ponto, mais relacionado ao exemplo em questão, é que, por fundamentar a separação das fontes em uma qualidade específica delas, a aplicação eficiente deste tipo de estratégia passa a estar condicionada à satisfação de hipóteses restritivas sobre os sinais e sistemas envolvidos.

Uma maneira de contornar essas dificuldades seria conceber novas estratégias tendo em mente a busca por um maior grau de generalidade, no sentido de se

assumir tão pouca informação quanto possível sobre a geração dos dados a serem processados. Este paradigma representa o princípio essencial do processamento não-supervisionado (cego) de sinais, e, conseqüentemente, da separação cega de fontes (BSS, *Blind Source Separation*), um dos principais temas de pesquisa nesta área.

Na separação cega de fontes, a recuperação dos sinais de interesse é feita com base apenas em suas misturas. Desde seu surgimento, na década de 1980, este problema vem recebendo uma atenção considerável de pesquisadores das mais diversas áreas. Um dos motivos desse interesse significativo diz respeito à generalidade presente em sua formulação, o que, por sua vez, proporciona uma ampla gama de aplicações. Alguns exemplos de problemas de separação de fontes podem ser encontrados em processamento de sinais biomédicos, telecomunicações e tratamento de sinais de áudio.

Inicialmente, o problema de separação cega de fontes foi abordado através da análise de componentes independentes (ICA - *Independent Component Analysis*), uma técnica em representação de dados que pode ser vista como uma extensão da clássica análise de componentes principais (PCA). Não obstante o fato de a ICA ter origens no problema de separação, esta técnica possui aplicações que não necessariamente estão ligadas a este problema. Ainda assim, esta abordagem é geralmente associada à BSS, certamente por prover as bases dos principais algoritmos desenvolvidos para tal problema.

1.1 Objetivos

Diante da relevância do problema de BSS, um primeiro objetivo do presente trabalho é realizar um estudo das principais técnicas existentes para a resolução deste problema, em particular daquelas baseadas na ICA. Em um primeiro momento, nosso estudo abrangeu os casos em que o processo de mistura é modelado por um sistema linear. Em seguida, abordamos uma importante extensão do problema de separação na qual a mistura das fontes possui um caráter não-linear.

Um outro objetivo que acompanhou o desenvolvimento deste trabalho foi a busca por contribuições originais para os casos estudados. Isto ocorreu para uma importante classe de modelos não-lineares, denominada modelos com não-

linearidade posterior. No caso, um novo algoritmo capaz de superar um dos principais empecilhos presentes nesta situação, a convergência para mínimos locais, foi proposto. Nossa solução teve como base o uso de um algoritmo evolutivo (opt-aiNet) na etapa de treinamento e de um estimador de entropia baseado em estatísticas de ordem.

Por fim, tendo em vista que a abrangência do assunto dificulta por vezes um estudo preliminar da literatura em BSS, um outro objetivo desta proposta foi a redação de um documento que auxilie novos pesquisadores na área.

1.2 Organização

O restante da presente dissertação está organizada em quatro capítulos, sendo que o conteúdo de cada um deles é apresentado a seguir:

- Capítulo 2: Neste capítulo, os aspectos básicos do problema de separação cega de fontes e algumas das principais aplicações deste assunto são descritas. Além disso, tratamos de um outro conceito de fundamental importância em BSS: a análise de componentes independentes. Por fim, mencionamos outros métodos, alternativos à ICA, que podem ser utilizados na tarefa de separação.

- Capítulo 3: A separação de fontes para o caso de misturas lineares é tratada neste capítulo. Inicialmente, os principais critérios de separação existentes, bem com os algoritmos desenvolvidos a partir deles são apresentados. Um outro ponto discutido se refere às relações existentes entre esses diversos critérios e à existência de mínimos espúrios nas superfícies de otimização associadas. Por fim, uma análise comparativa dos principais algoritmos, obtida através de simulações, é apresentada. No caso, consideramos algoritmos com operação em batelada e adaptativos.

- Capítulo 4: Neste capítulo, a extensão do problema de separação de fontes para o caso em que as misturas são não-lineares é abordada. Inicialmente, discute-se a viabilidade do uso da ICA neste problema. Em seguida, algumas técnicas destinadas ao caso geral, sem restrições sobre o sistema misturador,

são brevemente discutidas. Um outro ponto de grande relevância neste tema, o modelo com não-linearidade posterior, também é abordado neste capítulo. No tocante a esta classe de modelos, expomos algumas das soluções existentes e discutimos as principais dificuldades a serem superadas neste caso. Finalmente, apresentamos uma nova proposta para este caso em particular e os resultados de alguns experimentos realizados com o intuito de verificar a eficácia de tal solução.

- Capítulo 5: Encerra o documento um capítulo contendo as conclusões gerais da dissertação e as perspectivas para trabalhos futuros.

Capítulo 2

Separação Cega de Fontes

Este capítulo versa sobre os aspectos básicos referentes à Separação Cega de Fontes (BSS, *Blind Source Separation*). Inicialmente, descrevemos o problema e apresentamos um breve histórico sobre este tema. Em seguida, discorremos sobre algumas de suas principais aplicações e expomos a caracterização matemática do problema. Por fim, introduzimos os fundamentos de um assunto fortemente relacionado à BSS, a Análise de Componentes Independentes (ICA, *Independent Component Analysis*), e tecemos alguns comentários sobre outras abordagens alternativas à ICA em BSS.

2.1 Descrição do Problema

Consideremos a situação apresentada na Figura 2.1. No caso, um conjunto de N sinais fontes é submetido à ação de um sistema misturador, ou seja, um sistema cujas M saídas correspondem a misturas de tais fontes[1]. O problema de separação cega de fontes está relacionado à recuperação desses sinais através de amostras das misturas. A peculiaridade da BSS perante outros temas em filtragem é que, nesse caso, não há a necessidade de um conhecimento preciso do sistema misturador e das fontes[2], o que a torna particularmente útil em problemas de caráter não-supervisionado, ou

[1] Mais adiante, veremos como N e M podem estar relacionados.

[2] Por exemplo, veremos mais adiante que, na abordagem da BSS via ICA, é necessário apenas conhecer a estrutura do sistema misturador e que as fontes sejam estatisticamente independentes

seja, nas situações em que é inviável, ou mesmo impossível, utilizar qualquer tipo de sinal piloto no ajuste do sistema separador. O problema é dito *cego* justamente devido a esta falta de informação sobre as misturas e as fontes.

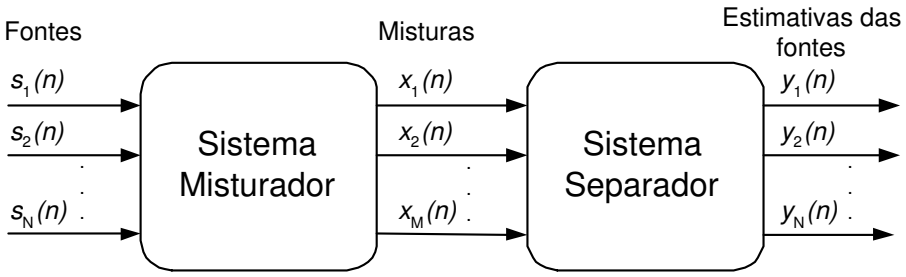


Figura 2.1: Modelo do Problema de Separação Cega de Fontes

2.2 Breve Histórico

O trabalho de Hérault, Jutten e Ans (Hérault, Jutten & Ans, 1985) é considerado o marco inicial da BSS. A motivação de tal trabalho proveio de um problema biológico relacionado à codificação empregada pelo sistema nervoso central para a ativação muscular. Mais especificamente, levando em conta que um único sinal codifica duas informações relevantes (no caso, deslocamento e velocidade angular do movimento) e que o cérebro consegue realizar esta tarefa com sucesso, tentou-se obter um método computacional capaz de distinguir essas duas informações. A despeito das limitações práticas do algoritmo desenvolvido (trataremos delas com mais detalhes na Seção 3.2.1), esse trabalho foi o primeiro a apontar para a necessidade da aplicação de estatísticas de ordem superior ao problema, o que, como veremos na Seção 2.5, mostrou-se fundamental na concepção de métodos eficientes de BSS. Além do mais, um dos alicerces do paradigma BSS/ICA, a modelagem algébrica dos sistemas misturador e separador, também foi originalmente introduzido por esse mesmo trabalho.

Apesar da importância das contribuições presentes em Héraul et al. (1985), foi somente no início da década de 1990 que a BSS passou a atrair uma atenção

entre si.

considerável dos pesquisadores, principalmente na Europa. Um dos principais responsáveis por esse feito foi Pierre Comon, que, a partir dos resultados obtidos na década de 1950 por Darmois, formalizou a idéia da ICA e mostrou como a independência estatística se insere no problema de separação de fontes (Comon, 1994). Essa contribuição teve papel fundamental no desenvolvimento de novos métodos de BSS e no estudo das relações entre as diversas estratégias posteriormente desenvolvidas.

Um outro pesquisador francês cuja contribuição foi fundamental para o desenvolvimento da BSS foi Jean-François Cardoso. Além dos estudos sobre o estimador de máxima verossimilhança em BSS (Cardoso, 1998a), Cardoso introduziu métodos tensoriais no problema (Cardoso & Souloumiac, 1993) e também desenvolveu o chamado gradiente relativo (Cardoso & Laheld, 1996), um método de otimização extremamente conhecido em BSS, que, como será visto na Seção 3.2.8, foi obtido independentemente por Amari (Amari, 1998), que o denominou de gradiente natural.

Podemos dizer que o trabalho de Bell e Sejnowski (Bell & Sejnowski, 1995) também foi crucial para a popularização da BSS, pois, além de estabelecer importantes ligações entre alguns estudos em codificação neural e a ICA, o algoritmo proposto por tais pesquisadores causou um certo furor na comunidade de processamento de sinais devido, sobretudo, à sua capacidade de separar um considerável número de fontes e à sua simplicidade de implementação, qualidades fundamentais em diversas aplicações das técnicas de BSS. Assim, essa contribuição serviu para mostrar que, apesar da complexidade inerente ao problema de separação, seria sim possível desenvolver técnicas implementáveis em cenários práticos.

Por fim, ainda no que diz respeito às contribuições centrais responsáveis pelo desenvolvimento do conjunto BSS/ICA, destacamos os trabalhos de três pesquisadores finlandeses: Karhunen, Oja e Hyvärinen. Os trabalhos de Karhunen e Oja (Karhunen, Pajunen & Oja, 1998) permitiram interpretar a ICA como uma extensão não-linear da consagrada técnica de Análise de Componentes Principais (PCA, *Principal Component Analysis*). Tal abordagem teve um papel fundamental no entendimento da ICA como um tema relevante em análise de dados, ou, ainda, em análise multivariável. Já Hyvärinen contribuiu para o desenvolvimento de critérios

baseados na maximização da não-gaussianidade (Hyvärinen, Karhunen & Oja, 2001), nos quais se baseia uma das técnicas mais utilizadas em separação de fontes, o FastICA.

Além das contribuições supracitadas, há ainda outros trabalhos da década de 1990 que foram importantes para a consolidação da BSS. Procuraremos nessa dissertação, especificamente no Capítulo 3, apresentar os aspectos básicos das principais abordagens desse período. Veremos que a grande maioria desses trabalhos foi desenvolvida levando em conta modelos simplificados de sistemas misturadores, essencialmente lineares e instantâneos (adentraremos a descrição desses modelos na Seção 2.4).

Já em um segundo período da BSS, cujo início se deu no final da década de 1990, as atenções dos pesquisadores voltaram-se para a extensão dos resultados previamente obtidos a casos mais complexos de sistemas misturadores, como, por exemplo, modelos não-lineares, dinâmicos e sub-determinados. Ainda nos dias atuais, essas vertentes correspondem aos principais temas de estudo em BSS.

Podemos também destacar outros dois assuntos que marcam o estágio atual dos estudos em BSS. O primeiro deles refere-se à dissociação entre a BSS e a ICA, que pode ser constatada pelo crescente número de trabalhos na literatura que abordam a BSS por meio de outras metodologias, alternativas à ICA, e, também, pelo surgimento de outras aplicações para a ICA. A outra tendência na área remete à consolidação de diversas aplicações em BSS, principalmente em processamento de sinais biomédicos, que pode ser verificada pela expressiva quantidade de artigos que versam sobre essa questão nas duas últimas edições da conferência mais representativa da área (Puntonet & Prieto, 2004; Rosca, Erdogmus, Principe & Haykin, 2006).

2.3 Aplicações

A generalidade presente na formulação do problema de BSS possibilita uma vasta gama de aplicações, compreendendo desde problemas envolvendo sinais biomédicos a problemas relacionados à econometria. Veremos nesta seção algumas aplicações de destaque das técnicas de BSS nessas diferentes áreas.

2.3.1 Processamento de sinais biomédicos

Em engenharia biomédica, é de grande interesse o desenvolvimento de métodos de aquisição de sinais biomédicos de um paciente que sejam não-invasivos e, ainda assim, confiáveis. O EEG (Eletroencefalograma) e o ECG (Eletrocardiograma) são dois exemplos bem conhecidos de técnicas que operam de acordo com este princípio. Todavia, tal tarefa é de extrema complexidade, tendo em vista a impossibilidade de captar, por meio de sensores posicionados em uma determinada região do corpo humano, apenas os sinais de interesse para um determinado exame, principalmente devido à interferência de sinais gerados pelos mais diversos tipos de atividade fisiológica. Em suma, esses procedimentos são, geralmente, caracterizados por uma baixa relação sinal-ruído (SNR, *Signal-to-Noise Ratio*).

Uma estratégia freqüentemente utilizada para diminuir a intensidade do ruído nas amostras obtidas fundamenta-se na repetição de diversas realizações do exame, de modo que seja possível levantar um comportamento médio dos dados de interesse. A despeito dos bons resultados atingidos, esse tipo de abordagem exige a execução de um elevado número de repetições, o que, em alguns casos, pode não ser um procedimento viável. Além disso, tal conduta pode causar fadiga nos indivíduos examinados, o que, por sua vez, acarreta alterações artificiais dos padrões obtidos, principalmente no monitoramento de sinais cerebrais.

O emprego de técnicas de BSS oferece uma alternativa eficiente a essa abordagem, posto que, neste caso, a recuperação dos sinais de interesse se dá através de estágios sofisticados de processamento conduzidos posteriormente à captação dos dados, o que requer a realização de apenas um experimento. Além disso, como veremos mais adiante, a ausência de modelos capazes de determinar quais sinais fisiológicos interferentes são captados, e, ademais, como eles se misturam, posiciona esse tipo de problema em uma condição extremamente favorável à aplicação dos métodos de BSS. Uma boa evidência dessa aplicabilidade pode ser comprovada pela expressiva quantidade de trabalhos de separação de sinais biomédicos, a tal ponto que podemos dizer que, atualmente, esta área corresponde ao principal domínio de aplicações técnicas de BSS. Na seqüência, apresentamos algumas das aplicações relacionadas a este tema.

ECG (Eletrocardiograma)

O ECG é um procedimento muito comum que permite o diagnóstico de diversas patologias cardíacas através do monitoramento da atividade elétrica do coração. Basicamente, os sinais elétricos captados nesse exame apresentam características temporais particulares como, por exemplo, a chamada onda P e o complexo QRS (Jung et al., 2000). Assim, a partir da observação de distorções nesses padrões característicos, um especialista é capaz de diagnosticar se há alguma patologia, e, dependendo de como esta distorção se dá, de determinar de qual mal se trata.

É de grande complexidade a situação em que se deseja monitorar a atividade cardíaca de um feto, dado que, nesta situação, além da presença de sinais interferentes, a atividade cardíaca da mãe sobressai nos sinais captados (Jung et al., 2000; Lathauwer, Moor & Vandewalle, 2000). Apesar de algumas técnicas serem capazes de obter a freqüência cardíaca fetal, geralmente os procedimentos práticos executados nestes métodos são significativamente mais complexos que os presentes no ECG (Lathauwer, Moor & Vandewalle, 2000). Por outro lado, mesmo um especialista encontraria dificuldades para analisar a atividade cardíaca do feto através dos sinais ECG, o que indica a necessidade de um estágio de processamento capaz de recuperar os sinais de interesse. Uma possibilidade neste caso é lançar mão de técnicas de BSS, que, conforme já mostrado na literatura (Lathauwer, Moor & Vandewalle, 2000; Barros, 2002), são capazes de gerar resultados satisfatórios a partir de algoritmos simples e eficientes.

As técnicas de BSS também podem ser aplicadas em outras duas questões referentes à análise de ECG. A primeira delas relaciona-se à remoção dos mais diversos tipos de sinais interferentes (usualmente chamados de artefatos), como, por exemplo, sinais provenientes das atividades respiratória e muscular (Barros, Mansour & Ohnishi, 1998). Uma outra possibilidade (Jung et al., 2000) sugere a utilização das ferramentas de BSS na tarefa de distinguir os estágios de repolarização e despolarização dos átrios e dos ventrículos.

EEG/MEG/fMRI

A análise de exames de monitoramento de atividade cerebral exerce uma papel fundamental no problema de mapear quais regiões do cérebro são responsáveis pela execução de uma determinada tarefa. Além do mais, esses exames auxiliam especialistas na realização de diagnósticos de determinadas patologias. Naturalmente, diante da complexidade existente em qualquer tipo de procedimento cirúrgico no cérebro, é fundamental que tais exames possam ser feitos de modo não-invasivo. O EEG (Eletroencefalograma) e o MEG (Magnetoencefalograma) (Jung et al., 2000) constituem dois exemplos de técnicas que operam de acordo com este princípio, ou seja, que captam as atividades cerebrais através de sensores posicionados no escalpo. Ao passo que o EEG é sensível às correntes elétricas induzidas em uma determinada região cerebral ativada, o MEG capta os campos magnéticos originados por esta mesma ativação.

Uma das mais utilizadas técnicas de mapeamento cerebral fundamenta-se na observação das atividades cerebrais de um paciente no decorrer de um experimento no qual um certo tipo de estímulo (auditivo, por exemplo) é apresentado diversas vezes. Uma dificuldade inerente a essa estratégia provém da existência de sinais interferentes gerados pelas atividades cerebrais de fundo e pelos diversos artefatos existentes, como por exemplo, os sinais elétricos resultantes do piscar dos olhos. Deste modo, os sinais EEG e MEG representam um conjunto de misturas, sendo que, assim como no caso do ECG, pouco se sabe sobre os sinais fontes e sobre o processo de mistura. É justamente neste contexto que as técnicas de BSS vêm sendo aplicadas com sucesso (Hyvärinen, Karhunen & Oja, 2001; Jung et al., 2000), mesmo em situações com baixa SNR.

Um outro método capaz de monitorar as funções cerebrais de modo não-invasivo é ressonância magnética funcional (fMRI, *Functional Magnetic Resonance Imaging*), cujo modo de operação está associado às diferenças entre as propriedades magnéticas do sangue oxigenado (diamagnético) e o desoxigenado (paramagnético). Quando uma determinada região do cérebro é ativada, há um aumento local do fluxo de sangue oxigenado, e, conseqüentemente, uma alteração das propriedades magnéticas no entorno de tal região. É exatamente esta alteração que é detectada pela fMRI. Assim como no caso do EEG e do MEG, a aplicação das técnicas de BSS também

pode contribuir para uma melhor interpretação das imagens adquiridas (Calhoun, Adali, Hansen, Larsen & Pekar, 2003).

2.3.2 Telecomunicações - BSS e equalização cega de canais

A aplicação da BSS em telecomunicações está fortemente relacionada a um tema de expressiva relevância em comunicações digitais: a equalização de canais. A seguir, faremos uma descrição sucinta deste assunto com o intuito de indicar as principais relações entre ambos os temas.

A idéia essencial de um sistema de comunicação é fazer com que a informação enviada por um transmissor possa ser obtida de maneira tão fiel ao original quanto possível por um receptor. Assim sendo, é primordial que o desenvolvimento de sistemas de comunicação leve em conta estratégias capazes de mitigar as distorções introduzidas pelo canal, elemento presente entre o transmissor e o receptor, na informação transmitida. Em uma das estratégias mais empregadas, a equalização de canal, utiliza-se um filtro (equalizador) no receptor de modo que este seja capaz de inverter a ação do canal. O esquema básico da equalização é apresentado na Figura 2.2. No caso, os sinais $s(n)$, $x(n)$ e $y(n)$ correspondem, respectivamente, ao sinal transmitido, ao sinal recebido e à estimativa do sinal transmitido.

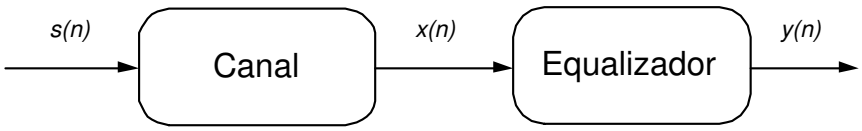


Figura 2.2: O Esquema de Equalização

Em essência, o desenvolvimento de técnicas de equalização está intimamente relacionado à concepção de critérios que guiem o ajuste dos parâmetros livres do equalizador de modo que se obtenha uma boa estimativa do sinal transmitido. Por exemplo, em um dos paradigmas mais conhecidos, adota-se como critério a minimização do erro quadrático médio entre a saída do equalizador e o sinal desejado, no caso, o sinal transmitido (Haykin, 1996).

No caso supracitado, chama a atenção o fato de que o critério adotado se apóia no conhecimento tanto do sinal recebido quanto de amostras do sinal transmitido.

Essa necessidade caracteriza o paradigma de equalização supervisionada (Haykin, 1996). Em contrapartida, os critérios presentes na equalização não-supervisionada (ou cega) utilizam, além dos sinais recebidos, apenas algumas informações estatísticas dos sinais transmitidos. Uma vantagem desta estratégia em relação ao paradigma supervisionado é a possibilidade de realizar o ajuste dos parâmetros concomitantemente com a transmissão dos dados. Por outro lado, a etapa de ajuste dos parâmetros no caso cego é significativamente mais complexa (Haykin, 1994).

Após essa descrição sucinta, sumariemos a essência da equalização cega: recuperar o sinal transmitido, através de um filtro no receptor, valendo-se apenas de amostras da saída do canal. Ora, à luz do que discutimos sobre o problema de BSS na Seção 2.1, fica patente a semelhança entre as formulações de ambos os casos. A diferença básica é que, originalmente, a equalização é definida em um cenário SISO e se baseia em filtragem temporal, ao passo que a BSS aborda sistemas MIMO e se fundamenta em filtragem espacial. Ainda assim, é possível formular o problema de equalização cega de canais SISO como um de BSS (Hyvärinen, Karhunen & Oja, 2001).

Motivados por essa formulação análoga, alguns trabalhos (H. H. Yang, 1998; Hyvärinen, Karhunen & Oja, 2001) abordam o problema de equalização a partir de técnicas consagradas em BSS. Em uma outra linha, um estudo apresentado em (Attux et al., 2006) indica as relações entre algumas idéias originalmente associadas à equalização cega, como, por exemplo, as consagradas abordagens de Shalvi-Weinstein e de Godard e o teorema de Benveniste-Goursat-Rouget (Haykin, 1994), e paradigmas típicos em BSS. Nesse estudo, observa-se que, apesar das relações entre a equalização cega e a BSS, o estudo de tais temas se deu de maneira quase que independente, haja visto o reduzido número de grupos que atuam em ambas as linhas. Diante disso, argumenta-se que um melhor aproveitamento das analogias entre esses problemas poderia acarretar em avanços importantes nessas duas áreas.

No que tange o problema de equalização cega de canais MIMO, podemos afirmar que, em um âmbito teórico, esta situação praticamente se confunde com a formulação da BSS. Neste contexto, merece destaque o trabalho de Cavalcante (Cavalcante, 2004), que tratou um tópico relacionado à equalização

MIMO, a detecção multiusuário, a partir de uma abordagem fundamentada em BSS. A particularidade nesta situação é que se trata de um canal de múltiplo acesso, ou seja, a transmissão das informações enviadas por diferentes usuários se dá num mesmo canal. Este compartilhamento de recursos é possível devido à implantação de um esquema de múltiplo acesso como, por exemplo, as conhecidas estratégias FDMA, TDMA e CDMA.

2.3.3 Separação de sinais de áudio - O *cocktail party-problem*

Um outro importante domínio de aplicações para técnicas de BSS se encontra em problemas típicos de tratamento de sinais de áudio. Nesta linha, merece destaque o *cocktail-party problem*, um dos problemas mais ilustrativos em BSS. Na seqüência, apresentaremos uma descrição sucinta deste assunto.

Imagine a seguinte situação: uma pessoa se encontra em uma sala onde há diversos grupos de pessoas conversando ao mesmo tempo, como, por exemplo, em uma festa ou em uma reunião. Além disso, há no recinto ruído de fundo gerado, por exemplo, por música ambiente e ecos. Apesar de todas essas interferências, o ser humano, através de uma complexa interação entre os sistemas nervoso central e auditivo, é capaz de distinguir a voz ou o som de interesse em um determinado momento. Essa habilidade é conhecida na literatura como *cocktail-party effect* (B. Arons, 1992), justamente pela analogia com o cenário descrito.

Essa destacada capacidade do cérebro humano motiva o seguinte questionamento: será possível a um sistema de processamento artificial alimentado apenas por gravações de microfones posicionados pela sala distinguir o sinal de voz de uma pessoa qualquer? Ao passo que o cérebro humano resolve com certa facilidade este problema, conhecido como *cocktail-party problem* (ilustrado na Figura 2.3), o desenvolvimento de sistemas automáticos para realizar tal tarefa ainda corresponde a um complexo desafio. Até o presente momento, as técnicas baseadas no paradigma BSS via ICA despontam como as mais indicadas para esse tipo de aplicação (Hyvärinen, Karhunen & Oja, 2001).

Uma outra aplicação da BSS em processamento de sinais de áudio é a transcrição musical automática (Plumbley et al., 2002), cujo objetivo é determinar em uma música quais instrumentos e quais notas são executadas em um certo momento.

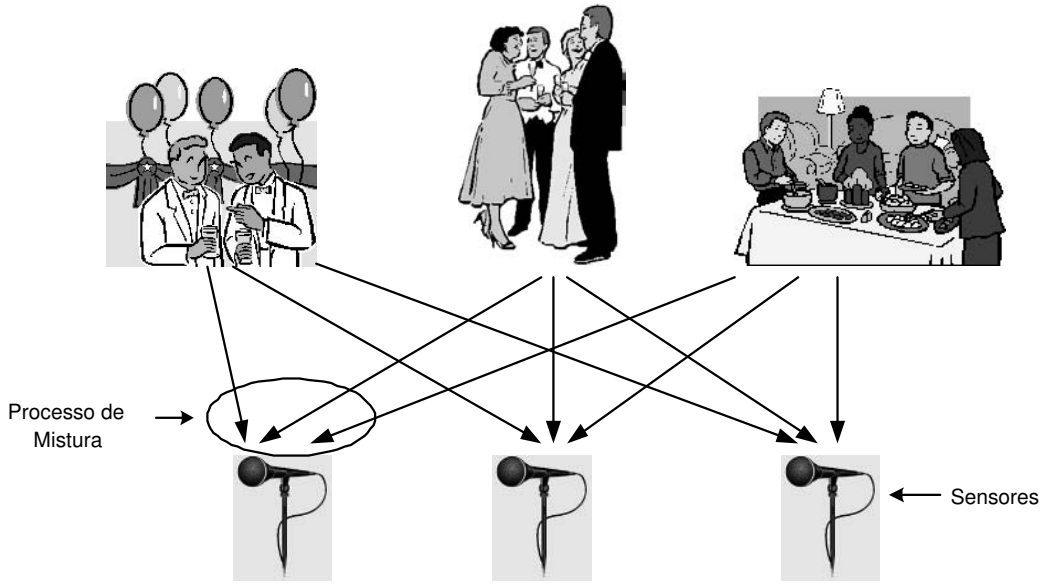


Figura 2.3: O *Cocktail-party Problem*

Também destacamos a utilização da BSS no reconhecimento automático de sinais de fala (F. Arons & Schuster, 1997).

2.3.4 Outras Aplicações

Além das aplicações descritas, há ainda outros problemas de BSS provenientes das mais diversas áreas. Eis alguns exemplos:

- Separação de imagens (Hosseini, Guidara, Deville & Jutten, 2006);
- Exploração geofísica (Hyvärinen, Karhunen & Oja, 2001);
- Arranjos de sensores químicos (Bermejo, Jutten & Cabestany, 2006);
- Cancelamento de reflexões (Hyvärinen, Karhunen & Oja, 2001).

Indicamos ao leitor interessado em outras aplicações as referências (Hyvärinen, Karhunen & Oja, 2001; Cichocki & Amari, 2002; Puntonet & Prieto, 2004; Rosca, Erdogmus, Principe & Haykin, 2006).

2.4 Descrição Matemática do Problema de BSS

Após termos visto as principais aplicações da BSS, além de um breve histórico deste assunto, passaremos a descrever como o problema em questão pode ser resolvido. Um primeiro passo, que será dado na presente seção, é buscar modelos matemáticos capazes de expressar as ações dos sistemas misturador e separador, de acordo com a idéia ilustrada na Figura 2.1.

De modo a obter uma descrição matemática do problema de BSS, considere que cada elemento do vetor $\mathbf{s}(n) = [s_1(n)\ s_2(n) \dots s_N(n)]^T$ corresponde a um sinal fonte. Analogamente, representemos os sinais misturados através do vetor $\mathbf{x}(n) = [x_1(n)\ x_2(n) \cdots x_M(n)]^T$. Em sua forma mais geral, o processo de mistura das fontes pode ser representado pela seguinte expressão:

$$\mathbf{x}(n) = \Im(\mathbf{s}(n),\ \mathbf{s}(n-1) \dots \mathbf{s}(n-L), \mathbf{n}(n)), \tag{2.1}$$

onde o mapeamento $\Im(\cdot)$ descreve a ação do sistema misturador, L corresponde ao número de amostras passadas levadas em conta no processo de mistura, ou seja, diz respeito à memória associada ao sistema, e o vetor de $\mathbf{n}(n)$ denota o ruído associado às próprias fontes (ruído de fonte) e/ou aos sensores (ruído de sensor).

É importante salientar que a formulação apresentada (2.1) possui um caráter eminentemente didático, haja visto a inexistência de técnicas de BSS capazes de lidar com todos os efeitos representados nessa expressão (ruído, memória, não-linearidade, etc). Em geral, as técnicas desenvolvidas em BSS são direcionadas para casos particulares, mais simplificados, da formulação apresentada. Assim, de modo a auxiliar o leitor na identificação dos diversos casos presentes na literatura, apresentamos na seqüencia como é a feita a classificação de um sistema misturador.

Sistemas Lineares e Não-Lineares Um sistema misturador é dito linear se o mapeamento $\Im(\cdot)$ atende ao princípio da superposição, ou seja, nas situações em que:

$$\Im(b_1\mathbf{s}_1(n) + b_2\mathbf{s}_2(n)) = b_1\Im(\mathbf{s}_1(n)) + b_2\Im(\mathbf{s}_2(n)), \tag{2.2}$$

para quaisquer constantes b_1 e b_2, e vetores de sinais $\mathbf{s}_1(n)$ e $\mathbf{s}_2(n)$. Caso contrário, o sistema misturador é dito não-linear.

Sistemas Instantâneos e Convolutivos Nas situações em que o processo de mistura depende de amostras passadas, ou seja, $L > 0$, o sistema misturador é dito convolutivo, ou com memória. Por outro lado, nas situações em que $L = 0$, o sistema é dito instantâneo.

Com Relação ao Número de Fontes e de Sensores Quando o número de sensores é maior que o número de fontes ($M > N$), tem-se o chamado caso sobre-determinado. Analogamente, o caso sub-determinado corresponde à situação em que ($M < N$). É interessante notar que os casos em que há sinais de ruído podem ser tratados como um caso específico de modelos sub-determinados, dado que é possível considerá-los como sendo fontes (Kofidis, 2001).

Já vimos que uma das características marcantes do problema de BSS é a falta de informação sobre o processo de mistura e sobre as fontes. Todavia, é fundamental ao menos um certo conhecimento sobre a estrutura do sistema misturador, pois, com base nessa informação, torna-se possível definir um sistema separador estruturalmente capaz de inverter a ação do processo de misturas (por exemplo, para inversão de misturas não-lineares, é imperativo o uso de uma estrutura de mesma natureza). Geralmente, esse tipo de conhecimento é obtido com base na aplicação de interesse. Como exemplo, é sabido que, em problemas de separação de sinais de áudio, o processo de mistura é notadamente convolutivo devido a efeitos de reverberação ou eco.

A maioria dos trabalhos em BSS abordam cenários com sistemas misturadores lineares, instantâneos e com o mesmo número de fontes e misturas. Nesta situação, o processo de mistura é descrito matematicamente (doravante omitiremos o índice temporal por questões de simplicidade) do seguinte modo:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.3}$$

sendo que, nesta situação, a matriz $\mathbf{A}$ de dimensão $N \times N$ é chamada de matriz de mistura. A despeito de sua simplicidade, esta classe de modelos é válida em uma vasta quantidade de problemas de BSS (Hyvärinen, Karhunen & Oja, 2001). Ademais, é possível, neste caso, recuperar as fontes através da Análise de Componentes Independentes, técnica esta que será o assunto da próxima seção.

2.5 Análise de Componentes Independentes

Em linhas gerais, a Análise de Componentes Independentes pode ser compreendida como uma técnica em análise de dados cujo objetivo é obter uma explicação alternativa para um determinado conjunto de dados a partir de um modelo generativo. A origem da ICA está diretamente ligada ao problema de BSS, sendo que o mesmo trabalho pioneiro em BSS de Hérault, Jutten e Ans (Hérault, Jutten & Ans, 1985) também é considerado o marco inicial da ICA.

Nesta seção, apresentaremos os aspectos básicos da ICA, situando-a em relação a alguns métodos semelhantes. Embora a ICA seja definida tanto para cenários lineares quanto para não-lineares, trataremos apenas o primeiro caso nesta seção, sendo que, no Capítulo 4, discutiremos em mais detalhes o segundo caso. Recomendamos ao leitor interessado em um estudo aprofundado sobre os aspectos teóricos da ICA a leitura do trabalho de Comon (Comon, 1994), responsável pela formalização matemática desse assunto, e as referências (Hyvärinen, Karhunen & Oja, 2001; Eriksson & Koivunen, 2004; Hyvärinen, 1999b).

2.5.1 Definição

A ICA é um tópico relativamente novo em análise de dados se comparada, por exemplo, com a PCA, que já é estudada há cerca de um século. Provavelmente, esse é o motivo pelo qual encontramos na literatura ao menos duas definições distintas para a ICA. De certo modo, essas duas definições se complementam, sendo que cada uma delas pode ser mais ou menos representativa dependendo da aplicação de interesse. Comecemos pela definição mais geral, estabelecida por Comon:

Definição 2.5.1 (ICA) *A ICA de um vetor aleatório* $\mathbf{x} = [x_1\ x_2 \cdots x_M]^T$ *consiste na determinação de uma transformação linear* $\mathbf{y} = \mathbf{W}\mathbf{x}$ *de tal maneira que os elementos do vetor aleatório* $\mathbf{y} = [y_1\ y_2 \cdots y_N]^T$ *otimizem uma função custo* $\Psi(\mathbf{y})$*, denominada função contraste, que expresse uma medida de independência.*

Uma primeiro ponto marcante dessa definição é a sua relação com a idéia de independência estatística entre as variáveis aleatórias, o que nos motiva a relembrar este conceito fundamental da teoria de probabilidade. As variáveis aleatórias

$x_1, x_2 \ldots x_N$ são estatisticamente independentes entre si quando a seguinte condição é estabelecida

$$p_{x_1,x_2,\ldots x_N}(x_1, x_2, \ldots x_N) = p_{x_1}(x_1)p_{x_2}(x_2)\ldots p_{x_N}(x_N), \tag{2.4}$$

onde $p_{x_1,x_2,\ldots x_N}(x_1, x_2, \ldots x_N)$ corresponde à função densidade de probabilidade conjunta das variáveis envolvidas e $p_{x_i}(x_i)$ representa a função densidade de probabilidade marginal de x_i.

Um outro conceito importante presente nessa definição é o de função contraste ou, simplesmente, contraste. Matematicamente, um contraste $\Psi(\mathbf{y})$ realiza um mapeamento entre um conjunto de funções densidades de probabilidade, associadas aos elementos do vetor aleatório $\mathbf{y}$, e o conjunto dos números reais[3], satisfazendo as seguintes propriedades

- $\Psi(\mathbf{y})$ deve ser invariante às permutações dos elementos de $\mathbf{y}$:

  $$\Psi(\mathbf{y}) = \Psi(\mathbf{P} \cdot \mathbf{y}), \tag{2.5}$$

  para qualquer matriz de permutação $\mathbf{P}$;

- $\Psi(\mathbf{y})$ deve ser invariante à escala:

  $$\Psi(\mathbf{y}) = \Psi(\mathbf{\Lambda} \cdot \mathbf{y}), \tag{2.6}$$

  para qualquer matriz diagonal Λ;

- Quando os elementos $\mathbf{y}$ forem independentes entre si, deve valer:

  $$\Psi(\mathbf{y}) \geq \Psi(\mathbf{A} \cdot \mathbf{y})^4, \tag{2.7}$$

  para qualquer matriz inversível $\mathbf{A}$;

[3]Note o leitor que o contraste não é uma simples função de uma variável aleatória. Ou seja, para mapear um conjunto de variáveis aleatórios em um número real, o contraste deve estar associado a um operador tal como a esperança estatística.

