Análise de componentes independentes com aplicações em séries temporais financeiras

Hugo Valesini Gegembauer

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística
Orientador: Prof. Dr. Pedro A. Morettin

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, Março de 2010

Análise de componentes independentes com aplicações em séries temporais financeiras

Este exemplar corresponde à redação
final da dissertação/tese devidamente corrigida
e defendida por Hugo Valesini Gegembauer
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Prof. Dr. Pedro Alberto Morettin (Presidente)- IME-USP.
- Profa. Dra. Clélia Maria de Castro Tóloi - IME-USP.
- Profa. Dra. Thelma Sáfadi - UFLA

"A matemática, vista corretamente, possui não apenas verdade, mas também suprema beleza - uma beleza fria e austera, como a da escultura."

(Bertrand Russel)

"Onde houver força, também o número se tornará senhor: ele tem mais força."

(Nietzsche)

Dedicatória

À minha Mãe Jonia,
aos meus irmãos Luis e Carla.

Agradecimentos

Agradeço a todos, que de forma direta ou indireta tornaram possível a conclusão deste trabalho. Em especial:

Primeiramente à minha mãe e meus irmãos que sempre me apoiaram incondicionalmente.

À minha avó Jonia pelo carinho e à Graça pelo cuidado de sempre.

Ao meu orientador Prof. Dr. Pedro A. Morettin pelo voto de confiança, paciência, disponibilidade, atenção e por me conduzir na execução e conclusão deste trabalho.

À Patrícia Murakami por ajudas inestimáveis em alguns momentos.

Ao Fábio Trevisan pela ajuda num momento crucial do desenvolvimento desta tese.

Ao Ivan Pagnossin por incontáveis ajudas com o Latex.

Ao Plínio Lucas por diversas ajudas ao longo do mestrado todo.

Ao sempre amigo, Flávio Augusto.

Finalmente, à CAPES pelo apoio financeiro.

Resumo

Assumindo que os dados financeiros são gerados por um conjunto de componentes independentes (CI), iremos usar o modelo ICA-GARCH, uma alternativa aos modelos multivariados. Neste modelo a análise de componentes independentes serve para procurar os fatores latentes com heterocedasticidade condicional. Usou-se o algoritmo FastICA para estimar os componentes independentes e a matriz de pesos ($\mathbf{A}$). Após estimar os CI's ajustou-se um modelo GARCH univariado para cada um deles, fazendo a previsão em seguida e comparando os resultados de diferentes números de CI's usados.

É fato que o modelo adotado reduz a complexidade de estimar modelos GARCH multivariados tranformando-o em um número pequeno de modelos de volatidade univariados.

Palavras-chave: Análise de Componentes Independentes (ICA), Modelos Fatoriais, Séries Temporais Financeiras, GARCH multivariado, Previsão da Volatilidade, Retornos.

Abstract

Assuming that the financial data are generated by a set of independent components (IC's), we will use the ICA-GARCH model, an alternative to the multivariate models. In this model, the independent component analysis search the conditionally heteroskedastic latent factors. We used the FastICA algorithm to estimate the independent components and the weight matrix ($\mathbf{A}$). After estimating the IC's we fit an univariate GARCH model to each of them, following by the forecast after that and comparing the results of different numbers of IC's used.

It is a fact that the adopted model reduces the complexity to estimate a multivariate GARCH model by transforming it into a small number of univariate volatility model.

Keywords: Independent Component Analysis (ICA), Factor Models, Financial Time Series, Multivariate GARCH, Forecasting Volatility, returns.

Sumário

Lista de Abreviaturas **xiii**

Lista de Figuras **xv**

Lista de Tabelas **xvii**

1 Introdução **1**

2 Preliminares **3**

2.1 Representação Linear de Dados Multivariados 4

2.2 "Separação Cega de Fontes" 6

2.3 Análise de Componentes Independentes 8

2.3.1 Aplicações 9

2.3.2 Como Encontrar os Componentes Independentes 9

3 Fundamentos Básicos **13**

3.1 Independência Estatística 13

3.2 Cumulantes e Momentos 13

3.3 Curtose 16

3.4 Não-correlação e Branqueamento 17

3.5 Gradientes e Métodos de Otimização 18

3.6 Máxima Verossimilhança 23

3.7 Entropia e Negentropia 24

3.8 Modelos ARCH e GARCH 27

3.9 Decorrelação Não-Linear 28

4 Análise de Componentes Independentes **31**

4.1 Definição 31

4.2 Restrições à ICA 31

4.3 Ambiguidades da ICA 33

4.4 Centrando as Variáveis 34

4.5 ICA pela Maximização da Não-Normalidade 34

4.5.1 Medindo Não-Normalidade pela Curtose 36

4.5.2 Medindo Não-Normalidade pela Negentropia 39
4.6 ICA pela Estimação de Máxima Verossimilhança 44
4.7 ICA por Outros Métodos 48
4.8 Algumas Considerações Práticas 49

5 ICA Aplicada à Finanças **51**

6 Aplicando ICA a Dados Reais **57**
6.1 Ponderando os Componentes Independentes 59
6.2 Identificando os Fatores 63

7 Considerações Finais **67**

Referências Bibliográficas **69**

Lista de Abreviaturas

AR	Auto-regressivo.
ARCH	Autoregressive Conditional Heteroscedasticity.
BOVESPA	Bolsa de Valores de São Paulo.
BSS	Blind Source Separaration (Separação Cega de Fonte).
CI	Componente Independente.
CUC-GARCH	Conditionally Uncorrelated Components-GARCH.
fdp	Função de Densidade de Probabilidade.
GARCH	Generalized Autoregressive Conditional Heteroscedasticity.
GICA	Generalized Independent Componente Analysis.
IBEX	Iberia Index.
IBOVESPA	Índice da Bolsa de Valores de São Paulo.
ICA	Independent Componente Analysis.
JADE	Joint Approximate Diagonalization of Eigenmatrices.
MdRAE	Median Relative Absolute Error (Erro Mediano Relativo Absoluto).
MGARCH	Multivariate GARCH.
O-GARCH	Ortoghonal-GARCH.
PCA	Principal Component Analysis.
SOBI	Second-Order Blind Identification.

Lista de Figuras

6.1 Preço diário de fechamento da ação da Vale de 2002 a 2007. 57
6.2 Retorno diário da Vale entre os anos de 2002 a 2007. 59
6.3 Variabilidade explicada individualmente pelos 20 componentes independentes ordenados por importância 61
6.4 Componentes principais escolhidos: (A) - CP11; (B) - CP12; (C) - CP18 e (D) - CP03 62

Lista de Tabelas

6.1 Setores, segmentos, empresas e códigos de negociação das ações 58
6.2 Curtose dos retornos de média zero 59
6.3 Componentes ordenados em função da variabilidade explicada 60
6.4 Coeficientes dos componentes escolhidos 62
6.5 Modelos ajustados e seus coeficientes 63
6.6 Média das MdRAE medida para os 20 retornos 65

Capítulo 1

Introdução

Desde que [11] introduziu o modelo ARCH (Autoregressive conditional heteroscedasticity) e [6] o generalizou propondo o modelo GARCH (Generalized Autoregressive conditional heteroscedasticity) muitos pesquisadores tem tido interesse em modelar a volatilidade em séries temporais financeiras. No caso de séries multivariadas, a volatilidade tende a se mover através dos vários mercados o que torna necessário o uso de modelos multivariados para entender esses movimentos.

Hoje em dia existe uma série de modelos propostos para se medir tais volatilidades. Dentre eles poderíamos citar o MGARCH (Multivariate GARCH), O-GARCH (Ortoghonal GARCH - um dos mais populares) que estima os fatores não observados usando componentes principais, dentre outros. Neste trabalho, seguiremos o que foi feito em [14], onde foi proposto um novo modelo para modelar volatilidades multivariadas assumindo que estas são uma combinação linear de alguns modelos GARCH univariados. O modelo foi chamado de ICA-GARCH, onde a análise de componentes independentes (ICA em inglês) pode ser vista como um modelo fatorial onde as componentes não observadas são não-gaussianas e mutuamente independentes.

O trabalho está organizado da seguinte maneira: No capítulo 2 apresentamos resumidamente do que se trata o modelo ICA bem como um rápido histórico e suas utilizações e como encontrá-los; da representação linear de dados multivariados e a definição e explicação do termo "separação cega de fontes". No capítulo 3 apresentamos todos os fundamentos matemático-estatísticos que são utilizados ao longo do texto, alguns como uma revisão outros mais importantes de forma mais aprofundada. O capítulo 4 é o cerne deste trabalho, o mais importante. Nele é feito de forma mais aprofundada as explicações relacionadas à ICA, tais como a definição, as restrições, as ambiguidades, métodos de maximização e medidas de não-normalidade e a derivação de alguns algoritmos. No capítulo 5 apresentamos alguns exemplos diversos de uso da ICA em dados financeiros. Já o capítulo 6 apresenta os resultados de uma aplicação da ICA a uma carteira de 20 ações da BOVESPA e por fim o capítulo 7 apresenta as conclusões deste trabalho bem como sugestões de possíveis estudos futuros utilizando a ICA.

Capítulo 2

Preliminares

A Análise de Componentes Independentes (ICA - sigla em inglês de "Independent Component Analysis") é uma técnica estatística para encontrar fatores ou componentes fundamentais vindos de dados estatísticos multivariados, os quais tipicamente são apresentados em uma grande base de dados.

Supomos os dados observados como sendo uma mistura linear ou não linear de algumas variáveis latentes desconhecidas e o sistema de mistura também é desconhecido. As variáveis latentes são supostas não-gaussianas e mutuamente independentes e são chamadas de componentes independentes dos dados observados. Estes componentes independentes, também chamados de fatores ou fontes, podem ser encontrados pela ICA.

O método ICA visa componentes que são tanto estatisticamente independentes como não-gaussianos. Pode ser visto como uma extensão da análise de componentes principais e análise fatorial, sendo uma técnica muito mais poderosa, capaz de encontrar fatores quando estes métodos clássicos falham completamente.

Os dados analisados pela ICA podem ser oriundos de várias de áreas de aplicação, incluindo imagens digitais e base de dados de documentos, bem como indicadores econômicos e medidas psicométricas. Em muitos casos, as medidas são dadas como um conjunto de sinais paralelos ou séries temporais; o termo Separação Cega de Fonte (BSS - sigla em inglês de "Blind Source Separation") é usado para caracterizar esse problema. Exemplos típicos são misturas de sinais de falas simultâneas que são capturados por vários microfones, ondas cerebrais gravadas por múltiplos sensores, sinais de interferência de rádio chegando de telefone celulares e séries temporais paralelas obtidas a partir de algum processo industrial.

A técnica de ICA é uma invenção relativamente nova. Foi introduzida pela primeira vez no início da década de 80 no contexto da modelagem de redes neurais. Em meados da década de 90 alguns novos algoritmos altamente bem sucedidos foram introduzidos por vários grupos de pesquisa, juntamente com demonstrações de problemas como o caso do "Cocktail-Party", onde num ambiente

várias pessoas conversam ao mesmo tempo e uma música toca ao fundo. O objetivo neste caso, consiste, através de algum método de separação, encontrar os sinais de voz de cada pessoa e da música, a partir dos sinais misturados captados pelos sensores. ICA se transformou num dos tópicos mais excitantes, tanto no campo das redes neurais, como em estatística avançada e processamento de sinal. Apareceram aplicações do ICA a problemas reais em processamento de sinais biomédicos, separação de sinais de áudio, telecomunicações, diagnóstico de falhas, análise de séries temporais financeiras e mineração de dados.

Muitos artigos sobre ICA foram publicados nos últimos 20 anos em um grande número de revistas e conferências em várias áreas como processamento de sinal, redes neurais artificiais, estatística, teoria da informação e outras.

2.1 Representação Linear de Dados Multivariados

Um problema em estatística e áreas relacionadas é como encontrar uma representação adequada de dados multivariados. Representação, aqui, significa que nós de alguma forma transformamos os dados para um estrutura essencial mais visível ou acessível.

Vamos supor que os dados consistem de um número de variáveis que nós observamos conjuntamente. Vamos denotar o número de variáveis por m e o número de observações por T. Podemos então denotar o dado por $x_i(t)$ onde o índice toma valores $i = 1, \ldots, m$ e $t = 1, \ldots, T$. As dimensões m e T podem ser muito grandes.

Uma formulação bastante geral do problema pode ser expressa como segue: Qual poderia ser a função de um espaço m-dimensional para um n-dimensional ($n < m$) tal que as variáveis transformadas dão a maior parte da informação que está escondida neste amplo conjunto de dados? Ou seja, as variáveis transformadas devem ser os fatores ou componentes ocultos que descrevem a estrutura essencial dos dados. É esperado que estes componentes correspondam a alguma causa física, econômica, social, etc que os envolveu no processo que gerou os dados.

Em muitos casos, consideramos apenas funções lineares porque a interpretação obtida é mais simples. Então, cada componente, digamos y_i, é expresso como uma combinação linear das variáveis observadas:

$$y_i(t) = \sum_{j=1}^{m} w_{ij} x_j(t), \text{ para } i = 1, \ldots, n, \tag{2.1}$$

onde os w_{ij} são alguns coeficientes que definem a representação. O problema pode então ser reformulado como o problema em determinar os coeficientes w_{ij}. Usando álgebra linear podemos

expressar a transformação linear na equação (2.1) como uma multiplicação de matrizes. Reunindo os coeficientes w_{ij} em uma matrix $\mathbf{W}$, a equação se transforma em:

$$\begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix} = \mathbf{W} \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{pmatrix} \tag{2.2}$$

Uma aproximação estatística básica consiste em considerar os $x_i(t)$ como um conjunto de T realizações das m variáveis aleatórias. Então, cada conjunto $\{x_i(t),\ t = 1,\ \ldots,\ T\}$ é uma amostra de uma variável aleatória, a qual denotaremos por x_i. Nessa estrutura, podemos determinar a matriz $\mathbf{W}$ através de propriedades estatísticas das componentes y_i transformadas. A seguir discutiremos algumas propriedades estatísticas que podem ser usadas; uma delas conduzirá à análise de componentes independentes.

Um princípio estatístico para escolher a matriz $\mathbf{W}$ é limitar o número de componentes y_i, talvez 1 ou 2, e determinar $\mathbf{W}$ de forma que y_i contenha a máximo de informação dos dados quanto possível. Isto nos leva a uma família de técnicas, como a análise de componentes principais ou análise fatorial.

Outro princípio que tem sido utilizado para determinar $\mathbf{W}$ é independência: os componentes y_i devem ser estatisticamente independentes. Isto significa que o valor de qualquer um destes componentes não nos dá informação nenhuma sobre os valores dos outros componentes.

De fato, na análise fatorial, frequentemente, é exigido que os fatores sejam independentes, mas isso é apenas parcialmente verdade, porque esta imposição supõe que os dados tem distribuição gaussiana multivariada. Se os dados são normais multivariados, é simples encontrar componentes que são independentes, porque para estes dados, componentes não correlacionados são sempre independentes.

Na realidade, os dados frequentemente não seguem uma distribuição gaussiana e a situação não é tão simples como os métodos supõem. Por exemplo, em muitos casos reais os dados apresentam distribuições "supergaussianas". Isto significa que as variáveis aleatórias tomam relativamente, de forma mais frequente, valores que são pertos de zero ou muito grandes. Em outras palavras, a densidade de probabilidade dos dados tem um pico no zero e caudas pesadas (muitos valores afastados do zero), quando comparados à densidade gaussiana de mesma variância. Este é ponto de partida da ICA. Queremos achar componentes estatisticamente independentes, nos casos gerais onde os dados não são gaussianos.

2.2 "Separação Cega de Fontes"

Vamos olhar para o mesmo problema de achar uma boa representação, a partir de diferentes pontos de vista. Este é um problema em processamento de sinais e que também mostra o fundo histórico da ICA.

Considere a situação onde existe um número de sinais emitidos por algum objeto físico ou fonte. Esta origem física pode ser, por exemplo, constituída por diferentes áreas do cérebro emitindo sinais elétricos; pessoas falando na mesma sala, assim emitindo sinais de fala; ou telefones celulares emitindo suas ondas de rádio. Suponha que haja vários sensores ou receptores que estão em diferentes posições, sendo assim cada mistura das fontes originais é gravada com pesos um pouco diferentes.

Com o objetivo de simplificar a exposição, digamos que existem três fontes que dão origem aos sinais, e também três sinais observados, denotados por $x_1(t)$, $x_2(t)$ e $x_3(t)$, os quais são as amplitudes dos sinais gravados no tempo t, e por $s_1(t)$, $s_2(t)$ e $s_3(t)$ os sinais originais. Os $x_i(t)$ são uma combinação linear dos $s_i(t)$ com coeficientes constantes a_{ij}, os quais dão as misturas dos pesos e que dependem da distância entre as fontes e os sensores e se supõe que sejam desconhecidos:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \tag{2.3}$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t)$$

$$x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t)$$

Não podemos conhecer os valores a_{ij} sem conhecer todas as propriedades do sistema de mistura, os quais podem ser extremamente complexos. A fonte do sinal s_i também é desconhecida, uma vez que não podemos gravá-la diretamente. O que gostaríamos de fazer é encontrar os sinais originais a partir das misturas $x_1(t)$, $x_2(t)$ e $x_3(t)$. Isto se chama de Separação Cega de Fontes. O termo "cega" significa que temos pouca ou nenhuma informação a respeito das fontes.

Podemos seguramente supor que os coeficientes de mistura a_{ij} são diferentes o suficiente, para tornar a matriz que formam invertível. Então, existe uma matriz $\mathbf{W}$ com coeficiente w_{ij}, tal que podemos escrever os s_i como:

$$s_1(t) = w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \tag{2.4}$$

$$s_2(t) = w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t)$$

$$s_3(t) = w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t)$$

Tal matriz $\mathbf{W}$ pode ser encontrada como a inversa da matriz que consiste dos coeficientes de mistura a_{ij} em (2.3) se conhecermos os coeficientes a_{ij}.

Vemos, agora, que de fato este problema é matematicamente similar ao que tínhamos quando queríamos encontrar uma boa representação dos dados $x_i(t)$ na equação (2.2). De fato, podemos considerar cada sinal $x_i(t), t = 1, \cdots, T$ como uma amostra de variáveis aleatórias x_i, assim que o valor da variável aleatória é dado pela amplitude do sinal gravado no tempo t.