[4]Esta condição representa a situação na qual a recuperação da independência é feita pela maximização do contraste. Alternativamente, é possível definir esta propriedade do seguinte modo: $\Psi(\mathbf{y}) \leq \Psi(\mathbf{A} \cdot \mathbf{y})$. Este segundo caso corresponde à situação em que a recuperação da independência está associada a um problema de minimização.

Essa última restrição ilustra bem o propósito de uma função contraste que é justamente quantificar o "nível de independência" entre os elementos de um vetor aleatório, dado que a igualdade nessa expressão ocorre, em ambas as restrições, apenas quando a matriz $\mathbf{A}$ não combina nenhuma das variáveis aleatórias. Nesse sentido, as duas primeiras condições são necessárias, tendo em vista que a independência entre variáveis aleatórias é insensível à ordem entre elas e à multiplicação delas por constantes.

Uma segunda definição está mais associada ao contexto de estimação e representação de dados.

Definição 2.5.2 (ICA) *A ICA de um vetor aleatório* $\mathbf{x} = [x_1\ x_2 \cdots x_M]^T$ *consiste em determinar o seguinte modelo generativo linear (comumente chamado de modelo ICA):*

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.8}$$

onde os elementos de $\mathbf{s} = [s_1\ s_2 \cdots s_N]^T$ *são estatisticamente independentes entre si e* $\mathbf{A}$ *corresponde a uma matriz constante de dimensão* $M \times N$ *.*

No caso em que o vetor aleatório $\mathbf{x}$ é gerado por um modelo ICA e $N = M$, as definições 2.5.1 e 2.5.2 são equivalentes. Por exemplo, sob a ótica da segunda definição, temos que $\mathbf{x} = \mathbf{A}\mathbf{s}$ e, portanto, $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$, o que está de acordo com a definição 2.5.1, dado que os elementos de $\mathbf{s}$ são estatisticamente independentes.

2.5.2 Aplicação da ICA ao problema de BSS

Com o objetivo de compreendermos como a ICA se insere no problema de separação, recapitulemos o modelo de sistema misturador mais simplificado, dado por $\mathbf{x} = \mathbf{A}\mathbf{s}$, onde $\mathbf{A}$ corresponde a uma matriz de mistura quadrada. Sob a hipótese de que os sinais fontes são estatisticamente independentes entre si (hipótese razoável em várias situações), fica claro, para uma matriz $\mathbf{A}$ que de fato combine as fontes, que os elementos de $\mathbf{x}$ não são mais independentes. O ponto fundamental da aplicação da ICA diz respeito exatamente a esta constatação, no sentido de que essa metodologia se propõe a separar as fontes a partir da recuperação da independência. Assim, no paradigma BSS/ICA, o sistema separador, no caso a matriz $\mathbf{W}$, é ajustado de modo

a gerar estimativas das fontes $\mathbf{y}$ de tal forma que os elementos deste vetor sejam os mais independentes possíveis entre si.

Porém, diante da discussão acima, surge a seguinte questão fundamental: tornar as estimativas das fontes independentes implica necessariamente na recuperação das fontes? Essa questão está relacionada com a separabilidade do modelo ICA. Foi Pierre Comon (Comon, 1994) quem respondeu esta pergunta, fornecendo, assim, todo respaldo matemático necessário para o desenvolvimento da ICA e, conseqüentemente, da BSS. A partir do teorema de Darmois, ele mostrou que, de fato, é possível separar as fontes com base na recuperação da independência estatística, desde que as fontes e o sistema separador satisfaçam algumas condições. Para entendermos melhor essa questão é interessante descrever em mais detalhes o conceito de separabilidade.

Separabilidade

O modelo de mistura (2.3) é dito *separável* se, para toda matriz $\mathbf{W}$ que resulte em um vetor $\mathbf{y}$ cujos elementos são estatisticamente independentes entre si, tem-se que $y = \mathbf{Wx} = \mathbf{\Lambda P s}$, onde $\mathbf{\Lambda}$ e $\mathbf{P}$ correspondem a matrizes diagonal e de permutação, respectivamente. Deste modo, em um modelo separável, é possível recuperar as fontes através da ICA. Note, no entanto, que há duas ambigüidades associadas a este procedimento, no sentido de que não é possível recuperar a ordem das fontes nem seus ganhos de escalas. De certo modo, essa limitação é bem intuitiva, pois, conforme discutido anteriormente, a independência estatística entre os elementos de um vetor não é alterada por permutações e escalas.

O seguinte teorema, provado por Comon (Comon, 1994), evidencia as condições necessárias para que um dado sistema misturador linear seja separável:

Teorema 2.5.1 (Separabilidade do Modelo ICA) *O modelo* $\mathbf{x} = \mathbf{As}$ *é separável se e somente a matriz* $\mathbf{A}$ *possuir posto completo, e, no máximo, um dos elementos do vetor aleatório* $\mathbf{s}$ *for gaussiano.*

Esse teorema indica uma outra restrição (já vimos que as fontes devem ser estatisticamente independentes entre si) da aplicação ICA ao problema de BSS, que é exatamente a incapacidade de recuperar fontes gaussianas. Essa limitação ficará

clara na Seção 2.5.3 quando tratarmos da utilização de estatísticas de segunda ordem para realizar a separação.

Os aspectos teóricos discutidos relativos à separabilidade demonstram que é possível resolver o problema de separação com base na independência estatística, através da aplicação da ICA. Na seção seguinte, veremos que, em contrapartida, o conceito de correlação, uma medida "mais fraca" que a independência, não é suficiente para prover a separação das fontes.

2.5.3 Aplicação de técnicas baseadas em estatísticas de segunda ordem à BSS

Até a década de 1980, a grande maioria das técnicas em filtragem estatística eram baseadas em estatísticas de segunda ordem, devido, principalmente, ao fato de que modelos gaussianos de sinais eram praticamente os únicos utilizados e à simplicidade matemática inerente a esse tipo de abordagem. Todavia, durante o desenvolvimento da teoria de filtragem não-supervisionada, primeiramente no conceito de identificação e equalização cega, constatou-se que a resolução dessa classe de problemas vai além das estatísticas de segunda ordem, exigindo informações sobre as chamadas estatísticas de ordem superior (HOS, *Higher-Order Statistics*) (Nikias & Petropulu, 1993). Em BSS, esse conhecimento também é indispensável.

Já vimos na Seção 2.5.2 que o problema de BSS pode ser resolvido com base na independência estatística, ou seja, através de HOS, posto que a independência pressupõe o conhecimento das densidades de probabilidade, e, conseqüentemente, de todas as estatísticas de uma variável aleatória. Nesta seção, por outro lado, apresentaremos o motivo pelo qual a abordagem da BSS utilizando somente informações de segunda ordem não é suficiente para a resolução do problema.

A despeito dessa limitação, o emprego de técnicas de segunda ordem resultou em contribuições significativas para a BSS, principalmente no desenvolvimento de estágios de pré-processamento. Motivados por isso, apresentaremos, em um primeiro momento, os fundamentos de uma consagrada técnica dessa classe: a PCA.

Análise por componentes principais (PCA)

A PCA (Hyvärinen, Karhunen & Oja, 2001), ou transformada discreta de Karhunen-Loève, é uma consagrada técnica em análise de dados usada principalmente em aplicações de compressão de dados e extração de características. Diferentemente da ICA, que utiliza a independência estatística como medida de redundância entre as componentes, a PCA emprega a correlação para realizar essa tarefa. Se, por um lado, o uso da correlação, uma medida "mais fraca" de informação se comparada com a independência, pode não ser suficiente para explorar toda a redundância entre variáveis aleatórias, pelo outro, permite o desenvolvimento de algoritmos simplificados e, não-obstante, eficazes nos casos em que os dados de interesse são gaussianos.

Em linhas gerais, a PCA pode ser entendida sob a ótica do problema de compressão de um vetor aleatório $\mathbf{x} = [x_1 \cdots x_M]^T$, sendo que há redundância entre os elementos desse vetor. Na PCA, a tarefa de compressão é realizada através da busca de uma transformação linear que, aplicada ao vetor $\mathbf{x}$, resulte em um vetor aleatório $\mathbf{y} = [y_1 \cdots y_N]^T$ (com $N < M$) cujos elementos, denominados componentes principais, sejam descorrelacionados entre si. Além disso, a busca é conduzida de modo que as projeções de $\mathbf{x}$ nesse novo espaço de coordenadas, ou seja, os elementos de $\mathbf{y}$, possuam a máxima quantidade de informação, que, no caso, é medida pela variância.

Matematicamente, a determinação das componentes principais pode ser realizada do seguinte modo: em um primeiro momento, busca-se um vetor $\mathbf{w_1} = [w_{11} \cdots w_{M1}]^T$, de norma euclidiana unitária, tal que a seguinte combinação linear possua máxima variância:

$$y_1 = \mathbf{w}_1^T \mathbf{x}. \tag{2.9}$$

No caso, diz-se que y_1 é a primeira componente principal de $\mathbf{x}$. O mesmo processo é feito para determinar a segunda componente principal y_2; porém, tendo em vista o espírito do processo de compressão, não deve existir correlação entre esse novo dado e y_1. É possível mostrar que a determinação de y_2 (Hyvärinen, Karhunen & Oja, 2001) do seguinte modo:

$$y_2 = \mathbf{w}_2^T \mathbf{x}, \tag{2.10}$$

sendo que $\mathbf{w}_1^T\mathbf{w}_2^T = 0$, garante que a correlação entre y_1 e y_2 seja nula. Finalmente, tal procedimento pode ser generalizado para a determinação da i-ésima componente principal y_i, com $i \leq M$. Assim,

$$y_i = \mathbf{w}_i^T\mathbf{x}, \tag{2.11}$$

com $\mathbf{w}_i^T\mathbf{w}_j = \delta_{ij}$, onde δ_{ij} corresponde à função delta.

Um outro modo de compreender o problema relativo à determinação das componentes principais baseia-se em uma formulação de erro quadrático médio mínimo (MMSE, *Minimum Mean Square Error*). Nesta situação, a determinação dos vetores base $\mathbf{w}_i$ é conduzida a partir da minimização do erro quadrático médio de compressão, descrito por:

$$J_{PCA} = E\{\|\mathbf{x} - \sum_{i=1}^{N}(\mathbf{w}_i^T\mathbf{x})\mathbf{w}_i\|^2\}, \tag{2.12}$$

com $\mathbf{w}_i^T\mathbf{w}_j = \delta_{ij}$. Diante desta formulação, nota-se que, quanto maior for o número de componentes principais considerados na compressão, menor será o erro de compressão.

Assumindo que os elementos do vetor $\mathbf{x}$ possuem média nula, é possível mostrar que a solução do problema de otimização descrito pela expressão (2.12) está relacionada aos autovetores da matriz de correlação $\mathbf{R}_x = E\{\mathbf{x}\mathbf{x}^T\}$ (considerando que $\mathbf{x}$ é um vetor de média nula). Mais especificamente, qualquer base ortonormal do sub-espaço gerado por esses autovetores é uma solução ótima do problema em questão. Além disso, os autovalores dessa matriz fornecem informações sobre a ordem das componentes principais, ou seja, o autovetor com maior autovalor associado corresponde à primeira componente principal e assim sucessivamente. Infelizmente, esta solução analítica do problema de PCA baseada em uma decomposição de autovalores e autovetores pode não ser viável em muitas aplicações, especialmente naquelas em que se exige um processamento em tempo real. Isso motivou o desenvolvimento de um considerável número de técnicas adaptativas de PCA, das quais muitas são fundamentadas em redes neurais artificiais (Diamantaras & Kung, 1996).

Recuperação das fontes via descorrelação

Ainda no que diz respeito à expressão (2.12), não há compressão dos dados na situação em que $N = M$. Todavia, este procedimento permite a determinação de uma transformação sobre $\mathbf{x}$ que resulte em um vetor aleatório descorrelacionado, de modo semelhante ao que ocorre na ICA. Este processo é comumente chamado de branqueamento espacial, posto que nenhum tipo de informação temporal é levada em conta. Na seqüência, mostraremos que não é possível separar as fontes realizando apenas tal processo.

Considerando o modelo ICA apresentado em (2.3), é possível obter a matriz de correlação entre as misturas, no caso dada por

$$\mathbf{R_x} = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{A}\mathbf{R_s}\mathbf{A}^T = \mathbf{A}\mathbf{A}^T, \tag{2.13}$$

onde $\mathbf{R_s}$ corresponde à matriz de correlação das, fontes, sendo que, diante da hipótese de independência das fontes e da consideração, sem perda de generalidade, de que as fontes possuem variância unitária, tal matriz equivale à identidade. Isso justifica a última igualdade desta expressão.

Após a atuação do sistema separador, a correlação entre as estimativas das fontes é dada por

$$\mathbf{R_y} = \mathbf{W}\mathbf{R_x}\mathbf{W}^T, \tag{2.14}$$

de tal forma que, para descorrelacionar (branquear) as saídas do sistema separador, é necessário determinar $\mathbf{W}$ tal que:

$$\mathbf{R_y} = \mathbf{I} \rightarrow \mathbf{W}\mathbf{R_x}\mathbf{W}^T = \mathbf{I}, \tag{2.15}$$

onde $\mathbf{I}$ é a matriz identidade.

Uma solução dessa equação pode ser obtida através da decomposição em autovalores e autovetores da matriz de correlação das misturas, dada por $\mathbf{R_x} = \mathbf{E}\mathbf{D}\mathbf{E}^T$, onde a matriz $\mathbf{E}$ é ortogonal, com colunas que correspondem aos autovetores de $\mathbf{R_x}$, e $\mathbf{D}$ é uma matriz diagonal contendo os autovalores de $\mathbf{R_x}$. No caso, é fácil verificar que o seguinte sistema separador satisfaz a condição expressa em (2.15):

$$\mathbf{W} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T. \tag{2.16}$$

Agora, consideremos um sistema separador dado por $\mathbf{QW}$, onde $\mathbf{Q}$ corresponde a uma matriz ortogonal. Substituindo esta solução na expressão (2.15), e lembrando que $\mathbf{QQ}^T = \mathbf{I}$ e $\mathbf{EE}^T = \mathbf{I}$ (matrizes ortogonais), obtém-se:

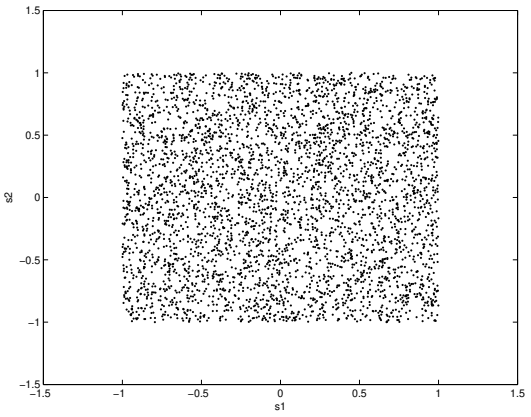
$$\mathbf{WR_x W}^T = \mathbf{QED}^{-1/2}\mathbf{E}^T\mathbf{EDE}^T\mathbf{ED}^{-1/2}\mathbf{E}^T\mathbf{Q}^T = \mathbf{I}. \tag{2.17}$$

Diante desse resultado, é possível concluir que a recuperação das fontes exclusivamente a partir da correlação apresenta uma indeterminação relacionada com um fator ortogonal (no caso, a matriz $\mathbf{Q}$), de uma maneira análoga ao problema da indeterminação da fase em outros casos de filtragem (Kofidis, 2001).

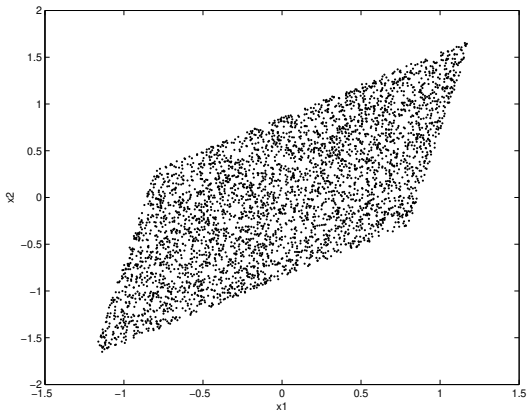
Na Figura 2.4, essa indeterminação é ilustrada a partir de uma caracterização geométrica. Com este intuito, são exibidas nas Figuras 2.4(a), 2.4(b) e 2.4(c) as distribuições conjuntas de duas fontes uniformemente distribuídas, suas misturas e suas estimativas obtidas pelo branqueamento dessas misturas, respectivamente. Sendo o efeito do sistema misturador linear modelado por uma matriz, as misturas são geradas a partir de escalonamentos e rotações das fontes. Apesar da recuperação via braqueamento conseguir recuperar as escalas das fontes, ela é incapaz de recuperar a rotação pois, como vimos, existe uma indeterminação referente a uma matriz ortogonal, cujo efeito é exatamente a rotação dos dados.

Essa ineficácia das estatísticas de segunda ordem nos auxilia no entendimento, de um modo mais intuitivo, de uma significativa limitação da ICA: a impossibilidade de recuperar fontes gaussianas. É sábido que apenas o conhecimento da média e da variância caracterizam uma variável aleatória gaussiana, ou seja, a definição desta não vai além das estatísticas de segunda ordem. Por outro lado, a eficácia da ICA pode ser encarada como um resultado da consideração de momentos de ordem superior. Diante dessas duas constatações, é de se esperar que, por não "possuir informações" desta sorte, as variáveis gaussianas não sejam discriminadas no decorrer da aplicação da ICA.

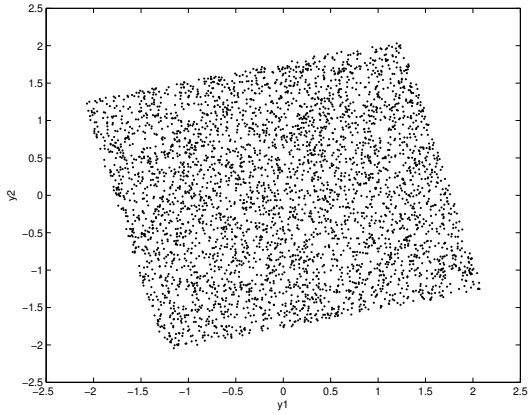
Por fim, enfatizamos que, a despeito da impossibilidade de recuperação das fontes utilizando somente um sistema MIMO branqueador, esta estratégia pode ser útil como uma etapa de pré-processamento para a BSS, pois, após a sua aplicação, resta apenas determinar uma matriz de rotação. Isto implica em uma simplificação do espaço de busca, possibilitando uma redução considerável do número de parâmetros



(a) Distribuição conjunta das fontes.



(b) Distribuição conjunta das misturas.



(c) Distribuição conjunta das estimativas obtidas a partir do braqueamento das misturas.

Figura 2.4: Tratamento da BSS considerando estatística de segunda ordem.

de problema de otimização associado. Além disso, em algumas técnicas de ICA, o branqueamento prévio das misturas é fundamental para atingir a separação.

2.5.4 Aplicações da ICA

A aplicação clássica da ICA é o problema de BSS. Aliás, é habitual na literatura, principalmente nos trabalhos mais antigos, o tratamento da BSS e da ICA praticamente como sinônimos, o que corresponde a uma descrição errônea,

dado que, rigorasamente, a ICA corresponde a uma metodologia capaz de resolver o problema de BSS. Isto fica evidente se observarmos o amplo número de aplicações da ICA que não necessariamente estão ligadas ao problema de separação. Eis alguns exemplos:

- Extração de características (Bell & Sejnowski, 1997);
- Compressão de dados (Guilhon, Medeiros & Barros, 2005);
- Econometria (Hyvärinen, Karhunen & Oja, 2001);
- Predição de séries temporais (Hyvärinen, Karhunen & Oja, 2001);
- Estágio de pré-processamento em sistemas classificadores (Sanchez-Poblador, Monte-Moreno & Solé-Casal, 2004).

2.6 Outras estratégias em Separação Cega de Fontes

A existência de cenários práticos nos quais as hipóteses fundamentais da ICA (independência e não-gaussianidade) não são válidas motivou a aplicação de outras abordagens ao problema de BSS. Nesta seção, faremos uma descrição sucinta de duas delas: a análise de componente esparsos (SCA, *Sparse Component Analysis*) e o tratamento Bayesiano.

2.6.1 Análise de componentes esparsos

A SCA vem se mostrando uma ferramenta eficaz no problema de BSS, sobretudo nos casos de modelos de mistura sub-determinados. Esta abordagem fundamenta-se na hipótese da esparsidade das fontes, isto é, assume-se que na maior parte do tempo as fontes assumem valores próximos a zero (Bofill, 2001) (exemplos de fontes esparsas são apresentados na Figura 2.5). Esta situação é típica em cenários com sinais de vozes e instrumentos musicais.

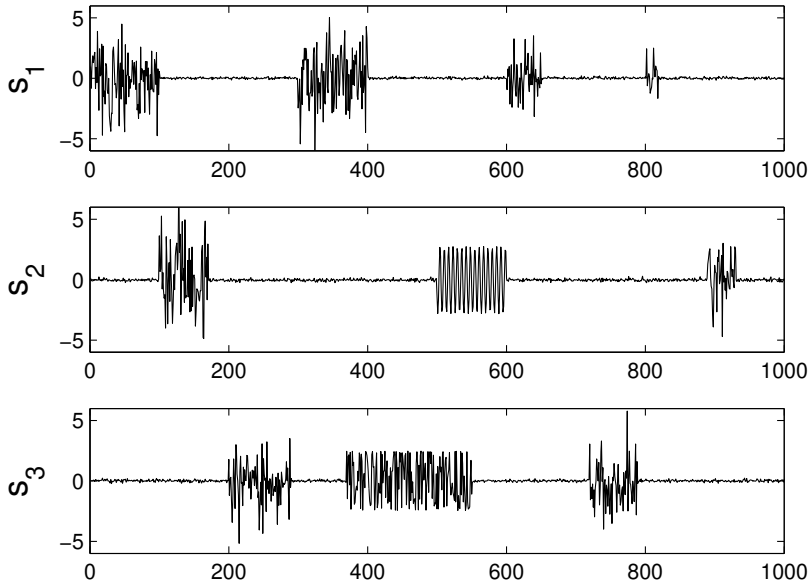
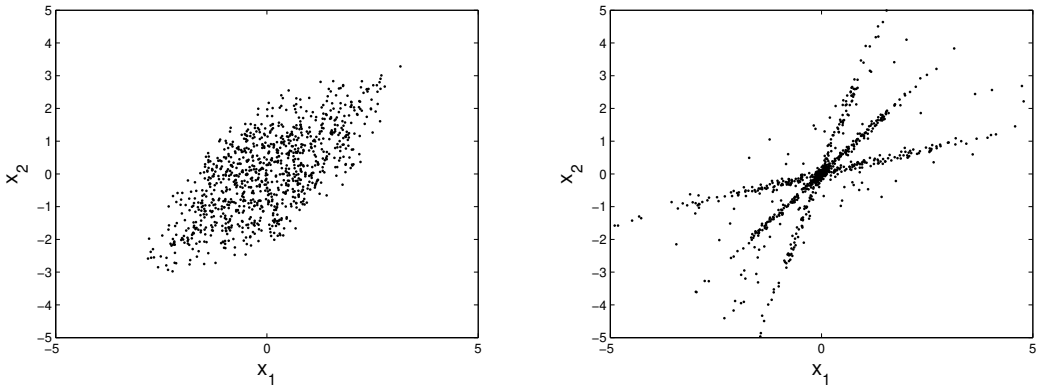


Figura 2.5: Exemplo de fontes esparsas

De modo a entendermos a utilidade da hipótese de esparsidade no caso sub-determinado, analisemos a Figura 2.6, que apresenta a distribuição conjunta das saídas de um sistema misturador sub-determinado (3 fontes e 2 sensores) em duas situações: fontes uniformemente distribuídas, logo não-esparsas (Figura 2.6(a)), e fontes esparsas apresentadas na Figura 2.5 (Figura 2.6(b)). Ao passo que no primeiro cenário nenhuma informação sobre a matriz de mistura pode ser obtida, na segunda situação, as direções dos vetores coluna desta matriz são salientadas, pois, como as fontes assumem valores quase nulos em boa parte do tempo, é provável que, em um determinado período, apenas uma das fontes esteja "ativa", e, conseqüentemente, o vetor de misturas $\mathbf{x} = [x_1 \quad x_2]^T$ nesse instante possua a mesma direção da coluna da matriz de mistura associada a tal fonte. Logo, é possível determinar todas as colunas dessa matriz a partir da estimação das direções no espaço definido pelas misturas que apresentam as maiores concentrações de amostras. Este processo pode ser feito através de técnicas de clusterização (Bofill, 2001).

Mesmo nos casos em que as fontes não são suficientemente esparsas no tempo,



(a) Distribuição das misturas de fontes não-esparsas.

(b) Distribuição das misturas de fontes esparsas.

Figura 2.6: Saídas de um sistema misturador sub-determinado (3 fontes e 2 sensores).

ainda é possível utilizar a SCA. Isso pode ser feito aplicando uma transformação linear (Fourier, por exemplo) ao conjunto de dados de modo que, neste novo domínio, as fontes sejam esparsas. Assim, após a etapa de separação, realizada no novo domínio, aplica-se a transformação inversa e obtêm-se as estimativas das fontes.

2.6.2 Abordagem Bayesiana

Como já vimos, o bom funcionamento das técnicas baseadas na ICA é garantido quando as fontes são estatisticamente independentes entre si, o que pode ser uma hipótese pouco realista em diversos problemas. Uma outra desvantagem relacionada à ICA é a dificuldade de incorporar informações *a priori* do problema em questão. Uma possível solução para ambos os problemas provém da abordagem Bayesiana de separação de fontes (Djafari, 1999).

A idéia essencial na abordagem Bayesiana é determinar uma expressão para a probabilidade *a posteriori* $P(\mathbf{A}, \mathbf{s}|\mathbf{x})$ em termos da verossimilhança do modelo $P(\mathbf{x}|\mathbf{A}, \mathbf{s})$ e das probabilidades *a priori* das fontes e da matriz de mistura, dadas por $P(\mathbf{s})$ e $P(\mathbf{A})$, respectivamente. No caso, isto é feito utilizando a regra de Bayes,

o que resulta na seguinte expressão:

$$P(\mathbf{A}, \mathbf{s}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{A}, \mathbf{s})P(\mathbf{A})P(\mathbf{s}). \tag{2.18}$$

As fontes e a matriz de mistura podem ser estimadas a partir da maximização conjunta ($\mathbf{A}$ e $\mathbf{s}$) desta probabilidade *a posteriori*. Os dois últimos termos dessa expressão representam o conhecimento *a priori* sobre o problema.

Uma outra vantagem da abordagem Bayesiana é a sua capacidade de operar em cenários ruidosos. O procedimento é basicamente o mesmo que o apresentado na expressão (2.18), sendo acrescentado apenas um modelo probabilístico para o ruído. Por fim, salientamos o fato que o tratamento Bayesiano também é aplicável na separação de misturas sub-determinadas.

2.7 Sumário

Neste capítulo, o problema de separação cega de fontes foi apresentado. Inicialmente, um breve histórico, algumas das principais aplicações e uma caracterização matemática desse assunto foram expostas. Em seguida, os aspectos básicos da ICA e o uso desta técnica no problema BSS foram tratados. Além disso, discutiu-se o porquê da ineficácia do uso em BSS de métodos baseados em estatísticas de segunda ordem. Por fim, os aspectos de outras duas técnicas que podem ser utilizadas no problema de BSS, a SCA e a abordagem Bayesiana, foram descritos.

Capítulo 3

Separação de Misturas Lineares

Neste capítulo, descrevemos o funcionamento das principais estratégias empregadas em BSS na situação em que o sistema misturador é linear e instantâneo e o número de fontes é o mesmo que o de sensores. Primeiramente, apresentamos as diversas propostas concebidas para este caso. Em seguida, fazemos algumas considerações sobre as relações existentes entre os diferentes critérios associados a tais abordagens. Além disso, tecemos alguns comentários sobre a eventual presença de mínimos locais em alguns dos critérios descritos. Finalmente, analisamos o desempenho dos algoritmos apresentados, considerando suas versões adaptativas e com modo de operação em batelada.

3.1 Etapas no Projeto de uma Técnica de BSS

No capítulo anterior, já tivemos uma noção das etapas fundamentais presentes no projeto de uma técnica de BSS ao vermos, por exemplo, que os algoritmos baseados na idéia da ICA estão associados a um problema de otimização, no qual a função custo em questão está diretamente relacionada ao conceito de independência estatística. Também discutimos um outro ponto crucial a ser levado em conta no projeto, que diz respeito ao conhecimento da estrutura do sistema misturador. É importante, antes de iniciarmos nossa exposição sobre as ferramentas existentes em BSS, que o leitor tenha em mente uma noção mais sistematizada das etapas presentes

na concepção de uma técnica de BSS.

Em essência, o desenvolvimento de uma técnica de BSS passa pelos seguintes estágios:

1. Definição da estrutura do sistema separador;
2. Estabelecimento de um critério que guie a busca por um bom conjunto de parâmetros do sistema separador;
3. Aplicação de uma técnica capaz de resolver o problema de otimização resultante.

É importante salientar aqui o caráter didático desta divisão, uma vez que pode haver, em algumas abordagens, uma sobreposição entre essas diferentes etapas. Ainda assim, todas as estratégias discutidas neste capítulo apresentam essa estrutura claramente definida.

3.2 Principais Abordagens em BSS - Caso Linear

Nesta seção, apresentaremos as diversas estratégias existentes em BSS linear, começando pela solução pioneira de Heráult e outros. Salientamos que a divisão das sub-seções está majoritariamente associada à idéia empregada no desenvolvimento do critério de separação, ou seja, ao segundo item de nossa discussão anterior. No entanto, em cada sub-seção, indicamos quais são as ferramentas de otimização comumente associadas a cada critério.

De acordo com a notação adotada no Capítulo 2, os sinais fontes, as misturas e as estimativas das fontes são representados, respectivamente, pelos vetores $\mathbf{s} = [s_1\ s_2 \dots s_N]^T$, $\mathbf{x} = [x_1\ x_2 \dots x_N]^T$ e $\mathbf{y} = [y_1\ y_2 \dots y_N]^T$. Além do mais, os sistemas misturador e separador são denotados pelas matrizes $\mathbf{A}$ e $\mathbf{W}$, respectivamente. Finalmente, em todas as técnicas apresentadas a seguir, assume-se que as fontes são estatisticamente independentes e não-gaussianas, em conformidade com as restrições da ICA.

3.2.1 A proposta de Heráult, Jutten e Ans

A solução de Hérault, Jutten e Ans (Hérault, Jutten & Ans, 1985) utiliza a seguinte estrutura como sistema separador

$$\mathbf{y} = \mathbf{x} - \mathbf{M}\mathbf{y}, \tag{3.1}$$

sendo que a matriz $\mathbf{M}$ possui tamanho $N \times N$, e considera-se que os elementos de sua diagonal são iguais a zero, o que significa que não existe auto-realimentação. Equivalentemente, poderíamos representar esta estrutura através de sua relação entrada-saída, ou seja, com $\mathbf{y} = \mathbf{W}\mathbf{x}$, onde $\mathbf{W} = (\mathbf{M} + \mathbf{I})^{-1}$.

Para entendermos a idéia associada à utilização dessa estrutura, consideremos a expressão (3.1) para uma dado elemento y_i

$$y_i = x_i - \sum_{j,j\neq i} m_{ij} y_j. \tag{3.2}$$

Sob a hipótese de que todas as misturas contém informações sobre todas as fontes, argumentou-se que, na situação em que boas estimativas das fontes y_j já estão disponíveis, seria possível recuperar a fonte y_i retirando de x_i a contribuição das estimativas y_j que, por hipótese, já estariam suficientemente próximas das fontes originais. Em um primeiro momento, tentou-se implementar esta idéia através da minimização do erro quadrático médio entre os dois termos da expressão (3.2), para todas as estimativas. Neste caso, aplicando o método do gradiente descendente para resolver esta tarefa, obtém-se a seguinte regra de atualização

$$m_{ij} \leftarrow m_{ij} - \mu E\{y_i y_j\}, \tag{3.3}$$

onde μ é o passo de adaptação.

Analisando esta regra de aprendizado notamos que a sua convergência em termos da média estatística requer que

$$E\{y_i y_j\} = 0 \qquad \forall i, j, \tag{3.4}$$

ou seja, para que este algoritmo convirja, é necessário que as estimativas das fontes sejam descorrelacionadas entre si (considerando fontes com média nula). Tendo em

vista a discussão presente na Seção 2.5.3, é de se esperar que esta condição não seja suficiente para prover a separação das fontes, pois leva em conta somente informações de segunda ordem. Foi justamente nesse ponto que a contribuição de Hérault et al foi fundamental para o desenvolvimento da BSS, pois, embora eles não tivessem uma idéia plena sobre a importância da independência estatística no problema em questão, eles já sabiam que sua solução iria além das estatísticas de segunda ordem.

Deste modo, buscou-se contornar o problema através de uma simples modificação na expressão (3.3), que consistiu na introdução de funções não-lineares ímpares $f(\cdot)$ e $g(\cdot)$, resultando em uma nova regra de ajuste, descrita por

$$m_{ij} \leftarrow m_{ij} - \mu E\{f(y_i)g(y_j)\}. \tag{3.5}$$

Neste caso, a convergência desta regra de atualização se dá quando

$$E\{f(y_i)g(y_j)\} = 0, \qquad \forall i, j, \tag{3.6}$$

ou seja, quando os sinais são, digamos, *não-linearmente descorrelacionados* entre si.

Os resultados obtidos em (Hérault, Jutten & Ans, 1985) sugeriram que, de fato, se considerássemos essa versão não-linear da correlação, seria possível recuperar as fontes. O papel decisivo das não-linearidades neste caso pode ser entendido a partir da expansão em séries de Taylor das funções $f(\cdot)$ e $g(\cdot)$. Lembrando que a esperança estatística $E\{\cdot\}$ é um operador linear, a condição (3.6) pode ser expressa da seguinte forma

$$\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} f_k g_l E\{y_i^k y_j^l\} = 0 \qquad \forall i, j, \tag{3.7}$$

onde f_k e g_l correspondem, respectivamente, ao k-ésimo e ao i-ésimo coeficiente da expansão em série de Taylor das funções $f(\cdot)$ e $g(\cdot)$. Assim, uma possível solução da equação (3.7) ocorre quando

$$E\{y_i^k y_j^l\} = 0 \qquad \forall i, j, k, l, \tag{3.8}$$

Devido ao fato das funções não-lineares introduzidas serem ímpares, k e l assumem apenas valores ímpares, o que torna a condição (3.8) plausível para o problema em questão, pois, se considerássemos momentos pares, esta condição só seria alcançada para sinais determinísticos.

A expressão (3.8) nos permite constatar uma interessante propriedade relacionada ao uso de funções não-lineares em critérios estatísticos. De modo implícito, essa estratégia introduz informações sobre algumas estatísticas de ordem superior das variáveis aleatórias envolvidas no problema. Evidentemente, não há garantia alguma de que a convergência da regra (3.5) implique na independência estatística entre as estimativas das fontes. Porém, é importante ressaltar que, ao considerar as estatísticas de ordem superior, essa estratégia utiliza uma medida mais sólida que a correlação. Quanto à sua aplicação prática, essa técnica opera satisfatoriamente apenas em cenários com reduzido número de fontes, necessariamente subgaussianas.