A questão agora é: como podemos estimar os coeficientes w_{ij} em (2.4)? Queremos obter um método geral que funcione em diferentes circunstâncias e que de fato forneça uma resposta ao problema bastante geral pelo qual começamos: encontrar uma boa representação de dados multivariados. Então, usamos propriedades estatísticas muito gerais. Nós observamos os sinais x_1, x_2 e x_3 e queremos achar uma matriz $\mathbf{W}$ de modo que a representação seja dada pelos sinais da fonte original s_1, s_2 e s_3.

Surpreendentemente, uma solução simples ao problema pode ser encontrada, justamente, considerando independência estatística dos sinais. De fato, se os sinais são não-gaussianos, isto é suficiente para determinar os coeficientes w_{ij} assim que os sinais

$$y_1(t) = w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \tag{2.5}$$

$$y_2(t) = w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t)$$

$$y_3(t) = w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t)$$

sejam estatisticamente independentes. Se os sinais y_1, y_2 e y_3 são independentes, então eles são iguais aos sinais originais s_1, s_2 e s_3. (Eles podem ser multiplicados por algum escalar constante, de qualquer forma, isto tem uma pequena significância.)

Usando apenas esta informação de independência estatística, podemos de fato estimar os coeficientes da matriz $\mathbf{W}$. Vimos que a partir de um conjunto de dados que aparentam ser apenas um ruído, somos capazes de estimar o sinal da fonte original, usando um algoritmo que usa apenas a

informação de independência.

2.3 Análise de Componentes Independentes

Vimos que o problema da separação cega de fonte se reduz a encontrar uma representação linear na qual os componentes são estatisticamente independentes. Em situações práticas, em geral, não podemos encontrar uma representação onde os componentes são realmente independentes, mas podemos pelo menos encontrar componentes que são tão independentes quanto possível.

Isto nos leva a seguinte definição da ICA, a qual será considerada em detalhes adiante e nos próximos capítulos. Dado um conjunto de observações de variáveis aleatórias $\big(x_1(t), x_2(t), \cdots, x_n(t)\big)$, onde t é o tempo, supõe-se que estas sejam geradas como uma mistura linear de componentes independentes:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} = \mathbf{A} \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{pmatrix} \tag{2.6}$$

onde $\mathbf{A}$ é alguma matriz desconhecida. A análise de componentes independentes consiste agora em estimarmos tanto a matriz $\mathbf{A}$ como os $s_i(t)$, onde apenas observamos os $x_i(t)$. Observe que assumimos aqui que o número de componentes independentes s_i é igual ao número de variáveis observadas; isto é uma simplicação que não é completamente necessária.

Alternativamente, poderíamos definir ICA como segue: encontre uma transformação linear dada pela matriz $\mathbf{W}$ como em (2.2), de tal modo que as variáveis aleatórias $y_1, i = 1, \cdots, n$ sejam tão independentes quanto possível. Esta formulação não é realmente muito diferente da anterior, visto que após estimar $\mathbf{A}$, sua inversa é $\mathbf{W}$.

Pode ser mostrado (ver mais adiante) que este problema é bem definido, isto é, o modelo em (2.6) pode ser estimado se e somente se os componentes s_i são não-gaussianos. Isto é uma necessidade fundamental que também explica a principal diferença entre ICA e análise fatorial, na qual a normalidade dos dados é levada em conta. De fato, ICA pode ser considerada como uma análise fatorial não-gaussiana, visto que nesta também modelamos os dados como uma mistura linear de alguns fatores latentes.

2.3.1 Aplicações

Devido a generalidade dos modelos ICA, existem aplicações em muitas áreas diferentes.

- Imagens cerebrais: frequentemente há diferentes fontes no cérebro emitindo sinais que são confundidos nos sensores localizados fora da cabeça, exatamente como no modelo de separação cega de fonte.

- Econometria: frequentemente temos séries temporais paralelas, e ICA pode decompô-las em componentes independentes que nos daria uma percepção da estrutura do conjunto de dados.

- Aplicações diferentes em extração de caracaterísticas em imagens, onde queremos encontrar características que sejam tão independentes quanto possível.

2.3.2 Como Encontrar os Componentes Independentes

Pode ser surpreendente que os componentes independentes possam ser estimados a partir de uma mistura linear com não mais suposições que a sua independência. Tentaremos explicar brevemente como isso é possível; para mais detalhes, veja o cápitulo 4.

Não-Correlação não é Suficiente

A primeira coisa que devemos notar é que independência é uma propriedade mais forte que não-correlação. Considerando o problema da separação cega de fonte podemos encontrar muitas representações diferentes não-correlacionadas dos sinais que não sejam independentes e não possam separar as fontes. Não-correlação por si só não é suficiente para separar os componentes. Este também é o motivo pelo qual a análise de componentes principais (PCA - Sigla em inglês) e análise fatorial não podem separar os sinais: eles dão componentes que são não-correlacionados.

Usando um método bem conhecido de decorrelação, podemos transformar qualquer mistura linear de componentes independentes em componentes não-correlacionados, neste caso a mistura é ortogonal. Então, o truque na ICA é estimar a transformação ortogonal que resta após a estimação. Isto é algo que os métodos clássicos não podem estimar porque eles são baseados, essencialmente, na mesma informação da covariância que a decorrelação.

Decorrelação Não-Linear é o método básico da ICA

Segundo [16], uma forma de expressar como independência é mais forte que não-correlação é dizer que independência implica não-correlação não-linear. Se s_1 e s_2 são independentes, então quaisquer transformações não-lineares $g(s_1)$ e $h(s_s)$ são não-correlacionadas (no sentido de que suas

covariâncias é zero.) Em contrapartida, para duas variáveis aleatórias que são meramente não correlacionadas, tal transformação não linear não tem covariância zero em geral.

Então podemos tentar executar ICA através de uma forma mais forte de decorrelação, encontrando uma representação onde os y_i são não-correlacionados mesmo após alguma transformação não-linear. Isto dá um simples princípio de estimação da matriz $\mathbf{W}$:

> **Princípio 1 da estimação da ICA**: Decorrelação não-linear. Encontre a matriz $\mathbf{W}$ que para qualquer $i \neq j$, os componentes y_i e y_j são não-correlacionados, e os componentes transformados $g(y_i)$ e $h(y_j)$ são não correlacionados, onde g e h são funções não-lineares adequadas.

Isto é uma aproximação válida para estimar ICA: se as não-linearidades são corretamente escolhidas, o método encontra os componentes independentes.

Embora esse método seja muito intuitivo, ele deixa aberto uma importante questão: Como devem ser escolhidas as não-linearidades g e h? Respostas a essas perguntas podem ser encontradas usando princípios da teoria da estimação e da teoria da informação. A teoria da estimação fornece o método mais clássico de estimação em qualquer modelo estatístico: O método da máxima verossimilhança. A teoria da informação fornece medida exata de independência, tal como a informação recíproca. Usando uma dessas teorias podemos determinar as funções não-lineares g e h de modo satisfatório.

Componentes Independentes são os Componentes maximamente Não-Gaussianos

Um outro princípio muito importante e intuitivo da estimação da ICA é a máxima não-normalidade. A ideia é que de acordo com o teorema central do limite, somas de variáveis aleatórias não-normais são mais próximas à normal do que a sua distribuição original. Então, se tomarmos uma combinação linear $y = \sum_i b_i x_i$ das variáveis (as quais, por conta do modelo linear, também é uma combinação linear dos componentes independentes), esta será maximamente não-normal se for igual a um dos componentes independentes. Isto ocorre, porque se fosse uma mistura real de dois ou mais componentes, estaria mais perto da distribuição gaussiana, devido ao teorema central do limite.

Então, o princípio pode ser expresso como segue:

> **Princípio 2 da estimação da ICA**: Máxima não-normalidade. Encontre o máximo local de não-normalidade de uma combinação linear $y = \sum_i b_i x_i$ sob a restrição de que a variância de y seja constante. Cada máximo local dá um componente independente.

Para medir não-normalidade na prática, podemos usar, por exemplo a curtose que é um cumulante de quarta ordem. Cumulantes tem interessantes propriedades algébricas e estatísticas que os fazem ter um papel importante na teoria da ICA.

Estimação da ICA necessita mais do que covariâncias

Existem muitos outros métodos para estimar o modelo ICA. O que todos tem em comum é que consideram algumas estatísticas que não estão contidas na matriz de covariância (a matriz que contém as covariâncias entre todos os pares x_i).

Usando a matriz de covariância podemos decorrelacionar os componentes no sentido linear ordinário, mas não mais forte. Então, todos os métodos ICA usam alguma forma de estatística de ordem mais alta do que 2, que especificamente tenha informação que não está contida na matriz de covariância. Já comentamos duas formas de informação de maior ordem: correlações não-lineares e curtose. Muitos outros tipos podem também ser usados.

Métodos Numéricos são Importantes

Além do princípio de estimação, tem-se que encontrar um algoritmo para executar os cálculos necessários. Como o princípio de estimação usa funções não-quadráticas, os cálculos necessários, usualmente, não podem ser expressos usando algebra linear simples, podendo ser extremamente complexos. Algoritmos numéricos são assim uma parte integral dos métodos de estimação ICA.

Os métodos numéricos são tipicamente baseados na otimização de algumas funções objetivas. O método básico de otimização é o método do gradiente. De interesse particular tem-se o algoritmo de ponto fixo chamado FastICA que tem sido usado para explorar uma particular estrutura dos problemas ICA. Por exemplo, podemos usar estes métodos para encontrar o máximo de não-normalidade como medido pelo valor absoluto do excesso de curtose.

Capítulo 3

Fundamentos Básicos

Neste capítulo apresentaremos alguns conceitos e resultados que serão úteis no entendimento do restante do texto.

3.1 Independência Estatística

Um conceito chave que constitui o fundamento da análise de componentes independentes é a independência estatística. Formalmente, independência estatística é definida em termos das densidades de probabilidade. São necessários 2 critérios para verificar independência. Para mais detalhes, consultar [18].

(a) Se $X_1, \ldots, X_n$ são independentes e possuem densidades $f_{X_1}, \ldots, f_{X_n}$, então a função

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), (x_1, \ldots, n_n) \in \mathbf{R}^n,$$

é densidade conjunta das variáveis aleatórias $X_1, \ldots, X_n$ i.e., $f = f_{X_1,\ldots,X_n}$.

(b) Reciprocamente, se $X_1, \ldots, X_n$ têm densidade conjunta f satisfazendo

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), (x_1, \ldots, n_n) \in \mathbf{R}^n,$$

onde $f_i(x) \geq 0$ e $\int_{-\infty}^{\infty} f_i(x)dx = 1$, $\forall i$, então $X_1, \ldots, X_n$ são independentes e f_i é a densidade de X_i, para $i = 1, \ldots, n$.

3.2 Cumulantes e Momentos

Seja X uma variável aleatória contínua de média zero e densidade de probabilidade $f_X(x)$. A função característica $\varphi(t)$ de X é definida como a transformada de Fourier da função de densidade $f_X(x)$:

$$\varphi(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} f_X(x)dx, \tag{3.1}$$

onde $i = \sqrt{-1}$ e t real. Toda distribuição de probabilidade é especificada de modo único por sua função característica e vice-versa. Expandindo a função característica $\varphi(t)$ em série de Taylor, temos:

$$\varphi(t) = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{\infty} \frac{x^k (it)^k}{k!} \right) f_X(x)dx = \sum_{k=0}^{\infty} E[X^k] \frac{(it)^k}{k!}. \tag{3.2}$$

Então os coeficientes desta expansão são os momentos centrais de X, $E[X^k] = \mu'_k$. Muitas vezes é desejável usar a função geradora de cumulantes que é definida como o logaritmo natural da função característica:

$$\phi(t) = ln[\varphi(t)] = ln\big[E[e^{itX}]\big] \tag{3.3}$$

Portanto, os cumulantes κ_k de X são definidos de uma maneira similar aos momentos como coeficientes da expansão em série de Taylor da função geradora de cumulantes:

$$\phi(t) = \sum_{k=0}^{\infty} \kappa_k \frac{(it)^k}{k!} \tag{3.4}$$

Igualando, $\phi(t) = ln[\varphi(t)]$, temos:

$$\exp\left\{ \sum_{k=0}^{\infty} \kappa_k \frac{(it)^k}{k!} \right\} = \sum_{k=0}^{\infty} E[X^k] \frac{(it)^k}{k!}. \tag{3.5}$$

Portanto, o k-ésimo momento μ'_k é o coeficiente de $\frac{(it)^k}{k!}$ em $\varphi(t)$ e o k-ésimo cumulante κ_k é o coeficiente de $\frac{(it)^k}{k!}$ em $\ln\big(\varphi(t)\big)$, onde este último pode ser obtido através das derivadas de $\phi(t)$:

$$\kappa_k = (-i)^k \left. \frac{d^k \phi(t)}{dt^k} \right|_{t=0}. \tag{3.6}$$

Para uma variável aleatória X de média zero, os quatro primeiros cumulantes são:

$$\kappa_1 = E[X] = 0,$$

$$\kappa_2 = E[X^2] - \big(E[X]\big)^2 = \mu_2',$$

$$\kappa_3 = E[X^3] - 3E[X^2]E[X] + 2\big(E[X]\big)^3 = \mu_3',$$

$$\kappa_4 = E[X^4] - 3\big(E[X^2]\big)^2 - 4E[X^3]E[X] + 12E[X^2]\big(E[X]\big)^2 - 6\big(E[X]\big)^4 = \mu_4' - 3(\mu_2')^2.$$

Logo, os três primeiros cumulantes são iguais aos respectivos momentos e o quarto cumulante é conhecido como curtose. Para mais detalhes, consultar [22], [26] e [30].

As definições anteriores podem ser estendidas ao caso multivariado. Porém, agora t é um vetor coluna tendo a mesma dimensão que $\underset{\sim}{X}$ e a integral é calculada sobre todos os componentes de $\underset{\sim}{X}$. Os momentos e cumulantes de $\underset{\sim}{X}$ são obtidos de maneira semelhante ao caso escalar. Para um vetor aleatório $\underset{\sim}{X}$ de média zero, os cumulantes de segunda, terceira e quarta ordem são dados por:

$$cum(X_i, X_j) = E[X_i X_j],$$

$$cum(X_i, X_j, X_k) = E[X_i X_j X_k],$$

$$cum(X_i, X_j, X_k, X_l) = E[X_i X_j X_k X_l] - E[X_i X_j]E[X_k X_l] - E[X_i X_k]E[X_j X_l] - E[X_i X_l]E[X_j X_k].$$

(Para mais detalhes veja [24]).

Então, vemos que o segundo cumulante é igual ao segundo momento $E[X_i X_j]$, o qual é a covariância c_{ij} entre as variáveis X_i e X_j. Similarmente, o terceiro cumulante é igual ao terceiro momento $E[X_i X_j X_k]$. Entretanto, o quarto cumulante, difere do quarto momento $E[X_i X_j X_k X_l]$ das variáveis aleatórias X_i, X_j, X_k e X_l.

Dissemos que muitas vezes é desejável usar a função geradora de cumulantes, ao invés da função geradora de momentos. Isso se deve ao fato dos cumulantes terem as seguintes propriedades, as quais os momentos não têm:

1. Sejam $\mathbf{x}$ e $\mathbf{y}$ vetores aleatórios estatisticamente independentes tendo a mesma dimensão, então os cumulantes da soma $\mathbf{z}= \mathbf{x}+\mathbf{y}$ são iguais à soma dos cumulantes de $\mathbf{x}$ e $\mathbf{y}$. Esta propriedade também estende-se à soma de mais de dois vetores aleatórios independentes.

2. Se a distribuição do vetor aleatório $\mathbf{x}$ é normal-multivariada, todos os cumulantes de ordem maior ou igual a três são nulos. Esta propriedade é bem útil, tornando possível o uso de cumulantes para extrair a parte de um sinal não gaussiano.

3.3 Curtose

O quarto momento $\mu_4' = E[x^4]$ é usado em alguns algoritmos ICA por causa da sua simplicidade. Em vez de usarmos o quarto momento central $\mu_4 = E\big[(x - \mu_x)^4\big]$, a estatística de quarta ordem chamado curtose é usualmente empregada, pois apresenta propriedades bastante úteis que o quarto momento central não têm. Já derivamos o resultado da curtose na seção anterior, mas discutiremos aqui com um pouco mais de detalhes por causa da sua simplicidade e importância na análise de componentes independentes e separação cega de fontes.

A curtose é definida no caso de média zero pela equação:

$$curt(x) = E[x^4] - 3\big(E[x^2]\big)^2. \tag{3.7}$$

Alternativamente, a curtose normalizada ou também conhecida como excesso de curtose, é dada por

$$\tilde{\kappa} = \frac{E[x^4]}{\big(E[x^2]\big)^2} - 3. \tag{3.8}$$

Para dados branqueados $E[x^2] = 1$ ambas as versões da curtose se reduzem a:

$$curt(x) = \tilde{\kappa} = E[x^4] - 3. \tag{3.9}$$

Isto significa que para dados brancos, o quarto momento $E[x^4]$ pode ser usado ao invés da curtose para caracterizar a distribuição de x. Curtose é basicamente uma versão normalizada do quarto momento.

Uma propriedade útil da curtose é a sua aditividade. Se x e y são duas variáveis aleatórias estatisticamente independentes, então segue que:

$$curt(x + y) = curt(x) + curt(y). \tag{3.10}$$

Vale ressaltar que esta propriedade não vale para o quarto momento, o que mostra mais uma vez a vantagem de se usar cumulantes no lugar dos momentos. Também, para qualquer escalar α,

$$curt(\alpha x) = \alpha^4 curt(x). \tag{3.11}$$

Uma outra característica importante da curtose é que ela é a quantidade estatística mais simples para indicar a não-normalidade de uma variável aleatória. Para distribuições gaussianas o excesso de curtose é zero e estas são conhecidas como mesocúrticas. Distribuições tendo excesso de curtose negativa são ditas subgaussianas ou platicúrticas e as distribuições com excesso de curtose positiva são chamadas de supergaussianas ou leptocúrticas.

Apesar da curtose ser usada frequentemente como uma medida quantitativa da não-normalidade de uma variável aleatória, alguns cuidados devem ser tomados. A razão disto, é que o excesso de curtose de variáveis supergaussianas podem ter um valor positivo muito grande (o máximo é infinito em príncipio), mas valores negativos do excesso de curtose das variáveis subgaussianas são restringidas pelo valor mínimo possível de -2 (quando a variância é normalizada à unidade). Então comparar a não normalidade de variáveis supergaussianas e subgaussianas uma com a outra, usando simplesmente a curtose, não é apropriado. Entretanto, a curtose pode ser usada como uma medida simples de não-normalidade se as variáveis comparadas são do mesmo tipo, sejam subgaussianas ou supergaussianas.