3.2.2 Minimização da informação mútua

Com base na discussão feita na Seção 2.5, podemos afirmar que a implementação prática da ICA depende de uma escolha apropriada da função contraste, que, por sua vez, pode ser compreendida, de uma maneira intuitiva, como um tipo de medida de independência estatística entre variáveis aleatórias. Neste contexto, levando em conta a definição de independência, apresentada em (2.4), uma possível solução seria utilizar, como contraste, alguma entidade matemática capaz de expressar uma idéia de distância entre a distribuição conjunta e as distribuições marginais das variáveis aleatórias em questão. Uma possibilidade é a divergência de Kullback-Leibler (Comon, 1994). Matematicamente, a divergência de Kullback-Leibler entre duas funções multidimensionais $f(\mathbf{r})$ e $g(\mathbf{r})$ é dada por

$$D(f(\mathbf{r}), g(\mathbf{r})) = \int f(\mathbf{r}) \log \frac{f(\mathbf{r})}{g(\mathbf{r})} d\mathbf{r}. \tag{3.9}$$

Um importante caso particular dessa medida ocorre quando uma das funções corresponde à densidade de probabilidade conjunta de um vetor aleatório $\mathbf{y}$ e a outra denota o produto das densidades marginais dos elementos deste vetor. Matematicamente, este caso é expresso do seguinte modo

$$D(p_{\mathbf{y}}(\mathbf{y}), p_{y_1}(y_1)p_{y_2}(y_2)\ldots p_{y_N}(y_N)) = \int p_{\mathbf{y}}(\mathbf{y}) \log \frac{p_{\mathbf{y}}(\mathbf{y})}{p_{y_1}(y_1)p_{y_2}(y_2)\ldots p_{y_N}(y_N)} d\mathbf{y}. \tag{3.10}$$

Apesar desta medida não ser uma distância de fato, uma vez que não é simétrica, ela sempre assume valores não-negativos, e é igual a zero somente quando as duas funções em questão forem iguais. Ou seja, somente no caso em que os elementos de $\mathbf{y}$ forem independentes. Nesta situação, o argumento do logaritmo equivale a um, e, conseqüentemente, tal expressão se anula, em consonância com a definição de contraste.

A expressão (3.10) é também a definição de uma importante grandeza em teoria da informação: a informação mútua entre os elementos de um vetor aleatório[1], que, após um simples desenvolvimento, pode ser representada do seguinte modo

$$D(p_{\mathbf{y}}(\mathbf{y}), p_{y_1}(y_1)p_{y_2}(y_2)\ldots p_{y_N}(y_N) = I(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{y}), \tag{3.11}$$

onde $H(\cdot)$ corresponde à medida de entropia diferencial. Deste modo, o ajuste do sistema separador $\mathbf{W}$ pode ser feito de modo a minimizar a informação mútua das estimativas das fontes $\mathbf{y} = \mathbf{W}\mathbf{x}$.

A principal dificuldade relacionada à minimização da informação mútua encontra-se na etapa de otimização. Em Babaie-Zadeh (2002), mostrou-se que a aplicação do método do gradiente descendente (*steepest descent*) fornece a seguinte regra de atualização

$$\mathbf{W} \leftarrow \mathbf{W} - \mu E\{\Psi(\mathbf{y})\mathbf{x}^T\} - (\mathbf{W}^T)^{-1}, \tag{3.12}$$

onde $\Psi(\mathbf{y}) = [\psi_{y_1}(y_1), \ldots, \psi_{y_N}(y_N)]$, de modo que $\psi_{y_i}(y_i) = (p_{y_i}(y_i)'/p_{y_i}(y_i))$. Diante dessa expressão, nota-se que o ajuste da matriz $\mathbf{W}$ requer, a cada iteração, o conhecimento das densidades de probabilidade das estimativas das fontes, o que acarreta num significativo custo computacional. Uma alternativa a este problema se dá através de uma aproximação da informação mútua por expressões que levam em conta apenas um reduzido número de estatísticas de ordem superior. Veremos mais adiante que este tipo de abordagem relaciona-se com outros critérios de BSS.

[1]No Apêndice A, apresentamos uma breve revisão dos principais conceitos da teoria da informação.

3.2.3 Estimação por máxima verossimilhança

Diferentemente do critério de minimização da informação mútua, que pode ser visto como uma aplicação direta da ICA, a maioria dos critérios no caso linear da BSS, embora também possuam ligações com a ICA, foram concebidos a partir de outras idéias, não necessariamente ligadas à independência estatística. Neste contexto, um dos exemplos mais conhecidos provém de uma consagrada técnica da teoria de estimação: a estimação por máxima verossimilhança. Nesta seção, descreveremos como a abordagem por máxima verossimilhança é desenvolvida no problema de separação. Porém, antes, revisaremos este importante conceito.

Primeiramente, recapitulemos o problema de estimação de parâmetros, a saber: determinar um estimador para os parâmetros $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]$ a partir de um conjunto de amostras $\mathbf{e} = [e(1), \ldots, e(J)]$ que contém informações sobre eles. Na estimação por máxima verossimilhança, as estimativas de $\boldsymbol{\theta}$, denotadas por $\widetilde{\boldsymbol{\theta}}$, são obtidas através da maximização da chamada função de máxima verossimilhança $L(\boldsymbol{\theta})$. Matematicamente, este procedimento é descrito do seguinte modo

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} p_{\mathbf{e}}(\mathbf{e} \mid \boldsymbol{\theta}), \tag{3.13}$$

ou seja, a abordagem por máxima verossimilhança busca o conjunto de parâmetros que maximiza a probabilidade condicional de $\mathbf{e}$ dado $\boldsymbol{\theta}$.

Normalmente, assume-se que as observações $\mathbf{e} = [e(1), \ldots, e(J)]$ são estatisticamente independentes entre si, e que, portanto, a função de máxima verossimilhança é dada por

$$L(\boldsymbol{\theta}) = p_{\mathbf{e}}(\mathbf{e} \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} p_e(e(j)|\boldsymbol{\theta}). \tag{3.14}$$

Há alguns aspectos teóricos interessantes no paradigma de máxima verossimilhança, como, por exemplo, o fato desta abordagem ser capaz de atingir assintoticamente o limitante inferior de Cramer-Rao (Kay, 1993).

Com relação ao emprego da estimação por máxima verossimilhança no problema de BSS, os parâmetros a serem determinados e os dados disponíveis são, respectivamente, os elementos da matriz de mistura $\mathbf{A}$ (ou, equivalentemente,

do sistema separador $\mathbf{W}$) e as amostras das misturas, representadas por $X = [\mathbf{x}(1), \ldots, \mathbf{x}(J)]$. Assim sendo, a função de máxima verossimilhança é dada por

$$L(\mathbf{A}) = \prod_{j=1}^{J} p_{\mathbf{x}}(\mathbf{x}(j)|\mathbf{A}). \tag{3.15}$$

Lembrando que (Papoulis, 1993)

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{s}}(\mathbf{A}^{-1}\mathbf{x})}{|\det(\mathbf{A})|}, \tag{3.16}$$

e considerando a hipótese de independência entre as fontes, a expressão (3.15) pode ser escrita do seguinte modo

$$L(\mathbf{A}) = \prod_{j=1}^{J} \frac{\prod_{n=1}^{N} p_{s_n}(\widetilde{\mathbf{a}}_n \mathbf{x}(j))}{|\det(\mathbf{A})|}. \tag{3.17}$$

onde $\widetilde{\mathbf{a}}_n$ corresponde à n-ésima linha da matriz inversa de $\mathbf{A}$. Equivalentemente, podemos descrever a função de máxima verossimilhança em termos da matriz $\mathbf{W}$. Nesta situação, temos que

$$L(\mathbf{W}) = \prod_{j=1}^{J} \prod_{n=1}^{N} p_{s_n}(\mathbf{w}_n \mathbf{x}(j)) |\det(\mathbf{W})|, \tag{3.18}$$

onde $\mathbf{w}_n$ denota a n-ésima linha da matriz $\mathbf{W}$.

Normalmente, considera-se o logaritmo da expressão (3.18), a chamada função de máxima verossimilhança logarítmica. Evidentemente, sendo o logaritmo uma função monotônica crescente, o máximo desta nova expressão coincide com o máximo da função de máxima verossimilhança. No contexto da BSS, esta simplificação, acrescida de uma normalização com respeito ao número de amostras, resulta na seguinte expressão

$$\frac{1}{J}\log(L(\mathbf{W})) = \frac{1}{J}\sum_{j=1}^{J}\sum_{n=1}^{N} \log(p_{s_n}(\mathbf{w}_n \mathbf{x}(j))) + \log(|\det(\mathbf{W})|). \tag{3.19}$$

Note que, em vez de basear-se em produtórios, esta expressão é descrita através de somatórios, o que a torna mais simples em um contexto prático.

Através da lei dos grandes números (Cardoso, 1998a), é possível obter uma variante probabilística da expressão (3.19), dada por

$$\frac{1}{J}\log(L(\mathbf{W})) = E\{\sum_{n=1}^{N}\log(p_{s_n}(\mathbf{w}_n\mathbf{x}(j)))\} + \log(|\det(\mathbf{W})|). \tag{3.20}$$

Em (Kofidis, 2001), a partir desta expressão, mostra-se que a abordagem por máxima verossimilhança em BSS também pode ser compreendida à luz da divergência de Kullback-Leibler. No caso, demonstrou-se que a estimação do sistema separador (ou misturador) pelo critério de máxima verossimilhança resulta no seguinte problema de otimização

$$\widetilde{\mathbf{W}} = \arg\max_{\mathbf{W}}(L(\mathbf{W})) \triangleq \arg\min_{\mathbf{W}}(D(p_{\mathbf{Wx}}(\mathbf{r}))) \mid p_{\mathbf{s}}(\mathbf{r})). \tag{3.21}$$

onde $D(\cdot|\cdot)$ corresponde a divergência de Kullback-Leibler, descrita em (3.9).

Essa nova representação da abordagem de máxima verossimilhança no problema de separação nos permite interpretá-la como uma critério de casamento de funções densidade de probabilidade. Mais especificamente, a idéia presente na expressão (3.21) baseia-se na determinação de uma matriz $\mathbf{W}$ de modo que as densidades das fontes $\mathbf{s}$ e de suas estimativas $\mathbf{Wx}$ sejam as mais próximas possíveis, no sentido da divergência de Kullback-Leibler[2].

Um outro ponto que fica claro a partir de (3.21) é que a função de máxima verossimilhança em BSS satisfaz os requisitos de um contraste, dado que o ótimo desta expressão ocorre somente quando houver um casamento entre as distribuições das estimativas e das fontes, situação esta que, assumindo as hipóteses de separabilidade, implica necessariamente na recuperação das fontes escalonadas e/ou permutadas. Todavia, é importante salientar que a estratégia de máxima verossimilhança possui uma importante limitação relacionada *à necessidade do conhecimento das densidades de probabilidades das fontes*, o que, por sua vez, contrasta com o caráter cego da separação.

A despeito dessa significativa limitação prática do estimador de máxima verossimilhança, ainda assim é possível realizar a busca por um sistema separador

[2]Em (Attux et al., 2006), discute-se as semelhanças entre essa abordagem e o teorema de Benveniste-Goursat-Rouget, um dos principais resultados em equalização cega.

adequado a partir de (3.21), desde que se empregue estimativas para as densidades das fontes. Veremos, na próxima seção, que tal conduta possui uma ligação direta com uma das principais abordagens em BSS: o critério Infomax.

3.2.4 O critério Infomax

O princípio Infomax (*Information Maximization*), desenvolvido por Linsker (Linsker, 1988), corresponde a um dos principais paradigmas de treinamento não-supervisionado de redes neurais baseados em elementos da teoria da informação. Em um primeiro momento, esse conceito foi aplicado a sistemas de processamento linear. Porém, em 1994, mostrou-se que a extensão não-linear do Infomax está diretamente relacionada ao princípio da redução de redundância de Barlow (Nadal & Parga, 1994), que, por sua vez, possui uma estreita ligação com a ICA. No seminal trabalho de Bell e Sejnowski (Bell & Sejnowski, 1995), no qual constatou-se, de modo independente, essa mesma relação, propôs-se uma técnica para resolver o problema de BSS fundamentada no critério Infomax. Na seqüência, apresentaremos os principais pontos relativos a esta idéia.

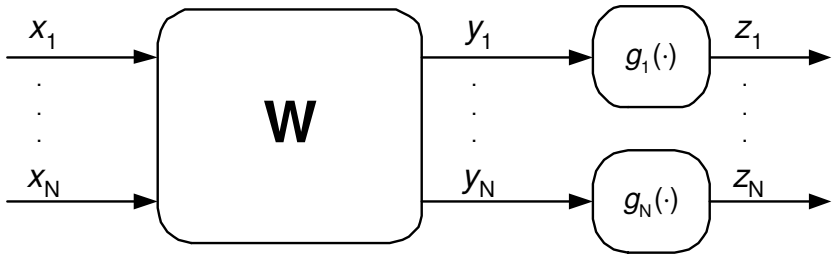


Figura 3.1: Estrutura do sistema separador no critério Infomax

A proposta de Bell e Sejnowski abordou o treinamento de uma rede neural constituída por um estágio linear, que representaremos pela matriz $\mathbf{W}$, seguido de um estágio não-linear caracterizado pelas funções de ativação (monotonicamente crescentes de 0 a 1) $\mathbf{g}(\cdot) = [g_1(\cdot) \ldots g_N(\cdot)]$, de acordo com a Figura 3.1. Matematicamente, o mapeamento entrada-saída desta estrutura é dado por:

$$\mathbf{z} = \mathbf{g}(\mathbf{y}) = \mathbf{g}(\mathbf{W}\mathbf{x}) = [\ g_1(\mathbf{w}_1\mathbf{x}) \quad \ldots \quad g_N(\mathbf{w}_N\mathbf{x})\ ]^T. \tag{3.22}$$

No trabalho em questão, define-se um sistema separador com base nesta rede neural, sendo que as estimativas das fontes, nesta situação, são fornecidas pelas saídas do estágio linear. Apesar de não participar diretamente no processamento dos dados, o estágio não-linear, como veremos mais adiante, possui fundamental importância no ajuste dos parâmetros da rede. Este tipo de função difere da maioria das aplicações de redes neurais em processamento de sinais, nas quais o caráter não-linear dessas estruturas é responsável pela capacidade de aproximação universal, desejável em problemas onde existe a necessidade de aproximação de mapeamentos, como identificação e equalização.

Na abordagem Infomax, o ajuste dos parâmetros da rede neural é feito de modo a *maximizar a transferência de informação* entre as entradas e as saídas de tal estrutura. Uma maneira de quantificar esse princípio provém da definição da informação mútua de Shannon, o que nos permite entender o objetivo do Infomax como sendo a maximização da informação mútua entre as entradas e as saídas da rede. Diante disso, e levando em conta somente a adaptação da matriz $\mathbf{W}$, a aplicação do critério Infomax na estrutura descrita em (3.22) resulta no seguinte problema de otimização

$$\max_{\mathbf{W}} I(\mathbf{z}, \mathbf{x}) = H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}), \tag{3.23}$$

onde $I(\mathbf{z}, \mathbf{x})$ corresponde à informação mútua entre $\mathbf{z}$ e $\mathbf{x}$. Devido ao mapeamento entrada-saída dessa estrutura ser determinístico, a entropia condicional $H(\mathbf{z}|\mathbf{x})$ não depende de $\mathbf{W}$, e, portanto, nesta situação, o critério Infomax é equivalente à maximização da entropia conjunta das saídas dessa rede.

Valendo-se da expressão da entropia de uma transformação (Apêndice A), é possível mostrar que a entropia conjunta das saídas pode ser expressa da seguinte maneira

$$H(\mathbf{z}) = H(\mathbf{x}) + E\{\sum_{i=1}^{N} \log(g_i'(\mathbf{w}_i\mathbf{x}))\} + \log(|\det(\mathbf{W})|), \tag{3.24}$$

onde $g_i'(\cdot)$ representa a derivada primeira da função $g_i(\cdot)$. Nesta expressão, apenas os dois últimos termos dependem de $\mathbf{W}$ e, portanto, o problema de otimização descrito

em (3.23) possui, equivalentemente, a seguinte formulação

$$\max_{\mathbf{W}} H(\mathbf{z}) \triangleq \max_{\mathbf{W}} E\{\sum_{i=1}^{N} \log(g_i^{'}(\mathbf{w}_i\mathbf{x}))\} + \log(|\det(\mathbf{W})|). \tag{3.25}$$

Essa nova formulação do critério Infomax em BSS aponta para uma forte correspondência entre esta abordagem e aquela baseada na estimação por máxima verossimilhança. Este resultado fica claro à luz da expressão (3.20), que difere de (3.25) apenas nas funções não-lineares empregadas em cada um dos critérios. Note ainda que, na situação em que as funções de ativação empregadas correspondem às distribuições cumulativas das fontes, isto é, $g_i(r) = \int_{-\infty}^{r} p_{s_i(s)} ds$, os dois critérios são equivalentes.

A equivalência entre essas duas abordagens foi demonstrada por Cardoso (Cardoso, 1997), que também investigou uma outra questão muito relevante relacionada a este assunto, a saber: dado que o Infomax, em um contexto de separação, pode ser considerado uma aproximação da solução de máxima verossimilhança, qual seria a influência da escolha de uma função de ativação qualquer? Note que, sob a ótica do estimador de máxima verossimilhança, esta situação corresponderia a assumir aproximações para as densidades de probabilidade das fontes. Mais adiante, trataremos desta questão em maiores detalhes. Antes, apresentaremos dois algoritmos capazes de ajustar o sistema separador de acordo com o paradigma Infomax/Máxima Verossimilhança.

Algoritmos BS e ACY

Uma maneira de se obter uma matriz de separação $\mathbf{W}$ que maximize a expressão (3.25) é realizar a busca a partir do método do gradiente descendente. Nesta situação, é possível mostrar (Hyvärinen, Karhunen & Oja, 2001) que o gradiente da entropia conjunta das saídas em relação à $\mathbf{W}$ é dado por

$$\frac{\partial H(\mathbf{z})}{\partial \mathbf{W}} = E\{\mathbf{G}(\mathbf{W}\mathbf{x})\mathbf{x}^T\} + (\mathbf{W}^T)^{-1}, \tag{3.26}$$

onde $\mathbf{G}(\cdot) = [G_1(\cdot) \ldots G_N(\cdot)]$ é um vetor de funções de modo que $G_i(x) = d\log(g_i^{'}(x))/dx$. Diante disso, obtém-se a seguinte regra de atualização para o ajuste

de $\mathbf{W}$

$$\mathbf{W} \leftarrow \mathbf{W} + \mu\{E\{\mathbf{G}(\mathbf{W}\mathbf{x})\mathbf{x}^T\} + (\mathbf{W}^T)^{-1}\}, \tag{3.27}$$

onde μ corresponde ao passo de adaptação. Na literatura, esta regra é conhecida como algoritmo Bell-Sejnowski (BS), e sua versão estocástica (adaptativa) pode ser obtida desconsiderando o operador de esperança presente nesta expressão.

Alternativamente, é possível tratar o problema de otimização dado pela expressão (3.25) empregando um método que se fundamenta em uma variante do gradiente, denominada gradiente relativo ou natural (Cardoso & Laheld, 1996; Amari, 1998). Na Seção 3.2.8, discutiremos os principais conceitos por trás dessa técnica. Por ora, adiantamos que, neste caso em particular, esta entidade relaciona-se com o gradiente do seguinte modo

$$\frac{\partial_N H(\mathbf{z})}{\partial \mathbf{W}} = \frac{\partial H(\mathbf{z})}{\partial \mathbf{W}}\mathbf{W}^T\mathbf{W}, \tag{3.28}$$

onde $\frac{\partial_N H(\mathbf{z})}{\partial \mathbf{W}}$ corresponde ao gradiente natural de $H(\mathbf{z})$ em relação à matriz $\mathbf{W}$. Substituindo esta expressão em (3.27), obtém-se a seguinte regra de atualização

$$\mathbf{W} \leftarrow \mathbf{W} + \mu(\mathbf{I} + E\{\mathbf{G}(\mathbf{y})\mathbf{y}^T\})\mathbf{W}, \tag{3.29}$$

Esta técnica é conhecida na literatura como algoritmo ACY (Amari, Cichocki e Yang). Diferentemente do algoritmo BS, essa técnica não requer inversão de matrizes, sendo esta uma grande vantagem com relação ao custo computacional requerido.

A influência da função de ativação

Tendo discutido os principais algoritmos de BSS baseados no critério Infomax, e, equivalentemente, no estimador de máxima verossimilhança, analisemos o importante ponto que diz respeito à influência das funções de ativação no desempenho das técnicas derivadas a partir desse paradigma. Os trabalhos de Amari (Amari, Chen & Cichocki, 1997) e Cardoso (Cardoso, 1998b) abordam essa questão através de uma análise de estabilidade local dos pontos estacionários relativos às regras de atualização descritas anteriormente. Na seqüencia, apresentaremos as idéias básicas presentes nesses dois trabalhos.

Primeiramente, de modo a simplificar a análise, façamos a seguinte modificação na regra de atualização da matriz de separação $\mathbf{W}$ presente no algoritmo ACY

$$\mathbf{W} \leftarrow \mathbf{W} + \mu(\mathbf{I} - E\{\mathbf{F}(\mathbf{y})\mathbf{y}^T\})\mathbf{W}, \tag{3.30}$$

onde $\mathbf{F}(\cdot) = -\mathbf{G}(\cdot)$.

Analisemos o que ocorre no algoritmo (3.30) quando $\mathbf{W}$ de fato corresponde à inversa da matriz de mistura, porém escalonada[3], ou seja,$\mathbf{W} = \mathbf{\Lambda}\mathbf{A}^{-1}$, de modo que $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$. As matrizes $\mathbf{\Lambda}\mathbf{A}^{-1}$ tal que $\mathbf{\Lambda}$ satisfaça a seguinte restrição

$$E\{F_i(\lambda_i s_i)\lambda_i s_i\} = 1, \quad i = 1, \ldots, n, \tag{3.31}$$

são pontos estacionários do algoritmo, pois, devido a hipótese de independência entre as fontes (considerando nulas suas médias), a seguinte relação é válida

$$E\{F_i(\lambda_i s_i)\lambda_j s_j\} = 0, \quad i \neq j, \tag{3.32}$$

e, assim, temos que $(\mathbf{I} + E\{\mathbf{F}(\mathbf{y})\mathbf{y}^T\}) = 0$. Chama a atenção nessa constatação que este fato *independe* das funções $\mathbf{F}(\cdot)$, pois a expressão (3.32) é satisfeita, sob as hipóteses mencionadas, para qualquer função não-linear inversível.

Em um primeiro momento tal observação pode parecer um tanto incoerente, pois, mesmo nos casos em que as funções $\mathbf{F}_i(\cdot)$ são lineares, as matrizes de separação $\mathbf{\Lambda}\mathbf{A}^{-1}$ são pontos estacionários do algoritmo, contradizendo a necessidade de estatísticas de ordem superior, as quais são introduzidas justamente pelas não-linearidades, no problema de BSS. No entanto, em se tratando de uma dinâmica, é necessário verificar se tais pontos estacionários são estáveis, pois, somente assim, estes corresponderiam a atratores potenciais para o processo de aprendizado.

Uma maneira simples de verificar a estabilidade de um ponto estacionário pode ser feita através de técnicas de linearização. Este procedimento consiste em linearizar a dinâmica não-linear em torno de um dado ponto estacionário e, posteriormente, verificar se o sistema linear resultante é estável. Obviamente, este tipo de análise é útil apenas no estudo de estabilidade local, dado que o comportamento do sistema linear resultante é capaz de aproximar o sistema não-linear apenas em uma certa vizinhança do ponto estacionário.

[3]Sem perda de generalidade, a ambigüidade relacionada às permutações não é considerada.

A partir dessa análise baseada em linearização, verificou-se (Cardoso, 1998b) que as condições que garantem a estabilidade local dos pontos estacionários $\mathbf{W} = \mathbf{\Lambda}\mathbf{A}^{-1}$ são as seguintes

$$\begin{aligned} &(1+\upsilon_i)(1+\upsilon_j) > 1, \quad 1 \leq i < j \leq N, \\ &(1+\upsilon_i) > 0, \quad 1 \leq i \leq N \end{aligned} \tag{3.33}$$

onde

$$\upsilon_i = E\{F_i'(y_i)\}E\{y_i^2\} - E\{F_i(y_i)y_i\}, \quad i = 1, \ldots, N. \tag{3.34}$$

Uma condição suficiente para a estabilidade é $\upsilon_i > 0$ para todas as fontes. No caso de uma fonte gaussiana, por exemplo, υ_i é nulo. Assim, em cenários em que há mais de uma fonte deste tipo, não é possível separá-las, dado que a primeira condição de estabilidade não é satisfeita para este par, o que confirma as limitações da ICA discutidas previamente.

Analisando as expressões (3.33) e (3.34), fica claro que a condição de estabilidade *depende* das funções $F_i(\cdot)$ e também das distribuições de probabilidade das fontes, pois devemos lembrar que nos pontos estacionários sob análise as estimativas das fontes correspondem às fontes escalonadas, ou seja, $\mathbf{y} = \mathbf{\Lambda}\mathbf{s}$.

No tocante à abordagem de máxima verossimilhança, onde $F_i(\cdot) = -\frac{p'_{s_i}(\cdot)}{p_{s_i}(\cdot)}$, argumenta-se que as condições (3.33) são sempre satisfeitas (Cardoso, 1998b). Por outro lado, há outras possibilidades de escolha para as funções não-lineares de modo que a estabilidade ainda seja garantida. Neste sentido, a partir da observação de que o sinal de υ_i está, para um bom número de funções não-lineares, diretamente relacionado ao fato das fontes serem sub-gaussianas (curtose negativa[4]) ou super-gaussianas (curtose positiva), é possível satisfazer esta condição de estabilidade empregando apenas duas classes de funções, sendo uma para cada tipo de fonte. Assim, a utilização de um estimador de máxima verossimilhança aproximado (o critério Infomax), de modo que as funções de ativação sejam definidas com base na gaussianidade das fontes, não implica em uma perda de desempenho em relação ao caso ideal, ao menos no sentido de que os pontos ótimos relacionados às matrizes de separação que estão presentes neste segundo esquema aparecem também no primeiro caso[5].

[4] A definição da curtose é expressa em (3.38).

[5] É importante frisar que a análise de estabilidade discutida, embora indique que os pontos

Geralmente, nas aplicações de BSS, as fontes envolvidas pertencem a uma mesma classe com relação à gaussianidade. Por exemplo, em separação de vozes é comum adotar modelos super-gaussianos para todas as fontes. Todavia, caso existam fontes de ambas as classes de modo que o número exato de cada uma delas seja desconhecido, já não é mais possível abordar o problema a partir das técnicas baseadas no critério Infomax, dado que, em sua versão original, a definição das funções de ativação, consideradas fixas durante todo o processo, não leva em conta as características das fontes. Na próxima seção, apresentaremos uma técnica capaz de superar esta dificuldade: o algoritmo Infomax estendido.

O algoritmo Infomax estendido

O motivo pelo qual o algoritmo Infomax estendido (T. W. Lee, Girolami & Sejnowski, 1999) é capaz de separar fontes de classes distintas é que, além da matriz $\mathbf{W}$, as funções de ativação da rede neural apresentada na Figura 3.1 também são ajustadas. No caso, esta adaptação fica restrita somente a duas possibilidades, pois, diante da discussão apresentada na seção anterior, é possível separar misturas de fontes sub-gaussianas e super-gaussianas utilizando somente duas classes de funções.

Na proposta original desta técnica, as seguintes funções de ativação são utilizadas: $g(y) = y - \tanh(y)$, para fontes sub-gaussianas, e $g(y) = y + \tanh(y)$, para fontes super-gaussianas. Deste modo, a atualização da matriz $\mathbf{W}$ (considerando um ajuste baseado no gradiente natural) é expressa do seguinte modo

$$\mathbf{W} \leftarrow \mathbf{W} + \mu(\mathbf{I} - E\{\mathbf{M}\tanh(\mathbf{y})\mathbf{y}^T - \mathbf{y}\mathbf{y}\} \tag{3.35}$$

onde $\mathbf{M} = diag(m_1, \ldots, m_n)$ é uma matriz diagonal tal que, idealmente, $m_i = 1$ para fontes super-gaussianas e $m_i = -1$ para fontes sub-gaussianas.

Assumindo que não há nenhum tipo de informação sobre as fontes, torna-se imperativo a adoção de uma estratégia para a determinação dos parâmetros

relacionados a matrizes de separação podem ser localmente estáveis (dependendo das funções não-lineares), não garante que tais pontos sejam os únicos estáveis. De fato, não há qualquer prova, para um caso geral, sobre a existência, ou inexistência, de soluções espúrias. Não obstante, a idéia de que é sempre possível recuperar as fontes com funções não-lineares aproximadas está amplamente difundida na comunidade, como pode ser visto em (Liu, Chiu & Xu, 2004).

m_i. Para tal fim, emprega-se a condição suficiente de estabilidade proveniente da expressão (3.33), isto é, $\upsilon_i > 0$. Neste caso em particular, este momento é dado por

$$\upsilon_i = m_i(E\{\text{sech}^2(y_i)\}E\{y_i^2\} - E\{\tanh(y_i)y_i\}) > 0 \quad i = 1, \ldots, n. \tag{3.36}$$

Assim, para garantir a positividade desta expressão é necessário que m_i tenha o mesmo sinal que $(E\{\text{sech}^2(y_i)\}E\{y_i^2\} - E\{\tanh(y_i)y_i\})$ e, portanto, a adaptação de m_i pode ser conduzida do seguinte modo

$$m_i = \text{sign}(E\{\text{sech}^2(y_i)\}E\{y_i^2\} - E\{\tanh(y_i)y_i\}), \tag{3.37}$$

onde sign$(\cdot)$ representa a função sinal.

As expressões (3.35) e (3.37) caracterizam o algoritmo Infomax estendido. Em virtude da existência do cálculo da expressão (3.37), ausente na proposta original do algoritmo BS, há, no Infomax estendido, um aumento da complexidade computacional.

3.2.5 Maximização da não-gaussianidade

Uma classe de critérios em BSS tem como idéia principal o ajuste do sistema separador de modo que as densidades de probabilidades de cada uma das estimativas sejam, em algum sentido, as mais distantes possível de uma variável gaussiana, num procedimento usualmente denominado de maximização da não-gaussianidade misturas. Um dos principais atrativos presentes nessa estratégia é a possibilidade de recuperar cada uma das fontes individualmente. Além disso, um dos algoritmos mais conhecidos em BSS, o FastICA, foi inicialmente desenvolvido com base neste critério. Abordaremos, na seqüência, os principais tópicos relacionados a esta estratégia, assim como o algoritmo FastICA.

A chave para entendermos a idéia presente na abordagem via maximização da não-gaussianidade está associada ao teorema central do limite (Papoulis, 1993), que, em linhas gerais, estabelece que o resultado da soma de um conjunto de variáveis aleatórias resulta em uma variável mais próxima de uma gaussiana do que qualquer uma pertencente a esse conjunto. Portanto, sob a ótica deste teorema, é de se esperar que as misturas sejam mais próximas a gaussianas se comparadas com as

fontes. Logo, uma tentativa de ajustar $\mathbf{W}$ fundamenta-se justamente na recuperação da não-gaussianidade das estimativas das fontes. Obviamente, essa justificativa é apenas ilustrativa. No entanto, como veremos mais adiante, alguns resultados indicam que essa abordagem possui relações com a ICA.

Uma das técnicas originadas da abordagem em questão está diretamente ligada ao conceito de curtose, ou cumulante de quarta ordem, cuja definição para uma variável aleatória de média zero, x, é dada por

$$\kappa_4 = E\{x^4\} - 3[E\{x^2\}]^2. \tag{3.38}$$

Uma interessante propriedade da curtose é que esta medida é não-nula para a grande maioria das variáveis aleatórias, sendo a gaussiana uma das poucas exceções desta regra. Deste modo, um critério de acordo com espírito da maximização não-gaussianidade pode ser obtido através da maximização do valor absoluto desta grandeza para cada uma das estimativas individualmente, caso se queira determinar apenas algumas fontes, ou conjuntamente, na situação em que todas as fontes são desejadas. Essa abordagem foi proposta em (Delfosse & Loubaton, 1995), onde também se propôs um esquema, denominado *deflation* (mais adiante veremos os aspectos básicos desta técnica), destinado à extração das fontes de uma maneira serial.

Também é possível entender o critério de maximização da não-gaussianidade em um contexto mais associado à teoria da informação. Um dos principais resultados desta área afirma que, no conjunto de todas as variáveis aleatórias de mesma variância, a variável gaussiana é aquela que possui a maior entropia (Cover & Thomas, 1991). Logo, uma possível estratégia de maximização da não-gaussianidade seria buscar a minimização das entropias marginais das estimativas das fontes.

Em BSS, essa idéia foi primeiramente introduzida a partir do conceito de negentropia, um tipo de entropia normalizada em relação a uma variável gaussiana. A negentropia de uma variável aleatória x é dada por[6]:

$$J(x) = H(x_g) - H(x), \tag{3.39}$$

onde x_g corresponde a uma variável aleatória gaussiana com a mesma variância que x. Em vista do que já discutimos sobre a entropia de uma gaussiana, é evidente

[6] Também é possível definir este conceito para um vetor aleatório.

que a negentropia é uma medida sempre não-negativa, sendo nula somente quando x for uma variável gaussiana. Um ponto favorável à negentropia, no que diz respeito ao seu uso como medida de gaussianidade, é que esta entidade é consideravelmente mais robusta a *outliers* se comparada à abordagem baseada na curtose.

A despeito desta maior robustez, há uma significativa dificuldade de ordem prática na aplicação direta da negentropia ao problema de BSS, relacionada à necessidade de etapas de estimação de entropia, haja visto que a definição (3.39) está diretamente ligada a este conceito. Felizmente, é possível obter uma boa estimativa da negentropia a partir de uma aproximação baseada nos chamados momentos polinomiais (Hyvärinen, 1999a), dada por:

$$J(y) = \alpha(E\{G(y)\} - E\{G(\nu)\})^2, \tag{3.40}$$

onde $G(\cdot)$ é uma função não-linear não-quadrática, α é uma constante e ν é uma variável aleatória gaussiana de média e zero e variância unitária. Como conseqüência da utilização de uma gaussiana normalizada neste caso, é necessário restringir, durante a etapa de adaptação, a potência de cada uma das estimativas das fontes, assumindo, por exemplo, que $E\{y_i\} = E\{\mathbf{w}_i^T\mathbf{x}\} = 1$.

Maximização da não-gaussianidade através do FastICA

Com o intuito de descrever como é feita a maximização da não-gaussianidade através do FastICA, considere a recuperação de uma fonte, isto é, o ajuste de uma das linhas da matriz $\mathbf{W}$, denotada por $\mathbf{w}_i^T$, de modo que $y_i = \mathbf{w}_i^T\mathbf{x}$ resulte numa estimativa satisfatória de uma certa fonte. Além disso, consideremos a maximização da negentropia aproximada, expressa em (3.40). Tal situação é descrita pelo seguinte problema de otimização

$$\tilde{\mathbf{w}}_i = \arg\max_{\mathbf{w}_i}(E\{G(y_i)\} - E\{G(\nu)\})^2, \tag{3.41}$$

com restrição que $E\{y_i\} = E\{\mathbf{w}_i^T\mathbf{x}\} = 1$.

É interessante notar que o máximo da expressão (3.41) é atingido para um certo ótimo de $E\{G(y_i)\}$, pois o termo $E\{G(\nu)\}$ é constante. Assim sendo tal problema de maximização e o de otimizar este primeiro termo são equivalentes. Através do

método de Lagrange, é possível verificar que este último problema de otimização é resolvido quando a seguinte condição é satisfeita

$$E\{\mathbf{x}G'(\mathbf{w}_i^T\mathbf{x})\} + \beta\mathbf{w}_i = 0, \tag{3.42}$$

onde β é uma constante que depende do valor ótimo $\tilde{\mathbf{w}}_i$.

O algoritmo FastICA, em sua versão destinada à otimização da negentropia, foi derivado com base nessa condição de otimalidade. Levando em conta algumas aproximações resultantes do fato que as misturas já estão branqueadas, aplicou-se o método de Newton para a resolução de (3.42), donde se obteve a seguinte regra de atualização

$$\begin{aligned} \mathbf{w}_i &\leftarrow E\{\mathbf{x}G'(\mathbf{w}_i^T\mathbf{x})\} - E\{G''(\mathbf{w}_i^T\mathbf{x})\}\mathbf{w}_i \\ \mathbf{w}_i &\leftarrow \mathbf{w}_i \;/\; \| \mathbf{w}_i \| \,. \end{aligned} \tag{3.43}$$

Para misturas branqueadas, a normalização feita garante a restrição sobre as variância da estimativa y_i. Com relação à função não-linear, uma boa escolha é dada por $G'(y) = \tanh(y)$ (Hyvärinen, 1999a).

Para extrair N fontes através da regra de ajuste (3.43), é necessário executá-la para N vetores $\mathbf{w}_i$. Todavia, torna-se necessário lançar mão de algum mecanismo que evite que essas execuções convirjam para um mesmo ótimo, o que representaria sempre recuperar a mesma fonte. Já vimos que, como uma conseqüência do branqueamento prévio das misturas, o sistema separador corresponde a uma matriz ortogonal, o que implica que suas linhas são necessariamente ortogonais. Assim sendo, uma estratégia viável para contornar esta dificuldade seria inserir alguma etapa no algoritmo que garanta a ortogonalidade dos vetores obtidos.

Uma maneira de fazer isso é aplicar o método de ortogonalização de Gram-Schimdt a cada execução. Em essência, este procedimento funciona do seguinte modo: feita a extração da primeira fonte, a recuperação da segunda é conduzida aplicando a regra de ajuste do FastICA, porém, a cada iteração, retira-se do vetor em processo de estimação a contribuição do vetor referente à primeira fonte, de modo que esses dois vetores sejam ortogonais. Para a extração de uma terceira fonte, deve-se, a cada iteração, retirar a contribuição dos dois vetores estimados, e assim sucessivamente. Este procedimento recebe o nome de *deflation*.

Um problema do procedimento acima descrito é que, caso haja algum erro na extração de uma dada fonte, este erro será acumulado na determinação das fontes

Tabela 3.1: Algoritmo FastICA com ortogonalização simétrica.

1. Centralizar e branquear dos dados.
2. Definir os valores iniciais de $\mathbf{w}_i$ (colunas of $\mathbf{W}$). Ortogonalizar $\mathbf{W}$ de acordo com o passo 4.
3. Para todo i, executar a regra de ajuste (3.43).
4. Realizar a ortogonalização simétrica de $\mathbf{W}$, que pode ser feita do seguinte modo
$$\mathbf{W} \leftarrow \left(\mathbf{W}\mathbf{W}^T\right)^{-1/2}\mathbf{W}.$$
5. Caso não convirja, voltar ao passo 3.

seguintes. Uma alternativa a tal empecilho é realizar em paralelo todas as execuções, e, a cada iteração finalizada, ortogonalizar a matriz $\mathbf{W}$ formada pelos vetores obtidos em cada iteração. Este procedimento é conhecido como ortogonalização simétrica, dado que nenhuma direção do espaço de busca é privilegiada neste caso. Uma maneira de realizar este procedimento de ortogonalização é apresentada na Tabela 3.1, mais especificamente no passo 4. Tal tabela apresenta a descrição do FastICA em conjunto com o procedimento de ortogonalização simétrica.

3.2.6 PCA não-linear

Já vimos, na Seção 2.5.3, que a recuperação das fontes não pode ser conduzida levando em conta somente as estatísticas de ordem dois, o que inviabiliza a aplicação da PCA a tal problema. Por outro lado, vimos também, na Seção 3.2.1, que uma maneira de introduzir estatísticas de ordem superior no problema de separação se dá através do emprego de funções não-lineares em critérios de separação que levam em conta a correlação entre as estimativas das fontes. As motivações por trás de uma outra abordagem de grande destaque em BSS, a PCA não-linear (NPCA, *Nonlinear Principal Component Analysis*)) (Hyvärinen, Karhunen & Oja, 2001),

estão relacionadas a essas duas questões. Ao introduzir elementos não-lineares na definição da PCA, tal estratégia contorna as limitações inerentes à aplicação desta clássica técnica ao problema de BSS. A seguir, discutiremos como isso é feito.

Levando em conta a discussão sobre a formulação da PCA como um problema de minimização do erro quadrático médio, apresentada na Seção 2.5.3, a aplicação desta técnica na estimação da matriz $\mathbf{W}$ poderia ser expressa pelo seguinte problema de otimização

$$\widetilde{\mathbf{W}} = \arg\min_{\mathbf{W}} E\{\|\mathbf{x} - \sum_{i=1}^{N}(\mathbf{w}_i^T\mathbf{x})\mathbf{w}_i\|^2\}, \tag{3.44}$$

sendo que $\mathbf{w}_i$ corresponde à i-ésima coluna de $\mathbf{W}$, e $\mathbf{w}_i^T\mathbf{w}_j = \delta_{ij}$ (Note que estamos considerando $\mathbf{w}_i$ um vetor coluna). Por outro lado, no caso da NPCA, a busca por $\mathbf{W}$ é conduzida da seguinte maneira

$$\widetilde{\mathbf{W}} = \arg\min_{\mathbf{W}} E\{\|\mathbf{x} - \sum_{i=1}^{N}(g_i(\mathbf{w}_i^T\mathbf{x}))\mathbf{w}_i\|^2\}, \tag{3.45}$$

onde $g_i(\cdot)$ é uma função necessariamente não-linear. Neste caso em específico, a não-linearidade é introduzida na projeção das bases $\mathbf{w}_i$ no conjunto de misturas $\mathbf{x}$, todavia, é importante ressaltar que há outras variantes não-lineares da PCA (Hyvärinen, Karhunen & Oja, 2001).

Em uma notação matricial, a função custo da expressão (3.45) pode ser reescrita da seguinte forma

$$J_{NPCA}(\mathbf{W}) = E\{\|\mathbf{x} - \mathbf{W}^T\mathbf{g}(\mathbf{W}\mathbf{x})\|^2\}, \tag{3.46}$$

onde $\mathbf{g}(\cdot) = [g_1(\cdot)\ldots g_N(\cdot)]$. Na abordagem via NPCA, é comum fazer com que as misturas passem por um estágio de branqueamento, que, como já vimos, pode ser conduzido pela aplicação da PCA. Neste caso, a matriz $\mathbf{W}$ a ser determinada é ortogonal, ou seja, $\mathbf{W}\mathbf{W}^T = \mathbf{I}$. Aplicando esta condição na descrição matricial do problema, e, após um breve desenvolvimento, obtém-se que

$$J_{NPCA}(\mathbf{W}) = \sum_{i=1}^{N} E\{[y_i - g_i(y_i)]^2\}. \tag{3.47}$$

Esta nova função custo apresenta uma similaridade patente com os critérios de Bussgang (Attux et al., 2006), uma classe de funções custo amplamente utilizadas

em equalização cega. Assim, mais uma vez, notamos uma grande similaridade entre este problema e o de separação de fontes.

Na abordagem via NPCA, assim como nas técnicas discutidas anteriormente, uma maneira de resolver o problema de otimização associado se dá através da aplicação do método do gradiente descendente. Aliás, o primeiro algoritmo baseou-se justamente nessa idéia (Hyvärinen, Karhunen & Oja, 2001). Todavia, a formulação de mínimo erro quadrático motivou a aplicação do consagrado algoritmo RLS (Recursive Least Squares) (Karhunen & Pajunen, 1997). De fato, essa estratégia pode ser considerada uma importante vantagem da NPCA perante às outras técnicas de separação. Na seqüência, apresentaremos uma breve descrição de tal abordagem.

Adaptação do NPCA utilizando o algoritmo RLS

Apesar de existir, no contexto da NPCA, estratégias que adaptam a matriz $\mathbf{W}$ em batelada, essa abordagem está preponderantemente associada a técnicas com operação em tempo real, dado que o atrativo deste paradigma reside em sua ligação com o algoritmo RLS, que, por sua vez, está estreitamente ligado a métodos adaptativos.

O algoritmo RLS é utilizado na minimização de um critério de mínimos quadrados, porém, modificado pela introdução de um fator de esquecimento. No caso da NPCA, esta formulação é expressa do seguinte modo

$$J_{NPCA}(\mathbf{W}(k)) = \sum_{k=1}^{K} \beta^{K-k} \parallel \mathbf{x}(k) - \mathbf{W}^T(k)\mathbf{z}(k) \parallel^2, \tag{3.48}$$

onde β corresponde ao fator de esquecimento, $\mathbf{z}(k) = \mathbf{g}(\mathbf{W}(k-1)\mathbf{x}(k))$ e k representa o índice temporal. Note que essa descrição também é adequável à PCA se considerarmos $\mathbf{z}(k) = \mathbf{W}(k-1)\mathbf{x}(k)$.

Em (B. Yang, 1995), um algoritmo do tipo RLS, denominado PAST, foi desenvolvido para a minimização da expressão (3.48) para o caso da PCA. A sua extensão para o caso não-linear foi feita em (Karhunen & Pajunen, 1997). Neste

trabalho, a seguinte regra de ajuste foi proposta

$$\begin{aligned}
&\mathbf{h}(k) = \mathbf{P}(k-1)\mathbf{z}(k) \\
&\mathbf{m}(k) = \mathbf{h}(k)/(\beta + \mathbf{z}^T(k)\mathbf{h}(k)) \\
&\mathbf{P}(k) = \beta^{-1}\Upsilon[\mathbf{P}(k-1) - \mathbf{m}(k)\mathbf{h}^T(k)] \\
&\mathbf{e}(k) = \mathbf{x}(k) - \mathbf{W}^T(k-1)\mathbf{z}(k) \\
&\mathbf{W}(k) = \mathbf{W}(k-1) + \mathbf{m}(k)\mathbf{e}^T(k),
\end{aligned} \tag{3.49}$$

sendo que $\Upsilon[\mathbf{J}]$ gera uma nova matriz cuja parte triangular superior é a mesma da matriz $\mathbf{J}$, e cuja parte triangular inferior é obtida pela transposição da parte superior. Em (Karhunen & Pajunen, 1997), argumenta-se que uma boa opção para valores iniciais $\mathbf{W}(0)$ e $\mathbf{P}(0)$ é atribuir matrizes unitárias.

Na NPCA, assim como no caso do critério de maximização da não-gaussianidade, é possível extrair as fontes de maneira seqüencial. Essa é uma conseqüência direta do modo de operação da PCA, que é baseada na extração individual de cada uma das componentes principais. Nesse caso, a recuperação de fontes distintas pode feita através da aplicação do algoritmo RLS em conjunto como uma técnica do tipo *deflation* (Karhunen & Pajunen, 1997), por exemplo.

3.2.7 Métodos algébricos - JADE

Na Seção 2.5.3, vimos que as informações de segunda ordem podem ser levadas em conta no problema de separação por meio de uma diagonalização da matriz de correlação (ou covariância) relativa ao vetor de misturas $\mathbf{x}$. Nesta mesma linha, uma outra abordagem em BSS incorpora as informações de ordem superior ao problema a partir de processos de diagonalização de entidades que contenham tais informações, como, por exemplo, o tensor de cumulantes e a matriz de cumulantes associada a um vetor aleatório. Esse procedimento é a essência de uma pioneira técnica em BSS denominada JADE (*Joint Approximate Diagonalization of Eigenmatrices*) (Cardoso & Souloumiac, 1993). Antes de discutirmos as idéias centrais desta técnica, relembraremos a definição de um cumulante e também de um importante conceito relacionado a esta entidade.

Dado o conjunto de variávies aleatórias $z_1, z_2, \ldots, z_p$, um cumulante cruzado de

ordem r deste conjunto é definido como (Nikias & Petropulu, 1993)

$$cum(z_1^{k_1}, z_2^{k_2}, \ldots, z_p^{k_p}) \triangleq (-j)^r \frac{\partial^r \Psi(\omega_1 \ldots, \omega_p)}{\partial w_1^{k_1} \ldots \partial w_p^{k_p}}\bigg|_{\omega_1=\ldots=\omega_p=0}, \tag{3.50}$$

onde $\Psi(\omega_1 \ldots, \omega_p)$ representa o logaritmo da função característica conjunta e $k_1 + k_2 + \ldots + k_p = r$. No JADE, explora-se um cumulante de quarta ordem, dado por:

$$\begin{aligned} cum(z_1, z_2, \ldots, z_p) = E\{z_1 z_2 z_3 z_4\} - E\{z_1 z_2\}E\{z_3 z_4\} \\ - E\{z_1 z_3\}E\{z_2 z_4\} - E\{z_1 z_4\}E\{z_2 z_3\}. \end{aligned} \tag{3.51}$$

Note que esta medida, diferentemente da clássica medida de correlação de segunda ordem, apresenta quatro argumentos. Deste modo, a definição de uma estrutura semelhante à matriz de correlação necessita do conceito de tensor, que, em linhas gerais, pode ser entendido como uma extrapolação do conceito de matriz, no sentido de que um tensor pode apresentar um número de entradas maior do que dois. Neste contexto, define-se o tensor de cumulante como um conjunto de elementos, sendo que o elemento $ijkl$ é dado por $cum(z_i, z_j, z_k, z_l)$, e com i, j, k, l variando de 1 a p.

Um outro conceito que utilizaremos é o de matriz de cumulante associada a um vetor aleatório $\mathbf{x} = [x_1, \ldots, x_N]$ e a uma matriz $\mathbf{M}$ de dimensão $N \times N$. No caso, o elemento ij da matriz de cumulante é dado por

$$[Q^{\mathbf{x}}(\mathbf{M})]_{ij} = \sum_{k,l=1}^{N} Cum(x_i, x_j, x_k, x_l) M_{kl}, \tag{3.52}$$

sendo que os índices i e j variam de 1 a N. A matriz de cumulante $Q^{\mathbf{x}}(\mathbf{M})$ pode ser entendida como o resultado da aplicação do tensor de cumulante de $\mathbf{x}$ à matriz $\mathbf{M}$, ou seja, $Q^{\mathbf{x}}(\mathbf{M}) = \Gamma(\mathbf{M})$, onde $\Gamma(\cdot)$ representa a transformação efetuada por um tensor.

Vejamos agora como essas medidas são consideradas no problema de separação. No caso em que o vetor $\mathbf{x}$ representa os sinais misturados após um etapa de branqueamento, ou seja, $\mathbf{x} = \mathbf{U}\mathbf{s}$, sendo $\mathbf{U}$ uma matriz ortogonal, é possível mostrar (Cardoso, 1999) que a matriz de cumulantes associada a $\mathbf{x}$ é dada por

$$Q^{\mathbf{x}}(\mathbf{M}) = \mathbf{U} \triangle(\mathbf{M}) \mathbf{U}^T, \tag{3.53}$$

onde $\triangle(\mathbf{M})$ é uma matriz diagonal cujos parâmetros dependem de $\mathbf{M}$ e das curtoses das fontes.

Multiplicando ambos os lados da expressão (3.53) por $\mathbf{U^T}$ à esquerda e por $\mathbf{U}$ à direita, e lembrando que $\mathbf{U}^T = \mathbf{U}^{-1}$ (devido à ortogonalidade de $\mathbf{U}$), temos que:

$$\mathbf{U}^T Q^{\mathbf{x}}(\mathbf{M})\mathbf{U} = \mathbf{U}^T\mathbf{U}\triangle(\mathbf{M})\mathbf{U}^T\mathbf{U} = \triangle(\mathbf{M}), \tag{3.54}$$

donde podemos concluir que a matriz $\mathbf{U}$ diagonaliza $Q^{\mathbf{x}}(\mathbf{M})$. Logo, uma estratégia plausível para a recuperação das fontes se daria a partir da diagonalização de $Q^{\mathbf{x}}(\mathbf{M})$.

Em (Cardoso & Souloumiac, 1993), argumenta-se que a diagonalização da matriz de cumulante (3.53), para uma matriz $\mathbf{M}$ arbitrária, garante a identificação de $\mathbf{U}$ desde que haja, no máximo, uma fonte com curtose nula (novamente a restrição sobre as fontes gaussianas) e que os autovalores de $Q^{\mathbf{x}}(\mathbf{M})$ sejam distintos. Infelizmente, os autovalores desta matriz dependem, além da matriz $\mathbf{M}$, de $\mathbf{U}$, que não conhecemos *a priori*. Ou seja, não é possível estabelecer qualquer tipo de garantia sobre esses autovalores. Além disso, mesmo quando a matriz $\mathbf{M}$ garante a existência de autovalores distintos, a diagonalização exata da matriz de cumulante pode ser inatingível em um cenário prático, posto que os cumulantes são obtidos a partir de estatísticas amostrais. A idéia fundamental presente no algoritmo JADE tem justamente como objetivo evitar esses problemas.

Nessa técnica, em vez de se realizar a diagonalização considerando apenas uma matriz de cumulante, adota-se um esquema de diagonalização conjunta de diferentes matrizes de cumulante, sendo cada uma delas definida por uma matriz $\mathbf{M}_i$. Matematicamente, a função custo a ser minimizada no algoritmo JADE é dada por

$$D(\mathbf{U}) = \sum_i \Omega(\mathbf{U}^T Q^{\mathbf{x}}(\mathbf{M_i})\mathbf{U}), \tag{3.55}$$

onde o operador $\Omega(\cdot)$ expressa a soma quadrática dos elementos que não estão na diagonal principal. No que diz respeito à escolha das matrizes $\mathbf{M}_i$, adotam-se as automatrizes relativas ao tensor de cumulante, ou seja, as N matrizes $\mathbf{M}_i$ tal que $Q^{\mathbf{x}}(\mathbf{M}_i) = \lambda\mathbf{M}_i$. Ao proceder desta forma, todas as informações relevantes do tensor de cumulante são, de certo modo, consideradas.

Para otimizar a expressão (3.55) é possível utilizar o método de Jacobi para diagonalização de matrizes. Inicialmente concebido para a diagonalização de uma

única matriz (Golub & Loan, 1989), este método foi estendido para prover a diagonalização conjunta de diversas matrizes em (Cardoso & Souloumiac, 1996). A idéia essencial presente nesta abordagem é representar a matriz $\mathbf{U}$ através de um produto de matrizes de rotação, dado por

$$\mathbf{U} = \sum_{i,j=1,i\neq j}^{N} \mathbf{R}_{ij}, \tag{3.56}$$

onde a matriz de rotação $\mathbf{R}_{ij}$, de dimensões $N \times N$, é idêntica a uma matriz identidade, com a exceção de seus elementos nas posições (i,i), (i,j), (j,i) e (j,j), dados por

$$r_{ii} = \cos\theta,\ r_{ij} = \sin\theta,\ r_{ji} = -\sin\theta,\ r_{jj} = \cos\theta. \tag{3.57}$$

O parâmetro θ é o ângulo de rotação associado à matriz $\mathbf{R}_{ij}$.

É possível, para cada uma das matrizes de rotação $\mathbf{R}_{ij}$, determinar analiticamente qual é o ângulo de rotação que minimiza a expressão (3.55), conforme descrito em (Cardoso & Souloumiac, 1996). Assim, a diagonalização conjunta das matrizes de cumulante pode ser feita a partir da obtenção, através de expressões analíticas, de cada um dos ângulos de rotação para todos os possíveis pares ij. Tal abordagem pode ser compreendida como uma possível iteração do algoritmo JADE, de modo que a condição de parada ocorra quando o ângulo de rotação obtido for menor que uma ângulo mínimo $\theta_{\min}$ previamente definido. A despeito do reduzido número de iterações necessárias para a convergência deste procedimento de diagonalização conjunta, tal técnica se torna consideravelmente ineficiente em cenários com um elevado número de fontes, devido ao aumento de complexidade presente na determinação analítica dos ângulos de rotação.

Uma dos principais atrativos presentes no JADE resulta do fato de que, tanto o desenvolvimento da função contraste (3.55), quanto o do processo de diagonalização conjunta são válidos para vetores aleatórios complexos (Cardoso & Souloumiac, 1993), ou seja, o JADE pode ser utilizado também na separação de fontes complexas.

3.2.8 EASI

Até agora, dedicamos cada seção para apresentar abordagens em BSS cuja inovação diz respeito, sobretudo, à introdução de critérios capazes de guiar o

processo de separação. Nesta seção, fugiremos um pouco dessa linha para apresentar uma técnica, denominada EASI (*Equivariant Adaptive Source Separation*), cujo desenvolvimento está preponderantemente relacionado ao processo de otimização do sistema separador.

A principal motivação presente no EASI é o desenvolvimento de uma técnica com operação em tempo real que apresente a chamada propriedade de desempenho uniforme, que, em linhas gerais, estabelece que o processo de estimação depende exclusivamente das características das fontes. Neste caso, para cenários com o mesmo conjunto de fontes, mudar a matriz de separação implica simplesmente em alterar a condição inicial do processo de adaptação (Cardoso & Laheld, 1996).

Com o intuito de alcançar a propriedade de desempenho uniforme, propôs-se um esquema de adaptação serial para a matriz $\mathbf{W}$. Nesta estratégia, pressupõe-se que a regra de adaptação possui a seguinte forma: $\mathbf{W} \leftarrow (\mathbf{I} - \lambda\mathbf{H}(\mathbf{y}))\mathbf{W}$, onde $\mathbf{H}(\cdot)$ mapeia um vetor em uma matriz e λ corresponde ao passo de adaptação.

Para determinar qual é a melhor escolha de $\mathbf{H}(\cdot)$, é interessante, primeiramente, que relembremos o significado da otimização via gradiente de uma dada função custo $\phi(\mathbf{W})$. Em essência, essa técnica de otimização busca por um incremento Γ tal que a função $\phi(\mathbf{W}+\Gamma)$ apresente um valor maior ou menor, caso se trate de uma minimização, levando em conta somente informações locais de primeira ordem. Isto é feito do seguinte modo: determina-se a expansão em série de Taylor (primeira ordem) da função custo em torno de $\mathbf{W}$, dada por

$$\phi(\mathbf{W}+\Gamma) \approx \phi(\mathbf{W}) + \tau(\frac{\partial\phi(\mathbf{W})}{\partial\mathbf{W}}^{T}\Gamma), \tag{3.58}$$

onde $\tau(\cdot)$ representa o traço de uma matriz. Esta expressão representa uma aproximação da função custo através de um hiperplano, e, portanto, é natural que se busque adaptar $\mathbf{W}$ na direção relativa à inclinação deste hiperplano, dada por $\frac{\partial\phi(\mathbf{W})}{\partial\mathbf{W}}$, que corresponde justamente ao gradiente da função custo em relação à matriz $\mathbf{W}$.

Este procedimento é igualmente aplicável ao paradigma de adaptação serial. No entanto, a forma imposta para esse tipo de regra de ajuste sugere uma modificação no conceito do gradiente, posto que, neste caso, busca-se um incremento Γ de tal forma que $\phi(\mathbf{W}+\Gamma\mathbf{W})$ seja maximizado ou minimizado. Deste modo, seguindo o mesmo

raciocínio apresentado no parágrafo anterior, é natural que este "novo gradiente" corresponda ao coeficiente linear relativo à seguinte expansão de primeira ordem

$$\phi(\mathbf{W}+\Gamma\mathbf{W}) \approx \phi(\mathbf{W}) + \tau(\frac{\partial_R \phi(\mathbf{W})}{\partial \mathbf{W}}^T \Gamma). \tag{3.59}$$

Desta expressão surge o conceito de gradiente relativo, dado por $\frac{\partial_R \phi(\mathbf{W})}{\partial \mathbf{W}}$. No contexto descrito, o gradiente relativo relaciona-se com o gradiente tradicional da seguinte maneira: $\frac{\partial_R \phi(\mathbf{W})}{\partial \mathbf{W}} = \frac{\partial \phi(\mathbf{W})}{\partial \mathbf{W}}\mathbf{W}^T$.

Para uma função custo dada por $\phi(\mathbf{W}) = E\{f(\mathbf{y})\}$, é possível mostrar (Cardoso & Laheld, 1996) que o gradiente relativo é dado por

$$\frac{\partial_R \phi(\mathbf{W})}{\partial \mathbf{W}} = E\{\mathbf{f}'(\mathbf{y})\mathbf{y}^T\}, \tag{3.60}$$

onde $\mathbf{f}(y) = [f'(y_1) \dots f'(y_N)]$. Deste modo, obtém-se a seguinte regra de adaptação serial que minimiza $\phi(\mathbf{W})$

$$\mathbf{W} \leftarrow \mathbf{W} - \lambda E\{\mathbf{f}'(\mathbf{y})\mathbf{y}^T\}\mathbf{W}. \tag{3.61}$$

Obtém-se a versão estocástica dessa regra de ajuste através da supressão do operador de esperança.

Apesar da expressão (3.60) corresponder a uma solução adequada ao problema de separação, haja visto, por exemplo, sua semelhança com o algoritmo ACY, há ainda uma dificuldade relacionada com o contraste $\phi(\mathbf{W}) = E\{f(\mathbf{y})\}$. Em alguns casos, como, por exemplo, em critérios de maximização de não-gaussianidade, deve-se levar em conta no processo de otimização a restrição de que as saídas do sistema separador são necessariamente brancas. Nos algoritmos que operam em batelada, essa restrição é satisfeita realizando o branqueamento prévio das misturas e restringindo a busca de um sistema separador que seja ortogonal, como no caso do FastICA. Entretanto, em algoritmos adaptativos, a execução dessas etapas de branquemento e ortogonalização pode constituir um significativo empecilho prático.

Uma maneira de contornar este problema é incorporar tal restrição diretamente na regra de atualização (3.61). Isto pode ser feito, de um modo analítico, considerando a projeção do gradiente relativo no espaço de matrizes ortogonais (Cardoso & Laheld, 1996), o que resulta na seguinte regra de atualização

$$\mathbf{W} \leftarrow \mathbf{W} - \lambda E\{\mathbf{y}\mathbf{y}^T - \mathbf{I} + \mathbf{f}'(\mathbf{y})\mathbf{y}^T - \mathbf{y}\mathbf{f}'(\mathbf{y})^T\}\mathbf{W}. \tag{3.62}$$

Esta é a regra de adaptação presente no algoritmo EASI. Note que esta expressão também é serial.

Na expressão (3.62), a escolha das funções não-lineares segue essencialmente o mesmo princípio presente no critério Infomax, e, assim sendo, também é necessário, no algoritmo o EASI, o conhecimento da gaussianidade das fontes. No caso em que todas as fontes são sub-gaussianas, é possível separá-las definindo $f(\mathbf{y}) = \sum_{i=1}^{N} y_i^4$, o que implica num vetor $\mathbf{f}$ cujo i-ésimo elemento é dado por $f_i' = 4y_i^3$. Veremos na próxima seção que esta função custo está diretamente ligada à minimização da informação mútua das estimativas das fontes, bem como à maximização da não-gaussianidade.

3.3 Sobre os Critérios de Separação

No decorrer da apresentação das principais estratégias em BSS linear, na Seção 3.2, mencionamos a existência de relações entre alguns dos critérios apresentados. Vimos, por exemplo, que o critério Infomax possui uma estreita ligação com o estimador de máxima verossimilhança em BSS. Na presente seção, abordaremos outras importantes relações entre esses critérios, as quais nos permitirão compreendê-los à luz de um paradigma unificado, que, no caso, é originário da idéia primordial da ICA: a recuperação da independência.

Após isso, trataremos de uma outra importante questão, relacionada aos critérios presentes em BSS linear, que remete à eventual presença de mínimos locais nas funções custos derivadas dessas estratégias.

3.3.1 Relações entre os critérios

Com o objetivo de compreender as diversas relações entre as estratégias apresentadas sob a ótica da ICA, adotaremos em nossa explanação, como critério de referência, a minimização da informação mútua, pois, conforme já discutimos, essa medida corresponde a um tipo de implementação direta da ICA, no sentido de que ela é capaz de discriminar a independência de um vetor aleatório.

Primeiramente, recapitulemos a expressão da informação mútua das saídas do

sistema separador, descrita em (3.11). Tendo em vista que $H(\mathbf{y}) = H(\mathbf{x}) + \log(|\det \mathbf{W}|)$, é possível reescrevê-la do seguinte modo

$$I(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{x}) - \log(|\det \mathbf{W}|). \tag{3.63}$$

No processo de otimização, que é feito com a relação a $\mathbf{W}$, não é necessário levar a entropia conjunta das misturas, $H(\mathbf{x})$, haja visto que este termo não depende dos parâmetros do sistema separador.

Uma outra simplificação da expressão (3.63) ocorre na situação em que se considera a separação de misturas previamente branqueadas. Neste caso, já vimos que as matrizes $\mathbf{W}$ que invertem a ação do processo de mistura residem em um conjunto de matrizes ortogonais ou de rotação. Considerando a simplificação do parágrafo anterior e lembrando que o determinante de uma matriz ortogonal é 1, a recuperação das fontes via minimização da informação mútua é, neste caso em particular, equivalente à minimização da seguinte expressão

$$C(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) \tag{3.64}$$

com respeito à matriz $\mathbf{W}$ e de tal forma que esta busca seja efetuada no espaço de matrizes ortogonais.

Ora, vimos, na Seção 3.2.5, que a minizimização de uma função custo desse tipo está diretamente relacionada ao critério de maximização da não-gaussianidade, especialmente ao processo de maximização das negentropias das saídas. Naquela ocasião, justificamos tal abordagem, de uma maneira mais intuitiva, a partir do teorema central do limite. Agora, fica claro que essa estratégia está, de fato, ligada à ICA. Além disso, diante da relação apresentada, fica claro o porquê da necessidade, na maximização da não-gaussianidade, da etapa de branqueamento e, conseqüentemente, da ortogonalização da matriz $\mathbf{W}$ a cada iteração.

O critério baseado na maximização das curtoses (valor absoluto) das estimativas também possui uma relação com o critério de minimização da informação mútua, o que reforça a ligação entre esta última abordagem e a maximização da não-gaussianidade. No caso, utilizando aproximações da entropia marginal baseadas em

expansões em séries, e considerando misturas previamente branqueadas, é possível obter (T.-W. Lee, Girolami, Bell & Sejnowski, 2000) a seguinte aproximação para a informação mútua

$$I(\mathbf{y}) \approx J(\mathbf{x}) - \frac{1}{48}\sum_{i=1}^{N}\kappa_4^2(i) \tag{3.65}$$

onde $J(\mathbf{x})$ e $\kappa_4(i)$ correspondem à negentropia conjunta das misturas branqueadas e à curtose da i-ésima fonte, respectivamente. Tendo em vista que o primeiro termo desta expressão não depende de $\mathbf{W}$, é evidente que a sua minimização equivale à maximização das curtoses das estimativas dos sinais fontes. Vale lembrar que esta aproximação é válida para matrizes de mistura ortogonais, o que, novamente, indica a necessidade de estágios de braqueamento na abordagem de maximização da não-gaussianidade.

Além da maximização da não-gaussianidade, o paradigma de máxima verossimilhança em BSS também está ligado à minimização da informação mútua. Para entendermos esta relação, reescrevamos a expressão (3.63), porém desconsiderando o termo de entropia conjunta, que não depende de $\mathbf{W}$, e expressando a entropia a partir de uma esperança, ou seja

$$C(\mathbf{y}) = \sum_{i=1}^{N} E\{\log(p_{y_i}(y_i))\} - \log(|\det \mathbf{W}|). \tag{3.66}$$

A observação desta expressão em conjunto com (3.20) revela a relação entre os critérios em questão. Apesar de as densidades de probabilidade presentes em ambas as expressões serem distintas, na situação em que as fontes são, de fato, recuperadas, as duas funções custos são equivalentes, posto que as densidades das fontes e de suas estimativas são idênticas nessa situação.

Levando em conta a relação de equivalência entre o estimador de máxima verossimilhança e o critério Infomax, é de se esperar que este último critério também esteja ligado à minimização da informação mútua. De fato, podemos compreendê-lo como uma aproximação da informação mútua, no sentido de que as densidades de probabilidade das estimativas são aproximadas por funções não-lineares neste caso.

Por fim, analisemos as relações entre o NPCA e outras propostas em BSS (em (Karhunen, Pajunen & Oja, 1998), é apresentada uma discussão detalhada

destas relações). Para tal, consideremos o critério de NPCA descrito em (3.47). No caso em que as funções $g_i(\cdot)$ são escolhidas do seguinte modo

$$g_i(y_i) = \begin{cases} y_i^2 + y_i & \text{se} \quad y \geq 0 \\ -y_i^2 + y_i & \text{se} \quad y < 0, \end{cases} \tag{3.67}$$

esse critério equivale a seguinte expressão

$$J_{NPCA}(\mathbf{W}) = \sum_{i=1}^{N} E\{y_i^4\}. \tag{3.68}$$

Assim, este simples desenvolvimento indica que há ligações entre o critério NPCA e a minimização dos momentos de quarta ordem das estimativas das fontes.

Lembrando que, na situação em que os dados são previamente branqueados e a matriz $\mathbf{W}$ é ortogonal, a minimização de um momento central de ordem superior é equivalente à minimização da curtose (ver expressão (3.38)), conclui-se que a NPCA está relacionada à minimização das curtoses das estimativas das fontes. Assim, considerando que todas as fontes possuem valores negativos de curtose (fontes sub-gaussianas), a NPCA é semelhante à abordagem de maximização da não-gaussianidade, e, conseqüentemente, à minimização da informação mútua.

A breve apresentação realizada na presente seção evidencia as relações entre os principais critérios em BSS. Vimos que essas abordagens relacionam-se, seja de modo direto ou indireto, à minimização da informação mútua entre as estimativas fontes, que, conforme já discutido, pode ser entendida como a implementação direta da ICA. Assim sendo, é possível compreender tais abordagens como técnicas que realizam a ICA, posto que o objetivo presente em cada uma delas está implicitamente ligado à recuperação da independência estatísticas das estimativas das fontes através do ajuste da matriz $\mathbf{W}$. Embora o objetivo de tais técnicas seja o mesmo, é importante frisar que o desenvolvimento desses diversos critérios permitiu, como visto no decorrer da Seção 3.2, a elaboração de algoritmos com modos de operação distintos.

3.3.2 Presença de pontos ótimos locais

Na seção anterior, vimos que a etapa de treinamento nas principais técnicas de separação linear se baseia em algoritmos de busca local, tais como o gradiente e

o método de Newton. Em superfícies de custo multimodais, que contam com a presença de ótimos locais, a aplicação deste tipo de solução se tornaria inviável, dado que haveria um novo risco relativo à convergência do algoritmo para esses pontos locais que não necessariamente corresponderiam a uma solução satisfatória do problema de separação. Diante disso, é razoável afirmar que um assunto de considerável importância em BSS, embora ainda pouco tratado na literatura, diz respeito ao estudo das funções custo associadas aos principais critérios de separação. Nesta seção, faremos um breve de sumário de algumas contribuições neste assunto.

Recentemente, uma série de trabalhos (Vrins & Verleysen, 2005b, 2005a; D. Pham & Vrins, 2005) estudou a presença de mínimos locais no critério de minimização da informação mútua entre as estimativas das fontes. Dada a dificuldade envolvida em tal análise, considerou-se um caso em que apenas duas fontes são misturadas.

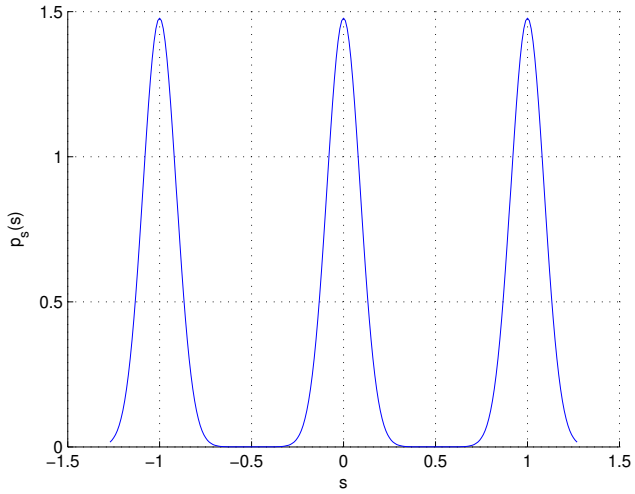
Uma outra hipótese assumida nesses trabalhos foi a ortogonalidade da matriz de mistura. Além disso, assumiu-se um sistema separador também ortogonal, o que permite descrever a relação entre as fontes e suas estimativas através da seguinte parametrização

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}. \tag{3.69}$$

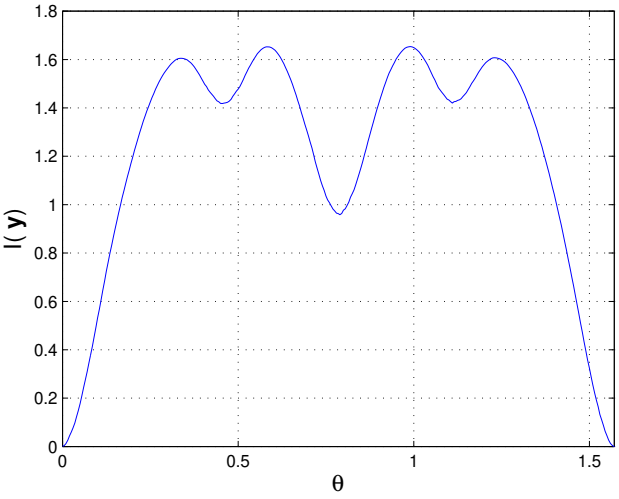
Este cenário busca representar o caso em que uma etapa de branqueamento é aplicada sobre as misturas, pois, como já vimos, a relação entre as misturas branqueadas e as fontes é dada por uma matriz de rotação. Porém, ficará claro mais adiante que este modelo é um pouco mais restrito, no sentido de que o sistema separador também é considerado ortogonal.

No cenário descrito, a existência de mínimos de locais depende das densidades de probabilidade das fontes. Em (Vrins & Verleysen, 2005b), demonstrou-se que a multimodalidade de funções custo como a informação mútua está diretamente ligada à multimodalidade das densidades de probabilidade das fontes. Por exemplo, considerando fontes gaussianas tri-modais (Figura 3.2(a)), a informação mútua em função do ângulo θ é apresentada na Figura 3.2(b). Note que, como esperado, esta medida se anula nos ângulos 0 e $\pi/2$ (recuperação sem, e com, inversão de fase, respectivamente), porém, há três pontos de mínimos locais, sendo um deles, o

ângulo $\pi/4$, um ponto estável.