3.4 Não-correlação e Branqueamento

Se a $Cov(X,Y) = E(XY) - E(X)E(Y) = 0$ dizemos que X e Y são não correlacionadas. Se X e Y são independentes então são não correlacionadas. Porém, não-correlação não implica em independência, ou seja covariância zero não implica necessariamente em independência, a não ser que (X,Y) tenham distribuição gaussiana bivariada. Para mais detalhes ver [18].

Uma propriedade um pouco mais forte que não-correlacionado é "brancura (whiteness)". Brancura de um vetor de média zero, $\mathbf{y}$, significa que seus componentes são não-correlacionados e suas variâncias são unitárias. Em outras palavras, a matriz de covariância (assim como a matriz de correlação) de $\mathbf{y}$ é igual a matriz identidade:

$$E[\mathbf{y}\mathbf{y}^T] = \mathbf{I}. \tag{3.12}$$

Consequentemente, branqueamento significa que transformamos linearmente o vetor de dados

observados $\mathbf{x}$ através de uma multiplicação linear com alguma matriz $\mathbf{V}$

$$\mathbf{z} = \mathbf{V}\mathbf{x}, \tag{3.13}$$

obtendo então um novo vetor $\mathbf{z}$ que é "branco". O branqueamento é sempre possível. Um método bastante popular é usando a decomposição de autovalores da matriz de covariâncias

$$E[\mathbf{x}\mathbf{x}^T] = \mathbf{E}\mathbf{D}\mathbf{E}^T, \tag{3.14}$$

onde $\mathbf{E}$ é a matriz ortogonal de autovetores de $E[\mathbf{x}\mathbf{x}^T]$ e $\mathbf{D}$ é a matriz diagonal de seus autovalores, $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$. Branqueamento pode agora ser feito pela matriz de branqueamento

$$\mathbf{V} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T. \tag{3.15}$$

A matriz de branqueamento calculada desta maneira é denotada por $E[\mathbf{x}\mathbf{x}^T]^{-1/2}$ ou $C^{-1/2}$. Para mais detalhes ver [16]

3.5 Gradientes e Métodos de Otimização

O principal objetivo na ICA é estimar a matriz $\mathbf{W}$ que irá fornecer os componentes independentes. Geralmente $\mathbf{W}$ não pode ser calculada em uma forma fechada, ou seja não podemos escrevê-la como uma função que possa ser avaliada diretamente. Portanto, muitas vezes os métodos de solução são baseados em funções de custo, onde as soluções $\mathbf{W}$ da ICA são encontradas maximizando-se ou minimizando-se estas funções. Minimização de funções multivariadas, possivelmente sob algumas restrições, é assunto da teoria de otimização.

Vetor Gradiente

Considere uma função $g : \mathbb{R}^m \to \mathbb{R}$,

$$g = g(w_1, \ldots, w_m) = g(\mathbf{w}).$$

Assumindo que g seja diferenciável, o vetor gradiente com respeito a $\mathbf{w}$ é o vetor coluna m-dimensional de derivadas parciais

$$\frac{\partial g}{\partial \mathbf{w}} = \begin{pmatrix} \frac{\partial g}{\partial w_1} \\ \vdots \\ \frac{\partial g}{\partial w_m} \end{pmatrix}. \tag{3.16}$$

Em alguns métodos de iteração também é necessário usar gradientes de segunda ordem,

$$\frac{\partial^2 g}{\partial \mathbf{w}^2} = \begin{pmatrix} \frac{\partial^2 g}{\partial w_1^2} & \cdots & \frac{\partial^2 g}{\partial w_1 w_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial^2 w_m w_1} & \cdots & \frac{\partial^2 g}{\partial w_m^2} \end{pmatrix}, \tag{3.17}$$

também conhecida como matriz Hessiana da função $g(\mathbf{w})$. Estes conceitos são generalizados para um vetor de funções de n elementos

$$g(\mathbf{w}) = \begin{pmatrix} g_1(\mathbf{w}) \\ \vdots \\ g_n(\mathbf{w}) \end{pmatrix} \tag{3.18}$$

cujos elementos $g_i(\mathbf{w})$ são eles mesmos funções de $\mathbf{w}$. A matriz Jacobiana com respeito a $\mathbf{w}$ é dada por:

$$\frac{\partial g}{\partial \mathbf{w}} = \begin{pmatrix} \frac{\partial g_1}{\partial w_1} & \cdots & \frac{\partial g_n}{\partial w_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial w_m} & \cdots & \frac{\partial g_n}{\partial w_m} \end{pmatrix}. \tag{3.19}$$

Matriz Gradiente

Em muitos algoritmos temos que considerar uma função escalar g avaliada nos elementos de uma matriz $\mathbf{W}_{m\times n} = (w_{ij})$:

$$g = g(\mathbf{W}) = g(w_{11}, \ldots, w_{ij} \ldots, w_{mn}).$$

Uma função típica deste tipo é o determinante de $\mathbf{W}$. De forma análoga ao vetor gradiente, a matriz gradiente é uma matriz de mesma dimensão que $\mathbf{W}$, cujo ij-ésimo elemento é a derivada

parcial de g em relação a w_{ij}

$$\frac{\partial g}{\partial \mathbf{W}} = \begin{pmatrix} \frac{\partial g}{\partial w_{11}} & \cdots & \frac{\partial g}{\partial w_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial g}{\partial w_{m1}} & \cdots & \frac{\partial g}{\partial w_{mn}} \end{pmatrix}. \tag{3.20}$$

Se $\mathbf{W}$ é uma matriz quadrada invertível temos que:

$$\frac{\partial}{\partial \mathbf{W}} \det \mathbf{W} = (\mathbf{W}^T)^{-1} \det \mathbf{W}, \tag{3.21}$$

onde $\det\mathbf{W}$ é o determinante de $\mathbf{W}$. Para a demonstração do resultado da equação (3.21) ver [16]. Este resultado implica em

$$\frac{\partial \log|\det \mathbf{W}|}{\partial \mathbf{W}} = \frac{1}{\det \mathbf{W}} \frac{\partial |\det \mathbf{W}|}{\partial \mathbf{W}} = (\mathbf{W}^T)^{-1}. \tag{3.22}$$

A equação (3.22) é um exemplo de matriz gradiente composta por funções log, valor absoluto e determinante e é necessária nas soluções dos problemas ICA através da estimação de máxima verossimilhança.

Gradiente de Descida

Muitos dos critérios na ICA tem a forma básica de minimizar a função de custo $\zeta(\mathbf{W})$ em relação a $\mathbf{W}$ ou a uma de suas colunas $\mathbf{w}$. Em muitos casos exitem também restrições que limitam o conjunto de soluções possíveis. Para os problemas irrestritos de minimização de uma função multivariada, a aproximação clássica é o gradiente de descida. Iremos considerar em mais detalhe o caso onde a solução é um vetor $\mathbf{w}$; no caso de ser uma matriz, a solução segue de maneira análoga.

No gradiente de descida, minimizamos a função $\zeta(\mathbf{w})$ iterativamente começando com um ponto inicial $\mathbf{w}(0)$, calculando o gradiente de $\zeta(\mathbf{w})$ neste ponto e então movendo na direção do gradiente negativo através de uma distância adequada. Repete-se então o mesmo procedimento no novo ponto, e assim sucessivamente. Para $t = 1, 2, \ldots$ temos a seguinte regra de atualização:

$$\mathbf{w}(t) = \mathbf{w}(t-1) - \alpha(t) \frac{\partial \zeta(\mathbf{w})}{\partial \mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}(t-1)}, \tag{3.23}$$

com o gradiente tomado no ponto w(t-1). O parâmetro $\alpha(t)$ fornece o comprimento do passo na direção do gradiente negativo, também chamado de *tamanho do passo* ou *taxa de aprendizado.* A iteração (3.23) continua até a convergência, a qual na prática acontece quando a distância euclidiana entre duas soluções subsequentes$||\mathbf{w}(t) - \mathbf{w}(t-1)||$ é menor que algum nível pequeno de tolerância.

Se não há nenhuma razão para enfatizar o tempo ou passo da iteração, uma notação mais conveniente é usada para as regras de atualização. Denotando a diferença entre o novo e o antigo valor por

$$\mathbf{w}(t) - \mathbf{w}(t-1) = \Delta\mathbf{w} \tag{3.24}$$

podemos escrever então (3.23) como

$$\Delta\mathbf{w} = -\alpha\frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}} \Rightarrow \Delta\mathbf{w} \propto \frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}}. \tag{3.25}$$

Em outros casos a taxa de aprendizado pode e deve ser dependente do tempo. Nestes casos, uma terceira forma muito conveniente de escrever tais regras de atualização em conformidade com as linguagens de programação é dada por:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha\frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}}, \tag{3.26}$$

onde o símbolo $\leftarrow$ significa substituição, ou seja, o valor do lado direito da equação é calculado e substituído em $\mathbf{w}$. A desvantagen desse método é que, a menos que $\zeta(\mathbf{w})$ seja muito simples ou suave, a descida mais íngreme ("steepest descent") nos levará perto de um mínimo local ao invés do mínimo global.

Na análise númerica um grande número de métodos mais eficientes são usados na minimização e maximização de funções, e estas podem ser aplicadas ao problema ICA. Sua vantagen se encontra na velocidade de convergência em termos de números de iterações requeridas, mas a desvantangem é um aumento, às vezes, na complexidade computacional por iteração. Vamos considerar o método de segunda ordem, o qual leva em consideração a informação contida nas derivadas de segunda ordem da função de custo.

Desenvolvendo a função de custo $\zeta(\mathbf{w})$ em série de Taylor em volta do ponto $\mathbf{w}$ como

$$\zeta(\mathbf{w}') = \zeta(\mathbf{w}) + \left[\frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}}\right]^T (\mathbf{w}' - \mathbf{w}) + \frac{1}{2}(\mathbf{w}' - \mathbf{w})^T \frac{\partial^2\zeta(\mathbf{w})}{\partial\mathbf{w}^2}(\mathbf{w}' - \mathbf{w}) + \cdots \tag{3.27}$$

na tentativa de minimizar a função $\zeta(\mathbf{w})$, qual seria a escolha do novo ponto $\mathbf{w}'$ que fornece a maior diminuição no valor de $\zeta(\mathbf{w})$? Podemos escrever $\mathbf{w}' - \mathbf{w} = \Delta\mathbf{w}$ e minimizar a função $\zeta(\mathbf{w}') - \zeta(\mathbf{w}) = \left[\frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}}\right]^T \Delta\mathbf{w} + \frac{1}{2}\Delta\mathbf{w}^T \frac{\partial^2\zeta(\mathbf{w})}{\partial\mathbf{w}^2}\Delta\mathbf{w}$ em relação a $\Delta\mathbf{w}$. O gradiente desta função com respeito a $\Delta\mathbf{w}$ é $\frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}} + \frac{\partial^2\zeta(\mathbf{w})}{\partial\mathbf{w}^2}\Delta\mathbf{w}$. Como a matriz Hessiana é simétrica, se também for positiva definida, então a função terá uma forma parabólica e o mínimo será dado pelo zero do gradiente. Ajustando o gradiente a zero, temos

$$\Delta\mathbf{w} = -\left[\frac{\partial^2\zeta(\mathbf{w})}{\partial\mathbf{w}^2}\right]^{-1} \frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}}, \tag{3.28}$$

de onde a seguinte regra de iteração de segunda ordem emerge:

$$\mathbf{w}' = \mathbf{w} - \left[\frac{\partial^2\zeta(\mathbf{w})}{\partial\mathbf{w}^2}\right]^{-1} \frac{\partial\zeta(\mathbf{w})}{\partial\mathbf{w}}. \tag{3.29}$$

O algoritmo (3.29) é chamado de *Método de Newton* e é uma das mais eficientes formas de minimizar funções.

Gradiente Natural e Gradiente Relativo

O espaço paramétrico nem sempre é Euclidiano, podendo ter uma estrutura métrica Riemanniana. Nestes casos a direção de maior inclinação é dada pelo gradiente natural. A regra do gradiente natural de descida para uma função de custo $\zeta(\mathbf{W})$ é dada por:

$$\Delta\mathbf{W} \propto -\frac{\partial\zeta}{\partial\mathbf{w}}\mathbf{w}^T\mathbf{w}. \tag{3.30}$$

Para mais detalhes a respeito de gradiente natural e da demonstração até chegar ao gradiente proposto ver [2].

Um resultado relacionado foi desenvolvido por Cardoso [8] a partir de um ponto de partida levemente diferente. O resultado não será demonstrado aqui, mas para mais detalhes ver [8] e [16].

Esse resultado desenvolvido foi chamado de gradiente relativo.

Método de Lagrange

A forma mais proeminente e largamente usada para levar em conta restrições é o método de Lagrange. A forma da função Lagrangeana é dada por:

$$\mathcal{L}(\mathbf{w}, \lambda_1, \ldots, \lambda_k) = \zeta(\mathbf{w}) + \sum_{i=1}^{k} \lambda_i H_i(\mathbf{w}), \tag{3.31}$$

onde $\lambda_1, \ldots, \lambda_k$ são chamados multiplicadores de Lagrange. O número k é o mesmo número das equações de restrições. O gradiente de $\mathcal{L}(\mathbf{w}, \lambda_1, \ldots, \lambda_k)$ em relação a λ_i é simplesmente a i-ésima função de restrição $H_i(\mathbf{w})$. Um ponto importante é que quando a forma do gradiente de $\mathcal{L}(\mathbf{w}, \lambda_1, \ldots, \lambda_k)$ em relação a $\mathbf{w}$ é zero,

$$\frac{\partial \zeta(\mathbf{w})}{\partial \mathbf{w}} + \sum_{i=1}^{k} \lambda_i \frac{\partial H_i(\mathbf{w})}{\partial \mathbf{w})} = 0 \tag{3.32}$$

muda-se o problema de minimização em um conjunto de equações que é muito mais fácil de ser resolvido.

Métodos de projeção

Muitas vezes os problemas de otimização tem restrições que são relativamente simples: de forma típica, requerem que a norma ou alguma forma quadrática de $\mathbf{w}$ seja constante. Nestes casos, podemos usar uma outra forma de otimização sob essas restrições, que se baseia em projeções sob o conjunto de restrições. Isto significa que podemos resolver o problema de minimização com uma regra de aprendizagem irrestrita, o qual pode ser uma iteração de Newton, descida mais íngreme, ou qualquer outra que seja adequada, mas depois de cada passo da iteração, a solução $\mathbf{w}$ é projetada ortogonalmente sobre o conjunto de restrições, satisfazendo-as.

3.6 Máxima Verossimilhança

Seja $(X_1, \ldots, X_n)$ uma amostra aleatória de tamanho n da variável aleatória X com função densidade $f(x|\theta)$, com $\theta \in \Theta$, onde Θ é o espaço paramétrico. A função de verossimilhança de θ correspondente à amostra aleatória observada é dada por

$$L(\theta; \mathrm{x}) = \prod_{i=1}^{n} f(x_i|\theta). \tag{3.33}$$

O estimador de máxima verossimilhança de θ é o valor $\hat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta; x)$. O logaritmo natural da função de verossimilhança de θ é denotado por

$$l(\theta; \mathrm{x}) = \log L(\theta; \mathrm{x}). \tag{3.34}$$

Como a função log é monótona estritamente crescente, o valor $\hat{\theta}$ que maximiza $\log L(\theta; \mathrm{x})$ é o mesmo que maximiza $L(\theta; \mathrm{x})$.

Se $L(\theta; \mathrm{x})$ satisfizer algumas condições de regularidade, for contínua, duplamente diferenciável e seu domínio não depender do parâmetro, o estimador de máxima verossimilhança é uma raiz da função escore definida por:

$$U(\theta) = \frac{\partial \log L(\theta; \mathrm{x})}{\partial \theta}. \tag{3.35}$$

Em situações mais complicadas, a solução da equação (3.35) será obtida por procedimentos númericos. Nas situações em que o estimador não pode ser obtido através da equação (3.35) o máximo é obtido a partir da inspeção da função de verossimilhança. Para mais detalhes a respeito de estimação e função de verossimilhança consultar [10] e [5].

3.7 Entropia e Negentropia

Na ICA a teoria da informação, juntamente com a teoria da estimação são as duas principais ferramentas teóricas.

Entropia

Entropia é o conceito básico da teoria da informação. Entropia H é definida para uma variável aleatória discreta como

$$H(X) = -\sum_i P(X = a_i) \log P(X = a_i), \tag{3.36}$$

onde a_i são os possíveis valores que a variável aleatória X assume. O resultado de $H(X)$ depende do valor da base do logaritmo, o qual, usualmente, é usado na base 2. A base, geralmente não é mencionada, pois altera somente a medida de escala, não sendo importante.

Definindo a função f como

$$f(p) = -p \log p, \; 0 \leq p \leq 1, \tag{3.37}$$

a entropia pode então ser definida como

$$H(X) = \sum_i f\big(P(X = a_i)\big). \tag{3.38}$$

Considerando a forma de f podemos observar que a entropia é menor se as probabilidades $P(X = a_i)$ estão perto de 0 ou 1, e maior para valores intermediários.

De fato, a entropia de uma variável aleatória pode ser interpretada como o grau de informação que a observação da variável fornece. Quanto mais "aleatório", ou seja, mais imprevisível e desestruturada a variável é, maior será sua entropia.

Entropia Diferencial

A definição de entropia no caso de uma variável aleatória discreta pode ser generalizada para o caso contínuo, tanto para variáveis como para vetores, e é comumente chamada de entropia diferencial H, a qual é definida como

$$H(X) = -\int f_X(x) \log f_X(x) dx = \int f\big(f_X(x)\big) dx \tag{3.39}$$

onde $f_X(x)$ é a densidade da variável aleatória X.

A entropia diferencial pode ser interpretada como uma medida da aleatoriedade da mesma maneira que a entropia.

Distribuições de Máxima Entropia

Uma classe importante de métodos que tem diversas aplicações é dada pelo método de máxima entropia. Suponha que a informação disponível da densidade $f_X(x)$ de uma variável aleatória X é da forma

$$\int f_X(x) F_i(x) dx = c_i, \; i = 1, \cdots, m, \tag{3.40}$$

o que na prática, significa que temos que estimar a esperança $E[F_i(X)]$ de m funções F_i.

A questão agora é qual é a função de probabilidade que satisfaz a restrição em (3.40) e tem a entropia máxima dentre tais densidades? O resultado básico do método de máxima entropia (para mais detalher ver [22]) nos diz que sob algumas condições de regularidade a densidade que satisfaz a restrição (3.40) e tem a máxima entropia dentre todas as densidade, é da forma

$$f_0(x) = A \exp\left(\sum_i a_i F_i(x)\right), \tag{3.41}$$

onde A e a_i são constantes que são determinadas a partir de c_i, usando as restrições em (3.40) e a restrição $\int f_0(x)dx = 1$. Isto, geralmente leva a um sistema de $n+1$ equações não lineares que podem ser de difícil solução, e muitas vezes, métodos numéricos são usados.

Considerando o conjunto de variáveis aleatórias que tomam valores sobre os números reais e tem média zero e uma variância fixa, digamos igual a 1, então a distribuição de máxima entropia para tais variáveis é a distribuição gaussiana. Isto se dá, porque devido a (3.41) a distribuição tem a forma

$$f_0(x) = A \exp(a_1 x^2 + a_2 x) \tag{3.42}$$

e por definição, todas as distribuições dessa forma são gaussianas. Isto nos fornece um resultado fundamental: uma variável gaussiana tem a maior entropia dentro todas as variáveis aleatórias de variância unitária. Então, podemos usar a entropia como uma medida de não-normalidade. Esta propriedade pode ser generalizada para variâncias arbitrárias, e mais importante, para espaços multidimensionais.

Negentropia

Como visto, a propriedade de máxima entropia pode ser usada para definir uma medida de não normalidade. Um medida que seja zero para variáveis gaussianas e não-negativa pode ser obtida da entropia diferencial e é chamada de negentropia J, a qual é definida como

$$J(\mathbf{x}) = H(\mathbf{x}_{normal}) - H(\mathbf{x}), \tag{3.43}$$

onde $\mathbf{x}_{normal}$ é um vetor aleatório normal de mesma matriz de covariância Σ que $\mathbf{x}$ e sua entropia pode ser calculada como

$$H(\mathbf{x}_{normal}) = \frac{1}{2}\log|\det\Sigma| + \frac{n}{2}[1 + \log 2\pi], \tag{3.44}$$

onde n é a dimensão de $\mathbf{x}$.

A negentropia ainda tem uma propriedade adicional bastante interessante que é ser invariante por transformações lineares invertíveis. Em particular, multiplicações de uma variável aleatória por uma constante não altera a sua negentropia.

O problema em usar negentropia se dá na grande dificuldade computacional. Para se usar entropia diferencial ou negentropia na prática, devemos calcular integrais que envolvem funções de densidade de probabilidade o que muitas vezes é complicado. Algumas aproximações então podem ser usadas, como as atráves de cumulantes ou de funções não-polinomiais. Para mais detalhes a respeito dessas aproximações deve-se consultar [16].

3.8 Modelos ARCH e GARCH

Modelos ARCH

Os modelos autoregressivos com heteroscedasticidade condicional, ou ARCH como são conhecidos, foram introduzidos por [11]. Segundo [21] a idéia básica é que o retorno r_t é não correlacionado serialmente, mas a volatilidade (variância condicional) depende de retornos passados por meio de uma função quadrática.

Um modelo $ARCH(m)$ é definido por

$$r_t = \sqrt{h_t}\epsilon_t, \tag{3.45}$$

$$h_t = \alpha_0 + \alpha_1 r_{t-1}^2 + \cdots + \alpha_m r_{t-m}^2, \tag{3.46}$$

onde ϵ_t *i.i.d.* com média zero, $\alpha_0 > 0$, $\alpha_i \geq 0$, $i = 1, \ldots, m-1$, $\alpha_m > 0$ e $h_t = Var(r_t|\mathcal{F}_{t-1})$ com $\mathcal{F}_{t-1}$ denotando a informação até o instante $t-1$(usualmente $r_{t-1}, \ldots, r_1$).

Modelos GARCH

Em [6] foi sugerido uma generalização dos modelos ARCH, chamado GARCH ("Generalized ARCH"). Segundo [21] um modelo GARCH pode ser usado para descrever a volatilidade com menos parâmetros do que o modelo ARCH.

Um modelo $GARCH(m,n)$ é definido por

$$r_t = \sqrt{h_t}\epsilon_t, \tag{3.47}$$

$$h_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i r_{t-1}^2 + \sum_{j=1}^{n} \beta_j h_{t-j}, \tag{3.48}$$

onde ϵ_t *i.i.d.* com média zero, $\alpha_0 > 0$, $\alpha_i \geq 0$, $i = 1, \ldots, m-1$, $\beta_j \geq 0$, $j = 1, \ldots, n-1$, $\alpha_m > 0$, $\beta_n > 0$, $\sum_{i=1}^{q}(\alpha_1 + \beta_1) < 1, q = max(m,n)$.

Para ambos os modelos é comum supor que os ϵ_t sejam normais ou seguem uma distribuição t de student ou uma distribuição de erro generalizada.

Para mais detalhes a respeito dos modelos ARCH e GARCH consultar [21].

3.9 Decorrelação Não-Linear

Um conceito que aparece algumas vezes ao longo do texto já tendo sido citado no capítulo 2 e que merece maior detalhamento é o da decorrelação não-linear.

A correlação está relacionada a independência de tal maneira que variáveis independentes são sempre não-correlacionadas. O contrário não é verdadeiro, logo váriaveis podem ser não-correlacionadas, mas ainda dependentes. Em alguns casos apenas a não-correlação implica em independência. O melhor exemplo destes se dá com a distribuição gaussiana.

Estendendo o conceito de correlação, podemos definir a correlação não-linear de duas variáveis X_1 e X_2 como a $E[f(X_1)g(X_2)]$. Aqui $f(X_1)$ e $g(X_2)$ são duas funções onde pelo menos uma é não-linear. Típicos exemplos podem ser polinômios de grau maior que 1 ou outra função mais complexa como tangente hiperbólica. Isto significa que uma ou as duas variáveis são inicialmente transformadas não-linearmente em duas novas variáveis $f(X_1)$ e $g(X_2)$ para então a correlação

linear usual entre duas variáveis ser considerada.

O problema agora se transforma em: Assumindo que X_1 e X_2 sejam decorrelacionadas não linearmente, ou seja $E\big[f(X_1)g(X_2)\big] = 0$, podemos dizer algo a respeito da independência? Devemos esperar que fazendo esse tipo de correlação não-linear igual a zero, a independência pode ser obtida sob algumas condições adicionais a serem especificadas.

Existe um teorema geral que determina que X_1 e X_2 são independentes se e somente se $E\big[f(X_1)g(X_2)\big] = E\big[f(X_1)\big]E\big[g(X_2)\big]$para funções contínuas f e g que são zero fora de um intervalo finito. Baseado nisto, parece ser muito difícil provar independência rigorosamente, isto porque as funções f e g são arbitrárias. Para mais detalhes a respeito da decorrelação não-linear consultar [16].

Capítulo 4

Análise de Componentes Independentes

4.1 Definição

O modelo básico da ICA é um modelo gerador, o que significa que descreve como os dados observados são gerados por um processo de mistura de componentes s_j, os quais são latentes, o que significa que não podem ser observados diretamente. Podemos defini-lo então, como um modelo estatístico de "váriaveis latentes". Observando n variáveis aleatórias $x_1, x_2, \cdots, x_n$ as quais são modeladas como uma combinação linear de n variáveis aleátorias $s_1, \cdots, s_n$:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n, \text{ para todo } i = 1, \cdots, n \tag{4.1}$$

onde $a_{ij}, i, j = 1, \cdots, n$ são coeficientes reais e pela definição os s_i são mutuamente independentes.

Observe que não foi usado na notação o índice para o tempo t, usado anteriormente. No modelo básico assumimos que cada mistura x_i assim como cada componente independente s_j é uma variável aleatória.

4.2 Restrições à ICA

Para termos certeza que o modelo ICA pode ser estimado devemos fazer algumas suposições e restrições a saber:

1- **Os componentes são supostos independentes estatisticamente.**

Este é o príncipio no qual ICA se baseia. Surpreendentemente, não é necessário muito mais do que essa suposição para nos certificarmos que o modelo pode ser estimado. Este é o motivo pelo qual ICA é um método poderoso com aplicações em diferentes áreas.

2- **Os componentes independentes devem ter distribuições não-gaussianas.**

Podemos dizer de forma intuitiva que as distribuições gaussianas são " simples demais". Seus cumulantes de ordem mais alta são nulos, mas as informações contidas nesses cumulantes são essenciais na estimativa do modelo ICA. Desta forma, é essencialmente impossível realizar as estimativas se as variáveis observadas tiverem distribuição gaussiana. Note que no modelo, não assumimos que conhecemos qual distribuição não-gaussiana os componentes independentes têm; se soubermos, o problema será consideravelmente simplificado.

O branqueamento também ajuda a entender porque variáveis gaussianas são proibidas na ICA. Assumindo que a distribuição conjunta de dois componentes independentes, s_1 e s_2 é gaussiana, temos que a função de densidade de probabilidade é dada por:

$$\begin{aligned} f(s_1, s_2) = & \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \\ & \times exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{s_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{s_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2\right.\right. \\ & \left.\left. -2\rho_{12}\left(\frac{s_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{s_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}, \end{aligned} \tag{4.2}$$

onde σ_{11}, σ_{22}, μ_1, μ_2 e ρ_{12} são respectivamente a variância de s_1 e s_2, a média de s_1e s_2 e a correlação entre s_1 e s_2.

Como já vimos, a média pode ser considerada igual a zero e após o branqueamento as variáveis se tornam não apenas não-correlacionadas, mas com variância igual a 1, então temos que a distribuição conjunta acima fica dada por:

$$f(s_1, s_2) = \frac{1}{2\pi}\exp\left(-\frac{s_1^2+s_2^2}{2}\right) = \frac{1}{2\pi}\exp\left(-\frac{||s||^2}{2}\right). \tag{4.3}$$

Usando a clássica fórmula do método jacobiano (ver 3.19) para transformação de funções de densidade de probabilidade (fdp) e notando que para matriz ortogonal $\mathbf{A}^{-1} = \mathbf{A}^T$ nós temos que a densidade conjunta das misturas x_1 e x_2 é dada por:

$$f(x_1, x_2) = \frac{1}{2\pi}exp\left(-\frac{||\mathbf{A}^T\mathbf{x}||^2}{2}\right)|det\mathbf{A}^T|. \tag{4.4}$$

Devido a ortogonalidade de $\mathbf{A}$, temos que $||\mathbf{A}^T\mathbf{x}||^2 = ||\mathbf{x}||^2$ e $|\det \mathbf{A}| = 1$. Temos então:

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{||x||^2}{2}\right). \tag{4.5}$$

Vemos então que a matriz de mistura ortogonal não muda a fdp. A distribuição original e misturada, são idênticas. Desta forma, não temos como inferir a matriz de mistura a partir das misturas.

3- **Para simplicidade, assumimos que a matriz de mistura desconhecida é quadrada.**

Em outras palavras, o número de componentes independentes é igual ao número de misturas observadas. Esta suposição pode as vezes ser relaxada, porém ela simplifica muito a estimação. Após estimar a matriz $\mathbf{A}$, queremos calcular sua inversa, digamos $\mathbf{B}$, e obter os componentes independentes simplesmente através de:

$$s = B\mathbf{x}. \tag{4.6}$$

Também assumimos aqui que a matriz de mistura é invertível. Se esse não for o caso, é porque existem misturas redundantes que podem ser omitidas e neste caso a matriz não será quadrada. Neste caso o número de misturas não é igual ao número de componentes independentes.

4.3 Ambiguidades da ICA

No modelo ICA, $\mathbf{x} = As$, é fácil verificar as seguintes incertezas ou indeterminações:

1- **Não podemos determinar as variâncias dos componentes independentes.**

Isto se deve ao fato de que tanto s como $\mathbf{A}$ sendo desconhecidos, qualquer escalar α_i multiplicado por alguma das s_i originais pode ser sempre cancelado dividindo a coluna de $\mathbf{A}$ correspondente, a_i, por α_i:

$$\mathbf{x} = \sum_i (\frac{1}{\alpha_i} a_i)(s_i \alpha_i). \tag{4.7}$$

Como consequência, também podemos completamente fixar a magnitude dos componentes independentes. Desde que sejam variáveis aleatórias, o meio mais natural de se fazer isto é assumir que cada um tem variância unitária: $E[s_i^2] = 1$. Então a matriz $\mathbf{A}$ será adaptada no método de solução ICA para levar em conta esta restrição. Observe que isto ainda deixa a ambiguidade do sinal. Podemos multiplicar um componente independente por -1 sem afetar o modelo. Felizmente, essa ambiguidade é insignificante na maioria das aplicações.

2- **Não podemos determinar a ordem dos componentes independentes.**

Novamente, como s e $\mathbf{A}$ são desconhecidos, podemos mudar livremente a ordem dos termos na soma $\mathbf{x} = \sum_i a_i s_i$ e chamar qualquer componente independente de primeiro.

4.4 Centrando as Variáveis

Sem perda de generalidade, podemos assumir que tanto as misturas como os componentes independentes tem média zero. Esta suposição simplifica bastante a teoria e os algoritmos. Se a suposição não for verdadeira, podemos fazer com que seja, centrando as variáveis observáveis, isto é, subtraindo sua média amostral. Isto significa que as misturas originais, ditas x', são transformadas através de

$$\mathbf{x} = \mathbf{x}' - \bar{\mathbf{x}} \tag{4.8}$$

antes de se realizar a ICA. Então, os componentes independentes tem média zero também, pois

$$\bar{s} = A^{-1}\bar{\mathbf{x}}. \tag{4.9}$$

Por outro lado, a matriz de mistura permanece a mesma após a mudança de posição. Portanto, podemos sempre fazer isto sem que a estimação da matriz de mistura seja afetada. Após a estimação da matriz de mistura e dos componentes independentes de média zero, a média subtraída pode ser simplesmente reconstruída pela adição $A^{-1}\bar{\mathbf{x}}$, ao componente independente de média zero.

4.5 ICA pela Maximização da Não-Normalidade

Nesta seção abordaremos um princípio simples e intuitivo para a estimação do modelo ICA o qual será baseado na maximização da não-normalidade.

Retomemos (4.1) na forma matricial:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{4.10}$$

isto é, uma mistura de componentes independentes. Vamos assumir também que todos os componentes independentes tem a mesma distribuição. A estimação dos componentes independentes pode ser executada achando as combinações lineares certas das variáveis de mistura, desde que possamos inverter a mistura como

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}. \tag{4.11}$$

Então, para estimar um dos componentes independentes podemos considerar uma combinação linear dos x_i. Denotando isto por $y = \mathbf{A}^T\mathbf{x} = \sum_i b_i x_i$, onde $\mathbf{b}$ é um vetor a ser determinado. Também temos que $y = \mathbf{b}^T\mathbf{A}\mathbf{s}$. Então, y é uma combinação linear dos s_i, com coeficientes dados por $\mathbf{b}^T\mathbf{A}$. Denotando este vetor por $\mathbf{q}$, temos:

$$y = \mathbf{b}^T\mathbf{x} = \mathbf{q}^T\mathbf{s} = \sum_i q_i s_i. \tag{4.12}$$

Se $\mathbf{b}$ for uma das linhas da inversa da matriz de mistura $\mathbf{A}$, esta combinação linear da Eq.(4.12) será igual a um dos componentes independentes. Neste caso, o correspondente $\mathbf{q}$ será tal que apenas um dos seus elementos é 1 e todos os outros zero. A questão agora se transforma em como podemos usar o teorema central do limite para determinar $\mathbf{b}$ de modo que seja igual a uma das linhas da inversa de $\mathbf{A}$. Na prática não podemos determinar $\mathbf{b}$ exatamente porque não temos conhecimento a respeito da matriz $\mathbf{A}$, mas podemos encontrar um estimador que dê uma boa aproximação.

Variando os coeficientes em $\mathbf{q}$ e vendo como a distribuição de $y = \mathbf{q}^T\mathbf{s}$ muda, usamos a ideia fundamental de que a soma de duas variáveis independentes se aproxima mais da distribuição gaussiana do que as variávies originais e y é usualmente mais próximo do que os s_i e torna-se menos próximo quando de fato iguala um dos s_i. Devemos notar aqui que isto é estritamente verdade se os s_i tem a mesma distribuição, como assumimos anteriormente. E neste caso, claramente apenas um dos elementos q_i de $\mathbf{q}$ não é zero.

Na prática não conhecemos os valores de $\mathbf{q}$, mas não precisamos saber, porque pela definição $\mathbf{q}^T\mathbf{s} = \mathbf{b}^T\mathbf{x}$. Devemos apenas variar $\mathbf{b}$ e observar a distribuição de $\mathbf{b}^T\mathbf{x}$.

Então, podemos tomar como $\mathbf{b}$ um vetor que maximiza a não-normalidade de $\mathbf{b}^T\mathbf{x}$. Tal vetor necessariamente corresponde a $\mathbf{q} = \mathbf{A}^T\mathbf{b}$, o qual tem apenas um elemento diferente de zero. Isto significa que $y = \mathbf{b}^T\mathbf{x} = \mathbf{q}^T\mathbf{s}$ é igual a um dos componentes independentes. Maximizar a não-normalidade de $\mathbf{b}^T\mathbf{x}$ nos dá assim um dos componentes independentes.

De fato, a otimização da não-normalidade no espaço n-dimensional de vetores $\mathbf{b}$ têm $2n$ locais de máximo, dois para cada componente independente, correspondendo a s_i e $-s_i$ (lembrando que os componentes independentes podem ser estimados somente a menos de um sinal multiplicativo).

De um ponto de vista prático, devemos responder agora as seguintes perguntas: Como pode ser medida a não-normalidade de $\mathbf{b}^T\mathbf{x}$? E como podemos calcular os valores de $\mathbf{b}$ que maximizam (localmente) tal medida de não-normalidade?