(a) Densidade de probabilidade das fontes.



(b) Informação mútua em função do ângulo de rotação θ.

Figura 3.2: Presença de mínimos locais em cenários com fontes multimodais.

Em um primeiro momento, a existência de mínimos locais em cenários desse tipo aponta para uma dificuldade considerável na aplicação das técnicas anteriormente estudadas à separação de fontes multimodais. No entanto, um ponto interessante

a ser destacado é que a presença desses mínimos ocorre somente quando o sistema separador é restrito a uma matriz ortogonal. Por exemplo, vimos que, por assumir uma etapa de branqueamento anterior à separação, o FastICA busca soluções no espaço de matrizes ortogonais. No caso, isto é feito através da normalização das linhas da matriz de separação e da aplicação de métodos de ortogonalização a cada iteração, e não por uma restrição estrutural de ortogonalidade. Assim, é importante ressaltar que este caso já não é mais equivalente ao cenário descrito pelo modelo (3.69). Justamente por manter a ortogonalidade de $\mathbf{W}$ apenas de "forma aproximada", o FastICA é capaz de evitar a convergência para mínimos locais. Em suma, enfatizamos que o estudo de (Vrins & Verleysen, 2005b) é válido para o caso restrito em que todas as matrizes posicionadas entre as fontes e suas estimativas são ortogonais.

Conforme já discutimos na seção anterior, a minimização da informação mútua para o modelo (3.69) é equivalente à minimização das entropias marginais de cada uma das estimativas das fontes. Logo, no cenário descrito, o estudo realizado em (Vrins & Verleysen, 2005b) sobre a existência de mínimos locais é válido também para este último critério, bem como para aquele baseado na maximização da negentropia (D. Pham & Vrins, 2005).

Com relação à presença de mínimos nas situações em que os sistemas misturador e separador não são ortogonais, não encontramos trabalhos que abordam esta questão de um modo analítico. Todavia, há alguns trabalhos que constataram situações de convergência ruim através da execução de simulações. Em (Tichavský, Koldovský & Oja, 2006), por exemplo, verificou-se, após a realização de um significativo número de simulações do algoritmo FastICA, que há convergência para mínimos locais num percentual entre $0,01\%$ e 1% dos casos. Em um outro trabalho (Cardoso, 2000), um exemplo apresentado indica uma situação particular do estimador de máxima verossimilhança em que há a possibilidade de convergência para mínimos locais.

3.4 Análise de Desempenho

Nesta seção, buscamos realizar uma análise comparativa das técnicas de BSS apresentadas até o presente momento. Inicialmente, abordamos os algoritmos que possuem modo de operação em batelada, ao passo que em um segundo momento, consideramos suas as versões adaptativas (ou de tempo real). Antes de darmos início a essas análises, é conveniente explicitar algumas diretrizes adotadas em nosso estudo.

Primeiramente, é necessário estabelecer um índice capaz de quantificar o desempenho das diferentes técnicas de BSS. Dentre as diversas possibilidades, adotamos como medida de desempenho o erro de Amari (Cichocki & Amari, 2002), dado por

$$E(B) = \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}|b_{ij}|}{\max_j |b_{ij}|} - 1\right) + \frac{1}{2n}\sum_{j=1}^{n}\left(\frac{\sum_{i=1}^{n}|b_{ij}|}{\max_i |b_{ij}|} - 1\right). \tag{3.70}$$

Esta métrica assume valores entre 0 e $(n-1)$, sendo que é igual a zero somente quando a matriz $\mathbf{B}$ for igual a uma matriz identidade permutada e/ou escalonada. Desta forma, ao definirmos como argumento desta função a matriz de transferência entre as fontes e suas estimativas $\mathbf{WA}$, obtemos um índice coerente com a idéia da ICA.

Além disso, em nossas simulações, as seguintes fontes foram consideradas:

- Ondas senoidais;
- Ondas quadradas;
- Onda dente-de-serra;
- Sinais aleatórios uniformemente distribuídos.

Note que todas essas fontes são sub-gaussianas.

3.4.1 Algoritmos com operação em batelada

Conforme já mencionamos anteriormente, os algoritmos com modo de operação em batelada caracterizam-se pela utilização de um conjunto de amostras no cálculo da atualização em uma dada iteração. Essa classe de técnicas é indicada para aplicações em cenários estáticos, e naqueles em que não existem restrições severas de memória.

Nossa análise abrangeu os seguintes algoritmos: BS, ACY, FastICA (versão negentropia), JADE e EASI. No caso dos algoritmos BS e ACY, descritos na seção 3.2.4, adotamos a seguinte função não-linear $g(y) = \tanh(y) - y$. No que diz respeito ao FastICA, consideramos a implementação apresentada na Tabela 3.1, sendo que função não-linear utilizada, relativa à expressão (3.43) foi $G'(y) = \tanh(y)$. Por fim, assumimos, no algoritmo EASI, a minimização da função custo $J = E\{\sum_{i=1}^{N} y_i^4\}$, e, como conseqüência, a função não-linear da expressão fica dada por $f_i(y_i) = y_i^3$.

Em nossa análise, buscamos enfatizar importantes questões de ordem prática, tais como o desempenho de uma dada técnica em função do aumento do número de fontes e a sua degradação na presença de ruído . Além disso, consideramos alguns quesitos que não foram levados em conta em outros trabalhos (Hyvärinen, Karhunen & Oja, 2001; Giannakopoulos, Karhunen & Oja, 1998) que também buscam realizar uma análise dos algoritmos do caso linear. Na seqüência, descrevemos cada um desses quesitos.

Sensibilidade ao aumento do número de fontes. O primeiro fator analisado foi a evolução do desempenho dos algoritmos em função do aumento do número de fontes. Nesta análise, consideramos um número de amostras do vetor de misturas igual a 1500, sendo que o número de iterações realizadas em cada algoritmo foi igual a 1000. Os passos de adaptação das técnicas baseadas no método do gradiente (BS, ACY e EASI) foram ajustados preliminarmente de modo a garantir uma convergência satisfatória em todas as simulações executadas.

Os resultados obtidos nesse cenário são apresentados na Figura 3.3, que representa a média de 50 experimentos. É possível observar nesta figura que os algoritmos BS, ACY e FastICA apresentam um desempenho próximo neste quesito,

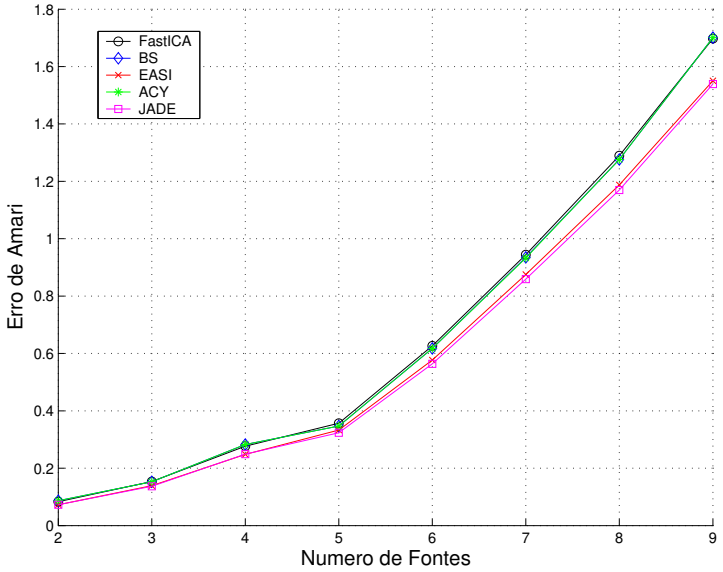


Figura 3.3: Erro de Amari em função do número de fontes.

sendo que os algoritmos EASI e JADE apresentam um ligeiro ganho de desempenho, em relação aos primeiros, a partir de 6 fontes.

Sensibilidade ao número de amostras. Em muitos problemas práticos de BSS, é possível que apenas um reduzido número de amostras das misturas esteja disponível. Nessas situações, é desejável que a técnica de BSS a ser utilizada seja capaz de operar com o mínimo número de amostras. Essa necessidade nos motivou a verificar a influência do número de amostras no desempenho dos algoritmos estudados.

O resultado apresentado na Figura 3.4 foi obtido em um cenário com 5 fontes, sendo que o número de iterações realizadas em cada algoritmo foi igual a 500. É possível notar que os algoritmos em estudo apresentam desempenhos próximos, com a exceção do FastICA que consegue separar as fontes de uma maneira satisfatória somente a partir de aproximadamente 150 amostras, ao passo que nas demais técnicas uma razoável recuperação das fontes pode ser alcançada com 100 amostras.

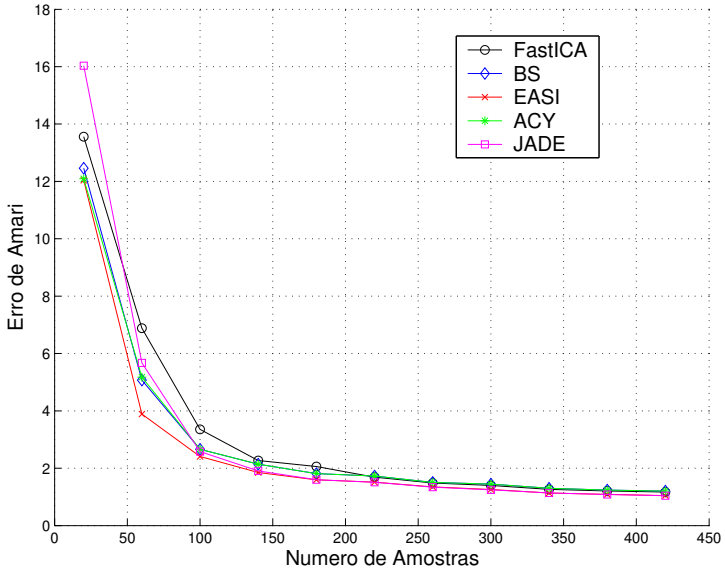


Figura 3.4: Erro de Amari em função do número de amostras.

Sensibilidade à gaussianidade das fontes. Uma das maiores limitações presentes no paradigma de separação de fontes baseado em ICA é a impossibilidade de operação em cenários com fontes gaussianas. Em termos teóricos, a não-gaussianidade das fontes é uma garantia suficiente para a eficácia da ICA. Todavia, em um âmbito prático, como conseqüência do uso de momentos amostrais, a presença de fontes com funções densidade de probabilidade próximas à de uma gaussiana podem dificultar a separação.

Com o objetivo de verificar a sensibilidade das técnicas de BSS à gaussianidade das fontes, considerou-se a separação de fontes caracterizadas pela seguinte função densidade de probabilidade:

$$p(y) = \frac{\alpha}{2\lambda\Gamma(\frac{1}{\alpha})} \exp(-|\frac{y}{\alpha}|^{\alpha}) \tag{3.71}$$

onde $\Gamma(x)$ corresponde à função gama. Essa distribuição, denominada distribuição gaussiana generalizada (DGG) (Cichocki & Amari, 2002), possui média nula e é caracterizada pelo parâmetro α. Quando $\alpha = 2$ tem-se uma distribuição gaussiana com média zero e variância unitária, enquanto que para valores maiores e menores que dois a variável aleatória em questão é sub-gaussiana e super-gaussiana,

respectivamente.

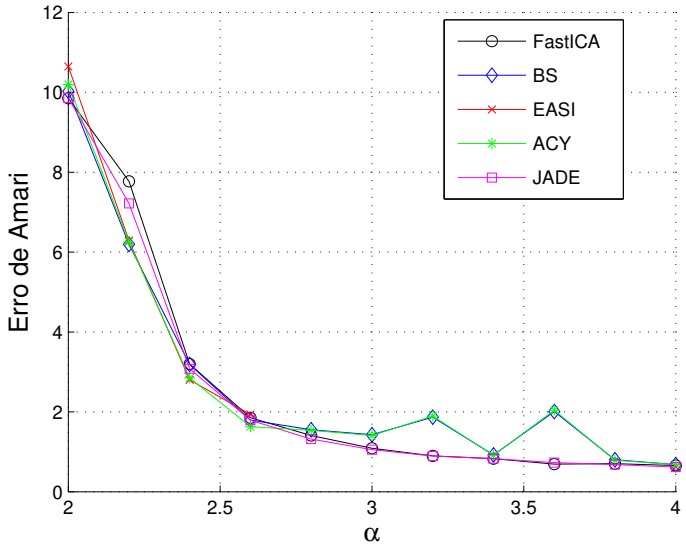


Figura 3.5: Erro de Amari em função de α.

Na Figura 3.5 (resultado médio de 100 experimentos), é possível observar a evolução do erro de Amari em função do parâmetro α em um cenário com 4 fontes (todas com DGG). Nota-se que há, de fato, uma dificuldade inerente às técnicas de BSS à medida que as fontes são próximas a gaussianas. Além disso, constata-se que, neste quesito, as técnicas de BSS estudadas apresentam desempenho semelhante.

Sensibilidade ao ruído. Existem algumas técnicas em BSS desenvolvidas especialmente para aplicações em que as misturas são corrompidas por ruído (Hyvärinen, Karhunen & Oja, 2001). No entanto, tais técnicas requerem um denso processamento computacional, fazendo com que se adote por muitas vezes os métodos clássicos nessas aplicações, ainda que estes sejam projetados para o modelo de mistura livre de ruído. Diante disso, buscamos analisar a deterioração do desempenho dos algoritmos analisados na presença de ruído.

Em um primeiro experimento realizado, consideramos a separação de três fontes (todas uniformes) a partir de misturas corrompidas por um ruído branco, aditivo e gaussiano (AWGN - *Additive White Gaussian Noise*). Na Figura 3.6, que representa

a média de 60 simulações executadas neste cenário, é apresentada a evolução do erro de Amari com o aumento da potência do ruído em cada sensor. Observando essa figura, fica evidente a influência do ruído no processo de separação, sendo que somente a partir de uma relação sinal-ruído (SNR) de aproximadamente 22 dB os algoritmos atingem um desempenho satisfatório, próximo ao caso sem ruído.

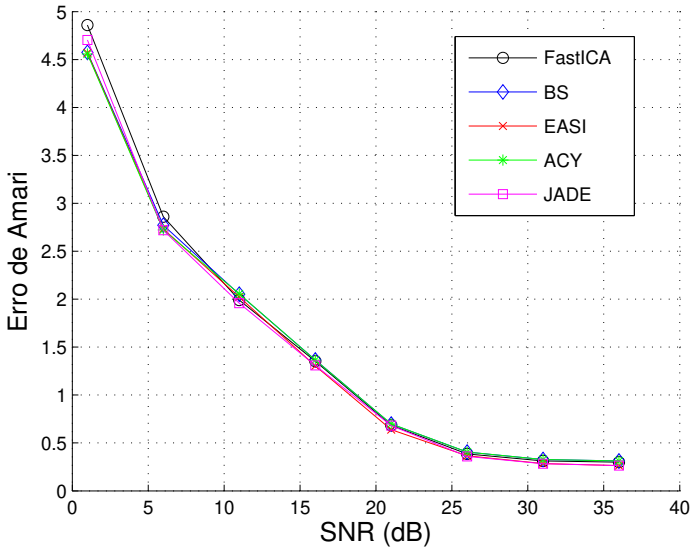


Figura 3.6: Robustez ao ruído (3 Fontes).

Também realizamos a mesma análise em um cenário com seis fontes. Novamente, fica visível a severa degradação imposta pelo ruído ao processo de recuperação das fontes, como mostra a Figura 3.7, que representa a média de 60 experimentos. Neste caso, uma separação satisfatória somente é atingida a partir de uma SNR de aproximadamente 18 dB.

Velocidade de convergência. Por fim, realizamos uma análise da velocidade de convergência das técnicas de BSS estudadas. Tendo em vista que, nos algoritmos baseados no método do gradiente, a escolha do passo de adaptação é determinante na convergência, buscamos, através de um grande número de simulações preliminares, determinar o maior valor possível que esse parâmetro poderia assumir de modo a se obter uma convergência estável.

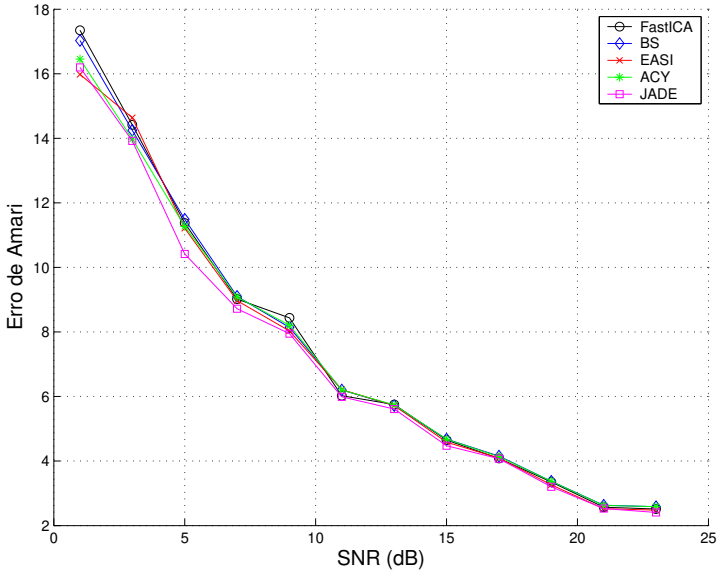


Figura 3.7: Robustez ao ruído (6 Fontes).

Em um primeiro cenário, estudou-se a convergência do erro de Amari na separação de duas fontes uniformemente distribuídas, situação esta apresentada na Figura 3.8, que representa a média de 50 simulações. Neste caso, foi possível constatar que o algoritmo FastICA convergiu consideravelmente mais rápido, se comparado às outras técnicas.

Em outros experimentos, foi possível verificar o mesmo comportamento dos algoritmos em relação à velocidade de convergência. Por exemplo, na Figura 3.9, a evolução do erro de Amari em um cenário com 5 fontes é apresentada.

Discussão. A dificuldade presente na comparação entre técnicas de BSS é decorrente principalmente da dependência destas em relação a um grande número de parâmetros, como, por exemplo, o número de dados, o número de iterações, o passo de adaptação, etc. No presente estudo, visando atenuar tais dificuldades, procuramos analisar isoladamente a influência de alguns fatores, que julgamos de extrema relevância prática, sobre o desempenho das técnicas de BSS.

Em quase todos os quesitos analisados, o desempenho dos algoritmos analisados foi semelhante. Certamente, isso é um resultado do procedimento adotado na

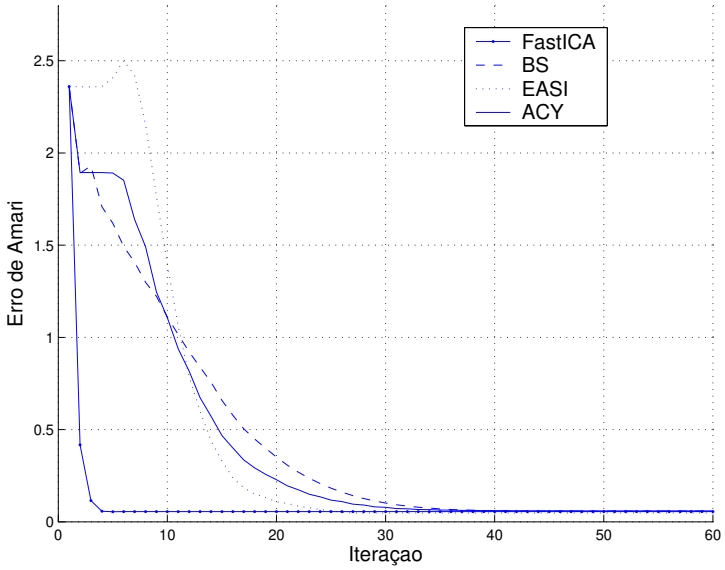


Figura 3.8: Velocidade de convergência (2 fontes).

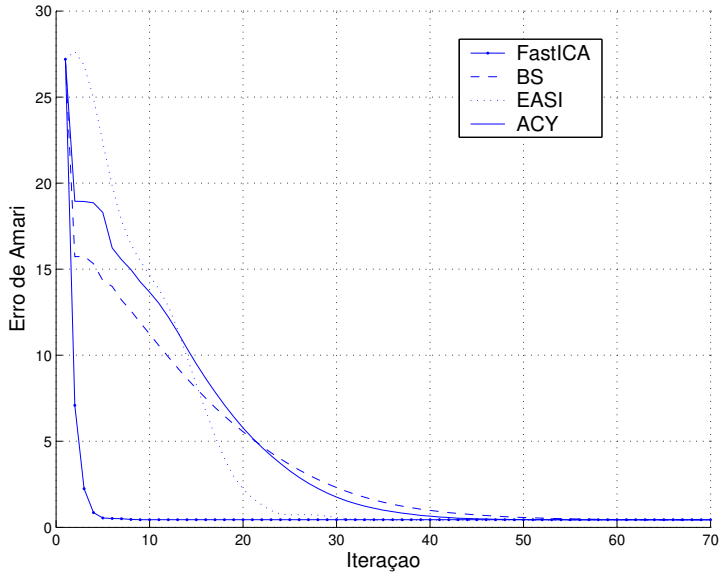


Figura 3.9: Velocidade de convergência (5 fontes).

determinação dos parâmetros de cada uma das técnicas, que buscou garantir a melhor convergência possível em cada uma das situações verificadas. Assim, levando

em conta que os quesitos analisados (com exceção da velocidade de convergência) dependem exclusivamente das estimativas finais de $\mathbf{W}$, nossa análise verificou, sobretudo, a eficácia das funções contraste envolvidas com cada um dos algoritmos, que, como vimos anteriormente, possuem relações entre si.

Apenas no estudo da velocidade de convergência das técnicas investigadas, constatamos uma discrepância entre as performances obtidas. Neste caso, as propriedades dos algoritmos, e não mais dos critérios, são responsáveis por um bom desempenho. Nesta situação, o algoritmo FastICA apresentou um desempenho consideravelmente superior aos outros. O principal motivo para este notável desempenho provém da estratégia de otimização empregada neste algoritmo. A otimização a partir do método de Newton permite, de certo modo, um ajuste dinâmico do passo de adaptação. Além disso, as simplificações presentes no desenvolvimento do FastICA permitem a obtenção de uma regra de ajuste em que não é necessária a inversão de matrizes, fazendo com que o tempo gasto por iteração também seja pequeno. De fato, em (Hyvärinen, 1999a), é mostrado que a velocidade de convergência do FastICA pode ser significativamente maior em relação às técnicas baseadas no gradiente.

3.4.2 Algoritmos adaptativos

Nos algoritmos adaptativos (*on-line*), a atualização do sistema separador é feita a cada nova amostra das misturas. Este tipo de estratégia é útil em aplicações que demandam processamento em tempo real e, outrossim, naquelas em que o sistema misturador é variante no tempo. Exemplos de problemas desta sorte são freqüentemente encontrados em telecomunicações.

Em nosso trabalho, analisamos o desempenho de quatro técnicas distintas: as versões estocásticas dos algoritmos BS, ACY e EASI, e o algoritmo de PCA não-linear com treinamento via RLS. A escolha das funções não-lineares para esses três primeiros algoritmos foi idêntica à apresentada na Seção 3.4.1. No que diz respeito ao NPCA-RLS, descrito em (3.49), consideramos a seguinte função não-linear $g(\cdot) = \tanh(\cdot)$. Em todas as técnicas, com exceção do EASI, o processo de separação foi feito após a realização de uma etapa de branqueamento das misturas.

Nossa análise abrangeu a verificação da evolução do desempenho dos algoritmos

frente ao aumento do número de fontes, bem como da robustez de cada uma delas ao ruído. No caso das técnicas adaptativas, diferentemente dos algoritmos em batelada, a verificação da velocidade de convergência constitui uma tarefa de considerável complexidade, dado que o caráter estocástico presente nesta situação dificulta a realização de uma análise nos mesmos moldes daquela apresentada na Seção 3.4.1. Entretanto, como veremos a seguir, na análise da sensibilidade das técnicas ao aumento do número de fontes, o modo como procedemos permitiu obter uma idéia da velocidade de convergência dos algoritmos estudados.

Sensibilidade ao aumento do número de fontes. Nesta análise, o ajuste dos parâmetros em cada algoritmo foi feito de acordo com o procedimento que descrevemos a seguir. Primeiramente, definimos um número fixo de iterações (o mesmo para todas as técnicas). Em seguida, através de ensaios preliminares, buscamos, para cada caso, por tentativa e erro, os parâmetros que forneciam o melhor desempenho no cenário definido.

No primeiro cenário analisado, consideramos um número de amostras igual a 2000. Na Figura 3.10, é apresentado o resultado deste primeiro ensaio. No caso, esta figura representa a média de 50 execuções. É possível observar nesta figura que, ao passo que nos algoritmos EASI e NPCA o desempenho cresce suavemente com o aumento do número de fontes, há uma acentuada tendência de crescimento na evolução do erro nos algoritmos ACY e BS. De certa forma, este pior desempenho está relacionado com a diferença na velocidade de convergência entre os algoritmos, pois, tendo em vista o procedimento de ajuste dos parâmetros adotado, foi necessário, para algoritmos mais lentos, definir passos de adaptação elevados, culminado num erro de Amari médio maior nesses casos.

Realizamos o mesmo procedimento em um outro cenário, agora considerando um número de iterações igual a 5000, e os resultados obtidos são exibidos na Figura 3.11. Neste caso, o algoritmos NPCA comportou-se consideravelmente melhor do que as outras técnicas analisadas. Novamente, tendo em vista o procedimento de ajuste dos parâmetros adotado, este notório desempenho é resultado da rápida velocidade de convergência associada a esse algoritmo.

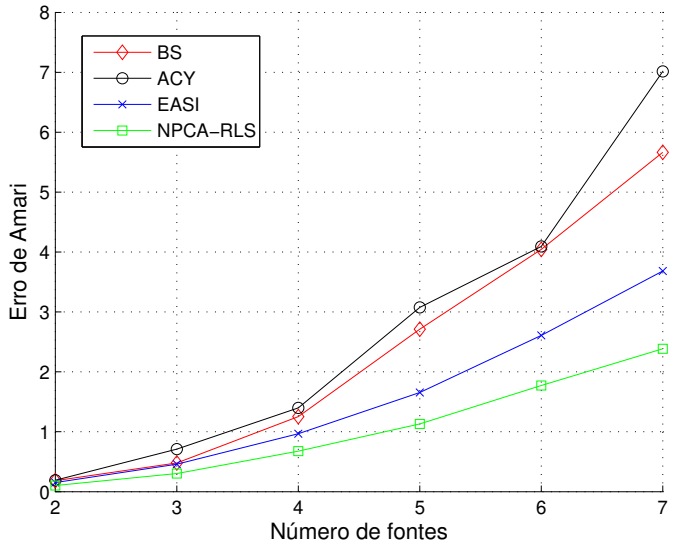


Figura 3.10: Erro de Amari em função do número de fontes (2000 amostras).

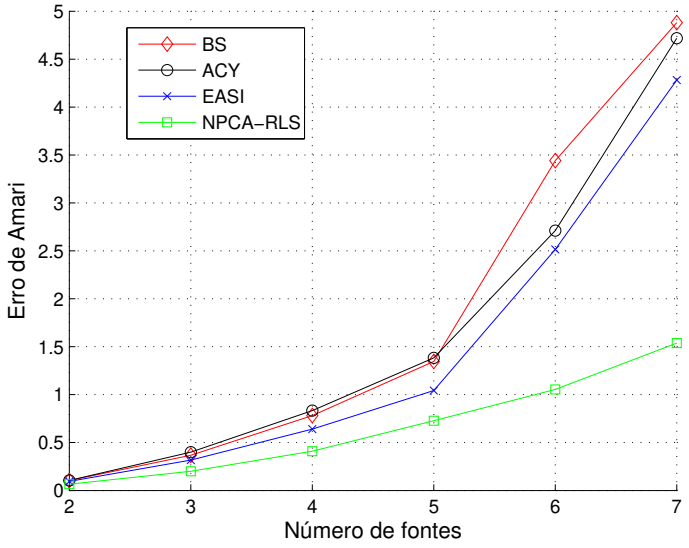


Figura 3.11: Erro de Amari em função do número de fontes (5000 amostras).

Sensibilidade ao ruído. Novamente, o ajuste dos parâmetros de cada algoritmo foi feito após a definição de um número fixo de iterações. No primeiro experimento

realizado, o número de fontes no problema foi igual a 4, e o número de iterações foi igual a 5000. Na Figura 3.12 são apresentados os resultados obtidos para esse caso.

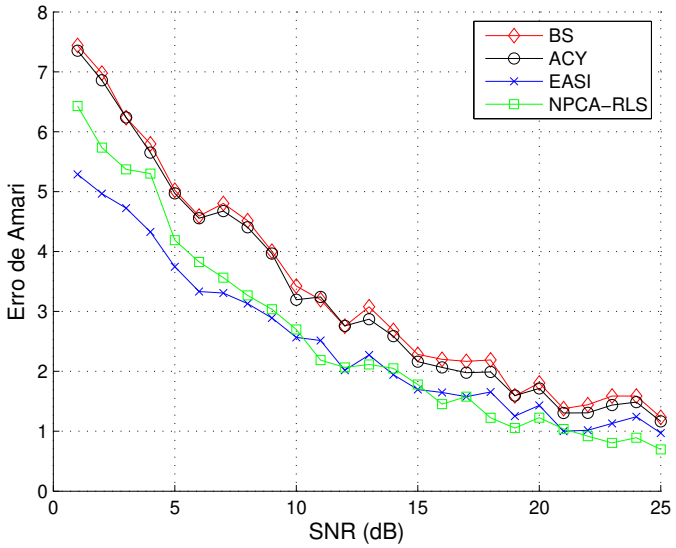


Figura 3.12: Robustez ao ruído (4 fontes).

Nessa situação, os algoritmos NPCA-RLS e EASI foram superiores aos algortimos BS e ACY. Observamos que, somente a partir do valor de aproximadamente $SNR = 12$ dB , os algoritmos são capazes de obter uma separação satisfatória das fontes. A partir desse valor, o algoritmo NPCA-RLS foi o melhor em relação a este critério, apresentando um desempenho levemente superior ao do obtido pelo pelo EASI.

Realizamos o mesmo ensaio considerando um número maior de fontes, no caso 7. Como é possível observar na Figura 3.13, assim como no caso anterior, os algoritmos EASI e NPCA-RLS apresentaram melhor desempenho em relação aos algoritmos BS e ACY.

Discussão. Na análise realizada, mereceu destaque o desempenho obtido pelo algoritmo NPCA-RLS. Vimos, ao verificar o erro de Amari em função do número de fontes, que essa técnica é capaz de prover valores significativamente menores em relação às outras estudadas. De fato, isso é uma conseqüência da rápida

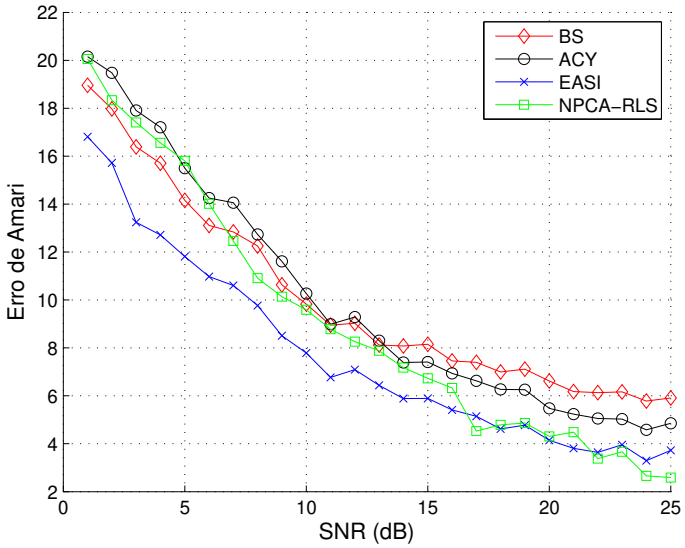


Figura 3.13: Robustez ao ruído (7 fontes).

velocidade de convergência desse algoritmo, o que, por sua vez, é resultado da técnica de treinamento empregada, no caso, o algoritmo RLS. Porém, é importante frisar que, nesta classe de algoritmos, embora a convergência seja rápida, há uma maior complexidade computacional por iteração em relação aos métodos baseados no gradiente estocástico, como é o caso dos algoritmos BS, ACY, e EASI.

Um outro ponto que chamou a atenção foi o desempenho do algoritmo EASI nos dois quesitos analisados. É importante ressaltar que, para esta técnica, não realizamos o branqueamento prévio das fontes, dado que esta etapa já esta embutida em sua regra de atualização, como vimos anteriormente. Em cenários práticos, especialmente naqueles com um elevado número de fontes, a realização do branqueamento pode adicionar um custo computacional significativo ao custo do algoritmo de BSS. Neste caso, torna-se atraente a adoção de uma regra de atualização que inclua tanto o branqueamento quanto a separação.

3.5 Sumário

Neste capítulo, em um primeiro momento, as principais técnicas para a resolução do problema de BSS linear foram apresentadas. Em seguida, realizou-se uma discussão acerca da equivalência entre algumas das abordagens discutidas, bem como sobre a existência de mínimos locais em certas situações. Por fim, apresentamos uma análise de desempenho dos algoritmos estudados em nossa revisão teórica. Consideramos tanto as técnicas adaptativas quanto as com modo de operação em batelada.

Capítulo 4

Separação de Misturas Não-lineares

Este capítulo trata do problema de separação de fontes na situação em que o sistema misturador é de natureza não-linear. Primeiramente, introduzimos tal problema, enfatizando suas relações com a ICA. Na seqüência, apresentamos algumas técnicas desenvolvidas para o caso em questão e abordamos um dos principais assuntos relacionados a esse tema, a separação de misturas com não-linearidade posterior (PNL, *Post-Nonlinear*). No caso, descrevemos importantes resultados téoricos inerentes aos modelos PNL, bem como algumas soluções já desenvolvidas para esta classe de sistemas misturadores.

Por fim, com o intuito de contornar um problema relativo à convergência para mínimos locais no treinamento de sistemas separadores PNL, propomos uma nova técnica. Os pilares de tal abordagem são o emprego de um algoritmo evolutivo na tarefa de otimização e o uso de um estimador de entropia fundamentado nas estatísticas de ordem. A viabilidade de nossa proposta é atestada a partir de simulações conduzidas em diversos cenários.

4.1 A Separação de Misturas Não-lineares e a ICA Não-linear

Uma extensão natural do problema de separação de misturas lineares ocorre quando o sistema misturador possui um caráter não-linear. Já vimos na Seção 2.4

que, neste caso, as misturas são dadas por

$$\mathbf{x} = \mathcal{F}(\mathbf{s}), \tag{4.1}$$

onde $\mathcal{F}(\cdot)$ corresponde a um mapeamento não-linear. Assim como no capítulo anterior, nos restringiremos ao caso em que o processo de mistura é instantâneo e o número de fontes é o mesmo que o de misturas.

Naturalmente, a solução do problema de separação cega de fontes não-linear (NBSS, *Nonlinear Blind Source Separation*) requer o ajuste de um sistema separador que também seja não-linear. Assim, as estimativas das fontes neste caso são dadas por

$$\mathbf{y} = \mathcal{G}(\mathbf{x}), \tag{4.2}$$

onde $\mathcal{G}(\cdot)$ corresponde ao mapeamento não-linear que representa a ação do sistema separador. Esta breve caracterização sinaliza uma primeira dificuldade no problema da NBSS. Ao passo que, no caso linear, não havia complicações referentes à escolha da estrutura do sistema separador, no caso não-linear esta tarefa é de considerável dificuldade, pois nem sempre é possível definir um sistema separador capaz de inverter perfeitamente a ação do mapeamento não-linear de mistura. Devido a esta limitação estrutural, é comum, neste caso, a existência de distorções residuais nas estimativas das fontes.

Em um primeiro momento, tendo em vista o sucesso desta abordagem no caso linear, as técnicas desenvolvidas para a NBSS fundamentaram-se na extensão não-linear da ICA. Basicamente, tal extensão pode ser compreendida à luz da primeira definição da ICA, apresentada na Seção 2.5.1; assim, na Análise de Componentes Independentes Não-linear (NICA, *Nonlinear Independent Component Analysis*), dado um vetor aleatório $\mathbf{x}$, busca-se um mapeamento não-linear $\mathcal{G}(\cdot)$ de modo que os elementos do vetor $\mathbf{y} = \mathcal{G}(\mathbf{x})$ sejam tão estatisticamente independentes quanto possível.

Em contraste com o caso linear, a aplicação da ICA ao problema de NBSS não mais garante, de um modo geral, a recuperação das fontes. Para um melhor entendimento do motivo dessa dissociação entre a NICA e a NBSS, é conveniente definir o conceito de mapeamento trivial.