4.5.1 Medindo Não-Normalidade pela Curtose

Existem variáveis aleatórias não-gaussianas cujo excesso de curtose é zero, mas estas podem ser consideradas muito raras. Tipicamente, o excesso de curtose é zero para distribuições normais, positiva para as distribuições supergaussinas (leptocúrticas) e negativas para as distribuição subgaussianas (platicúrticas). Para ilustrar em um simples exemplo a otimização da curtose e como encontrar os componentes independentes, considere o modelo bidimensional $\mathbf{x} = \mathbf{A}\mathbf{s}$. Assumindo que os componentes independentes s_1 e s_2 tem curtose igual a $\text{curt}(s_1)$ e $\text{curt}(s_2)$, respectivamente e ambas diferentes de zero, temos para $y = q_1 s_1 + q_2 s_2$

$$curt(y) = curt(q_1 s_1) + curt(q_2 s_2) = q_1^4 curt(s_1) + q_2^4 curt(s_2). \tag{4.13}$$

Na seção (4.3) pelo pré-processamento os cálculos foram realizados para variáveis com média zero e variância unitária, logo variância de y é igual a 1. Isto implica na seguinte restrição sobre $\mathbf{q}$: $E(y^2) = q_1^2 + q_2^2 = 1$. Geometricamente, isto significa que o vetor $\mathbf{q}$ está restrito ao círculo unitário no plano bidimensional.

O problema de otimização agora é encontrar o máximo da função $|curt(y)| = |q_1^4 curt(s_1) + q_2^4 curt(s_2)|$ sobre o círculo unitário. Podemos verificar que o máximo ocorre exatamente onde um dos elementos do vetor $\mathbf{q}$ é zero e o outro não. Por causa das restrições do círculo unitário, o elemento não nulo deve ser igual a 1 ou -1. Mas estes pontos são exatamente aqueles quando y é igual à componente independente $\pm s_i$ e assim o problema está resolvido. Para mais detalhes ver [16].

Para dados $\mathbf{z}$ que passaram pelo pré-processamento de branqueamento procuramos por uma combinação linear $\mathbf{w}^T\mathbf{z}$ que maximiza a não-normalidade. Desde que se tenha $\mathbf{q} = (\mathbf{VA})^T\mathbf{w}$,

então:

$$||q||^2 = (\mathbf{w}^T\mathbf{VA})(\mathbf{A}^T\mathbf{V}^T\mathbf{w}) = ||\mathbf{w}||^2. \tag{4.14}$$

Isto significa que restringir $\mathbf{q}$ à esfera unitária é equivalente a restringir $\mathbf{w}$ à mesma esfera. Sendo assim, basta maximizar o valor absoluto da curtose de $\mathbf{w}^T\mathbf{z}$ sob a simples restrição que $||\mathbf{w}|| = 1$. Também, após o branqueamento, as combinações lineares $\mathbf{w}^T\mathbf{z}$ podem ser interpretadas como projeções sobre a linha transposta pelo vetor $\mathbf{w}$. Cada ponto sob o círculo unitário corresponde a uma projeção. Isto mostra a utilidade do branqueamento.

Algoritmo do Gradiente usando a Curtose

Na prática, para maximizar o valor absoluto da curtose, poderiamos começar a partir de algum vetor $\mathbf{w}$, calcular a direção na qual a valor absoluto da curtose de $y = \mathbf{w}^T\mathbf{z}$ está crescento mais fortemente, baseado numa amostra disponível $\mathbf{z}(1), \cdots, \mathbf{z}(T)$ de misturas do vetor $\mathbf{z}$, e então mover o vetor $\mathbf{w}$ nesta direção. Esta idéia é implementada no método do gradiente e sua extensões.

O gradiente do valor absoluto da curtose de $\mathbf{w}^T\mathbf{z}$ pode ser calculado como:

$$\frac{\partial|curt(\mathbf{w}^T\mathbf{x})|}{\partial\mathbf{w}} = 4sinal\big(curt(\mathbf{w}^T\mathbf{z})\big)\Big[E\big\{z(\mathbf{w}^T\mathbf{z})^3\big\} - 3\mathbf{w}||\mathbf{w}||^2\Big], \tag{4.15}$$

pois para dados branqueados $E\big[(\mathbf{w}^T\mathbf{z})^2\big] = ||\mathbf{w}||^2$. Como queremos otimizar esta função sobre a esfera unitária $||\mathbf{w}||^2 = 1$, o método do gradiente deve ser complementado pela projeção de $\mathbf{w}$ sobre o círculo unitário após cada etapa. Isto pode ser feito dividindo $\mathbf{w}$ por sua norma.

Para facilitar a simplicação do algoritmo, podemos omitir o segundo termo entre colchetes, pois pode ser simplesmente mudado a norma de $\mathbf{w}$ no algoritmo do gradiente e não a sua direção. Isto se deve ao fato de apenas a direção de $\mathbf{w}$ ser interessante, e qualquer mudança na norma é insignificante, pois a norma é normalizada à unidade de qualquer maneira.

Deste modo, obtemos o seguinte algoritmo do gradiente:

$$\Delta\mathbf{w} \propto sinal\big[curt(\mathbf{w}^T\mathbf{z})\big]E\big[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3\big], \tag{4.16}$$

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{||\mathbf{w}||}.$$

Uma versão adaptativa deste algoritmo pode ser obtida também. Isto é possível tomando os valores instantâneos de $\mathbf{z}$ ao se omitir a esperança do segundo termo:

$$\Delta\mathbf{w} \propto sinal\big[curt(\mathbf{w}^T\mathbf{z})\big]\mathbf{z}(\mathbf{w}^T\mathbf{z})^3, \tag{4.17}$$

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{||\mathbf{w}||}.$$

Então cada obervação $\mathbf{z}(t)$ pode ser usada no algoritmo uma vez.

Algoritmo de Ponto Fixo usando Curtose

Na seção anterior usamos o método do gradiente. A vantagem desse método é que os "inputs" $\mathbf{z}(t)$ podem ser usados no algoritmo uma vez, permitindo uma adaptação rápida em ambientes não estacionários. Porém, a convergência é lenta e depende de uma boa escolha da velocidade de aprendizado das sequências. Uma escolha ruim, na prática, destrói a convergência. Então, algumas maneiras de tornar o aprendizado mais rápido e mais confiável podem ser necessários. Os algoritmos de iteração de ponto fixo são algumas alternativas.

Para derivar uma iteração de ponto fixo mais eficiente, o gradiente deve apontar na direção de $\mathbf{w}$, ou seja, o gradiente deve ser igual a $\mathbf{w}$ multiplicado por algum escalar constante. Isto significa que após normalização à norma unitária, o valor de $\mathbf{w}$ não muda, exceto possivelmente pela mudança em seu sinal. Equacionando o gradiente da curtose em (4.15) com $\mathbf{w}$, isto significa que devemos ter:

$$\mathbf{w} \propto \Big[E\{z(\mathbf{w}^T\mathbf{z})^3\} - 3||\mathbf{w}||^2\mathbf{w}\Big]. \tag{4.18}$$

Esta equação sugere um algortimo de ponto fixo onde primeiro calculamos o lado direito que nos dá um novo valor para $\mathbf{w}$:

$$\mathbf{w} \leftarrow E\{z(\mathbf{w}^T\mathbf{z})^3\} - 3\mathbf{w} \tag{4.19}$$

Após cada iteração de ponto-fixo, $\mathbf{w}$ é dividido por sua norma para permanecer sob as restrições ($||\mathbf{w}|| = 1$ sempre, por isso podemos omitir na Eq.(4.18).). O vetor $\mathbf{w}$ final dá um dos componenentes independentes como combinação linear de $\mathbf{w}^T\mathbf{z}$. Na prática as esperanças em (4.19) devem ser substituidas por suas estimativas.

Atualmente um algoritmo que trabalha muito bem, convergindo muito rápido e de forma bastante confiável é conhecido como FastICA. Este algoritmo tem duas propriedades que o tornam claramente superior ao algoritmo do gradiente na maioria dos casos. Primeiramente, a convergência desse algoritmo é cúbica, o que a torna muito rápida. E em segundo lugar, contrariamente ao algortimo do gradiente não há taxa de aprendizado ou outros parâmetros a serem ajustados o que o torna muito mais fácil de usar e confiável. O algoritmo do gradiente parece ser preferível somente em casos onde uma adaptação rápida em ambientes em mundança é necesssário.

4.5.2 Medindo Não-Normalidade pela Negentropia

Na seção anterior vimos como medir a não-normalidade pela curtose. Entretanto, a curtose tem algumas desvantagens na prática. O principal problema é que a curtose é muito sensível a outliers. Isso significa que o valor da curtose pode depender somente de poucas observações nas caudas das distribuições, as quais podem ser observações erradas ou irrelevantes. Em outras palavras, curtose não é uma medida robusta de não-normalidade.

Portanto, outras medidas podem ser melhores que a curtose em algumas situações. Nesta seção será considerada a negentropia cujas propriedades são de várias maneiras opostas às da curtose. É uma medida robusta, mas computacionalmente complicada.

A vantagem de se usar negentropia como medida de não normalidade é que esta é justificada pela teoria estatística. De fato, a negentropia é em algum sentido o estimador ótimo de não-normalidade. O problema em usar a negentropia é que é computacionalmente muito difícil. Estimar a negentropia usando a definição poderá requerer uma estimação (possivelmente não-paramétrica) da fdp. Entretanto, aproximações simples da negentropia são muito úteis.

Segundo [16], na prática somente precisamos de aproximações unidimensionais da (neg)entropia, então iremos considerar o caso escalar somente.

O método clássico de aproximação da negentropia é usando cumulantes de ordem mais alta, usando a expansão polinomial da densidade e o resultado dado pela aproximação é:

$$J(Y) \approx \frac{1}{12}\left[E(Y^3)\right]^2 + \frac{1}{48}\left[curt(Y)\right]^2. \tag{4.20}$$

Supõe-se que a variável aleatória y tem média zero e variância unitária. Esta aproximação muitas vezes conduz ao uso da curtose como visto na seção anterior. Isto ocorre porque o primeiro termo do lado direito da igualdade de (4.20) é zero no caso de variáveis aleatórias com distribuções simétricas, que são muito comum. Neste caso a aproximação (4.20) é equivalente ao quadrado da curtose. Maximização do quadrado da curtose é equivalente a maximização de seu valor absoluto. Esta aproximação, em particular sofre do mesmo problema de falta de robustez encontrada na curtose.

Uma aproximação útil é generalizar os cumulantes de ordem mais alta e então usar as esperanças de funções gerais não-quadráticas. Para mais detalhes, ver [16]. Em geral podemos substituir funções polinomiais y^3 e y^4 por qualquer outra função G_i, possivelmente mais que duas. O método então fornece uma forma simples de aproximação da negentropia baseado nas esperanças $E[G_i(y)]$. Como um caso especial simples, podemos tomar qualquer função não-quadrática G_1 e G_2 que sejam ímpar e par, respectivamente, obtendo a seguinte aproximação:

$$J(y) \approx k_1\Big[E\{G_1(y)\}\Big]^2 + k_2\Big[E\{G_2(y)\} - E\{G_2(\nu)\}\Big]^2 \tag{4.21}$$

onde k_1 e k_2 são constantes positivas e ν é uma variável gaussiana de média zero e variância unitária. Para a variável y supõe-se ter média zero e variância unitária e podemos notar que mesmo em casos onde esta aproximação não é muito acurada, (4.21) pode ser usada para construir uma medida de não-normalidade que seja consistente no sentido de que será sempre não-negativa e igual a zero se y tiver distribuição gaussiana.

No caso onde se usa apenas uma função não-quadrática G, a aproximação se transforma em:

$$J(y) \propto \Big[E\{G(y)\} - E\{G(\nu)\}\Big]^2, \tag{4.22}$$

para praticamente qualquer função G não-quadrática. Mas a questão aqui é que escolhendo G de forma certa, obtemos aproximações melhores de negentropia do que aquelas dadas por (4.20). Em particular, escolhendo uma G que cresce rapidamente, obtemos estimadores mais robustos. As seguintes escolhas de G tem se provado muito úteis:

$$G_1(y) = \frac{1}{a_1}\log\cosh(a_1 y), \tag{4.23}$$

$$G_2(y) = -\exp(-y^2/2)$$

onde $1 \leq a_1 \leq 2$ é alguma constante adequada, muitas vezes assumindo valores iguais a 1.

Assim obtemos aproximações da negentropia que está de acordo com as propriedades destas duas medidas clássicas de não-normalidade. São conceitualmente simples, rápidas de calcular, e ainda tem apelo de propriedades estatísticas, especialmente a robustez. Portanto, devemos usar estas funções em nossos métodos ICA. Para mais detalhes sobre o assunto, ver [16].

Algoritmo do Gradiente usando Negentropia

Como feito anteriormente para curtose, podemos derivar um algoritmo do gradiente para maximizar a negentropia. Tomando o gradiente da aproximação de negentropia em (4.22) com respeito a $\mathbf{w}$, e tomando a normalização $E\big[(\mathbf{w}^T\mathbf{z})^2\big] = ||\mathbf{w}||^2 = 1$ obtemos o seguinte algoritmo:

$$\Delta\mathbf{w} \propto \gamma E\big[\mathbf{z}g(\mathbf{w}^T\mathbf{z})\big], \tag{4.24}$$

$$\mathbf{w} \leftarrow \mathbf{w}/||\mathbf{w}||,$$

onde $\gamma = E\big[G(\mathbf{w}^T\mathbf{z})\big] - E\big[G(\nu)\big]$, ν sendo definido como anteriormente (variável aleatória normal padronizada). A normalização é necessária para projetar $\mathbf{w}$ sobre a esfera unitária para manter a variância de $\mathbf{w}^T\mathbf{z}$ constante. A função g é a derivada da função G usada na aproximação da negentropia. A constante γ, que fornece ao algoritmo uma tipo de qualidade de "auto-adaptação", pode ser estimada como segue:

$$\Delta\gamma \propto \Big[G(\mathbf{w}^T\mathbf{z}) - E\big\{G(\nu)\big\}\Big] - \gamma \tag{4.25}$$

a qual corresponde ao sinal da curtose na Eq.(4.15).

Para as funções g, podemos usar as derivadas em (4.23) que fornecem uma aproximação robusta da negentropia. Alternativamente, podemos usar a derivada correspondente à quarta potência como na curtose. Podemos escolher a partir de:

$$g_1(y) = tanh(a_1 y), \tag{4.26}$$

$$g_2(y) = y \exp(-y^2/2),$$

$$g_3(y) = y^3.$$

Algoritmo de Ponto Fixo usando Negentropia

Aqui também podemos usar um método de maximização da negentropia mais rápido do que o algoritmo do gradiente e este novamente, pode ser achado usando um algoritmo de ponto fixo. O algortimo FastICA resultante encontra uma direção, ou seja, um vetor unitário $\mathbf{w}$, tal que a projeção $\mathbf{w}^T\mathbf{z}$ maximiza a não-normalidade a qual aqui é medida pela aproximação da negentropia $J(\mathbf{w}^T\mathbf{z})$ dada em (4.22). Devemos lembrar que a variância de $\mathbf{w}^T\mathbf{z}$ deve ser restrita à unidade; para dados branqueados, isto é equivalente a restrição da norma de $\mathbf{w}$ ser unitária.

Olhando para o método do gradiente em (4.24) podemos sugerir a seguinte iteração de ponto fixo:

$$\mathbf{w} \leftarrow E\big[\mathbf{z}g(\mathbf{w}^T\mathbf{z})\big], \tag{4.27}$$

a qual pode ser seguida por uma normalização de $\mathbf{w}$. O coeficiente γ pode ser omitido porque de qualquer forma será eliminado pela normalização. Entretanto, a iteração em (4.27) não tem as mesmas propriedades ótimas de convergência do FastICA usando curtose. Isto se deve ao fato de que momentos não-polinomiais não têm as mesmas propriedades algébricas dos cumulantes. Então, a iteração em (4.27) tem que ser modificada, sendo possível adicionando a $\mathbf{w}$ uma constante α multiplicada em ambos os lados de (4.27) sem modificar os pontos fixos. Assim, temos:

$$\mathbf{w} = E\big[\mathbf{z}g(\mathbf{w}^T\mathbf{z})\big] \tag{4.28}$$

$$\Leftrightarrow$$

$$(1+\alpha)\mathbf{w} = E\big[\mathbf{z}g(\mathbf{w}^T\mathbf{z})\big] + \alpha\mathbf{w}, \tag{4.29}$$

e devido a subsequente normalização de $\mathbf{w}$ à norma unitária, a equação (4.27) fornece a iteração de ponto fixo que tem os mesmos pontos. Desta forma, escolhendo α de modo correto é possível obter um algoritmo que convirja tão rápido quanto o algoritmo de ponto fixo usando curtose.

O coeficiente α apropriado e, consequentemente, o algoritmo FastICA podem ser encontrados usando uma aproximação do método de Newton, que quando aplicado sobre o gradiente fornece um método de otimização que, frequentemente, converge em um número pequeno de passos. O problema deste método é que geralmente ele requer uma inversão de matriz a cada passo, o que faz com que o esforço computacional possa ser maior do que nos métodos do gradiente. Totalmente surpreendente é o fato de que usando propriedades especiais do problema da ICA, podemos encontrar uma aproximação do método de Newton que não necessita de uma matriz inversa a cada passo mas ainda converge, aproximadamente, com o mesmo número de iterações do método real de Newton (pelo menos na teoria). Este método de aproximação fornece um algoritmo de ponto fixo da forma (4.29).

Para derivar a aproximação do método de Newton, primeiro devemos notar que os máximos da aproximação da negentropia de $\mathbf{w}^T\mathbf{z}$ são tipicamente obtidos em determinadas situações ótimas de $E\big[G(\mathbf{w}^T\mathbf{z})\big]$, a qual de acordo com as condições de Lagrange tem seu ótimo sob a restrição $E\big[(\mathbf{w}^T\mathbf{z})^2\big] = ||\mathbf{w}||^2 = 1$ obtidos em pontos onde o gradiente Lagrangeano é zero:

$$E\big[\mathbf{z}g(\mathbf{w}^T\mathbf{z})\big] + \beta\mathbf{w} = 0. \tag{4.30}$$

Denotando a função do lado esquerdo da Eq.(4.30) por F, obtemos seu gradiente (que é a derivada segunda do Lagrangeano) como

$$\frac{\partial F}{\partial \mathbf{w}} = E\big[\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T\mathbf{z})\big] + \beta\mathbf{I}. \tag{4.31}$$

Para simplificar a inversão desta matriz, devemos aproximar o primeiro termo da Eq.(4.31). Uma aproximação razoável parece ser $E\big[\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T\mathbf{z})\big] \approx E\big[\mathbf{z}\mathbf{z}^T\big]E\big[g'(\mathbf{w}^T\mathbf{z})\big] = E\big[g'(\mathbf{w}^T\mathbf{z})\big]\mathbf{I}$. Assim, o gradiente se torna diagonal e pode ser facilmente invertido. Então, obtemos a seguinte aproximação para a iteração de Newton:

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{E\big[\mathbf{z}g(\mathbf{w}^T\mathbf{z})\big] + \beta\mathbf{w}}{E\big[g'(\mathbf{w}^T\mathbf{z})\big] + \beta}. \tag{4.32}$$

Este algoritmo ainda pode ser simplificado pela multiplicação por $\beta + E\big[g'(\mathbf{w}^T\mathbf{z})\big]$ em ambos os lados de (4.32), o que fornece após uma simplificação algébrica o seguinte resultado:

$$\mathbf{w} \leftarrow E\Big[\mathbf{z}g(\mathbf{w}^T\mathbf{z}) - E\big[g'(\mathbf{w}^T\mathbf{z})\big]\mathbf{w}\Big]. \tag{4.33}$$

A Eq.(4.33) é a iteração básica de ponto fixo no algoritmo FastICA.

4.6 ICA pela Estimação de Máxima Verossimilhança

Não é muito difícil derivar a verossimilhança de um modelo ICA livre de ruído. De acordo com o método do Jacobiano para transformações de variáveis, a densidade f_X do vetor de mistura

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{4.34}$$

pode ser escrita como

$$f_X(x) = |\det \mathbf{B}| f_X(s) = |\det \mathbf{B}| \prod_i f_i(s_i), \tag{4.35}$$

onde $\mathbf{B} = \mathbf{A}^{-1}$ e f_i denota as densidades dos componentes independentes. Esta também pode ser expressa como uma função de $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_n)^T$ e $\mathbf{x}$, dando:

$$f_X(x) = |\det B| \prod_i f_i(\mathbf{b}_i^T\mathbf{x}). \tag{4.36}$$

Assumindo que temos T observações de $\mathbf{x}$, denotadas por $\mathbf{x}(1),\cdots,\mathbf{x}(\mathrm{T})$, a verossimilhança pode ser obtida pelo produto desta densidade avaliada nos T pontos. Denotando-a por L e considerada como uma função de $\mathbf{B}$, temos

$$L(\mathbf{B}) = \prod_{t=1}^{T} \prod_{i=1}^{n} f_i\big(\mathbf{b}_i^t \mathbf{x}(t)\big) |\det B|. \tag{4.37}$$

Então a log-verossimilhança fica dada por:

$$\log L(\mathbf{B}) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i\big(\mathbf{b}_i^t \mathbf{x}(t)\big) + T \log |\det B|. \tag{4.38}$$

Para simplificar a notação e torná-la consistente com o que foi usado nas seções anteriores, vamos denotar a soma sobre a amostra indexada em t através da esperança e dividir a verossimilhança por T, obtendo:

$$\frac{1}{T} \log L(\mathbf{B}) = E\left[\sum_{i=1}^{n} \log f_i\big(\mathbf{b}_i^t \mathbf{x}\big)\right] + \log |\det B|. \tag{4.39}$$

A esperança aqui não é a teórica, mas uma média calculada a partir da amostra observada. Desta forma, expressamos a verossimilhança como uma função dos parâmetros do modelo, os quais são os elementos da matriz de mistura.

Neste ponto, nos deparamos com um problema. Ainda temos que estimar a densidade dos componentes independentes, e isto torna o problema muito mais complicado, porque em geral, a estimação de densidades é um problema não-paramétrico, o que significa que não podemos reduzir a estimação a um conjunto finito de parâmetros. De fato, o número de parâmetros a ser estimado é infinito, ou na prática, extremamente grande, sendo assim a estimação dos modelos ICA tem uma parte não-paramétrica, por isso a estimação é chamada às vezes de "semiparamétrica".

A estimação não-paramétrica de densidades é conhecida por ser um problema difícil. Este é o motivo pelo qual gostariamos de evitar esse tipo de estimação na ICA; e existem duas maneiras de fazê-lo. Primeiramente, em alguns casos é possível conhecer as densidades dos componentes independentes, usando algum conhecimento a priori sobre os dados. Neste caso, podemos simplesmente usar essas densidades a priori na verosimilhança e esta seria somente uma função de $\mathbf{B}$. Se pequenos erros na especificação da densidade a priori tem pequena influência sobre os estimadores, o procedimento dará resultados razoáveis. A segunda opção seria aproximar as densidades dos componentes independentes por uma família de densidades que são especificadas por um número limitado de parâmetros. Se o número de parâmetros na família de densidades necessitar ser muito grande, não ganhamos muito com a aproximação. Entretanto, se for possível usar uma família

simples de densidades para estimar o modelo ICA para qualquer densidade f_i, teremos conseguido uma solução simples. Felizmente, podemos usar uma simples parametrização de f_i, consistindo na escolha entre duas densidades, isto é, um único parâmetro binário. Para mais detalhes a respeito deste assunto, bem como o enunciado do teorema e sua prova, deve-se consultar [16].

Algoritmos para Estimação de Máxima Verossimilhança

Para executar a estimação de máxima verossimilhança precisamos de um algoritmo que apresente a maximização numérica da verossimilhança. O algoritmo mais simples é obtido pelo método do gradiente. Para isto, necessitamos derivar o gradiente estocástico da log-verossimilhança em (4.39) como

$$\frac{1}{T}\frac{\partial \log L(\mathbf{B})}{\partial \mathbf{B}} = [\mathbf{B}^T]^{-1} + E\big[\mathbf{g}(\mathbf{B}\mathbf{x})\mathbf{x}^T\big], \tag{4.40}$$

onde $\mathbf{g}(\mathbf{y}) = \big(g_i(y_i), \cdots, g_n(y_n)\big)$ é um componente que consiste em funções escore g_i das distribuições de s_i, definidas por:

$$g_i = (\log p_i)' = \frac{f_i'}{f_i}. \tag{4.41}$$

De onde segue o algoritmo para estimação de máxima verossimilhança (EMV):

$$\Delta\mathbf{B} \propto [\mathbf{B}^T]^{-1} + E\big[\mathbf{g}(\mathbf{B}\mathbf{x})\mathbf{x}^T\big]. \tag{4.42}$$

Uma versão estocástica desse algoritmo também pode ser usada. Isto significa que a esperança é omitida e em cada passo do algoritmo somente um ponto dos dados é usado:

$$\Delta\mathbf{B} \propto [\mathbf{B}^T]^{-1} + \mathbf{g}(\mathbf{B}\mathbf{x})\mathbf{x}^T. \tag{4.43}$$

Este algoritmo é muitas vezes chamado de algoritmo de *Bell-Sejnowski*. O algoritmo em (4.42) converge muito lentamente, principalmente devido a inversão da matriz $\mathbf{B}$ que é necessária em cada passo. A convergência pode ser melhorada pelo branqueamento dos dados e especialmente pelo uso do gradiente natural.

O método do gradiente natural simplifica a maximização da verossimilhança consideravelmente. O príncipio é baseado na estrutura geométrica do espaço paramétrico e está relacionado ao princípio do gradiente relativo, que usa a estrutura dos grupos dos problemas ICA. No caso básico da ICA, ambos os princípios são alcançados multiplicando o lado direito da Eq.(4.42) por $\mathbf{B}^T\mathbf{B}$, obtendo-se:

$$\Delta\mathbf{B} \propto [\mathbf{I} + E[\mathbf{g}(\mathbf{y})\mathbf{y}^T]\mathbf{B}. \tag{4.44}$$

Este algoritmo também pode ser interpretado como uma decorrelação não-linear. O princípio é que o algoritmo convirja quando $E[\mathbf{g}(\mathbf{y})\mathbf{y}^T] = \mathbf{I}$, o que significa que y_i e $g_j(y_j)$ são não-correlacionados para $i \neq j$. Para mais detalhes a respeito, bem como as densidades usadas para os casos supergaussianos e subgassianos, veja [16].

A verossimilhança também pode ser estimada por um algoritmo de ponto fixo e atualmente o FastICA pode ser aplicado diretamente à maximização da verossimilhança. Este algoritmo foi derivado nas seções anteriores para a otimização da $E[G(\mathbf{w}^T\mathbf{Z})]$ sob a restrição da norma unitária de $\mathbf{w}$. De fato, a maximização da verossimilhança dá um problema de otimização quase idêntico se restringirmos os estimadores dos componentes independentes a serem brancos. Isto significa que o termo $\log|\det\mathbf{B}|$ é constante e então a verossimilhança consiste basicamente da soma de n termos da forma otimizada pelo FastICA. Sendo assim, podemos usar diretamente o mesmo tipo de derivação da iteração de ponto fixo da seção anterior.

Na eq.(4.32) tinhamos a forma do algoritmo FastICA para dados branqueados, onde β pode ser calculado a partir da eq.(4.30) com $\beta = -E[y_i g(y_i)]$. Escrevendo em forma matricial, temos

$$\mathbf{W} \leftarrow \mathbf{W} + \mathrm{diag}(\alpha_i)\Big[\mathrm{diag}(\beta_i) + E\{\mathbf{g}(\mathbf{y})\mathbf{y}^T\}\Big]\mathbf{W}, \tag{4.45}$$

onde α_i é definido como $-1/\Big[E\{g'(\mathbf{w}^T\mathbf{z}) + \beta_i\}\Big]$ e $\mathbf{y} = \mathbf{W}\mathbf{z}$. Para dados não-branqueados podemos substituir $\mathbf{W}$ por $\mathbf{B}$, desde que tenhamos $\mathbf{W}\mathbf{z} = \mathbf{W}\mathbf{V}\mathbf{x}$ o que implica em $\mathbf{B} = \mathbf{W}\mathbf{V}$. Assim, obtemos a iteração básica do FastICA como:

$$\mathbf{B} \leftarrow \mathbf{B} + \mathrm{diag}(\alpha_i)\Big[\mathrm{diag}(\beta_i) + E\{\mathbf{g}(\mathbf{y})\mathbf{y}^T\}\Big]\mathbf{B}, \tag{4.46}$$

onde $\mathbf{y} = \mathbf{B}\mathbf{x}, \beta_i = -E[y_i g(y_i)]$ e $\alpha_i = -1/\Big[\beta_i + E\{g'(y_i)\}\Big]$. Após cada passo, a matriz $\mathbf{B}$ deve ser projetada sobre o conjunto de matrizes brancas. Isso pode ser executado pelo método clássico

envolvendo raízes quadradas de matrizes,

$$\mathbf{B} \leftarrow (\mathbf{B}\mathbf{C}\mathbf{B}^T)^{-1/2}\mathbf{B}, \tag{4.47}$$

onde $\mathbf{C}= E[\mathbf{x}\mathbf{x}^T]$ é a matriz de correlação dos dados. A inversa da raiz quadrada é obtida como em (3.15).

Podemos observar assim que FastICA pode ser considerado como uma versão otimizada do algoritmo do gradiente, pois sua convergência é rápida e otimizada pela escolha das matrizes diagonais α_i e β_i.

Outra vantagem do FastICA é que ele pode estimar componentes independentes tanto super como subgaussianos sem passos adicionais. A razão para isso é que a matriz diagonal α_i contém estimativas da natureza (super ou subgaussianas) dos componentes independentes. Por outro lado, a matriz diagonal β_i pode ser considerada como uma escala de não-linearidades, desde que possamos reformular $\Big[\text{diag}(\beta_i) + E\{\mathbf{g}(\mathbf{y})\mathbf{y}^T\}\Big] = \text{diag}(\beta_i)\Big[\mathbf{I} + E\{\mathbf{g}(\mathbf{y})\mathbf{y}^T\}\Big]$. Podemos dizer então que FastICA usa uma parametrização mais rica de densidades do que as citadas anteriormente, quando comparávamos apenas duas densidades.

As saídas y_i do FastICA são decorrelacionadas e normalizadas à variância unitária depois de cada passo. Tais operações não são necessárias no algoritmo do gradiente, porém se omitidas no FastICA o tornam instáveis, tornando assim o espaço da otimização ligeiramente reduzido. Nesta versão nenhum branqueamento preliminar é feito. Porém, na prática, muitas vezes é altamente recomendado fazer um pré-branqueamento, possivelmente combinando com a redução de dimensão via análise de componentes principais.

4.7 ICA por Outros Métodos

Existem outros métodos para estimarmos os componentes independentes a saber:

- **ICA pela minimização da Informação Mútua** - É uma importante aproximação para a estimação ICA inspirada na teoria da informação. A motivação deste caso se dá porque pode não ser muito realistica em muitos casos supor que os dados seguem um modelo ICA. Entretanto, gostariamos de desenvolver uma aproximação que não faz suposição nenhuma a respeito dos dados. O que queremos ter é uma proposta geral da medida de dependência dos componentes de um vetor aleatório. Usando tal medida, podemos definir ICA como uma decomposição linear que minimiza esta medida de dependência. Uma das principais utilidades da informação mútua é que serve como um sistema unificador de muitos princípios de

estimação, em particular a estimação de máxima verossimilhança e a maximização da não-normalidade. Esta aproximação fornece uma justificativa rigorosa do princípio heurístico de não-normalidade.

- **ICA por Métodos Tensoriais** - Uma aproximação para estimação da ICA consiste no uso de cumulantes tensores de alta ordem. Tensores podem ser considerados uma generalização de matrizes ou operadores lineares. Cumulantes tensores são então uma generalização da matriz de covariância, que é cumulante tensor de segunda ordem e o tensor de quarta ordem é definido pelo cumulante de quarta ordem $\big(cum(X_i, X_j, X_k, X_l)\big)$. Podemos usar a decomposição dos autovalores da matriz de covariâncias para branquear os dados. Como uma generalização deste princípio, podemos usar o cumulante tensor de quarta ordem para fazer os cumulantes de quarta ordem iguais a zero, ou pelo menos tão pequenos quanto possível. Este tipo de aproximação de decorrelação de ordem maior fornece uma classe de métodos de estimação para ICA

- **ICA por Decorrelação Não-Linear e Análise de Componentes Principais Não-Linerar** - A decorrelação não-linear pode ser vista como uma extensão dos métodos de segunda ordem tais como o braqueamento e a análise de componentes principais. Este método fornece componentes que são combinações lineares não-correlacionadas das variáveis originais. Os componentes independentes podem em muitos casos ser encontrados como combinações lineares não-correlacionadas *não-linearmente*. Uma outra aproximação a ICA esta relacionada com Análise de Componentes Principais (PCA) e é também chamada de PCA não-linear. Uma representação não-linear que minimize o critério do erro de mínimos quadrados é procurada para os dados originais. Para o caso linear os componentes principais são obtidos.

Estes métodos não serão abordados neste trabalho. Para mais informações a respeito, deve-se consultar [16] e [25].

4.8 Algumas Considerações Práticas

Ao aplicar algoritmos ICA na prática algumas considerações devem ser levadas em conta devido aos problemas que podem aparecer. O sucesso da ICA depende crucialmente da performance desses passos de pré-processamento. Alguns já foram considerados anteriormente, porém existem outros que não são necessários na teoria, mas que podem muitas vezes serem úteis na prática.

Quando os dados forem variáveis aleatórias obervadas de uma série temporal, pode ser muito útil filtrar esses dados, seja utilizando médias móveis, filtro passa-baixo ou filtro passa-alto. Outro pré-processamento que pode ser utilizado é via análise de componentes principais com o objetivo

de reduzir a dimensão dos dados.

Existem outras escolhas para serem feitas, como o algoritmo a ser usado, o número de componentes principais que devem ser estimados, dentre outros. Para uma discussão mais detalhada a respeito desses assunto ver [16].

Capítulo 5

ICA Aplicada à Finanças

Segundo [27] a globalização econômica e a internet aceleraram a integração dos mercados financeiros mundiais nos últimos anos. Movimentos no preço de um mercado podem difundir-se rapidamente e instantaneamente para outro. Por esta razão, mercados financeiros são mais dependentes uns dos outros do que antes e deve-se considerá-los conjuntamente para compreender melhor a estrutura dinâmica do mercado financeiro global. Um mercado pode contagiar outro mercado sob certas circunstâncias, contudo o relacionamento pode ser invertido se tais circunstâncias mudam. Consequentemente, conhecer como os mercados estão interrelacionados é de grande importância em finanças. Similarmente, para um investidor ou instituição financeira, a dinâmica de relação entre os retornos tem um importante papel na tomada de decisão.

Um dos métodos usados para estudar conjuntamente uma série múltipla de retornos pertence à análise de séries temporais multivariada a qual consiste de múltiplas séries simples referidas como componentes. Como dito nos capítulos anteriores, a ICA é uma técnica estatística para encontrar componentes fundamentais vindos de dados multivariados. Ou seja, aqui vemos uma situação na qual, claramente, a ICA pode ser utilizada. Existem muitas situações na qual séries temporais financeiras aparecem, que podem ter algum fator fundamental em comum. O que conduz os movimentos de uma série temporal finaceira? Isto certamente é uma questão do interesse de muitos, abragendo de pesquisadores que desejam entender o mercado finaceiro aos negociadores que se beneficiam desse conhecimento. A ICA pode revelar alguns mecanismos de condução que de outra maneira permaneceriam ocultos.

Segundo Back e Weigend a primeira vez que a ICA teve uma aplicação significativa em problemas financeiros ou econométricos foi em [3]. Neste trabalho o objetivo foi verificar se a ICA poderia fornecer alguma indicação de estruturas fundamentais do mercado de ações. Para isso foram utilizados os preços das ações das 28 maiores empresas da bolsa de Tóquio entre os anos de 1986 e 1989. Neste artigo, o algoritmo utilizado foi o JADE (Joint Approximate Diagonalization of Eigenmatrices, para mais detalhes ver [9]. O FastICA só foi apresentado em 1997 no artigo de Hyvärinen e Oja, para mais detalhes ver [17]). Verificou-se que apenas poucos componentes independentes contribuem para a maioria dos movimentos nos retornos das ações e que esses componentes podem

ser classificados em duas categorias a saber: (i) Choques raros, mas grandes, responsáveis pelas maiores mudanças nos preços das ações, e (ii) flutuações pequenas, mas frequentes, contribuindo pouco ao nível total das ações. Encontrou-se também que a ICA é uma ferramenta complementar a PCA, permitindo que a estrutura fundamental dos dados fosse observada mais prontamente. Encontrar as misturas maximamente independentes das ações originais, ou seja, do portfólio, pode ajudar a minimizar o risco em uma estratégia de investimentos.