Definição 4.1.1 (Mapeamento Trivial) *Um mapeamento* $\mathcal{H}(\cdot) = [\mathcal{H}_1(\cdot), \cdots, \mathcal{H}_n(\cdot)]$ *é trivial se este transforma qualquer vetor* $\mathbf{r}= [r_1, \cdots, r_n]$ *com elementos estatisticamente independentes em um vetor cujos elementos também são independentes entre si.*

É possível mostrar que um dado mapeamento $\mathcal{H}(\cdot)$ é trivial se, e somente se, a seguinte condição for satisfeita (Jutten & Karhunen, 2003):

$$\mathcal{H}_i(r_1, \cdots, r_n) = h_i(r_{\rho(i)}), \quad i = 1, \cdots, n \tag{4.3}$$

sendo que $h_i(\cdot)$ são funções arbitrárias e ρ corresponde a um operador de permutação sobre o conjunto $\{1, 2, \cdots, n\}$. Assim sendo, podemos concluir que um mapeamento é trivial se, e somente se, cada elemento do vetor resultante for uma função exclusiva de um dos elementos r_i.

Tendo essa definição em mente, voltemos ao problema de separação. A partir de (4.1) e (4.2), obtém-se a seguinte relação entre as fontes e suas estimativas:

$$\mathbf{y} = \mathcal{G}(\mathcal{F}(\mathbf{s})) = (\mathcal{G} \circ \mathcal{F})(\mathbf{s}), \tag{4.4}$$

onde o operador "$\circ$" corresponde à composição de duas funções. Diante disto e da definição 4.1.1, um primeiro ponto a ser ressaltado é que, para este caso, uma possível solução da NICA ocorre quando o mapeamento $\mathcal{G} \circ \mathcal{F}$ é trivial. Logo, nesta situação, a recuperação das fontes apresentaria, além da ambiguidade relacionada à permutação da ordem como no caso linear, uma distorção residual, possivelmente não-linear, caracterizada pelas funções $h_i(\cdot)$.

A despeito dessa nova ambiguidade, a recuperação das fontes baseada na determinação de $\mathcal{G}$ visando um mapeamento conjunto $\mathcal{G} \circ \mathcal{F}$ trivial pode ser considerada uma abordagem satisfatória, pois, ainda assim, cada estimativa recuperada é função de apenas uma fonte, de acordo com (4.3). Entretanto, a aplicação da NICA nem mesmo garante a obtenção de mapeamentos conjuntos triviais, o que dificulta consideravelmente a resolução deste problema. O seguinte exemplo, apresentado em (Jutten & Karhunen, 2003), ilustra bem a limitação da ICA no cenário não-linear.

Exemplo 4.1.1 *Considere duas fontes* s_1 *e* s_2 *com as seguintes distribuições:* $p_{s_1}(s_1) = s_1 exp(-s_1^2/2)$ *(Rayleigh) e* $p_{s_2}(s_2) = 2/\pi$ *quando* $s_2 \in [0, \pi/2)$ *(uniforme).*

Considerando que tais fontes são estatisticamente independentes, a densidade de probabilidade conjunta é dada por

$$p_{s_1 s_2}(s_1, s_2) = \begin{cases} \frac{2}{\pi} s_1 \exp\left(\frac{-s_1^2}{2}\right), s_2 \in [0, \pi/2); \\ 0, s_2 \in (-\infty, 0) \text{ ou } s_2 \in [\pi/2, +\infty). \end{cases} \tag{4.5}$$

Considere agora o seguinte mapeamento entre as fontes e suas estimativas

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathcal{H}(\mathbf{s}) = \begin{bmatrix} s_1 \cos(s_2) \\ s_1 \sin(s_2) \end{bmatrix}. \tag{4.6}$$

É interessante notar que os elementos de $\mathbf{y}$ *são misturas dos elementos de* $\mathbf{s}$*, haja visto que o jacobiano desta transformação não é diagonal. Contudo, vejamos o que ocorre com a densidade de probabilidade conjunta do vetor* $\mathbf{y}$*. De acordo com a expressão clássica de transformação de uma variável aleatória (Papoulis, 1993), essa densidade é dada por*

$$p_{y_1 y_2}(y_1, y_2) = \frac{p_{s_1 s_2}(s_1, s_2)}{|\det \mathbf{J}_{\mathcal{H}}(\mathbf{y})|}, \tag{4.7}$$

onde $\mathbf{J}_{\mathcal{H}}(\mathbf{y})$ *corresponde ao jacobiano do mapeamento* $\mathcal{H}(\cdot)$[1] *No caso, tem-se que*

$$|\det \mathbf{J}_H(\mathbf{y})| = \left|\det\left(\begin{bmatrix} \cos(s_2) & \sin(s_2) \\ s_1 \sin(s_2) & -s_1 \cos(s_2) \end{bmatrix}\right)\right| = \\ = |-s_1(\cos^2(s_2) + \sin^2(s_2))| = |s_1|. \tag{4.8}$$

Substituindo (4.5) e (4.8) em (4.7), e levando em conta que $s_1^2 = y_1^2 + y_2^2$*, temos que*

$$p_{y_1 y_2}(y_1, y_2) = \frac{2}{\pi} \exp(\frac{-(y_1^2 + y_2^2)}{2}) = (\sqrt{\frac{2}{\pi}} \exp(\frac{-y_1^2}{2}))(\sqrt{\frac{2}{\pi}} \exp(\frac{-y_2^2}{2})). \tag{4.9}$$

Note que esta densidade é fatorável em um produto de funções exclusivas de y_1 *e* y_2*, ou seja, os elementos de* $\mathbf{y}$ *são estatisticamente independentes entre si, embora eles ainda sejam uma mistura de* s_1 *e* s_2*.*

[1] Note que, no intervalo delimitado por $\mathbf{s}$, este mapeamento é inversível, o que permite obter o jacobiano em função $\mathbf{y}$.

Foi Darmois, em 1951, no âmbito da análise por fatores não-linear (Taleb & Jutten, 1999), quem primeiro verificou esta peculiaridade presente em sistemas não-lineares. Ele estabeleceu que existem mapeamentos não-triviais que transformam as fontes em vetores independentes, e, ainda assim, as misturam. No contexto da ICA, este resultado foi apresentado por Hyvärinen e Pajunen (Hyvärinen & Pajunen, 1999), em um trabalho que demonstrou como construir famílias de soluções não-triviais que resultam em saídas independentes.

De certa forma, essa dificuldade existente no tratamento da NBSS via NICA pode ser encarada como um conseqüência do fato de que as funções não-lineares apresentam um enorme grau de flexibilidade, tendo, por exemplo, a capacidade de misturar dois vetores e ainda assim torná-los independentes. Logo, ante a ineficácia da ICA na instância não-linear do problema de separação, como, então, recuperar as fontes neste caso?

Basicamente, a solução adotada para este problema de NBSS opera justamente no sentido de diminuir esse grau de flexibilidade, ora construindo funções objetivo que, além de considerar a independência, levam em conta outras informações do modelo (de uma maneira análoga à regularização no treinamento de redes neurais), ora restringindo os modelos de sistemas misturadores a uma classe de funções não-lineares separáveis, em uma tentativa de utilizar exclusivamente o ferramental da ICA para resolução do problema.

No que se refere à abordagem da NBSS via introdução de termos de regularização na função custo, a grande vantagem é a possibilidade de solucionar o problema para uma classe mais ampla de funções não-lineares. No entanto, a inclusão desses termos requer conhecimentos adicionais sobre o cenário em questão, o que descaracteriza ainda mais o caráter cego do trabalho. Além disso, geralmente, a inclusão desses termos é feita de uma maneira *ad hoc*, o que pode não garantir a eficácia de tal estratégia em cenários com não-linearidades severas.

No tocante à estratégia que considera uma classe restrita de sistemas misturadores, tentou-se, em um primeiro momento, utilizar a idéia da ICA em cenários com mapeamentos não-lineares suaves. Entretanto, mesmo para um mapeamento deste tipo, é possível que a recuperação da independência não implique na separação das fontes (Babaie-Zadeh, 2002). Desse modo, fez-se necessário uma

nova estratégia, na qual buscou-se por modelos que fossem, de fato, separáveis. Nesta linha, podemos citar Taleb & Jutten (1999), Taleb (2002), Hyvärinen & Pajunen (1999), Eriksson & Koivunen (2002) e Babaie-Zadeh (2002).

Em um desses trabalhos (Taleb & Jutten, 1999), o chamado modelo com não-linearidade posterior (PNL) foi introduzido. Desde então, o estudo de tal classe de modelos vem sendo um dos principais assuntos em NBSS. Na Seção 4.3, apresentaremos os principais aspectos deste caso. Antes, porém, exporemos uma descrição sucinta de dois paradigmas de resolução da NBSS que não necessariamente operam com modelos restritos de sistemas separadores.

4.2 Técnicas para o Caso Geral

Dedicamos esta seção à apresentação de algumas técnicas destinadas ao problema de geral de NBSS, no qual nenhum conhecimento sobre a estrutura do sistema separador é assumido. Inicialmente, descrevemos sucintamente uma das primeiras técnicas concebidas para a separação de misturas não-lineares, a abordagem via mapas auto-organizáveis. Em seguida, apresentamos um paradigma alternativo baseado no treinamento bayesiano de redes neurais. Este tipo de abordagem apóia-se na identificação do sistema misturador não-linear.

4.2.1 Aplicação de Mapas Auto-Organizáveis em NBSS

Uma das primeiras idéias para a resolução do problema de NBSS foi a aplicação de um mapa auto-organizável (Pajunen, Hyvärinen & Karhunen, 1996), um paradigma consagrado de rede neural cujo aprendizado é feito de maneira competitiva e não-supervisionada. Em linhas gerais, tal estrutura realiza um mapeamento de um espaço original, onde residem os dados de entrada, para um espaço definido por um arranjo de neurônios em que são previamente definidas relações de vizinhança. Cada um desses neurônios é representado por um vetor de referência no espaço de entrada. O mapeamento de um dado vetor de entrada é feito através da determinação do neurônio cujo vetor de referência é o mais próximo do vetor de entrada em questão, de acordo com alguma função de distância.

O treinamento do mapa, isto é, o ajuste dos vetores de referência (ou vetor de pesos) do mapa, pode ser feito através da execução dos seguintes passos:

1. Inicialização dos vetores de pesos;
2. Dada uma amostra de treinamento (os dados de treinamento são compostos por um conjunto de dados de entrada), verifica-se qual vetor de pesos, denominado vencedor, está mais próximo (norma euclidiana);
3. Calcula-se a diferença entre o vetor de pesos mais próximo e a amostra de treinamento;
4. Atualiza-se os vetor de pesos mais proximo somando-o com o vetor diferença multiplicado por um passo η.

Matematicamente a expressão de atualização do vetor de pesos mais próximo pode ser descrita desta forma

$$\mathbf{c}_{k,T+1} \leftarrow \mathbf{c}_{k,T} + \eta(\mathbf{x}_T - \mathbf{c}_{k,T}), \tag{4.10}$$

Onde $\mathbf{c}_{k,T}$ e $\mathbf{x}_T$ correspondem, respectivamente, à estimativa do k-ésimo vetor de pesos e a amostra de treinamento, ambos no instante T. Dependendo das relações de vizinhança, os vetores que estiverem mais próximos do neurônio vencedor também são atualizados, porém, com um valor menor de η.

Na aplicação dos mapas auto-organizáveis em BSS, as misturas correspondem aos dados de entrada do mapa. A propriedade que justifica esta aplicação reside no fato de que, após o treinamento, a distribuição do vetor de pesos é aproximadamente a mesma da distribuição conjunta dos dados de entrada. Ou seja, os vetores de pesos se distribuem de acordo com a distribuição das misturas, e, como conseqüência, cada um desses vetores de pesos vencem, aproximadamente, o mesmo número de vezes. Ou seja, temos nesta situação uma distribuição conjunta próxima a de uma uniforme no espaço de saída. No caso em que o mapa é retangular (Pajunen, Hyvärinen & Karhunen, 1996), isto implica que as coordenadas obtidas no espaço de saída são aproximadamente independentes entre si.

Em síntese, após o treinamento do mapa, a estimativa de uma dada fonte é obtida pelas coordenadas do neurônio cujo vetor de referência está mais próximo

ao vetor de mistura. Este procedimento implica numa quantização das estimativas das fontes, a qual pode ser atenuada através da realização de uma interpolação no mapa.

Há três problemas significativos nessa abordagem. O primeiro deles é que se busca separar as fontes através da recuperação da independência. Vimos que, no domínio não-linear, não há garantias de que este procedimento seja eficaz. Além disso, um segundo problema é que, neste método, a independência estatística é decorrente do fato de que o espaço de saídas do mapa é uniformemente distribuído. Por este motivo, as fontes devem possuir distribuições próximas à uniforme. Finalmente, tendo em vista que o arranjo de neurônios geralmente é bi-dimensional, ou seja, a dimensão do espaço de saída é bi-dimensional e, como conseqüência, este método opera somente em cenários com duas fontes.

4.2.2 Ensemble Learning

No contexto da análise de fatores não-lineares (NFA - *Nonlinear Factor Analysis*), desenvolveu-se uma nova proposta (Valpola, 2000) para o problema de NBSS, com modo de operação fundamentado em paradigmas bayesianos de treinamento de redes neurais. Em contraposição aos métodos de NBSS que buscam adaptar sistemas separadores visando contrabalançar a ação do sistema misturador, a abordagem desenvolvida investe, de uma maneira explícita, na tarefa de identificação do sistema separador, ou seja, de buscar explicações e modelos plausíveis para as misturas.

Uma importante questão relacionada à explicação de um conjunto de dados a partir de modelos previamente definidos diz respeito à capacidade de extrair informações de todo este conjunto utilizando somente suas amostras. Quando um modelo flexível, como, por exemplo uma rede neural MLP, é utilizado para essa tarefa, esta generalização dos dados a partir de amostras pode não ser facilmente obtida, uma vez que o alto grau de flexibilidade do modelo pode contribuir para a geração de uma solução excessivamente ajustada às amostras, de tal maneira que, perante todo o conjunto, tal solução não seja capaz de prover explicações coerentes. Por outro lado, a utilização de modelos simples, pouco flexíveis, pode não ser suficiente para explicar o conjunto.

Um procedimento interessante para contornar este problema consiste em explicar

o conjunto de dados em questão não apenas a partir de um único modelo, mas sim levando em conta vários modelos diferentes. Esta é a principal idéia presente no treinamento bayesiano de redes neurais. Neste caso, diferentemente de treinamentos convencionais que buscam determinar numericamente os parâmetros da rede neural, o objetivo é obter as probabilidades *a posteriori* desses parâmetros após as observações dos dados. Note o leitor que a descrição dos parâmetros desta maneira é equivalente a considerar várias explicações para os dados, de tal maneira que, para cada uma delas, exista uma probabilidade associada.

No trabalho em que se propôs a aplicação do aprendizado bayesiano à NBSS (Valpola, 2000), o seguinte modelo de sistema misturador foi considerado

$$\mathbf{x} = \mathbf{f}(\mathbf{s}) + \mathbf{n}, \tag{4.11}$$

onde $\mathbf{x}$ e $\mathbf{s}$ representam as misturas e as fontes, respectivamente, e $\mathbf{n}$ é um vetor com elementos gaussianos que representa o ruído.

Como explicação para as misturas $\mathbf{x}$, utilizou-se uma rede MLP, cujo mapeamento entrada-saída é dada por

$$\mathbf{f}(\mathbf{s}) = \mathbf{B}\mathbf{g}(\mathbf{A}\mathbf{s} + \mathbf{a}) + \mathbf{b}, \tag{4.12}$$

onde $\mathbf{A}$ e $\mathbf{a}$ correspondem, respectivamente, à matriz de pesos sinápticos e ao vetor de polarização da camada intermediária da rede. Já a matriz $\mathbf{B}$ e o vetor $\mathbf{b}$ representam, respectivamente, os pesos sinápticos e a polarização da camada de saída da rede.

Tendo em vista que o objetivo da abordagem em questão é determinar as probabilidades *a posteriori* de todos os componentes estatísticos do modelo (4.12), é necessário definir as suas probabilidades *a priori*. Uma possibilidade é realizar isto de modo paramétrico, utilizando variáveis gaussianas, que permitem um tratamento computacional mais simples. Neste caso, as distribuições *a priori* dos parâmetros do modelo (4.12) são dadas por:

$$\mathbf{x} \sim N(\mathbf{f}(\mathbf{s}), e^{2v_x}), \tag{4.13}$$

$$\mathbf{s} \sim N(0, e^{2v_s}), \tag{4.14}$$

$$\mathbf{A} \sim N(0, 1), \tag{4.15}$$

$$\mathbf{B} \sim N(0, e^{2v_B}), \tag{4.16}$$

$$\mathbf{a} \sim N(m_a, e^{2v_a}), \tag{4.17}$$

$$\mathbf{b} \sim N(m_b, e^{2v_b}), \tag{4.18}$$

onde $N(\mu, \sigma^2)$ representa uma variável gaussiana de média μ e variância σ^2. Note que, neste modelo, considera-se que o vetor $\mathbf{s}$ também é gaussiano, o que o caracteriza como um modelo típico de NFA e inviabiliza, em um primeiro momento, a sua aplicação ao problema de BSS, visto que, em problemas práticos, as fontes possuem as mais diversas densidades. Para superar este contratempo, Valpola (Valpola, 2000) estendeu este modelo ao caso em que $\mathbf{s}$ é caracterizado como uma mistura de gaussianas, tornando-o assim adequado para representação de sistemas misturadores práticos.

Denotando por Θ todos os parâmetros das distribuições estabelecidas, sua probabilidade *a posteriori* conjunta é dada por $P(\Theta|X)$. Na abordagem bayesiana, tal probabilidade expressa todo nosso conhecimento do sistema, sendo que, caso seja desejado obter as médias de cada um dos parâmetros, ou mesmo determinar um modelo que otimize uma dada função custo, é preciso realizar integrais sobre $P(\Theta|X)$, que eventualmente podem resultar em uma difícil tarefa, o que ocorre de fato no modelo apresentado.

Frente a esta dificuldade de cálculo das probabilidades *a posteriori*, uma alternativa viável é determinar uma aproximação paramétrica para essas probabilidades, utilizando funções que resultem em um simplificado tratamento matemático e que sejam computacionalmente eficientes. O aprendizado bayesiano variacional (também conhecido como *ensemble learning*) é fundamentado nesta idéia, sendo que, neste caso, a aproximação das probabilidades *a posteriori* é expressa por

$$\mathbf{q}(\Theta|X) = \prod_j q_j(\Theta_j|X) \tag{4.19}$$

onde $q_j(\Theta_j|X)$ corresponde à probabilidade *a posteriori* marginal do j-ésimo parâmetro. É assumido que as distribuições na expressão (4.19) são gaussianas, com exceção das referentes às fontes. Assim sendo, para determinar a aproximação, basta encontrar as médias e variâncias de cada uma das gaussianas presentes nessa expressão.

Em *ensemble learning*, a busca por esses parâmetros é feita a partir da minimização da divergência de Kullback-Leibler entre a aproximação $\mathbf{q}(\Theta|X)$ e a probabilidade *a posteriori* de fato $P(\Theta|X)$. Uma dificuldade inerente a este procedimento é a necessidade de determinar a divergência de Kullback-Leibler analiticamente.

Além da dificuldade analítica existente na aplicação do aprendizado variacional em NBSS, há outras desvantagens importantes. Uma delas está relacionada ao elevado número de aproximações realizadas em todas as etapas no desenvolvimento do algoritmo. Por exemplo, não há critério algum para a escolha do modelo paramétrico utilizado na identificação do sistema misturador, o que pode resultar em opções ruins, capazes de inviabilizar todo o procedimento. Além disso, o custo computacional referente a este algoritmo é considerável. Por outro lado, em virtude do desenvolvimento baseado em identificação, o problema de separabilidade não é tão premente como nos outros algoritmos para o caso geral (Valpola, 2000).

4.3 Separação de Misturas com Não-linearidade Posterior

Um sistema misturador PNL é essencialmente constituído por um estágio linear seguido por uma seção não-linear. Este tipo de estrutura pode ser útil no modelamento de sistemas nos quais o processo de mistura é de natureza linear, porém, com sensores que apresentam uma resposta não-linear. Esta situação ocorre, por exemplo, em sistemas de comunicação com estágios amplificadores na recepção (Taleb & Jutten, 1999) e em arranjo de sensores inteligentes de substâncias químicas e de gases (Bermejo, Jutten & Cabestany, 2006).

Matematicamente, a i-ésima saída de um sistema misturador PNL (ilustrado na Figura 4.1) é dada por

$$x_i = f_i(a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{iN}s_N), \tag{4.20}$$

onde $f_i(\cdot)$ corresponde a uma função não-linear. Utilizando uma notação matricial, obtém-se uma expressão para todas as misturas do seguinte modo

$$\mathbf{x} = \mathbf{f}(\mathbf{As}), \tag{4.21}$$

sendo que $\mathbf{f}(\cdot) = [f_1(\cdot) \dots f_N(\cdot)]$. Note que a ação das funções não-lineares se dá de modo individual, ou seja, apenas considerando uma mistura linear.

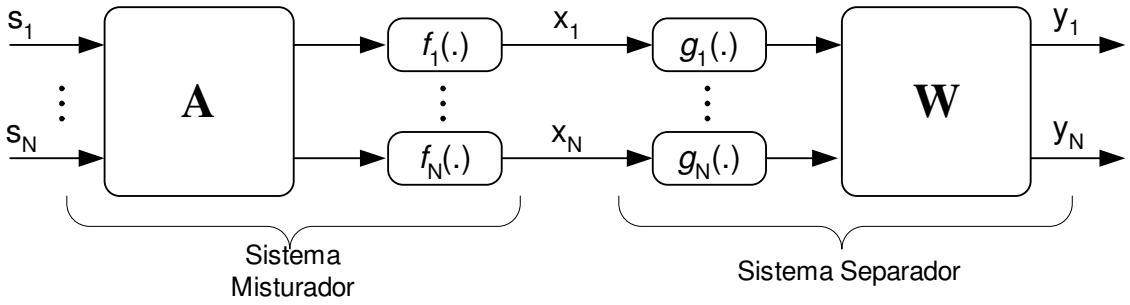


Figura 4.1: Modelo PNL

Levando em conta que o sistema separador deve ser capaz de inverter a ação do sistema misturador, é fundamental, no cenário PNL, que a estrutura do primeiro seja uma versão espelhada em relação ao segundo, de acordo com a Figura 4.1. Neste caso, a i-ésima saída de um sistema misturador PNL equivale a

$$y_i = w_{i1}g_1(x_1) + w_{i2}g_2(x_2) + \ldots + w_{iN}g_N(x_N), \tag{4.22}$$

sendo que $g_i(\cdot)$ é uma função não-linear. Em notação matricial, o processo de separação é dado por

$$\mathbf{y} = \mathbf{W}\mathbf{g}(\mathbf{x}), \tag{4.23}$$

sendo que $\mathbf{g}(\cdot) = [g_1(\cdot) \dots g_N(\cdot)]$.

4.3.1 Separação via recuperação da independência estatística

Ao passo que, no caso linear da BSS, ajusta-se apenas a matriz $\mathbf{W}$, no modelo PNL, há também a necessidade de adaptar cada uma das funções não-lineares do sistema separador. Analisando o mapeamento conjunto dos sistemas misturador e separador, dado por

$$\mathbf{y} = \mathbf{W}\mathbf{g}(\mathbf{f}(\mathbf{A}\mathbf{s})), \tag{4.24}$$

é evidente que as estimativas das fontes são estatisticamente independentes entre si na situação em que a composição $\mathbf{g} \circ \mathbf{f}$ corresponde a um vetor de funções lineares e

$\mathbf{WA} = \mathbf{\Lambda P}$ (sendo $\mathbf{\Lambda}$ e $\mathbf{P}$ matrizes diagonal e de permutação, respectivamente). Para que o modelo PNL seja separável, é necessário que o caminho reverso da constatação anterior também seja verdadeiro, ou seja, que, sendo $\mathbf{y}$ estatisticamente independente, tanto a seção não-linear quanto a linear devam se resumir a ganhos de escala e/ou permutações das fontes (no segundo caso).

A primeira prova de separabilidade do modelo PNL foi feita no seminal trabalho de Taleb e Jutten (Taleb & Jutten, 1999). Posteriormente, outras demonstrações dessa propriedade também foram apresentadas (Babaie-Zadeh, 2002; Achard & Jutten, 2005). As condições necessárias para a separabilidade do modelo PNL são explicitadas no seguinte teorema.

Teorema 4.3.1 (Separabilidade do Modelo PNL) *Considere as seguintes hipóteses*

- *A matriz* $\mathbf{A}$ *é inversível, e, de fato, mistura as fontes, ou seja, há ao menos dois elementos não-nulos em cada coluna e linha desta matriz;*
- *As funções* $\mathbf{f}(\cdot)$ $\mathbf{g}(\cdot)$ *são monotônicas, e, conseqüentemente,* $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ *também o é;*
- *Há, no máximo, uma fonte gaussiana;*
- *A função densidade de probabilidade conjunta das fontes é diferenciável e sua derivada é contínua em todo o seu suporte.*

Nessas condições, se os elementos de $\mathbf{y}$ *forem estatisticamente independentes, então todos os elementos do vetor* $\mathbf{h}(\cdot)$ *corresponderão a funções lineares e* $\mathbf{WA} = \mathbf{\Lambda P}$.

Um aspecto interessante deste teorema é que, para prová-lo, basta demonstrar que a independência estatística do vetor $\mathbf{y}$ implica na linearidade do mapeamento conjunto $\mathbf{h}$, pois, neste caso, têm-se exatamente as mesmas condições presentes no resultado de separabilidade obtido por Comon para o caso linear. De fato, todas as demonstrações seguiram esta linha.

4.3.2 O algoritmo de Taleb-Jutten

O resultado que acabamos de apresentar indica uma certa correspondência entre problema linear da BSS e a separação de misturas PNL, no sentido de que, neste último caso, também é possível separar as fontes via recuperação da independência. Entretanto, há uma diferença crucial entre essas duas situações que torna a resolução do cenário PNL significativamente mais complexa. Ao passo, que no caso linear, conforme já vimos, as diferentes abordagens existentes (não-gaussianidade, Infomax, NPCA etc.) estão ligadas à independência estatística, tais relações, no modelo PNL, não se mantêm, fazendo com que seja necessário empregar critérios capazes de quantificar "diretamente" o nível de independência entre as estimativas das fontes, como, por exemplo, a informação mútua entre elas.

O problema dessa abordagem é que qualquer medida diretamente relacionada com a independência estatística requer o conhecimento das densidades de probabilidade das variáveis aleatórias envolvidas. Esta exigência dificulta, sobretudo, a etapa de treinamento de um sistema separador PNL. Lembremos que, normalmente, nas diversas abordagens existentes para o caso linear, os critérios baseados em momentos de ordem superior, ou mesmo o desenvolvimento algébrico no Infomax, que permite expressar a entropia conjunta das saídas em função do jacobiano da transformação, deixavam o processo de otimização consideravelmente mais simplificado em relação à abordagem de informação mútua.

Infelizmente, essas alternativas não são mais viáveis em modelos PNL, e medidas diretas de independência devem ser utilizadas. Neste contexto, merece destaque o critério de minimização da informação mútua das estimativas das fontes. Para um sistema PNL, é possível mostrar, aplicando a expressão da transformação da entropia (Apêndice A), que a informação mútua de $\mathbf{y}$ é dada por

$$I(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{x}) - \log|\det \mathbf{W}| - E\left\{\log \prod_{i=1}^{N} |g_i'(x_i)|\right\}. \tag{4.25}$$

Note que, no decorrer do processo de treinamento, a estimação dos três últimos termos pode ser facilmente obtida. Entretanto, o primeiro termo corresponde ao somatório das entropias marginais das estimativas das fontes, as quais, por sua vez, dependem das distribuições de probabilidade das fontes.

Tabela 4.1: Algoritmo de Taleb e Jutten.

1. Defina os passos referentes as seções não-linear (μ_n) e linear (μ_l).
2. Adaptação da seção não-linear de acordo com a expressão (4.29).
$$\mathbf{\Theta}_i \leftarrow \mathbf{\Theta}_i + \mu_n \frac{\partial I(\mathbf{y})}{\partial \mathbf{\Theta}_i}. \tag{4.27}$$
3. Cálculo da nova saída da seção não-linear: $\mathbf{e} = \mathbf{g}(\mathbf{\Theta}, \mathbf{x})$.
4. Adaptação da seção linear de acordo com a expressão (4.26).
$$\mathbf{W} \leftarrow \mathbf{W} + \mu_l \frac{\partial I(\mathbf{y})}{\partial \mathbf{W}}. \tag{4.28}$$
5. Estimativa das fontes: $\mathbf{y} = \mathbf{W}\mathbf{e}$.
6. Caso não convirja, voltar ao passo 2.

No mesmo trabalho (Taleb & Jutten, 1999) em que demonstraram a separabilidade do modelo PNL, Taleb e Jutten propuseram um algoritmo baseado na otimização de (4.25) a partir do método do gradiente descendente. Naturalmente, essa abordagem requer a expressão do gradiente em relação tanto aos parâmetros da seção linear quanto aos da parte não-linear. Para o estágio linear, tal expressão é dada por

$$\frac{\partial I(\mathbf{y})}{\partial \mathbf{W}} = -E\{\Psi(\mathbf{y})\mathbf{e}^T\} - (\mathbf{W}^T)^{-1}, \tag{4.26}$$

onde $\Psi(\mathbf{y}) = [\psi_{y_1}(y_1), \ldots, \psi_{y_N}(y_N)]$, de modo que $\psi_{y_i}(y_i) = (p_{y_i}(y_i)'/p_{y_i}(y_i))$. Note que esta regra é semelhante à do caso linear (expressão (3.12)), com a diferença de que o vetor $\mathbf{e} = [e_1, \ldots, e_N]$ denota as saídas da seção não-linear do sistema separador. O ajuste da seção linear também pode ser feito a partir do gradiente natural da função em questão. Neste caso, basta multiplicar (4.26) por $\mathbf{W}^T\mathbf{W}$.

Com o intuito de verificar como é feito o ajuste da seção não-linear, representemos as funções não-lineares paramétricas do sistema separador por $g_i(\mathbf{\Theta}_i, x_i)$, $i =$

$1, \ldots, N$, onde $\mathbf{\Theta}_i$ corresponde ao vetor de parâmetros desta função. Deste modo, mostrou-se que

$$\frac{\partial I(\mathbf{y})}{\partial \mathbf{\Theta}_i} = -E\left\{ \frac{\partial \log |g_i'(\mathbf{\Theta}_i, x_i)|}{\partial \mathbf{\Theta}_i} \right\} - E\left\{ \Big(\sum_{k=1}^{N} \psi_{y_k}(y_k) w_{ki} \Big) \frac{\partial g_i(\mathbf{\Theta}_i, x_i)}{\partial \mathbf{\Theta}_i} \right\}, \qquad (4.29)$$

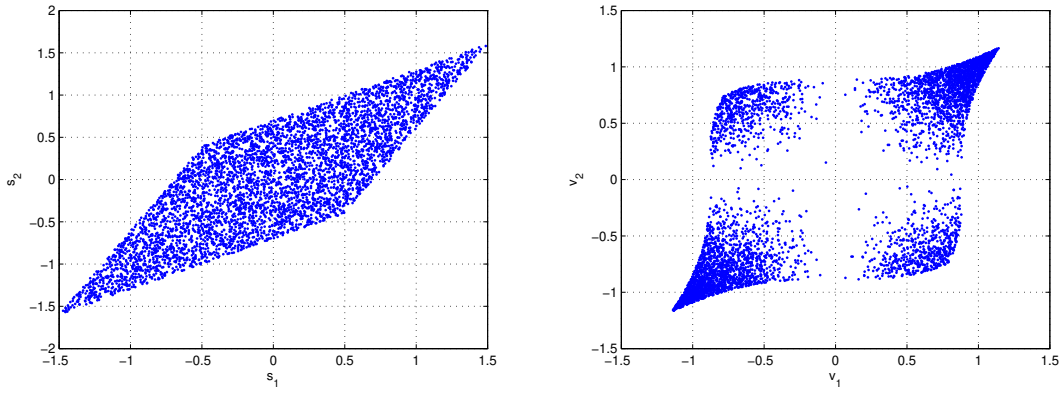
onde w_{ki} corresponde ao elemento (k, i) da matriz $\mathbf{W}$.

As expressões (4.26) e (4.29) constituem a base do algoritmo de Taleb e Jutten (TJ). Em princípio, o ajuste de cada uma das seções poderia ser feito paralelamente, isto é, calculando, a cada iteração, os valores de $\mathbf{e}$ e $\mathbf{y}$ obtidos com a configuração atual, para então atualizar as duas seções utilizando as regras de ajuste mencionadas. Todavia, no algoritmo TJ, adota-se um procedimento de ajuste serial. No caso, em uma dada iteração, atualiza-se primeiramente a seção não-linear, de modo que os novos valores de $\mathbf{e}$ já são levados em conta na atualização da matriz $\mathbf{W}$. Este procedimento está sumariado na Tabela 4.1.

4.3.3 Outras técnicas para a separação de misturas PNL

Há outras técnicas, além do algoritmo TJ, destinadas à separação de misturas PNL. Por exemplo, em (Babaie-Zadeh, 2002), propôs-se uma abordagem geométrica para o problema em questão. Com o intuito de entendermos essa proposta, consideremos um cenário com duas fontes, com distribuição uniforme, misturadas por um sistema PNL. Na Figura 4.2, apresentamos as distribuições conjuntas dos sinais provenientes da saída da seção linear do sistema separador (Figura 4.2(a)) e da saída da seção não-linear (Figura 4.2(b)), ou seja, as misturas captadas pelos sensores. Note que, nesta situação, o efeito das funções não-lineares (no caso, utilizamos $f_i(r_i) = \sqrt[3]{r_i}$) fica evidente nas distribuições das misturas.

Na solução geométrica, a adaptação das seções não-linear e linear é feita separadamente, em contraposição ao algoritmo TJ. No caso, as funções não-lineares do sistema separador são ajustadas de forma a transformar a distribuição conjunta das misturas (Figura 4.2(b) em nosso exemplo) em um paralelogramo. Naturalmente, um primeiro passo a ser executado é estimar as bordas da densidade, para, em seguida, determinar, através de um processo iterativo, quais funções não-lineares mapeiam a forma geométrica obtida na estimação em um paralelogramo.



(a) Distribuição conjunta das saídas da seção linear.

(b) Distribuição conjunta das saídas da seção não-linear.

Figura 4.2: Distribuições conjuntas em um sistema misturador PNL.

Feita esta etapa, o problema resultante é semelhante a um do tipo linear, e, assim sendo, é possível utilizar qualquer uma das técnicas lineares apresentadas anteriormente.

A possibilidade de realizar separadamente o ajuste das seções não-linear e linear é, certamente, o principal atrativo da abordagem geométrica, tendo em vista o atual estágio da BSS linear. Por outro lado, há algumas limitações consideráveis deste método. A principal delas é a impossibilidade de operar em cenários com mais de duas fontes. Além disso, é necessário que as densidades de probabilidade das fontes sejam de suporte finito. Por fim, mesmo para fontes limitadas, o processo de estimação das bordas pode se tornar extremamente complexo quando as funções de densidade de probabilidade das fontes estiverem concentradas em torno de seu ponto médio.

Em um outro trabalho (Achard, 2003) sobre separação de misturas PNL, considerou-se uma medida de independência estatística alternativa à informação mútua, denominada medida de dependência quadrática. Para um vetor aleatório $\mathbf{y} = [y_1 \dots y_N]$, esta grandeza é definida do seguinte modo

$$Q(y_1, \dots, y_N) = \frac{1}{2} \int D_{\mathbf{y}}(Y_1, \dots, Y_N)^2 dY_1 \dots dY_N, \tag{4.30}$$

de forma que

$$D_{\mathbf{y}}(Y_1,\ldots,Y_N) = E\Big\{\prod_{n=1}^{N}\mathcal{K}\Big(Y_n - \frac{y_n}{\sigma_{y_n}}\Big)\Big\} - \prod_{n=1}^{N}E\Big\{\mathcal{K}\Big(Y_n - \frac{y_n}{\sigma_{y_n}}\Big)\Big\}, \qquad (4.31)$$

onde σ_{y_n} é uma constante e $\mathcal{K}(\cdot)$ corresponde a um kernel. Uma das propriedades interessantes da dependência quadrática é que esta medida é nula se e somente se as variáveis aleatórias forem estatisticamente independentes entre si (Achard, 2003).

A aplicação da medida de dependência quadrática à separação de misturas PNL é capaz de contornar o problema associado à estimação da informação mútua. Contudo, esta vantagem se dá às custas de uma nova dificuldade: a necessidade de calcular numericamente a integral presente na expressão (4.30). Um outro ponto ainda em aberto nesta abordagem diz respeito à escolha do kernel $\mathcal{K}(\cdot)$. Apesar do desempenho dos algoritmos desenvolvidos ser fortemente dependente do kernel adotado, a escolha desta função é feita de uma maneira *ad hoc*.

Em (Solé-Casals, Babaie-Zadeh, Jutten & Pham, 2003; Zhang & Chan, 2005), propõe-se uma estratégia para a separação de misturas PNL baseada no chamado processo de "gaussianização" de variáveis aleatórias (Chen & Gopinath, 2000). Assim como o critério de maximização da não-gaussianidade em BSS linear, esta estratégia pode ser compreendida pela óptica do teorema central do limite. Assumindo que as fontes são independentes, este teorema garante que, para um número suficientemente grande de fontes, as saídas da seção linear do sistema misturador PNL tendem a variáveis aleatórias gaussianas. Sob esta hipótese, é de se esperar que as saídas da seção não-linear não mais sejam gaussianas, posto que a transformação de uma variável gaussiana por uma função não-linear resulta numa variável não-gaussiana. Logo, na abordagem via gaussianização, busca-se justamente adaptar a seção não-linear de tal sorte que cada uma de suas saídas seja novamente mais próxima de uma gaussiana quanto possível.

O processo de gaussianização de uma variável aleatória pode ser realizado analiticamente (Chen & Gopinath, 2000). No contexto do problema de separação, dada uma mistura x_i, cuja distribuição cumulativa de probabilidade equivale a $F_{x_i}(x_i)$, a variável $e_i = g_i(x_i)$ apresenta distribuição normal na situação em que

$$g_i(x_i) = \Phi^{-1}(F_{x_i}(x_i)), \qquad (4.32)$$

onde Φ^{-1} representa a inversa da função cumulativa de uma variável aleatória gaussiana normalizada. Note que, considerando um sistema misturador estático, é necessário estimar as cumulativas das fontes uma única vez.

Após a determinação das funções não-lineares de acordo com a expressão (4.32), a seção linear pode ser determinada através de qualquer uma das técnicas lineares. Logo, assim como na solução geométrica, a abordagem via gaussianização permite dividir o treinamento de um sistema separador PNL em duas tarefas independentes. Um aspecto inusitado desta técnica é que, quanto maior o número de fontes, melhor será a inversão da seção não-linear e, conseqüentemente, o desempenho global da solução. Essa característica decorre do fato de que a hipótese assumida de gaussianidade das misturas lineares é tão mais verdadeira conforme aumenta-se o número de fontes. Na situação em que essas misturas distam de variáveis gaussianas, pode haver uma significativa distorção não-linear residual nas fontes recuperadas. Frente a esta limitação, este método é visto, sobretudo, como uma etapa de pré-processamento para a aplicação posterior de uma outra técnica de separação de misturas PNL (Solé-Casals, Babaie-Zadeh, Jutten & Pham, 2003).

Por fim, mencionamos um trabalho (Rojas, Puntonet, Rodríguez-Álvarez, Rojas & Martín-Clemente, 2004) em misturas PNL cujo principal objetivo é contornar o problema de convergência para ótimos locais no decorrer do ajuste do sistema separador. Embora não exista nenhuma prova teórica da existência desses mínimos, observou-se, em alguns trabalhos (Achard, 2003; Babaie-Zadeh, 2002), que, freqüentemente, a convergência de algoritmos de adaptação baseados no método do gradiente não é satisfatória, o que seria uma conseqüência da multimodalidade da função custo inerente ao problema. Diante disso, em (Rojas, Puntonet, Rodríguez-Álvarez, Rojas & Martín-Clemente, 2004), métodos de busca heurística, tais como os algoritmos evolutivos (no caso, o algoritmo genético padrão) e a técnica de *simulated annealing*, são considerados. De fato, um dos principais atrativos dessa classe de algoritmos é a capacidade de lidar com problemas de otimização multimodais.

Com relação à função custo adotada em tal trabalho, utilizou-se uma aproximação da informação mútua baseada nas séries de Gram-Charlier (Comon, 1994). No caso, a expressão resultante depende apenas de alguns momentos de ordem superior. Se, por uma lado, esta aproximação acarreta uma diminuição do

tempo gasto da etapa de estimação, por outro, é sabido que (Taleb & Jutten, 1999), diferentemente do caso linear, a estimativa grosseira das densidades de probabilidade das fontes recuperadas no caso não-linear, necessárias no cálculo da informação mútua, pode comprometer seriamente o desempenho do processo de separação. Uma das motivações de nossa proposta para o caso PNL, que será apresentada na seção seguinte, é justamente elaborar um critério de separação em que seja possível obter uma boa estimativa da informação mútua num tempo que, ademais, não seja proibitivo.

4.4 Uma Nova Proposta para a Separação de Misturas PNL

No início do Capítulo 3, foram descritos os passos presentes no desenvolvimento de uma técnica BSS. No decorrer deste mesmo capítulo, vimos que a instância linear da BSS não apresenta dificuldades significativas em nenhuma dessas etapas. Por exemplo, a definição da estrutura do sistema separador é direta nesta situação (basta escolher uma matriz). Além disso, a existência de critérios associados à independência estatística, nos quais nenhum tipo de estimação de densidade de probabilidade é exigida, torna o problema deveras simplificado. E, finalmente, a presença de ótimos locais apenas em alguns casos particulares abre caminho para a utilização de etapas de treinamento baseadas em métodos de busca local, tais como os do gradiente, do gradiente natural e os métodos de segunda ordem.

Por outro lado, na separação de misturas PNL, há dificuldades patentes em cada uma dessas etapas, a começar pela definição da estrutura do sistema separador. Neste caso, de modo geral, a situação de perfeita inversão do sistema misturador já não é mais atingível, posto que não há mais garantias de que as funções adotadas no processo de separação serão estruturalmente capazes de prover a inversão daquelas do sistema misturador, a menos que se conheça a estrutura da seção não-linear de tal sistema. Em uma primeira análise, uma solução para este problema seria adotar estruturas não-lineares capazes de aproximar uma ampla gama de mapeamentos,

como, por exemplo, redes neurais artificiais[2]. Todavia, deve-se lembrar que há uma restrição premente de monotonicidade sobre essas funções não-lineares na separabilidade do modelo PNL, e, assim sendo, soluções dessa sorte devem levar em conta um intrincado compromisso entre a flexibilidade e a monotonicidade das funções envolvidas.

No tocante ao critério de separação, conforme mencionado anteriormente, é fundamental adotar alguma medida que esteja diretamente ligada à independência estatística, tal como a informação mútua. Neste caso, torna-se necessário a execução de etapas de estimação dessa grandeza, ou mesmo de sua diferencial, o que dificulta a execução do passo seguinte: o treinamento da estrutura de separação. Diante disso, a busca por métodos eficientes de estimação de tais grandezas passa a ser um ponto relevante em separação de modelos PNL.

Também mencionamos que, no treinamento do sistema separador PNL, há ainda uma dificuldade relacionada à existência de ótimos locais na função custo derivada da informação mútua. Neste caso, a eficácia desta etapa está fortemente associada à capacidade do algoritmo de otimização de realizar uma busca das soluções ótimas. Logo, é de se esperar que a aplicação das técnicas de busca local não seja tão eficiente como no caso linear.

Em nosso entendimento, a presença de mínimos locais na função é, certamente, um dos principais problemas a serem superados no treinamento de modelos PNL, haja visto que tal complicador afeta a maioria das soluções até então propostas. Mesmo em (Rojas, Puntonet, Rodríguez-Álvarez, Rojas & Martín-Clemente, 2004), onde a otimização é conduzida através de um algoritmo de busca global, uma recuperação satisfatória das fontes não é garantida devido, principalmente, às aproximações de informação mútua empregadas. Além do mais, existem técnicas em computação evolutiva, como, os sistemas imunológicos artificiais (Castro & Zuben, 2002), que apresentam um potencial maior de busca global se comparados com o algoritmo genético padrão, utilizado em tal trabalho.

No presente trabalho, a elaboração de uma nova técnica capaz de contornar as dificuldades associadas à presença dos mínimos locais foi considerada. Para tal fim, atuamos tanto na aplicação de uma técnica de otimização multimodal, no caso um

[2]Em (Taleb & Jutten, 1999), isto foi implementado a partir de redes MLP

algoritmo oriundo dos sistemas imunológicos artificiais, quanto na busca de uma estimador para a informação mútua, que fundamentou-se nas chamadas estatísticas de ordem. Na seqüência, daremos início à explanação de nossa proposta, começando pela descrição desses dois conceitos-chave.

4.4.1 Estimação da entropia através de estatísticas de ordem

Primeiramente, vejamos os aspectos básicos do emprego de estatísticas de ordem, conceito este relacionado com o ordenamento de variáveis aleatórias. Para tal, considere um conjunto de variáveis aleatórias $R_1, R_2, \ldots, R_T$ e seja uma dada realização deste conjunto representada por $r_1, r_2, \ldots, r_T$.

Em algumas situações, pode ser interessante observar o comportamento estatístico do valor máximo referente às realizações desse conjunto, dado por $\max(r_1, r_2, \ldots, r_T)$. É importante observar que este máximo define uma nova variável aleatória. Analogamente, poderíamos estar interessado nos valores mínimos, ou mesmo no segundo maior valor desse conjunto. A generalização desta idéia nos leva ao conceito de estatística de ordem, ou seja, a partir do conjunto de variáveis aleatórias $R_1, R_2, \ldots, R_T$, é possível definir um novo conjunto de T variáveis aleatórias ordenadas do seguinte modo

$$R_{(1:T)} \leq R_{(2:T)} \leq \ldots \leq R_{(T:T)}, \tag{4.33}$$

onde $R_{(t:T)}$ corresponde à t-ésima estatística de ordem, e representa a variável aleatória associada à t-ésima posição do conjunto original após a ordenação.

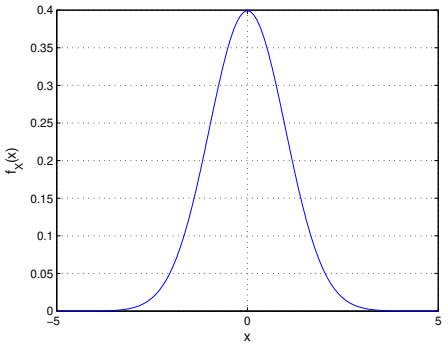
O conceito de estatística de ordem também pode ser definido para uma única variável aleatória (Even, 2003). De fato, cada realização de um conjunto de T variáveis aleatórias com distribuições de probabilidade idênticas pode ser vista como T realizações de uma única variável aleatória. Para nossos fins, apenas a definição para uma variável será necessária, pois estamos interessados nas estimativas das entropias marginais de sinais aleatórios. Logo, neste caso, o número de realizações T corresponde ao número de amostras disponíveis de um dado sinal.

Um interessante propriedade das estatísticas de ordem é a sua relação com a função quantil. Esta função corresponde à inversa da função cumulativa de probabilidade (Papoulis, 1993) de uma dada variável aleatória, e, de modo a ilustrar

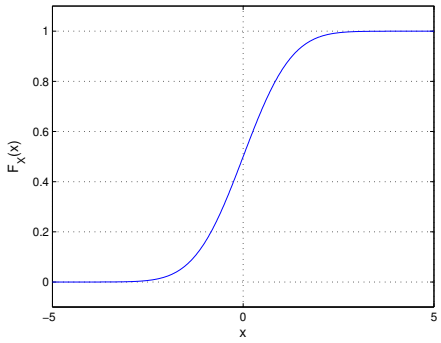
este conceito, a Figura 4.3 apresenta a densidade, a cumulativa e a função quantil de uma variável aleatória gaussiana. Conforme descrito em (Even & Moisan, 2005), a relação entre as estatísticas de ordem e a função quantil de uma variável aleatória R, representada por $Q_R(\cdot)$, é dada por

$$E\{R_{(t:T)}\} = Q_R\Big(\frac{t}{T+1}\Big), \quad t = 1, \ldots, T. \tag{4.34}$$

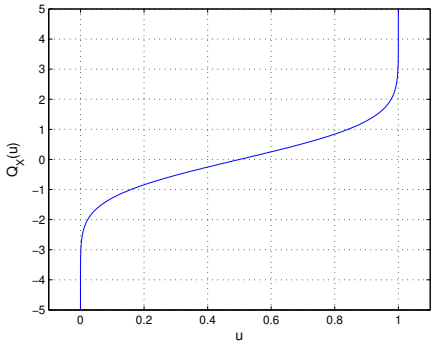
Esta expressão indica que as estatísticas de ordem carregam informações sobre a lei estatística que rege uma determinada variável aleatória. Essa constatação é a base do método de estimação de entropia adotado em nosso trabalho.



(a) Função densidade de probabilidade.



(b) Cumulativa.



(c) Função quantil.

Figura 4.3: Caracterização estatística de uma variável aleatória gaussiana.

Em (D.-T. Pham, 2000), demonstra-se que a entropia de uma variável aleatória Y pode ser expressa através da função quantil $Q_Y(\cdot)$, de acordo com a seguinte

expressão

$$H(Y) = \int_0^1 \log Q'_Y(u) du. \tag{4.35}$$

A partir de uma resolução numérica desta integral, obtém-se que

$$H(Y) \approx \sum_{l=2}^{L} \log \left[\frac{Q_Y(u_l) - Q_Y(u_{l-1})}{u_l - u_{l-1}} \right] \frac{u_l - u_{l-1}}{u_L - u_1}, \tag{4.36}$$

onde $\{u_1, \ldots, u_L\}$ é um conjunto de números crescentes no intervalo[3] $]0, 1[$. Esta expressão, conjuntamente com (4.34), sugere que é possível estimar a entropia através das estatísticas de ordem.

No caso em que há apenas uma realização de T amostras da variável aleatória Y, a estimação da função quantil pode ser feita com base na realização correspondente das estatísticas de ordem, ou seja

$$Y_{(t:T)} \approx Q_Y\Big(\frac{t}{T+1}\Big), \quad t = 1, \ldots, T. \tag{4.37}$$

Além do mais, é possível estimar a função quantil em pontos que não pertencem ao conjunto $\frac{t}{T+1}$ realizando uma interpolação linear, por exemplo. A inserção dessas duas aproximações na expressão (4.36) resulta num estimador de entropia simplificado cujo modo de operação se baseia em processos de ordenação e interpolação de dados.

Em comparação com outros métodos, a estimação da entropia via estatísticas de ordem apresenta um bom compromisso entre acurácia e complexidade computacional. Por exemplo, a estimação via métodos de *kernel* (Silverman, 1986) é capaz de fornecer resultados consideravelmente precisos, mas, requer um denso processamento computacional. Além disso, há, nesta estratégia, uma dificuldade relativa à determinação dos parâmetros dos *kernels* utilizados. No caso das estatísticas de ordem, é necessário apenas definir o passo de integração referente à expressão (4.36)[4]. Um outro método de estimação de entropia fundamenta-se

[3] Consideramos também que $u_i - u_{i-1}$ é constante para todo i.

[4] No decorrer de nosso trabalho, constatamos que a estimação via estatísticas de ordem está relacionada com o clássico método de histograma, porém com intervalos variáveis. No Apêndice B, mostramos que há uma relação entre o passo de integração definido e o tamanho de cada um desses intervalos.

na expansão em séries de Gram-Charlier de uma variável aleatória (Comon, 1994). Neste caso, a estimativa depende apenas dos momentos de terceira e quarta ordem, o que torna este processo extremamente rápido. Todavia, esta simplificação prática surge às custas de uma perda de precisão no processo de estimação.

A Figura 4.4 ilustra algumas das observações do parágrafo anterior. Neste gráfico, as estimativas da entropia de uma variável aleatória uniforme no intervalo $[-\sqrt{3}, \sqrt{3}]$, provenientes dos métodos discutidos, são apresentadas. No caso da estimação via métodos de *kernel*, consideramos *kernels* gaussianos com desvio padrão representado por σ. Fica evidente que, neste caso, há um compromisso entre o número de amostras requerido e a polarização da estimativa, que, por sua vez, está diretamente relacionado à escolha do desvio padrão. Com relação à estimação via séries de Gram Charlier, nota-se que, de fato, apenas um número reduzido de amostras é necessário. Porém, a estimativa obtida apresenta uma significativa polarização.

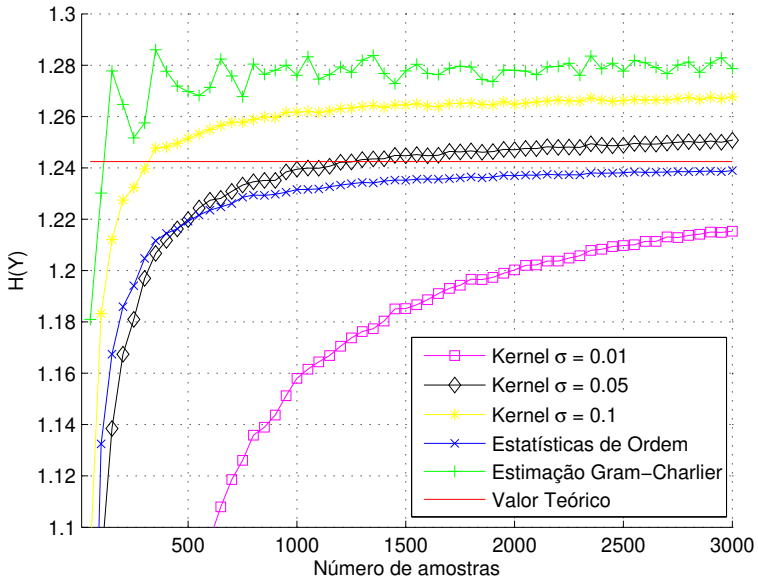


Figura 4.4: Estimação da entropia de uma variável aleatória uniforme.

4.4.2 Otimização a partir de uma rede imunológica artificial

O outro pilar de nossa proposta consiste na realização da etapa de treinamento do sistema separador através de um tipo de rede imunológica artificial propícia para tarefas de otimização, denominada opt-aiNet (*Artificial Immune Network for Optimization*) (Castro & Timmis, 2002). Este algoritmo, cujo modo de operação é inspirado no funcionamento do sistema imunológico das espécies superiores, pertence à classe dos algoritmos evolutivos (Bäck, Fogel & Michalewicz, 2000), e é dotado de elementos de busca local e global.

Em Attux (2005), realizou-se um amplo estudo sobre a aplicabilidade dos algoritmos evolutivos, particularmente da opt-aiNet, a alguns problemas típicos de processamento de sinais. Neste trabalho, verificou-se que, ao tornar menor a probabilidade de convergência para mínimos locais ruins em relação às técnicas de busca local, o uso desse ferramental proporciona melhorias significativas em cenários nos quais a etapa de treinamento culmina num problema de otimização multimodal. Além do mais, alguns resultados presentes neste trabalho sugerem que a opt-aiNet, por apresentar um maior potencial de busca global se comparada ao consagrado algoritmo genético padrão, apresenta, no mínimo, um desempenho equivalente ao desse clássico método.

Antes de prosseguirmos com a descrição do algoritmo opt-aiNet, é conveniente apresentar alguns conceitos típicos em computação evolutiva. O primeiro deles é o de população. Uma população corresponde a um conjunto de indivíduos, de modo que cada indivíduo representa uma solução candidata do problema de otimização em questão. Diz-se que os algoritmos evolutivos realizam uma busca populacional, pois, diferentemente de métodos como o do gradiente, diversas soluções candidatas são ajustadas a cada iteração, ou, no jargão desta área, a cada geração. Um outro conceito fundamental na computação evolutiva é o de *fitness* de um indivíduo. Esta medida tem como objetivo avaliar a qualidade de um determinado indivíduo, o que é fundamental para determinar o seu futuro na população. Assim sendo, é de se esperar que o *fitness* esteja diretamente ligado à função custo a ser otimizada.

Um dos motivos que explicam o diminuto risco de convergência sub-ótima no algoritmo opt-aiNet é a existência de um mecanismo de manutenção da diversidade da população, isto é, busca-se evitar uma população na qual os indivíduos sejam

"semelhantes" entre si. Na otimização de uma função custo com acentuado caráter multimodal, uma população semelhante poderia ser incapaz de explorar todos os diversos picos e vales existentes no problema. Uma outra peculiaridade da opt-aiNet é a existência de um mecanismo de controle do tamanho da população. Este ajuste dinâmico, caracterizado por podas e inserções de indivíduos, permite a obtenção de uma solução parcimoniosa em termos do tamanho da população final. Essas características estão descritas na Tabela 4.2, que apresenta um resumo do algoritmo opt-aiNet.

Em suma, há duas etapas principais na opt-aiNet. A primeira delas, relativa ao passo 2 da Tabela 4.2, é caracterizada pela realização de uma busca local cujo objetivo é melhorar a qualidade da população através de procedimentos de clonagem, mutação e seleção. O passo 3 diz respeito aos mecanismos que enfatizamos no parágrafo anterior. O controle do tamanho da população e a busca por um bom nível de diversidade são feitos através da supressão de indivíduos redundantes e da inserção de novos indivíduos na população.

No algoritmo opt-aiNet, há alguns parâmetros que devem ser previamente definidos: o número de clones (Nc), o limiar de supressão (σ_s), e o parâmetro β relacionado à mutação. Não há uma regra objetiva para o ajuste desses parâmetros, de modo que, neste trabalho, o fizemos após a realização de algumas simulações preliminares nos cenários analisados. Com relação ao tamanho da população inicial, o ajuste deste valor não é tão decisivo quanto o dos outros mencionados, em vista da existência do controle dinâmico do tamanho da população. Por fim, enfatizamos que, no presente trabalho, adotamos como critério de parada a realização de um número predefinido de iterações.

4.4.3 Comentários sobre a solução proposta

Feita a apresentação dos dois elementos fundamentais de nossa proposta, voltemos nossas atenções para algumas questões de ordem prática. Um primeiro ponto diz respeito à representação dos indivíduos. Dado que, em nosso problema, estamos interessados em ajustar os parâmetros das seções linear (matriz $\mathbf{W}$) e não-linear do sistema misturador, um indivíduo corresponde a um vetor de números reais contendo todos esses elementos.

Tabela 4.2: Descrição do algoritmo opt-aiNet.

1. Inicialização: geração aleatória da população;
2. Processo iterativo: até que o critério de parada seja atingido, faça:
   (a) Determinação do *fitness* de cada indivíduo;
   (b) Clonagem: produza um número (N_C) de cópias (clones) para cada indivíduo da indivíduo;
   (c) Mutação: introduza, em cada clone, um distúrbio (mutação) inversamente proporcional ao *fitness* de seu seus "indivíduos pais" (estes permanecem inalterados). A mutação segue a seguinte regra

$$c' = c + \alpha N(0,1), \text{ with } \alpha = \beta^{-1} \exp\left(-f^*\right), \tag{4.38}$$

   onde $N(0,1)$ corresponde a uma variável aleatória normal, c' e c representam o clone modificado pela mutação e o original, respectivamente; β é uma parâmetro de controle, e f^* é o *fitness* do indivíduo pai.
   (d) Determinação do *fitness* de todos os clones modificados pela mutação;
   (e) Seleção: Selecione o melhor indivíduo de cada grupo formado pelo indivíduo original e seus clones modificados (formação de uma nova população);
   (f) Convergência local: calcule o *fitness* médio da população. Caso não haja uma variação significativa deste *fitness* médio em relação à geração anterior, então vá em frente. Caso contrário, volte para o passo 2;
3. Interações na população;
   (a) Determine a afinidade (grau de similaridade medido pela distância euclidiana) entre todos os indivíduos;
   (b) Suprima todos os indivíduos exceto aqueles que, dentro de um certo par com afinidade menor que uma valor σ_s (limiar de supressão), tenham maior *fitness*. Determine o tamanho da nova população;
   (c) Introduza novos indivíduos na população, gerados aleatoriamente, e retorne ao passo 2

Ainda relativo à representação dos indivíduos, não podemos esquecer que existe uma restrição de monotonicidade das funções não-lineares do sistema separador. Em nosso trabalho, utilizamos somente funções polinomiais, e, de modo a assegurar a validade de tal restrição, consideramos apenas polinômios com potências ímpares. Além disso, é necessário garantir que os coeficientes desses polinômios sejam todos positivos, o que foi feito através da aplicação de uma função exponencial aos elementos de um dado indivíduo relativos à seção não-linear.

No algoritmo opt-aiNet, busca-se a maximização do *fitness*. Porém, em nosso problema, estamos interessados na minimização da informação mútua. Além disso, é desejável que o *fitness* assuma sempre valores positivos. Estas considerações podem ser levadas em conta definindo o *fitness* do seguinte modo

$$f^* = \exp\left(\frac{1}{\sum_{i=1}^{N} H(y_i) - \log|\det \mathbf{W}| - E\{\log\prod_{i=1}^{N}|g_i'(x_i)|\}}\right). \tag{4.39}$$

O denominador desta expressão provém de 4.25, porém é desconsiderada a entropia conjunta das misturas, já que esta medida não depende dos parâmetros do sistema separador. Com base na discussão da Seção 4.4.1, as entropias marginais de y_i são estimadas a partir das expressões (4.36) e (4.34).

4.4.4 Resultados

Com o intuito de avaliar o desempenho da técnica proposta, realizamos, em um primeiro momento, simulações em dois cenários distintos. Na seqüência, apresentamos e discutimos os resultados obtidos nesses casos.

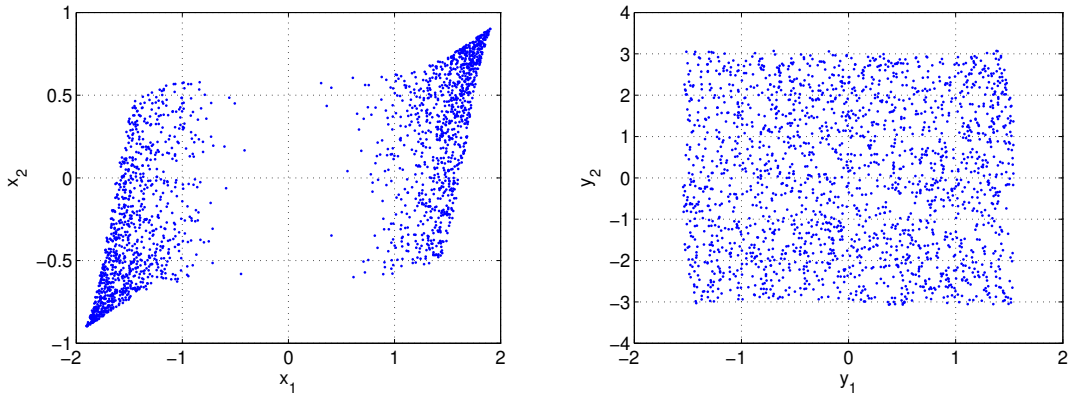
Primeiro cenário

Num primeiro cenário, consideramos a separação de duas fontes uniformemente distribuídas no intervalo $[-1, 1]$, e misturadas pelo seguinte sistema PNL

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 \\ 0.5 & 1 \end{bmatrix} \text{ e } \begin{array}{rcl} f_1(e_1) & = & \tanh(2e_1) \\ f_2(e_2) & = & 2\sqrt[5]{e_2} \end{array}. \tag{4.40}$$

Na seção não-linear do sistema separador, utilizamos polinômios do tipo $g(x_i) = ax_i^5 + bx_i^3 + cx_i$. Com relação aos parâmetros do algoritmo opt-aiNet, o seguinte conjunto foi definido: $N_C = 7$, $\beta = 60$ and $\sigma_s = 2$.

Na Figura 4.5(a), a distribuição conjunta das misturas é apresentada. Para esta situação, considerou-se, na etapa de treinamento, um conjunto de 2000 amostras dos sinais misturados. Após a realização do treinamento do sistema separador através da opt-aiNet, considerando 10000 gerações, foi possível obter boas estimativas das fontes, fato este que pode ser constatado na Figura 4.5(b), que mostra que a distribuição conjunta das estimativas obtidas ficou próxima à de uma uniforme. Ainda nesta figura, é possível observar uma certa distorção residual não-linear. Todavia, este fenômeno é, sobretudo, decorrente da limitação estrutural inerente ao problema, dado que é impossível inverter a ação da tangente hiperbólica através de um polinômio.



(a) Distribuição conjunta das misturas.

(b) Distribuição conjunta das estimativas obtidas.

Figura 4.5: Resultados - Primeiro cenário.

Segundo cenário

Em um segundo cenário, consideramos a tarefa de separar três fontes uniformemente distribuídas no intervalo $[-1, 1]$. No caso, consideramos o seguinte

sistema misturador PNL

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 & 0.5 \\ 0.5 & 1 & 0.4 \\ 0.4 & 0.6 & 1 \end{bmatrix} \quad \text{e} \quad \begin{aligned} f_1(e_1) &= 2\sqrt[3]{e_1} \\ f_2(e_2) &= 2\sqrt[3]{e_2} \\ f_3(e_3) &= 2\sqrt[3]{e_3} \end{aligned} . \tag{4.41}$$

Tanto os polinômios utilizados na seção não-linear quanto os parâmetros da opt-aiNet foram os mesmos do experimento anterior. Novamente, obtivemos boas estimativas das fontes após a realização do treinamento proposto para o sistema separador. Na Tabela 4.3, os valores de EQM de cada fonte e de sua respectiva estimativa (y_i) são apresentados.[5] Na Figura 4.6, é exibida uma janela temporal do sinal fonte e de sua respectiva estimativa (após normalização), ambas associadas ao maior EQM.

Tabela 4.3: EQM - segundo cenário

EQM($\times 10^{-2}$)	y_1	y_2	y_3
opt-aiNet	3,35	1,10	1,50

4.4.5 Uma crítica à solução proposta

Os resultados apresentados na seção anterior indicam que a técnica proposta é capaz de prover uma separação satisfatória das fontes. No entanto, há um problema que ainda não mencionamos, relacionado ao custo computacional de nossa solução. Infelizmente, em um algoritmo evolutivo como a opt-aiNet, a robustez à convergência para mínimos locais ocorre às custas de um aumento significativo da carga computacional exigida, tanto em termos de memória quanto de processamento. Naturalmente, este problema se torna ainda mais complicado com o aumento da dimensão do espaço de busca do problema de otimização a ser resolvido.

No problema de separação de misturas PNL, se considerarmos que as funções da seção não-linear do sistema separador são todas iguais, de modo que o número de parâmetros de cada uma delas seja C_p, a dimensão do espaço de busca é dada

[5] Devido à ambiguidade de ganho de escala, é necessário, na estimação do EQM, normalizar a variância dos sinais envolvidos.

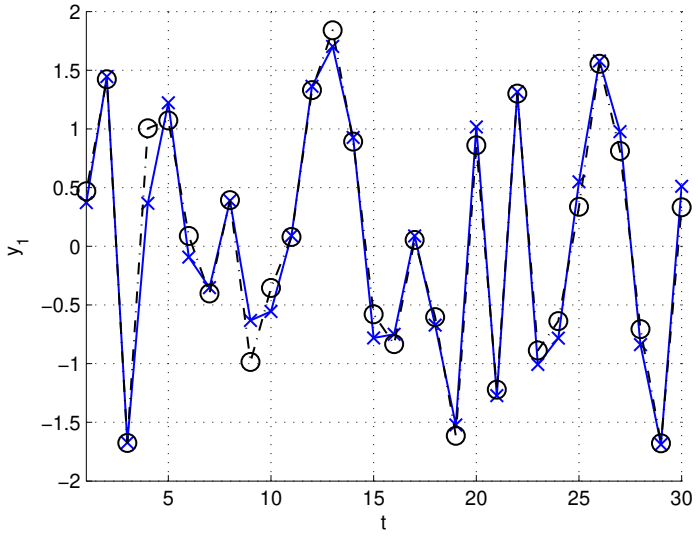


Figura 4.6: Segundo cenário - fonte $(-)$ e sua estimativa $(\cdot - \cdot)$.

por $N^2 + N \cdot C_p$. Esta relação evidencia que, para situações em que o valor de C_p é baixo, como nos casos estudados na Seção 4.4.4, o crescimento do espaço de busca com o aumento ao número de fontes está majoritariamente relacionado ao termo quadrático, que, por sua vez, diz respeito aos elementos da matriz $\mathbf{W}$. Por exemplo, caso desejássemos separar quatro fontes conservando os parâmetros adotados no segundo cenário analisado, teríamos um significativo aumento do espaço de busca de 18 para 28 parâmetros, sendo que apenas 3 são os novos parâmetros acrescidos na seção não-linear.

Essa discussão sugere uma nova estratégia na qual o ajuste das seções não-linear e linear seja conduzido separadamente, haja visto que esta última etapa poderia, em tese, ser realizada de modo eficiente através de algum algoritmo de BSS linear. De fato, já mencionamos que há algumas propostas que operam de acordo com este princípio, como é o caso da abordagem geométrica e da gaussianização. A principal dificuldade neste tipo de esquema refere-se à concepção de uma função custo para o ajuste da seção não-linear que não dependa dos parâmetros da matriz $\mathbf{W}$. O problema no caso é que a condição de separabilidade do modelo está ligada às saídas do sistema separador, e não há como saber, de uma maneira sistemática,

como esta condição se refletiria nas saídas da seção não-linear.

De modo a melhorar o desempenho da solução proposta, buscamos, motivados pelas razões expostas acima, dividir as etapas de ajuste das seções do sistemas separador, porém, mantendo uma função custo "global", fundamentada na informação mútua entre as estimativas $\mathbf{y}$. Nesta nova versão, o ajuste das funções não-lineares continua sendo realizado pelo algoritmo opt-aiNet. Por outro lado, a matriz $\mathbf{W}$ é adaptada através da aplicação do algoritmo FastICA[6] [7]. Vejamos como isto é feito.

Primeiramente, com relação à Seção 4.4.3, há uma modificação referente à representação dos parâmetros de otimização. Nesta nova estratégia, um indivíduo da população contém apenas os parâmetros das funções não-lineares. Além disso, há uma pequena alteração no cálculo do *fitness*. Dado um indivíduo (um conjunto de funções não-lineares), aplica-se o algoritmo FastICA sobre as saídas da seção não-linear (gerada deste conjunto) e, em seguida, calcula-se o valor do *fitness* de acordo com (4.39), porém, considerando as saídas geradas pelo algoritmo FastICA. Este procedimento está resumido na Tabela 4.4.

Na estratégia apresentada, o número de parâmetros do espaço de busca passa a ser $N \cdot C_p$. Evidentemente, esta redução se dá às custas de uma maior complexidade computacional por iteração, pois, nesta nova situação, cada cálculo do *fitness* exige a execução do algoritmo FastICA. Contudo, como veremos na próxima seção, esta nova estratégia acarreta em melhorias significativas de desempenho, principalmente em cenários com um maior número fontes.

[6] Optamos por este algoritmo em virtude de sua rápida convergência e de seu baixo custo computacional, como pudemos verificar no Capítulo 3. A versão utilizada está descrita na Tabela 3.1.

[7] Após o desenvolvimento de nossa solução, verificamos que um método proposto recentemente (Górriza et al., 2006) fundamenta-se numa idéia semelhante à nossa. Em tal trabalho, utiliza-se uma técnica baseada na conjunção de um algoritmo genético padrão e do FastICA. Todavia, há uma diferença relevante no que diz respeito ao método de estimação de entropia empregado.

Tabela 4.4: Modificação no cálculo do *fitness*.

1. Para um dado indivíduo (conjunto de parâmetros das funções não-lineares), calcule as saídas da seção não-linear $\mathbf{e}$.
2. Aplique o algoritmo FastICA nos sinais $\mathbf{e}$;
3. O *fitness* de um dado indivíduo obtido pela expressão (4.39) calculada para as saídas fornecidas pelo FastICA.

4.4.6 Avaliação da solução modificada

De modo a avaliar os ganhos obtidos com a introdução do algoritmo FastICA como elemento de busca local (doravante, denominaremos esta estratégia por opt-aiNet/FastICA), realizamos simulações em dois cenários distintos. Assim como na Seção 4.4.4, adotamos polinômios do tipo $g(x_i) = ax_i^5 + bx_i^3 + cx_i$ na seção não-linear do sistema separador.

Primeiro cenário

Em um primeiro cenário, consideramos o problema de separação de 3 fontes, uniformemente distribuídas entre $[-1, 1]$, misturadas pelo seguinte misturador PNL

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 & 0.5 \\ 0.5 & 1 & 0.4 \\ 0.4 & 0.6 & 1 \end{bmatrix} \quad \text{e} \quad \begin{array}{rcl} f_1(e_1) & = & \sqrt[3]{e_1} \\ f_2(e_2) & = & \sqrt[3]{e_2} \\ f_3(e_3) & = & \sqrt[3]{e_3} \end{array} . \tag{4.42}$$

No processo de separação, consideramos 1200 amostras das misturas. Além disso, para a versão opt-aiNet/FastICA, o seguinte conjunto de parâmetros foi definido $N_C = 5$, $\beta = 50$ e $\sigma_s = 3$. Também buscamos verificar neste cenário o desempenho da solução baseada na "opt-aiNet pura" (a qual denominaremos por opt-aiNet). No caso, o seguinte conjunto $N_C = 7$, $\beta = 60$ e $\sigma_s = 3$ foi considerado para esta técnica.