Em [19] a ICA foi aplicada a séries temporais financeiras paralelas representando o fluxo de caixa simultâneo de 40 lojas pertencentes à mesma cadeia de varejo medidas ao longo de 140 semanas. Os objetivos compreendiam encontrar fatores fundamentais em comum a todas as lojas que afetam o fluxo de caixa; o efeito de fatores específicos no fluxo de caixa em qualquer loja em particular, isto é o efeito de ações tomadas em lojas individuais e em seus ambientes locais; e baseado nesses fatores encontrar grupos de lojas com comportamento similar. Este problema é frequentemente encontrado na prática quando se tenta quantificar o efeito de ações gerenciais, ou seja, neste exemplo, as mudanças no fluxo de caixa são resultados dessas ações gerenciais, resultados de outras ações tais como aquelas dos "competidores" ou apenas flutuações estatísticas?

Para o problema do fluxo de caixa foi usado como algoritmo o FastICA, o qual encontrou 4 componentes independentes que representam 77% da força retida dos sinais originais. Um dos componentes representa claramente os feriados (mais importante sendo o Natal) e um outro variações sazonais mais lentas tais como as férias de verão. O terceito componente representa uma variação muito lenta, tal como uma tendência. O quarto e último componente não tem uma representação clara, mas dentre outras interpretações é possível que represente a posição relativa da rede de lojas na competição com seus concorrentes.

As transformações na ICA tendem a produzir sinais de componentes $s_j(t)$ que podem ser comprimidos com menos bits do que os sinais originais, $x_i(t)$. Eles são mais estruturados e regulares. Isto fornece motivação para tentar predizer os sinais $x_i(t)$ primeiramente indo ao espaço da ICA, fazendo a predição lá e então transformando de volta à série temporal original, como sugerido por [23] *apud* [16]. A predição pode ser feita separadamente e por métodos diferentes para cada componente, dependendo de sua estrutura temporal.

Em [20] são sugeridos os seguintes procedimentos:

1. Após subtrair a média de cada série temporal e o branqueamento (após o qual cada série temporal tem média zero e variância unitária), os componentes independentes $s_j(t)$ e a matriz de mistura $\mathbf{A}$, são estimados usando o algoritmo FastICA. O número de componentes independentes pode ser variável.

2. Para cada componente $s_j(t)$, um filtro não-linear adequado é aplicado para reduzir os efeitos de ruído - suavização para os componentes que contém frequências muito baixas e filtro passa-alto para componentes contendo altas frequências e/ou choques repentinos. A suavização não-linear é feita pela função de suavização f_j aplicada no sinal original $s_j(t)$,

$$s_j^s(t) = f_j\big[s_j(t+r), \ldots, s_j(t), \ldots, s_j(t-k)\big]. \tag{5.1}$$

3. Cada componente independente suavizado é predito separadamente, usando algum modelo auto-regressivo (AR). A predição é feita para um número de passos para o futuro. Isto é feito aplicando a função de predição, g_j, no sinal original suavizado, $s_j^s(t)$:

$$s_j^p(t+1) = g_j\big[s_j^s(t), s_j^s(t-1), \ldots, s_j^s(t-q)\big]. \tag{5.2}$$

   Os próximos passos são previstos deslocando a janela de comprimento q sobre os valores medidos e previstos do sinal suavizado.

4. As predições para cada componente independente são combinadas ponderando-os com os coeficientes de mistura, a_{ij}, obtendo então as predições, $x_i^p(t)$, para a série temporal original, $x_i(t)$:

$$\mathbf{x}^p(t+1) = \mathbf{A}\mathbf{s}^p(t+1) \tag{5.3}$$

   e similarmente para $t+2, t+3, \ldots$.

Os quatro procedimentos acima foram testados numa série de 10 taxas cambiais relevantes, retornando 5 componentes independentes (como exemplo dos fatores fundamentais das taxas de câmbio, podemos supor alguns indicadores econômicos, taxas de juros e fatores psicológicos). Os resultados foram promissores, uma vez que as predições realizadas através da ICA tiveram melhor performance do que as predições diretas. Segundo [20] usando suavização não-linear otimizada para cada componente independente separadamente, a predição destes tem uma performance mais acurada e o ruído é fortemente reduzido, permitindo uma melhor predição dos fatores fundamentais.

Tanto a predição ICA como AR são técnicas lineares, embora a primeira tenha uma regra de aprendizagem não-linear, entretanto a suavização é responsável pela não-linearidade do modelo. Para séries temporais geradas por um modelo linear o algoritmo tem uma performance similar em relação às técnicas lineares clássicas de análise de séries temporais. Entretanto alguns componentes

independentes podem ser úteis em analisar os impactos de diferentes fenômenos externos na taxa de câmbio.

E como último exemplo dado das aplicações da ICA em finanças temos o [14] no qual um novo modelo GARCH multivariado é proposto, chamado modelo GICA-GARCH. Desde que Engel em 1982 no seu artigo em [11] introduziu o modelo ARCH e 4 anos mais tarde Bollerslev em [6] propôs o modelo GARCH, muitos pesquisadores tem tido interesse em modelar volatilidades em finanças. O primeiro modelo GARCH multivarido, MGARCH, foi proposto em [7] como uma extensão do modelo GARCH e outras especificações deste modelo foram propostas na literatura, porém em todos eles, o número de parâmetros a serem estimados pode ser muito grande e as restrições que garantem que a matriz de covariância condicional seja positiva definida são difícies de implementar. Uma solução para esse problema é usar modelos fatoriais.

Uma gama enorme de modelos foram propostos, dentre os quais podemos citar o GARCH Ortogonal (O-GARCH, [1]), o GARCH Ortogonal generalizado (GO-GARCH, [29]) e o GARCH usando componentes condicionalmente não-correlacionados (CUC-GARCH, [12]). Todos esses modelos utilizam a análise de componentes principais para encontrar os fatores fundamentais. Diante disso, em [14] foi proposto um modelo GARCH onde os componentes fundamentais são encontrados via ICA.

Nestre trabalho foram usadas 19 ações, todas compondo o IBEX 35 (índice principal da Bolsa de Madri que compreende as 35 empresas com maior volume negociado) de 2000 a 2004, divididas em seis setores a saber: óleo e energia, tecnologia/telecomunicações, serviços financeiros, Materiais/indústria/construção, serviços de consumo (turismo, publicidade,etc) e outras mercadorias de consumo. O objetivo foi encontrar os componentes independentes e fazer as previsões da volatilidade dos retornos comparando os resultados com os obtidos via componentes principais.

Para tal comparação foram utilizados três algoritmos para encontrar os componentes independentes, FastICA, JADE e SOBI (Second-Order Blind Identification, para mais detalhes ver [4] que é uma extensão do trabalho feito em [28]) os quais tiveram os resultados comparados entre si, além de comparados com um método de componentes principais realizados através do modelo O-GARCH. Os resultados mostraram que os métodos via ICA tiveram melhor performance do que o método feito atráves dos componentes principais, independente do algoritmo utilizado, tanto na estimação dos componentes heteroscedásticos e/ou não-lineares bem como na previsão da volatilidade dos retornos. Já os resultados entre os diferentes tipos de algoritmos da ICA foram mistos, variando pouco dependendo do número de componentes a serem escolhidos. Porém, de maneira geral o FastICA e JADE apresentaram perfomances muito semelhantes e um pouco melhor do que o SOBI. Além disso, os componentes principais encontrados estavam principalmente associados com os setores elétrico, bancário e de construção.

Exposto isso, no próximo capítulo será apresentado um estudo detalhado da aplicação da ICA num conjunto de séries temporais financeiras.

Capítulo 6

Aplicando ICA a Dados Reais

Neste capítulo iremos aplicar a ICA a um conjunto de dados reais. Em primeiro lugar serão descritos os dados usados; em segundo explicaremos o procedimento para estimar os componentes e por fim apresentaremos os resultados usando componentes independentes para prever a volatilidade dos retornos.

Foram usados os preços diários de fechamento (veja Figura 6.1 para o gráfico de uma delas) de 20 ações entre jan/02 e dez/07 ($T = 1488$ é o número de observaçoes), todas elas compondo o índice BOVESPA (carteira teórica para o quadrimestre setembro/dezembro de 2009) e representando algo por volta de 56% do volume de negócios composto pelo índice. O índice BOVESPA (IBOVESPA) é o mais importante indicador do desempenho médio das cotações do mercado de ações brasileiro. É formado pelas ações com maior volume negociado nos últimos meses. É o valor atual, em moeda corrente, de uma carteira teórica de ações constituída em 02/01/1968 (valor-base: 100 pontos), a partir de uma aplicação hipotética. Supõe-se não ter sido efetuado nenhum investimento adicional desde então, considerando-se somente os ajustes efetuados em decorrência da distribuição de proventos pelas empresas emissoras (tais como reinversão de dividendos recebidos e do valor apurado com a venda de direitos de subscrição, e manutenção em carteira das ações recebidas em bonificação). Dessa forma, o índice reflete não apenas as variações dos preços das ações, mas também o impacto da distribuição dos proventos, sendo considerado um indicador que avalia o retorno total de suas ações componentes.

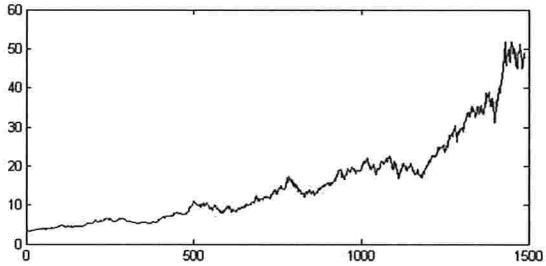

Figura 6.1: Preço diário de fechamento da ação da Vale de 2002 a 2007.

A Tabela 6.1 apresenta um resumo das 20 ações utilizadas.

Setor	Segmento	Empresa	Cód. Negociação
Matérias Básicas	Siderurgia	Gerdau	GGBR4
		Usiminas	USIM5
		Sid. Nacional	CSNA3
	Petroquímico	Braskem	BRKM5
	Papel e Celulose	Aracruz	ARCZ6
	Mineração	Vale	VALE5
Petróleo	Exploração e Refino	Petrobras	PETR4
Financeiro	Banco	Bradesco	BBDC4
		ITAUSA	ITSA4
		Banco do Brasil	BBAS3
Utilidade Pública	Energia Elétrica	CEMIG	CMIG4
		ELETROBRAS	ELET3
	Gás	Comgás	CGAS5
Consumo não Cíclico	Cigarros e Fumo	Souza Cruz	CRUZ3
	Bebidas	AMBEV	AMBV4
	Alimentos Processados	Sadia	SDIA4
Telecomunicações	Móvel	TIM	TCSL4
	Fixa	Telemar	TNLP4
Consumo Cíclico	Comércio	Lojas Americanas	LAME4
Bens Industriais	Material Aeronáutico	EMBRAER	EMBR3

Tabela 6.1: Setores, segmentos, empresas e códigos de negociação das ações

O retorno diário (veja a Figura 6.2 para o retorno da série da Figura 6.1) da i-ésima empresa é calculado como:

$$r_{it} = \log(p_{it+1}) - \log(p_{it}), i = 1, \ldots, 20, \tag{6.1}$$

onde p_{it} é o preço de fechamento da i-ésima ação no tempo t. Assim, temos uma matriz de retornos multivariado de 20×1487 (o algoritmo FastICA utilizado funciona com os retornos em linha e as colunas representando o tempo) o qual é denotado por $\mathbf{r}_t$, cujas colunas são os valores dos retornos das 20 ações em 1488 dias de negócios entre os anos de 2002-2007.

Retirando a média dos retornos, $\mathbf{x}_t = \mathbf{r}_t - \bar{\mathbf{r}}_t$, ou seja, transformando em uma série de média zero (primeiro passo para o branquemento dos dados) calculamos as curtoses dos coeficientes $\mathbf{x}_t$ e os resultados se encontram na Tabela 6.2.

Pelos valores encontrados podemos observar que os dados são supergaussianos, ou seja, tem as

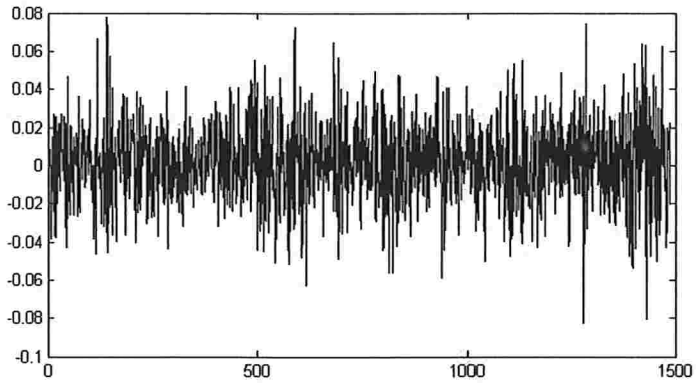

Figura 6.2: Retorno diário da Vale entre os anos de 2002 a 2007.

VALE5	PETR4	CMIG4	USIM5	BBDC4	ITSA4	SDIA4
3,6472	5,2481	3,6621	3,6457	3,8535	3,4422	3,8615
CGAS5	**AMBV4**	**ARCZ6**	**BBAS3**	**CSNA3**	**ELET3**	**GGBR4**
5,0000	8,5260	4,0729	4,4950	3,9946	3,7955	3,6997
BRKM5	**EMBR3**	**LAME4**	**CRUZ3**	**TNLP4**	**TCSL4**	
4,6436	4,7101	5,2558	3,8586	4,2014	3,5494	

Tabela 6.2: Curtose dos retornos de média zero

caudas mais pesadas que a da distribuição normal.

6.1 Ponderando os Componentes Independentes

Para estimar os componentes independentes utilizou-se um algoritmo de ponto fixo, FastICA, proposto por [15], que é uma versão mais robusta da versão apresentada em [17], pois usa uma aproximação de negentropia ao invés da curtose para medir a não-normalidade destes componentes. Após a estimação faz-se necessário definir quais destes são mais importantes para explicar a estrutura que fundamenta os dados. Como os componentes independentes não são classificados com respeito a nenhuma ordem, seguindo [3] classificaremos os componentes independentes em termos de sua variabilidade explicada.

Já vimos que a i-ésima variável observada é dada por $x_{it} = \sum_{j=1}^{20} a_{ij} s_{jt}$, e sua variância é dada por

$$var(x_{it}) = \sum_{j=1}^{20} a_{ij}^2, \forall i = 1, \ldots, 20. \tag{6.2}$$

Para cada x_{it}, como $i = 1, \ldots, m$, [3] definem os pesos dos componentes independentes em termos dos elementos da i-ésima linha de $\mathbf{A}$ como $s_t^{w(i)} = diag(a_{i1}, a_{i2}, \ldots, a_{im})$. Isto é, para cada x_{it}, o j-ésimo componente independente ponderado é dado por $s_{jt}^{w(i)} = a_{ij} s_{jt}$, e sua variância é dada por:

$$var\left(s_{jt}^{w(i)}\right) = a_{ij}^2, \forall i, j = 1, \ldots, m. \tag{6.3}$$

Então a partir de (6.2) e (6.3), a variância de x_{it} que é explicada por $s_{jt}^{w(i)}$, é calculada como:

$$\nu_j^i = \frac{a_{ij}^2}{\sum_{j=1}^m a_{ij}^2}. \tag{6.4}$$

A variância total de $\mathbf{x}_t$ explicada pela j-ésimo componente independente é dado por:

$$\vartheta_j = \frac{\sum_{i=1}^m \nu_j^i}{\sum_{j=1}^m \left(\sum_{i=1}^m \nu_j^i\right)}. \tag{6.5}$$

Assim, após obtermos $(\vartheta_1, \vartheta_2, \ldots, \vartheta_m)$, podemos ordenar os componentes independentes em termos de suas variabilidades. O componente mais importante será aquele que explica o máximo da variância de $\mathbf{x}_t$.

A Tabela 6.3 apresenta os componentes independentes ordenados em função da variabilidade explicada.

ϑ_j	%	ϑ_j	%	ϑ_j	%	ϑ_j	%
ϑ_{11}	14,20	ϑ_5	5,24	ϑ_4	3,51	ϑ_1	3,07
ϑ_{12}	12,32	ϑ_{20}	5,09	ϑ_6	3,48	ϑ_{14}	2,88
ϑ_{18}	8,23	ϑ_7	4,13	ϑ_2	3,46	ϑ_{13}	2,87
ϑ_3	6,06	ϑ_{10}	3,82	ϑ_{19}	3,22	ϑ_{15}	2,83
ϑ_{16}	5,96	ϑ_{17}	3,82	ϑ_9	3,07	ϑ_8	2,73

Tabela 6.3: Componentes ordenados em função da variabilidade explicada

Utilizando a Tabela 6.3 podemos construir o gráfico apresentado na Figura 6.3, a qual mostra a variabilidade explicada pelo componente estimado, para decidir o número ótimo de componentes a ser escolhido.

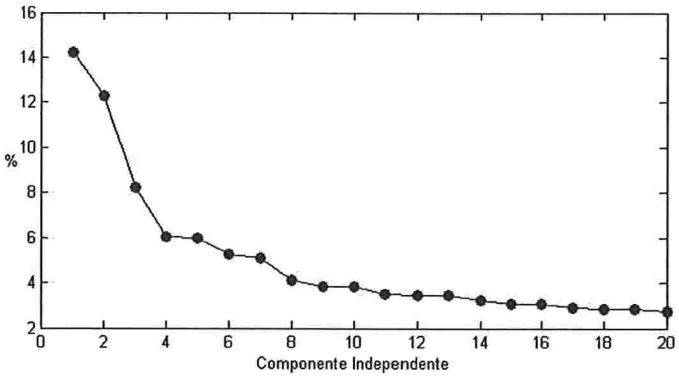

Figura 6.3: Variabilidade explicada individualmente pelos 20 componentes independentes ordenados por importância

Optamos por considerar os quatro primeiros componentes da Tabela 6.3, que respondem por 40,81% do total da variabilidade explicada.

Estamos interessados em investigar quais ações são mais importantes para definir cada componente independente. Como já visto, os componentes podem ser escritos como combinação linear dos retornos, $s_{it} = \sum_{j=1}^{20} w_{ij} x_{jt}$, onde w_{ij} representa o efeito do j-ésimo retorno sobre o i-ésimo componente, e os maiores pesos correspondem às ações mais importantes.