Com o intuito de comparar essas duas estratégias, realizamos 20 experimentos. Diante da diferença entre as dimensões do espaço de busca em cada caso,

Tabela 4.5: Velocidade de convergência - primeiro cenário

	opt-aiNet	opt-aiNet/FastICA
Velocidade de convergência (iteração)	10500	1400
Tempo médio por iteração (ms)	80	182
Tempo médio de convergência (min)	14	4,5

consideramos, como critério de parada, a execução de 6000 iterações para a opt-aiNet/FastICA e de 15000 para a opt-aiNet. Mesmo com este maior valor, observamos que, em 5 realizações, a opt-aiNet não convergiu satisfatoriamente (não consideraremos estes experimentos na comparação entre as técnicas). Por outro lado, a opt-aiNet/FastICA proporcionou bons sistemas separadores em todas as realizações.

Na Tabela 4.5, alguns tempos relativos à convergência dos dois algoritmos são apresentados[8]. Note o leitor que, como havíamos mencionado, a introdução do FastICA realmente implica num aumento do tempo médio de cada iteração. No entanto, a redução do número de parâmetros de busca propiciado por esta estratégia faz com que o número de iterações necessárias para a sua convergência seja consideravelmente menor em relação à opt-aiNet. Deste modo, em um balanço final, a opt-aiNet/FastICA apresenta, para o caso analisado, um tempo de convergência menor que o da opt-aiNet.

Também analisamos o EQM entre as fontes e as estimativas obtidas, apresentados na Tabela 4.6. Notamos que a técnica opt-aiNet/FastICA é capaz de obter boas estimativas das fontes, sendo estas melhores que as fornecidas pela opt-aiNet. Ressaltamos que seria possível melhorar o desempenho da opt-aiNet considerando um maior número de iterações. Porém, isso requereria um tempo ainda maior de convergência.

[8] Todas as simulações foram realizadas em um Athlon64 3000+ com 1GB de memória RAM. Além disso, o sistema operacional utilizado foi o Windows XP e os algoritmos foram implementados no Matlab 7.1.

Tabela 4.6: EQM - primeiro cenário.

EQM ($\times 10^{-3}$)	$\hat{s}_1$	$\hat{s}_2$	$\hat{s}_3$
opt-aiNet	7.33	9.50	6.45
opt-aiNet/FastICA	2.19	1.58	1.73

Tabela 4.7: Média dos EQMs no segundo cenário

EQM ($\times 10^{-3}$)	$\hat{s}_1$	$\hat{s}_2$	$\hat{s}_3$	$\hat{s}_4$
opt-aiNet/FastICA	1.59	2.18	1.69	2.00

Segundo cenário

Num segundo cenário, analisamos o desempenho da opt-aiNet/FastICA numa tarefa mais complicada, relativa à separação de 4 fontes, uniformemente distribuídas entre $[-1, 1]$. O sistema misturador PNL considerado neste caso foi o seguinte

$$\mathbf{A} = \begin{bmatrix} 1 & 0.7 & 0.5 & 0.7 \\ 0.5 & 1 & 0.7 & 0.4 \\ 0.4 & 0.6 & 1 & 0.7 \\ 0.8 & 0.7 & 0.5 & 1 \end{bmatrix} \quad \text{e} \quad \begin{aligned} f_1(e_1) &= \sqrt[3]{e_1} \\ f_2(e_2) &= \sqrt[3]{e_2} \\ f_3(e_3) &= \sqrt[3]{e_3} \\ f_4(e_4) &= \sqrt[3]{e_4} \end{aligned} . \tag{4.43}$$

Neste caso, o seguinte conjunto de parâmetros da opt-aiNet foi adotado: $N_c = 5$, $\beta = 50$ e $\sigma_s = 3$.

Após a realização de 20 experimentos, sendo o número de iterações em cada um foi de 8000, verificamos que em apenas um deles não se obteve uma convergência satisfatória. Para as 19 simulações que convergiram, a média dos EQM associado à recuperação de cada uma das fontes é apresentada na Tabela 4.7. Novamente, a aplicação do algoritmo opt-aiNet/FastICA foi capaz de prover boas estimativas das fontes. É importante frisar que neste cenário, a aplicação da opt-aiNet pura seria impraticável, dado o elevado número de parâmetros a ser ajustados neste problema.

4.5 Sumário

No presente capítulo, tratou-se o problema de separação cega de fontes misturadas por um sistema não-linear. Inicialmente, foi visto que, de modo geral, a

aplicação da ICA não mais garante a recuperação das fontes. Na seqüência, alguns algoritmos para o caso geral do problema de NBSS foram brevemente descritos. Um importante assunto relacionado a este problema, a separação de misturas PNL foi especificamente abordado neste capítulo, e algumas soluções para este caso particular foram discutidas.

Através do emprego da opt-aiNet e de um estimador de entropia baseado nas estatísticas de ordem, foi possível desenvolver um algoritmo para o treinamento de sistemas PNL capaz de contornar problemas relativos à convergência para mínimos locais. Além disso, realizamos uma modificação na solução proposta com o objetivo de torná-la menos complexa e, conseqüentemente, mais apta para operação em cenários com um maior número de fontes. Essa melhoria foi obtida através do desenvolvimento de uma técnica híbrida, na qual a seção linear é ajustada pelo algoritmo FastICA e a seção não-linear é adaptada pela opt-aiNet. Através da execução de simulações, verificamos a viabilidade de nossa proposta.

Capítulo 5

Conclusões e Perspectivas

Nesta dissertação, buscamos investigar as principais estratégias existentes no problema de separação cega de fontes. Nosso estudo contemplou os casos em que o sistema misturador é linear ou não-linear, sem memória e com igual número de sensores e fontes.

Em um primeiro momento, descrevemos o problema em questão e apresentamos algumas de suas principais aplicações. Além disso, constatamos que as soluções para o caso linear são as mais utilizadas em problemas práticos, o que certamente é uma conseqüência do grau de maturação atingido neste caso. Contudo, observamos um interesse crescente da comunidade pelos casos em que o processo de mistura é modelado por sistemas mais complexos, tais como aqueles que têm características não-lineares. Aliás, foi exatamente esta tendência que nos motivou a incluir esse tema em nosso estudo.

No que concerne ao nosso estudo dos principais critérios em BSS linear, vimos que existem relações diretas entre eles. Isso fica ainda mais evidente ao analisarmos os algoritmos gerados com base nesses critérios. Mais precisamente, é sempre possível identificar nesses algoritmos um elemento linear, a matriz que representa o sistema separador, e um elemento não-linear, útil na etapa de treinamento. Em (Kofidis, 2001; T.-W. Lee, Girolami, Bell & Sejnowski, 2000), por exemplo, fala-se em um paradigma universal de critérios e algoritmos para o caso de BSS linear. Um outro ponto interessante é que essa configuração universal parece motivar diversos autores

a classificar essas técnicas como algoritmos neurais, certamente devido à semelhança entre essa estrutura e uma rede neural (lembremos que, no critério Infomax, isso é patente).

Com relação à análise de desempenho dos algoritmos de BSS linear com modo de operação em batelada, verificamos que, na grande maioria dos quesitos considerados, os algoritmos testados apresentaram um comportamento parecido. A única exceção foi o algoritmo FastICA, que apresentou uma velocidade de convergência consideravelmente maior que a das outras técnicas. No caso dos algoritmos adaptativos, o algoritmo RLS desenvolvido para a otimização do critério de NPCA foi a técnica que apresentou a maior velocidade de convergência.

No contexto da NBSS, discutimos sobre a impossibilidade, de um modo geral, do uso da ICA neste problema. Por outro lado, apresentamos a classe de modelos PNL, na qual ainda é possível separar as fontes através da recuperação da independência. Este caso apresenta algumas dificuldades que não estão presentes nas técnicas lineares, tais como a necessidade de esquemas de estimação da informação mútua e a possibilidade de convergência para mínimos locais.

Como contribuição original, foi proposta no presente trabalho uma nova técnica para separação de misturas PNL, concebida com o propósito de prover um algoritmo robusto em relação à convergência para mínimos locais ruins. Para tal, utilizamos como técnica de treinamento um tipo de algoritmo evolutivo inspirado no funcionamento do sistema imunológico, a opt-aiNet, tendo em vista sua capacidade de realizar a otimização em superfícies multimodais.

Num primeiro momento, realizamos o treinamento do sistema separador PNL utilizando exclusivamente opt-aiNet. A despeito dos bons resultados obtidos, há uma significativa limitação nesta proposta no que se refere à sua complexidade computacional elevada. De fato, a robustez de uma técnica evolutiva surge às custas de uma maior demanda computacional, em comparação com os algoritmos baseados no gradiente, por exemplo. Assim, vimos que, para cenários com um número de fontes elevado, a aplicação desta técnica pode ser inviável. Em nosso trabalho, buscamos atenuar esta dificuldade através de uma nova versão, na qual a busca pela seção linear do sistema separador é feita pelo algoritmo FastICA. Através de algumas simulações, constatamos que, de fato, a introdução deste elemento acarreta

um significativo ganho de desempenho.

É importante frisar que mesmo a versão híbrida opt-aiNet/FastICA ainda apresenta um custo computacional significativo. Contudo, há de se ressaltar que a separação de misturas PNL requer processos de estimação da informação mútua entre as saídas do sistema separador, o que, neste caso particular, culmina num problema de estimação de entropia. Assim sendo, mesmo numa técnica baseada no gradiente, tal como o algoritmo de Taleb e Jutten, um denso processamento computacional é requerido, haja visto a necessidade de se estimar as entropias marginais a cada iteração. Além disso, ressaltamos que a ampla maioria das aplicações em BSS não requerem um processamento em tempo real, o que abre caminho para o uso da técnica proposta em tais cenários práticos.

Na concepção de nossa proposta, buscamos um estimador de entropia que apresentasse um bom compromisso entre acurácia e simplicidade computacional. No caso, adotamos um método de estimação baseado nas estatísticas de ordem. Tal método já havia sido aplicado ao problema de separação de fontes linear. Uma questão interessante que constatamos após a derivação de nossa técnica foi que este tipo de estimador se assemelha aos métodos baseados na estimação por histograma, porém com a diferença de que, no estimador utilizado, os intervalos dos *bins* são variáveis, conforme pode ser visto no Apêndice B.

Para finalizar esta dissertação, gostaríamos de expor algumas perspectivas para trabalhos futuros. Primeiramente, em um visão mais geral, acreditamos que, além do caso não-linear da BSS, as outras extensões do problema de separação (sistemas com memória e modelos sub-parametrizados) constituem temas interessantes de pesquisa na área. Em nosso entendimento, estes assuntos ainda não atingiram o mesmo nível de aprofundamento existente do caso linear, principalmente no que tange à existência de métodos eficientes e à exploração de possíveis aplicações.

No que diz respeito ao caso de misturas PNL e à solução proposta, vislumbramos as seguintes perspectivas:

- No desenvolvimento de nossa solução, não nos preocupamos com o problema estrutural existente na separação de misturas PNL, haja visto que, na grande maioria das simulações, definimos cenários nos quais era possível atingir uma inversão perfeita das funções não-lineares do sistema misturador. Neste

contexto, merece investigação a aplicação de estruturas não-lineares capazes de inverter uma maior gama de mapeamentos, e que, ainda assim, satisfaçam as restrições de monotonicidade presentes no problema.

- Investigar outros métodos de estimação de informação mútua e entropia, dada a importância desta etapa no problema em questão.

- Analisar a aplicabilidade de elementos de busca local na determinação dos parâmetros da seção não-linear de um sistema separador PNL. Além disso, uma outra perspectiva interessante seria investigar métodos capazes de determinar boas condições iniciais para esses parâmetros. Essas estratégias poderiam aumentar a velocidade de convergência da técnica proposta, mantendo a sua robustez a convergências sub-ótimas.

- Verificar o desempenho da solução proposta em cenários práticos.

Apêndice A

Alguns Conceitos Básicos em Teoria da Informação

Neste apêndice, descrevemos sucintamente alguns conceitos da teoria da informação utilizados com freqüência em separação cega de fontes. Sugerimos, por exemplo, a leitura da referência (Cover & Thomas, 1991) ao leitor interessado num estudo aprofundado do assunto.

Primeiramente, consideremos a definição de entropia de uma variável aleatória discreta X, dada por

$$H(X) = -\sum_{x} p_X(x) \log(p_X(x)), \tag{A.1}$$

onde $p_X(x)$ corresponde à distribuição de probabilidade de X. Esta grandeza tem como objetivo quantificar o nível de informação presente em X, ou, numa visão mais intuitiva, verificar o quão incerta esta variável é. Neste contexto, é de se esperar que uma variável aleatória associada ao evento do jogar de uma moeda sem viés tenha uma entropia maior do que a do caso em que a probabilidade de obter cara como resultado é, por exemplo, $0,7$, dado que a primeira situação é aquela com o maior grau de incerteza associado. O conceito de entropia também é definido para o caso multidimensional.

A entropia diferencial corresponde à extensão do conceito de entropia para

variáveis aleatória contínuas, e, para uma variável x, é dada por

$$H(x) = -\int_x p_x(x)\log(p_x(x))dx, \tag{A.2}$$

onde $p_x(x)$ corresponde à função densidade de probabilidade de x. Assim como no caso discreto, esta grandeza expressa a incerteza associada a uma variável aleatória. Porém, há algumas diferenças conceituais que merecem uma certa atenção especial (Cover & Thomas, 1991). Para um vetor aleatório $\mathbf{x}$, a entropia diferencial é dada por

$$H(\mathbf{x}) = -\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\log(p_{\mathbf{x}}(\mathbf{x}))d\mathbf{x}, \tag{A.3}$$

onde $p_{\mathbf{x}}(\mathbf{x})$ corresponde à função densidade de probabilidade conjunta dos elementos do vetor $\mathbf{x}$.

Um ponto extremamente útil na concepção de algoritmos em separação de fontes se refere à entropia de uma transformação. No caso, dado um vetor aleatório $\mathbf{x}$, a entropia da transformação $\mathbf{y} = \mathbf{f}(\mathbf{x})$ é expressa por

$$H(\mathbf{y}) = H(\mathbf{x}) + E\{\log|\det(J_{\mathbf{f}})|\}, \tag{A.4}$$

onde $J_{\mathbf{f}}$ corresponde ao jacobiano da transformação. Por exemplo, para uma transformação linear $\mathbf{y} = \mathbf{A}\mathbf{x}$, temos que

$$H(\mathbf{y}) = H(\mathbf{x}) + \log|\det\mathbf{A}|. \tag{A.5}$$

Um outro conceito muito utilizado em separação de fontes é o de informação mútua entre os elementos de um vetor aleatório $\mathbf{x} = [x_1\ x_2\ \ldots\ x_n]$, descrita por

$$I(\mathbf{x}) = \sum_{i=1}^{n} H(x_i) - H(\mathbf{x}). \tag{A.6}$$

Esta medida expressa, em termos simples, uma noção do nível de informação que se pode adquirir sobre um dos membros de um conjunto de variáveis aleatórias a partir do conhecimento dos demais. Esta grandeza sempre assume valores não-negativos, sendo que a única situação em que ela se anula ocorre quando os elementos do vetor são estatisticamente independentes entre si.

Apêndice B

Sobre o Estimador de Entropia Baseado nas Estatísticas de Ordem

Após o desenvolvimento de nossa solução para a separação de misturas PNL, verificamos uma interessante relação entre o estimador de entropia via estatísticas de ordem e uma técnica baseada no clássico método de estimação de função densidade de probabilidade (fdp) a partir de histogramas. Dedicamos este apêndice para apresentar esta relação.

Primeiramente, recapitulemos o método do histograma para estimação de densidades (Silverman, 1986). Para o caso unidimensional, dado um conjunto de amostras de uma variável aleatória, realiza-se uma divisão do espaço de valores possíveis dessa variável em intervalos (*bins*) de mesmo tamanho. Assim, a estimativa da densidade de probabilidade em cada um dos intervalos é dada pela razão entre a fração de amostras existentes no intervalo e o seu tamanho. Matematicamente, a estimativa da fdp de uma variável x, num dado intervalo i, é expressa por

$$\tilde{p_i}(x) = \frac{K_i}{Nd}, \tag{B.1}$$

onde K_i, N e d correspondem respectivamente, ao número de amostras em tal *bin*, ao número total de amostras e ao comprimento do *bin*. Note que os parâmetros N e d são constantes em todos os intervalos, e que ambos são estabelecidos *a priori*.

Um dos problemas existente nesse método é que, caso o número de amostras não seja suficientemente grande, a estimativa obtida nas regiões com baixa densidade

de pontos torna-se consideravelmente ruidosa. Uma alternativa a tal empecilho é considerar uma divisão com intervalos variáveis, de modo que, em cada um deles, exista uma mesma fração de amostras. Neste caso, a estimativa da fdp no intervalo i é dada por

$$\tilde{p_i}(x) = \frac{K}{Nd_i}, \tag{B.2}$$

onde d_i é corresponde ao tamanho do i-ésimo *bin*. Note que, agora, a fração de amostras no intervalo K/N é constante em todos os intervalos e deve ser definida previamente ao processo de estimação.

É possível estabelecer um estimador de entropia com base na expressão (B.2). Para tal, podemos considerar que a fdp de x pode ser aproximada, em cada um dos intervalos, por uma fdp uniforme com valor dado pelo histograma a ele associado. Aplicando esta simplificação na definição da entropia, apresentada em (A.2), obtemos a seguinte estimativa

$$\tilde{H}(x) = -\sum_{i=1}^{I} \frac{K}{Nd_i} \log\left(\frac{K}{Nd_i}\right) d_i = \sum_{i=1}^{I} \frac{K}{N} \log\left(\frac{Nd_i}{K}\right), \tag{B.3}$$

onde I corresponde ao número de *bins* considerados no processo de estimação.

Feita essa breve descrição, recapitulemos a expressão do estimador de entropia baseado na função quantil, proposto por Pham (D.-T. Pham, 2000)

$$\tilde{H}(x) = \sum_{l=2}^{L} \frac{u_l - u_{l-1}}{u_L - u_1} \log\left(\frac{Q_x(u_l) - Q_x(u_{l-1})}{u_l - u_{l-1}}\right), \tag{B.4}$$

onde $Q_x(\cdot)$ corresponde à função quantil de x, e $\{u_1, \ldots, u_L\}$ é um conjunto de números crescentes no intervalo $]0, 1[$. Realizando uma simples mudança de índices e considerando que $u_L - u_1 \approx 1$, podemos reescrever esta expressão do seguinte modo

$$\tilde{H}(x) = \sum_{l=1}^{L-1} (u_{l+1} - u_l) \log\left(\frac{Q_x(u_{l+1}) - Q_x(u_l)}{u_{l+1} - u_l}\right). \tag{B.5}$$

Em nosso trabalho assumimos um *grid* uniforme para u, ou seja, $u_{l+1} - u_l$ é constante para todo l. Assim, é possível observar uma relação direta entre as expressões (B.3) e (B.5). Nesta equivalência, temos que $u_{l+1} - u_l = K/N$, o que faz sentido, dado que ambos os membros dessa igualdade se encontram entre 0 e 1.

Um outro ponto interessante nesta relação diz respeito à função quantil. Lembrando que esta função é a inversa da cumulativa, podemos interpretar a diferença $Q_x(u_l) - Q_x(u_{l-1})$ do seguinte modo. Dada uma probabilidade acumulada u_{l-1} associada a um ponto $x_{p_{l-1}}$, esta expressão revela qual o valor do ponto x_{p_l} tal que a probabilidade de x estar entre $x_{p_{l-1}}$ e, x_{p_l} seja igual a $u_l - u_{l-1}$. Ora, isto é exatamente o que se deseja no método do histograma com intervalo variável: estimar um intervalo, no caso $[x_{p_{l-1}}, x_{p_l}]$, tal que sua probabilidade seja igual a um valor pré-definido, no caso $u_l - u_{l-1}$.

Essa breve discussão indica que, em essência, o estimador de entropia proposto por Pham é equivalente ao método baseado no histograma com intervalos variáveis. No caso, o modo como o comprimento de cada intervalo é calculado depende da função quantil, o qual, por sua vez, é aproximada através das estatísticas de ordem da variável aleatória em questão.

Referências

Achard, S. (2003). *Mesures de Dépendance pour La Séparation Aveugle de Sources.* Tese de Doutorado, Universite Joseph Fourier (França). citado na(s) página(s): 98, 99, 100

Achard, S., & Jutten, C. (2005). Identifiability of Post-Nonlinear Mixtures. *IEEE Signal Processing Letters*, *12*(5), 423-426. citado na(s) página(s): 94

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*, 251-276. citado na(s) página(s): 7, 44

Amari, S.-I., Chen, T.-P. & Cichocki, A. (1997). Stability analysis of learning algortihms in blind source separation. *Neural Networks*, *10*(8), 1345-1351. citado na(s) página(s): 44

Arons, B. (1992). A Review of the Cocktail Party Effect. *Journal of the American Voice I/O Society*, *12*, 35–50. citado na(s) página(s): 14

Arons, F., & Schuster, H. G. (1997). Blind Separation of Convolutive Mixtures and an Application in Automatic Speech Recognition in Noisy Environment. *IEEE Transactions on Signal Processing*, *45*(10), 2608–2619. citado na(s) página(s): 15

Attux, R. R. F. (2005). *Novos Paradigmas para Equalização e Identificação de Canais Baseados em Estruturas Não-lineares e Algoritmos Evolutivos.* Tese de Doutorado, Universidade Estadual de Campinas (Brasil). citado na(s) página(s): 107

Attux, R. R. F., Neves, A., Duarte, L. T., Suyama, R., Junqueira, C. C. M., Rangel, L. E. P. et al. (2006). On the Relationships between Blind Equalization and Blind Source Separation - Part II: Relationships. *Submetido ao Journal of the Brazilian Telecommunications Society.* citado na(s) página(s): 13, 40, 53

Babaie-Zadeh, M. (2002). *On Blind Source Separation in Convolutive and Nonlinear Mixtures.* Tese de Doutorado, Institut National Polytechnique de Grenoble (França). citado na(s) página(s): 37, 86, 87, 94, 97, 100

Bäck, T., Fogel, D. B. & Michalewicz, Z. (Eds.). (2000). *Evolutionary Computation 1:Basic Algorithms and Operators*. Institute of Physics Publishing. citado na(s) página(s): 107

Barros, A. K. (2002). Extracting the Fetal Heart Rate Variability using a Frequency Tracking Algorithm. *Neurocomputing*, *49*, 279–288. citado na(s) página(s): 10

Barros, A. K., Mansour, A. & Ohnishi, N. (1998). Removing Artifacts from Electrocardiographic Signals using Independent Components Analysis. *Neurocomputing*, *22*(1-3), 173–186. citado na(s) página(s): 10

Bell, A. J., & Sejnowski, T. J. (1995). An Information Maximisation Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, *7*(6), 1129–1159. citado na(s) página(s): 7, 41

Bell, A. J., & Sejnowski, T. J. (1997). The "Independent Components" of Natural Scenes are Edge Filters. *Vision Research*, *37*, 3327–3338. citado na(s) página(s): 28

Bermejo, S., Jutten, C. & Cabestany, J. (2006). ISFET Source Separation: Foundations and Techniques. *Sensors and Actuators B - Chemical*, *13*, 222–233. citado na(s) página(s): 15, 92

Bofill, P. (2001). Undetermined Blind Source Separation Using Sparse Representations. *Signal Processing*, *81*, 2353–2362. citado na(s) página(s): 28, 29

Calhoun, V. D., Adali, T., Hansen, L. K., Larsen, J. & Pekar, J. J. (2003). ICA of Functional MRI Data: An Overview. In *Proceedings of the Fourth International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2003* (p. 281-288). Nara, Japão. citado na(s) página(s): 12

Cardoso, J. F. (1997). Infomax and Maximum Likelihood of Blind Source Separation. *IEEE Signal Processing Letters*, *4*(4), 112-114. citado na(s) página(s): 43

Cardoso, J. F. (1998a). Blind signal separation: Statistical principles. *Proceedings*

of the IEEE, *86*(10), 2009-2025. citado na(s) página(s): 7, 40

Cardoso, J. F. (1998b). On the stability of some source separation algorithms. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing.* Cambridge, Inglaterra. citado na(s) página(s): 44, 46

Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, *11*(1), 157-192. citado na(s) página(s): 56

Cardoso, J. F. (2000). *Unsupervised Adaptive Filtering Vol. 1: Blind Source Separation* (S. Haykin, Ed.). Wiley. citado na(s) página(s): 67

Cardoso, J. F., & Laheld, B. H. (1996). Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, *44*(12), 3017-3030. citado na(s) página(s): 7, 44, 59, 60

Cardoso, J. F., & Souloumiac, A. (1993). Blind Beamforming for non-Gaussian Signals. *IEE Proceedings Radar and Signal Processing*, *140*(6), 362-370. citado na(s) página(s): 7, 55, 57, 58

Cardoso, J. F., & Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, *17*(1), 161–164. citado na(s) página(s): 58

Castro, L. N. de, & Timmis, J. I. (2002). *Artificial immune systems: A new computational intelligence approach.* Springer-Verlag. citado na(s) página(s): 107

Castro, L. N. de, & Zuben, F. J. V. (2002). Learning and Optimization Using the Clonal Selection Principle. *IEEE Transactions on Evolutionary Computation, Special Issue on Artificial Immune Systems*, *6*(3), 239–251. citado na(s) página(s): 102

Cavalcante, C. C. (2004). *Sobre Separação Cega de Fontes: Proposicões e Análise de Estratégias para Processamento Multi-Usuário.* Tese de Doutorado, Universidade Estadual de Campinas (Brasil). citado na(s) página(s): 13

Chen, S. S., & Gopinath, R. A. (2000). Gaussianization. In *Proc. Neural Information*

Processing Systems (NIPS) (p. 423-429). Denver, EUA. citado na(s) página(s): 99

Cichocki, A., & Amari, S.-I. (2002). *Adaptive blind signal and image processing: Learning algorithms and applications.* John Wiley & Sons. citado na(s) página(s): 15, 68, 71

Comon, P. (1994). Independent Component Analysis, a New Concept? *Signal Processing*, *36*(6), 287-314. citado na(s) página(s): 7, 18, 21, 36, 100, 106

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory* (2 ed.). John Wiley and Sons, Inc. citado na(s) página(s): 49, 123, 124

Delfosse, N., & Loubaton, P. (1995). Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, *45*, 59–83. citado na(s) página(s): 49

Diamantaras, K. I., & Kung, S. Y. (1996). *Principal Component Neural Networks: Theory and Applications.* John Wiley. citado na(s) página(s): 24

Djafari, A. M. (1999). A Bayesian Approach to Source Separation. In *Proceedings of the 19th International Worskhop on Bayesian and Maximum Entropy Methods (MaxEnt 1999).* Boise, EUA. citado na(s) página(s): 30

Eriksson, J., & Koivunen, V. (2002, setembro). Blind identifiability of class of nonlinear instantaneous ICA models. In *Proceedings of the XI European Signal Processing Conference (EUSIPCO 2002)* (Vol. 2, p. 7-10). Toulouse, France. citado na(s) página(s): 87

Eriksson, J., & Koivunen, V. (2004). Identifiability, Separability, and Uniqueness of Linear ICA Models. *IEEE Signal Processing Letters*, *11*(7), 601–604. citado na(s) página(s): 18

Even, J. (2003). *Contributions à la Separation de Sources à L'aide de Statistiques D'ordre.* Tese de Doutorado, Universite Joseph Fourier (França). citado na(s) página(s): 103

Even, J., & Moisan, E. (2005). Blind source separation using order statistics. *Signal*

Processing, *85*, 1744–1758. citado na(s) página(s): 104

Giannakopoulos, X., Karhunen, J. & Oja, E. (1998). Experimental comparison of neural ICA algorithms. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN'98)* (pp. 651–656). Skövde, Suécia. citado na(s) página(s): 69

Golub, G. H., & Loan, C. F. van. (1989). *Matrix Computations* (2 ed.). Johns Hopkins University Press. citado na(s) página(s): 58

Górriza, J. M., Puntonet, C. G., Rojas, F., Martin, R., Hornillo, S. & Lang, E. W. (2006). Optimizing Blind Source Separation with Guided Genetic Algorithms. *Neurocomputing*, *69*, 1442-1457. citado na(s) página(s): 114

Guilhon, D., Medeiros, E. & Barros, A. K. (2005). ECG Data Compression by Independent Component Analysis. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing.* Mystic, EUA. citado na(s) página(s): 28

Haykin, S. (1994). *Blind deconvolution.* Prentice-Hall. citado na(s) página(s): 13

Haykin, S. (1996). *Adaptive Filter Theory* (3 ed.). Prentice-Hall. citado na(s) página(s): 12, 13

Hérault, J., Jutten, C. & Ans, B. (1985). Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Actes du Xème Colloque GRETSI* (p. 1017-1022). Nice, France. citado na(s) página(s): 6, 18, 34, 35

Hosseini, S., Guidara, R., Deville, Y. & Jutten, C. (2006). Markovian Blind Image Separation. In *Proceedings of the Sixth International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2006* (p. 106-114). Charleston, EUA. citado na(s) página(s): 15

Hyvärinen, A. (1999a). Fast and robust fixed-point algorithmsfor independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626-634. citado na(s) página(s): 50, 51, 76

Hyvärinen, A. (1999b). Survey on Independent Component Analysis. *Neural Computing Surveys*, *2*, 94–128. citado na(s) página(s): 18

Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis.* Wiley. citado na(s) página(s): 8, 11, 13, 14, 15, 17, 18, 23, 28, 43, 52, 53, 54, 69, 72

Hyvärinen, A., & Pajunen, P. (1999). Nonlinear Independent Component Analysis: Existence and Uniqueness Results. *Neural Networks*, *3*, 429-439. citado na(s) página(s): 86, 87

Jung, T. P., Makeig, S., Lee, T. W., McKeown, M., Brown, G., Bell, A. J. et al. (2000). Independent Component Analysis of Biomedical Signals. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2000* (p. 633-644). Helsinki, Finlândia. citado na(s) página(s): 10, 11

Jutten, C., & Karhunen, J. (2003). Advances in Nonlinear Blind Source Separation. In *Proceedings of the Fourth International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2003.* Nara, Japão. citado na(s) página(s): 84

Karhunen, J., & Pajunen, P. (1997, abril). Blind Source Separation Using Least-Squares Type Adaptive Algorithms. In *Proceedings of the IEEE 1997 Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'97)* (p. 3361-3364). Munique, Alemanha. citado na(s) página(s): 54, 55

Karhunen, J., Pajunen, P. & Oja, E. (1998). The Nonlinear PCA Criterion in Blind Source Separation: Relations with Other Approaches. *Neurocomputing*, *22*, 5-20. citado na(s) página(s): 7, 63

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice-Hall. citado na(s) página(s): 38

Kofidis, E. (2001). *Blind source separation: Fundamentals and recent advances* (Tech. Rep.). Mini-curso no XIX Simpósio Brasileiro de Telecomunicações (SBrT2001). citado na(s) página(s): 17, 26, 40, 119

Lathauwer, L. D., Moor, B. D. & Vandewalle, J. (2000). Fetal Electrocardiogram Extraction by Blind Source Subspace Separation. *IEEE Transactions on Biomedical Engineering*, *47*(5), 567–572. citado na(s) página(s): 10

Lee, T.-W., Girolami, M., Bell, A. J. & Sejnowski, T. J. (2000). A unifying information-theoretic framework for independent component analysis. *Computers and Mathematics with Applications*, *39*(11), 1-21. citado na(s) página(s): 63, 119

Lee, T. W., Girolami, M. & Sejnowski, T. J. (1999). Independent Component Analysis using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources. *Neural Computation*, *11*, 417-441. citado na(s) página(s): 47

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer*, *21*(3), 105-117. citado na(s) página(s): 41

Liu, Z.-W., Chiu, K.-C. & Xu, L. (2004). One-Bit-Matching Conjecture for Independent Component Analysis. *Neural Computation*, *16*, 383-399. citado na(s) página(s): 47

Nadal, J.-P., & Parga, N. (1994). Nonlinear Neurons in the Low-noise Limit: a Factorial Code Maximises Information Transfer. *Network: Computation in Neural Systems*, *5*(4), 565-581. citado na(s) página(s): 41

Nikias, C., & Petropulu, A. P. (1993). *Higher-Order Spectra Analysis*. Prentice-Hall. citado na(s) página(s): 22, 56

Pajunen, P., Hyvärinen, A. & Karhunen, J. (1996). Nonlinear Blind Source Separation by Self-organizing Maps. In *Proceedings of the International Conference on Neural Information Processing* (p. 1207-1210). Hong Kong. citado na(s) página(s): 87, 88

Papoulis, A. (1993). *Probability, Random Variables, and Stochastic Processes* (3 ed.). McGraw Hill. citado na(s) página(s): 39, 48, 85, 103

Pham, D., & Vrins, F. (2005). Local Minima of Information-Theoretic Criteria in

Blind Source Separation. *IEEE Signal Processing Letters*, *12*(11), 788–791. citado na(s) página(s): 65, 67

Pham, D.-T. (2000). Blind Separation of Instantenaous Mixtures of Sources Based on Order Statistics. *IEEE Transactions on Signal Processing*, *48*(2), 363–375. citado na(s) página(s): 104, 126

Plumbley, M. D., Abdallah, S. A., Bello, J. P., Davies, M. E., Monti, G. & Sandler, M. B. (2002). Automatic Music Transcription and Audio Source Separation. *Cybernetics and Systems*, *33*(6), 603–627. citado na(s) página(s): 14

Puntonet, C. G., & Prieto, A. (Eds.). (2004). *Independent Component Analysis and Blind Signal Separation Fifth International Conference, ICA 2004*. Springer Lecture Notes in Computer Science. citado na(s) página(s): 8, 15

Rojas, F., Puntonet, C. G., Rodríguez-Álvarez, M., Rojas, I. & Martín-Clemente, R. (2004). Blind Source Separation in Post-Nonlinear Mixtures Using Competitive Learning, Simulated Annealing, and a Genetic Algorithm. *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, *34*(4), 407–416. citado na(s) página(s): 100, 102

Rosca, J., Erdogmus, D., Principe, J. C. & Haykin, S. (Eds.). (2006). *Independent Component Analysis and Blind Signal Separation Sixth International Conference, ICA 2006*. Springer Lecture Notes in Computer Science. citado na(s) página(s): 8, 15

Sanchez-Poblador, V., Monte-Moreno, E. & Solé-Casal, J. (2004). ICA as a Preprocessing Technique for Classification. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2004* (pp. 1165–1172). Granada, Espanha. citado na(s) página(s): 28

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRCs. citado na(s) página(s): 105, 125

Solé-Casals, J., Babaie-Zadeh, M., Jutten, C. & Pham, D.-T. (2003). Improving Algorithm Speed in PNL Mixture Separation and Wiener System Inversion. In

Proceedings of the Fourth International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2003. Nara, Japão. citado na(s) página(s): 99, 100

Taleb, A. (2002). A generic framework for blind source separation in structured nonlinear models. *IEEE Transactions on Signal Processing*, *50*(8), 1819-1930. citado na(s) página(s): 87

Taleb, A., & Jutten, C. (1999). Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, *47*(10), 2807-2820. citado na(s) página(s): 86, 87, 92, 94, 96, 101, 102

Tichavský, P., Koldovský, Z. & Oja, E. (2006). Performance Analysis of the FastICA Algorithm and Cramer-Rao Bounds for Linear Independent Component Analysis. *IEEE Transactions on Signal Processing*, *54*(4), 1189–1203. citado na(s) página(s): 67

Valpola, H. (2000). Nonlinear Independent Component Analysis Using Ensemble Learning: Theory. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2000* (p. 251-256). Helsinki, Finlândia. citado na(s) página(s): 89, 90, 91, 92

Vrins, F., & Verleysen, M. (2005a). Information Theoretic versus Cumulant-Based Contrasts for Multimodal Source Separation. *IEEE Signal Processing Letters*, *12*(3), 190–193. citado na(s) página(s): 65

Vrins, F., & Verleysen, M. (2005b). On the entropy minimization of a linear mixture of variables for source separation. *Signal Processing*, *85*, 1029–1044. citado na(s) página(s): 65, 67

Yang, B. (1995). Projection Approximation Subspace Tracking. *IEEE Transactions on Signal Processing*, *43*(1), 95–107. citado na(s) página(s): 54

Yang, H. H. (1998). On-line Blind Equalization via On-line Blind Separation. *Signal Processing*, *68*(3), 271–281. citado na(s) página(s): 13

Zhang, K., & Chan, L.-W. (2005). Extended Gaussianization Method for Blind

Separation of Post-Nonlinear Mixtures. *Neural Computation*, *17*(2), 425–452. citado na(s) página(s): 99

Índice de Autores