Observando a matriz $\mathbf{W}$ (tabela 6.4) verificamos que para os quatro componentes escolhidos não é possível fazer uma associação clara com algum setor. Ambos apresentam uma contribuição diversa entre as ações. Para a componente 11 as ações de maior peso são Braskem, ITAUSA e Gerdau (pesos semelhantes entre si). Para a componente 12 observamos que há uma grande contribuição da ação da Aracruz e num segundo momento das ações da ITAUSA, Telemar, TIM e Petrobras. Já a componente 18 é basicamente explicada pela ação da Vale. E por fim a componente 3 sofre grande influência das ações da Gerdau, ITAUSA e Vale além de uma contribuição considerável da Aracruz, CSN e TIM. Ou seja, os componentes independentes não podem ser vistos como um índice do mercado, mas sim como associados principalmente às ações descritas separando os retornos em termos de sua variabilidade individual explicada. A Tabela 6.4 apresenta os maiores coeficientes da matriz $\mathbf{W}$ para os componentes escolhidos.

CP11	CP12	CP18	CP03
Braskem - 25,151	Aracruz - 24,179	Vale - 37,124	Gerdau - 28,624
ITAUSA - 25,127	ITAUSA - 19,477	Embraer - 16,423	ITAUSA - 24,649
Gerdau - 22,867	Telemar - 18,811	Cemig - 15,012	Vale - 20,979
TIM - 11,122	TIM - 16,938	Aracruz - 12,406	Aracruz - 18,873
Telemar - 8,6327	Petrobras - 13,941	Souza Cruz - 12,384	CSN - 14,963
Bradesco - 7,0596	Braskem - 11,554	Usiminas - 12,092	Tim - 12,421
Souza Cruz - 6,9894	Comgás - 10,944	Gerdau - 11,006	Ambev - 8,7932
Cemig - 4,8665	Gerdau - 10,122	Comgás - 10,459	Sadia - 7,8157
Sadia - 4,5928	Vale - 10,092	Bradesco - 10,351	Braskem - 6,7779
Aracruz - 3,8861	Ambev - 9,8743	Eletrobras - 9,1754	Cemig - 6,4464
BB - 3,6095	Eletrobras - 7,9189	Tim - 7,8748	BB - 5,2088
Petrobras - 3,4416	Bradesco - 6,8929	Ambev - 6,1275	Souza Cruz - 5,0397
Ambev - 3,0161	CSN - 5,5489	BB - 2,7889	Usiminas - 3,3905
Vale - 2,8131	Usiminas - 5,2032	Petrobras - 1,9334	Embraer - 3,2827
Comgás - 1,5811	Embraer - 4,907	Sadia - 1,3859	Petrobras - 3,1551
CSN - 1,508	BB - 4,8098	Braskem - 1,0562	Lojas Ame. - 2,6693
Eletrobras - 0,7778	Lojas Ame. - 3,7955	CSN - 0,96037	Bradesco - 1,8108
Usiminas - 0,37879	Sadia - 2,936	ITAUSA - 0,83702	Comgás - 1,5549
Lojas Ame. - 0,35448	Cemig - 1,7648	Telemar - 0,5409	Eletrobras - 1,3869
Embraer - 0,23515	Souza Cruz - 1,6436	Lojas Ame. - 0,2064	Telemar 0,15594

Tabela 6.4: Coeficientes dos componentes escolhidos

A Figura 6.4 apresenta o gráficos dos 4 componentes escolhidos.

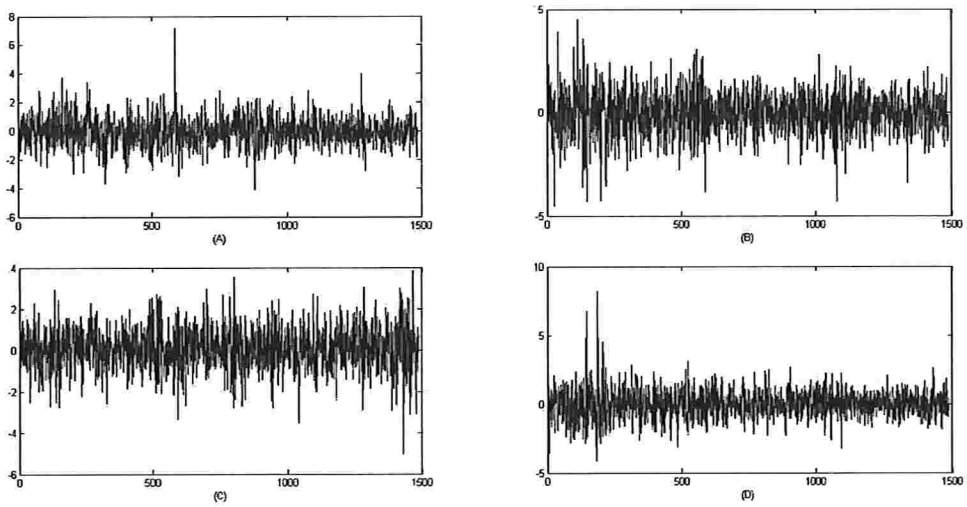

Figura 6.4: Componentes principais escolhidos: (A) - CP11; (B) - CP12; (C) - CP18 e (D) - CP03

6.2 Identificando os Fatores

Após estimar o modelo e ponderar os componentes independentes, estes foram ordenados em termos da variabilidade total explicada, sendo os primeiros os mais importantes. Assim, escolhemos $s_t^{(1)} = (s_{it}, \ldots, s_{rt})'$ como os r componentes para representar os dados e $s_t^{(2)} = (s_{r+1t}, \ldots, s_{mt})'$ consideramos os $m - r$ componentes como ruído. No nosso caso, $m = 20$ e $r = 4$, pois a partir do 4º componente a diferença entre as variabilidades explicadas é muito pequena. A partir de agora o foco se dá somente nos 4 componente selecionados para os quais serão ajustados modelos univariados $ARMA(p,q) - GARCH(p',q')$. Estima-se, então, a volatilidade univariada de cada componente independente e gera-se a matriz de covariância condicional de $s_t^{(1)}$, $\mathbf{H}_t$. Finalmente temos a matriz de covariância condicional dos dados observados e seu i-ésimo termo da diagonal, $\gamma_{it}^2 = \sum_{j=1}^{4} h_{jt} a_{ij}^2$, é a variância condicional de x_{it} para $i = 1, \ldots, 20$. É importante notar que a performance do modelo depende do método aplicado para estimar os componentes independentes.

Como nosso objetivo está na previsão da volatilidade dos retornos, podemos ordenar o procedimento de previsão nos seis seguintes passos:

1. Utilizando os dados entre jan/02 a dez/07 (1488 obervações) estimamos a matriz $\mathbf{A}$ e os componentes não observados. Então, ordenamos os componentes e $r = 4$ são fixados.

2. Usando as 1487 observações ajustamos um modelo $ARMA(p,q) - GARCH(p',q')$ para cada componente $\hat{s}_{jt}$, com $j = 1, \ldots, 4$. A Tabela 6.5 apresenta os modelos estimados bem como os coeficientes de acordo com a equação (3.48).

Componente	Modelo	Coeficientes
CP11	GARCH(1,1)	$\alpha_0 = 0,072339$ $\alpha_1 = 0,065365$ $\beta_1 = 0,861581$
CP12	AR(4)-GARCH(1,1)	$\phi_3 = -0,06065$; $\phi_4 = -0,07035$ $\alpha_0 = 0,008099$; $\alpha_1 = 0,020639$ $\beta_1 = 0,970502$
CP18	GARCH(1,1)	$\alpha_0 = 0,059991$ $\alpha_1 = 0,04817$ $\beta_1 = 0,892026$
CP03	AR(5)-GARCH(1,1)	$\phi_1 = 0,065112$; $\phi_2 = -0,057627$ $\phi_5 = -0,054906$; $\alpha_0 = 0,020139$ $\alpha_1 = 0,046746$; $\beta_1 = 0,930999$

Tabela 6.5: Modelos ajustados e seus coeficientes

3. Geramos a previsão de um passo a frente para a volatilidade de cada $\hat{s}_{jt}$,

$$\hat{h}_{j,1488|1487} = V\left[\hat{S}_{j1487}|\mathbf{I}_{1486}\right], j = 1, \ldots, 4. \tag{6.6}$$

Então, fazendo a previsão para $t = 1488, \ldots, 1609$, temos:

$$\hat{\mathbf{H}}_{t|1487} = diag(\hat{h}_{1,t|1487}, \ldots, \hat{h}_{4,t|1487}), t = 1488, \ldots, 1609, \tag{6.7}$$

a qual é a matriz de covariância condicional de $\hat{s}_t = (\hat{s}_{1t}, \ldots, \hat{s}_{4t})'$ no tempo t.

4. Temos que $\Omega_t = V(\mathbf{x}_t|\mathbf{I}_{t-1}) = \mathbf{A}\mathbf{H}_t\mathbf{A}'$ é a matriz de covariância condicional de $\mathbf{x}_t$, onde $\mathbf{H}_t = diag(h_{1t}, \ldots, h_{4t})$ é a matriz$_{4\times4}$ de covariância condicional de s_t no tempo t. Assim, a variância condicional de $\mathbf{x}_i$ no tempo t é dado pelo i-ésimo termo da diagonal de Ω_t,

$$\hat{\gamma}^2_{i,t|1487} = \sum_{j=1}^{4} \hat{h}_{j,t|1487} a^2_{ij}, i = 1, \ldots, 20, \ t = 1488, \ldots, 1609. \tag{6.8}$$

Então fazemos a previsão da variância condicional dos retornos a partir da volatilidade predita dos componentes não-observados.

5. Para avaliar a performance da previsão do modelo ICA-GARCH nós precisamos comparar a volatilidade prevista com a real. Entretanto, volatilidade não pode ser observada. De acordo com [13] podemos medir a "verdadeira volatilidade" do i-ésimo retorno no tempo t por

$$v_{it} = (x_{it} - \bar{x}_i)^2, i = 1, \ldots, 20, \ t = 1488, \ldots, 1609, \tag{6.9}$$

onde $\bar{x}_i$ é a média do retorno x_i sobre as 1487 observações. Então, o erro de previsão de um passo a frente é dado por:

$$\epsilon_{it} = v_{it} - \hat{\gamma}^2_{i,t|1487}, i = 1, \ldots, 20, \ t = 1488, \ldots, 1609. \tag{6.10}$$

6. Para avaliar a acurácia do modelo dividimos o erro de predição (6.10) por algum referencial alternativo obtido usando outro método padrão de previsão. Neste caso assumimos que o método do referencial prevê a volatilidade do retornos por sua variância marginal. Então, definimos a razão:

$$RE_{it} = \frac{\epsilon_{it}}{\epsilon^*_{it}}, i = 1, \ldots, 20, \ t = 1488, \ldots, 1609. \tag{6.11}$$

onde ϵ^*_{it} é o erro de previsão do i-ésimo retorno obtido pelo método de referência, isto é,

$$\epsilon^*_{it} = \upsilon_{it} - \hat{\sigma}^2_i, i = 1, \ldots, 20, \ t = 1488, \ldots, 1609. \tag{6.12}$$

onde $\hat{\sigma}^2_i$ é a variância marginal do i-ésimo retorno até o tempo t. Para minimizar o impacto de outliers quando analizamos a performance da previsão da volatilidade no modelo, usamos o Erro Mediano Relativo Absoluto (MdRAE - sigla em inglês):

$$MdRAE(RE_{it}) = median(|RE_{it}|). \tag{6.13}$$

A média de (6.13) para $i = 1, \cdots, 20$ e $t = 1488, \cdots, 1609$ foi de 0,736415. Lembrando que esse valor é para o caso de $r = 4$. Feito isso, repetimos os passos de 1 a 6 acima considerando apenas 2 e 3 componentes principais afim de verificar se o resultado se altera e quanto. O objetivo é verificar se perdemos muito com menos componentes. A Tabela 6.6 apresenta as médias de (6.13) para r de 2 a 4.

Número de Componentes (r)	$MdRAE(RE_{it})$
2	0,833938
3	0,771903
4	0,736415

Tabela 6.6: Média das MdRAE medida para os 20 retornos

Da Tabela 6.6 podemos verificar que quanto maior o número de componentes escolhidos (r) menor é a medida (6.13), ou seja, melhor é o modelo de previsão. Porém com apenas esses 3 resultados, podemos observar também que as diferenças entre os valores obtidos de (6.13) diminuem conforme r aumenta (diminui 0,062035 passando de 2 para 3 componentes e 0,035488 passando de 3 para 4 componentes).

Capítulo 7

Considerações Finais

Neste trabalho apresentamos um método alternativo para modelar volatilidades multivariadas. Podemos considerar o modelo como um ICA-GARCH o qual assume que os movimentos conjuntos de um vetor de dados financeiros são guiados por um número pequeno de componentes independentes os quais são estimados pela ICA, permitindo estimar um modelo GARCH multivariado usando um número pequeno de fatores independentes e com heteroscedasticidade condicional. Como este modelo reduz o número de parâmetros a serem estimados, ele fornece uma representação parcimoniosa dos dados.

Apesar da utilização da ICA não ser muito recente (os primeiros usos datam do ínicio da déc. 80), a sua utilização em dados finaneiros tem pouco tempo, todavia, a ICA tem se mostrado muito promissora, mais simples e com melhores resultados do que alguns modelos (tais como os MGARCH e modelos ortogonais que utilizam componentes principais).

Após apresentarmos um breve histórico da utilização da ICA em dados financeiros, foi feita a estimação e previsão para uma carteira contendo 20 ações da BOVESPA utilizando de 2 a 4 componentes independentes. Podemos concluir que quanto maior o número de componentes escolhidos menor é a medida das médias da $MdRAE(RE_{it})$, portanto melhor é o modelo de previsão. Observa-se também que as diferenças diminuem conforme r aumenta. Possíveis estudos futuros relacionados a ICA, podem envolver tanto métodos alternativos na escolha dos componentes principais bem como a escolha do número ótimo de componentes.

Uma outra possibilidade futura é a comparação dos resultados usando diferentes algoritmos para os dados financeiros brasileiros. Isso já foi feito para outros lugares, tal como com dados da bolsa de Madri (IBEX 35). Porém não sabemos se para os nossos dados o FastICA se adequa melhor ou seria algum outro algoritmo tais como JADE ou SOBI. Pode ser também que não haja diferenças entre eles como no caso do IBEX 35.

Referências Bibliográficas

[1] Alexander, C.(2001) *Market Models: A Guide to Financial Data Analysis.* Chichester, U.K.:Wiley

[2] Amari, S-I.(1998) Natural gradient works efficiently in learning. *Neural Computations*, 10(2):251-276, MIT.

[3] Back, A. D and Weigend, A. S.(1997) A first application of independent component analysis to extracting structure form stock returns. *International Journal of Neural Systems* 8, 473-484.

[4] BELOUCHRANI, A., ABED MERAIN, K., CARDOSO, J.-F., and MOULINES, E.(1997) A blind source separation tecnique based on second order statistics. *IEEE Transactions on Signal Processing* 45, 434-444.

[5] Bolfarine, H. e Sandoval, M. C.(2001) *Introdução à Inferência Estatística.* Sociedade Brasileira de Matemática, IMPA.

[6] Bollerslev, T.(1986) Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 31, 307-327.

[7] Bollerslev, T., Engle, R. F. and Wooldridgee, J. W.(1988) A capital asset pricing model with time-varying covariances. *Journal of Political Economy* 96, 116-131.

[8] Cardoso, J-F. e Laheld, B. H.(1996) Equivariant adaptative source separation. *IEEE Transactions on Signal Processing*, 44(12):3017-3030

[9] Cardoso, J. and Souloumiac, A.(1993) Blind beamforming for non-gaussian signals. *IEE Proc. F.* 140(6):771-774.

[10] Casella, G. e Berger, R. L.(2002) *Statistical Inference.* 2nd Ed. Duxbury Advanced Series.

[11] Engel, R. F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987-1008.

[12] Fan, J., Wang, M. and Yao, Q.(2008) Modelling multivariate volatilities via conditionally uncorrelated components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(4), 679-702.

[13] Franses, Ph. e Van Dijk, D.(1996) Forecasting stock market volatility using nonlinear GARCH models. *Journal of Forecasting*15, 229-235.

[14] García-Ferrer, A., González-Prieto, E. and Peña, D.(2008) A Multivariate generalized independent factor GARCH model with an application to financial stock returns. Universidad Carlos III de Madrid. Working Papers 08-75. *Statistics and Econometrics Series* 28.

[15] Hyvärinen, A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE. Transaction on Neural Networks* 10, 626-634.

[16] Hyvärinen, A., Karhunen, J. e Oja, E.(2001) *Independent Component Analysis.* New York: John Wiley & Sons.

[17] Hyvärinnen, A. and Oja, E.(1997) A fast fixed-point algorithm for independent component analysis. *Neural Computation* 9, 1483-1492.

[18] James, B. R.(1981) *Probabilidade: Um Curso em Nível Intermediário.* Rio de Janeiro: CNPq-IMPA. Projeto Euclides.

[19] Kiviluoto, K. and Oja, E.(1998) Independent component analysis for parallel financial time series. *Inc Proc. Int. Conf. on Neural Information Processing*, vol.2,pp.895-898.Tokyo

[20] Malaroiu, S., Kiviluoto, K. and Oja, E.(2000) Time series prediction with independent component analysis. *Technical Report.* Helsinki University Of Technology. Helsinki.

[21] Morettin, P. A.(2008) *Econometria Financeira - Um Curso em Séries Temporais Financeiras.* Editora Blucher.

[22] Papoulis, A.(1991) *Probability, Random Variables, and Stochastic Processes.* 3rd edition. McGraw-Hill.

[23] Pawelzik, K., Müller, K.-R and Kohlmorgen, J.(1996) Predicition of mixtures. *In Proc. Int. Conf. on Artificial Neural Networks (ICANN*'96), pages 127-132. Springer.

[24] Rosenblatt, M.(1985) *Stationary Sequences and Random Fields.* Birkhauser Boston Inc.

[25] Stone, J. V.(2004) *Independent Component Analysis: A Tutorial Introduction.* MIT.

[26] Stuart, A., e Ord, K.(2006) *Kendall's Advanced Theory of Statistics - Distribution Theory.* 6th. edition. New York: Oxford University Press.

[27] Tsay, R. S.(2005) *Analysis of Financial Time Series.* Second Edition. New York:Wiley

[28] Tong, L., Liu, R.-W., Soon, V. C., Huang, Y.-F. and Liu, R.(1990) AMUSE: A new blind identification algorithm. *Proceedings of IEEE ISCAS.*

[29] Van Der Weide, R.(2002) GO-GARCH: a multivariate generalized orthogonal GARCH model. *Journal of applied Econometrics* 17, 549-564.

[30] Whittle, P.(2000) *Probability via Expectation.* 4th. edition. Springer Texts in Statistics.